Final Project Description

By Animesh

Why did I choose this question ?

Diabetes is a major issue that happens among people who are aging where the blood sugar level drastically increase due to factors like age,obesity, and amount of sugar they consume they consume on a daily basis like chocolates and sweets. In cases like diabetes majority of the people depend on medicines and drugs like insulin to control their sugar levels and reduce it. The data set that have been provided brainstorm the diabetic patients of different ages on their glucose levels, insulin they have taken,pregnency ,bmi,etc. Therefore I have chosen this question to validate and see whether insulin really controls the sugar levels in this case glucose levels or not for people of different ages based on the data set that have already been provided. This will thus show proof and confirm to all people whether it actually does or not.

How did I answer this question ?

I answered this question with the help of data set that have already been provided in the csv file.However since there was a long set of data , I just considered taking just the first 5 data. In order to make it more simpler for me to read and interpret the data faster, I plotted the scatter plot for the 5 data sets to come up with my predictions that I can see. I also needed my individual research , because the glucose levels were present when you had taken the insulin. There was no data provided on basis of how much was the initial glucose level of people before taking the insulin. Therefore with the help of internet I had to research of the initial glucose levels of diabetic patients of their particular age and predict that as the initial data to make such assumptions. Though this is not a reliable method , since different people have different glucose levels however certain things were missing in the dataset and one can only research further in depth to understand the dataset better.

Links that I have researched in : https://www.mayoclinic.org/diseases-conditions/diabetes/diagnosis-treatment/drc-20371451#:~:text=A%20fasting%20blood%20sugar%20level,Glucose%20tolerance%20test.

https://www.healthline.com/health/diabetes/blood-sugar-level-chart#recommended-ranges

Why did I choose this particular approach ?

I chose this particular approach of making a scatter plot and researching on network so that I can visualize and interpret the data I can see faster. Plotting a scatter plot will also help me understand on how well are variables like insulin and glucose correlated, whether insulin really impacts the glucose level by a ton or not. It's important to research as well so that you can understand the things and make assumptions faster with such evidences, otherwise without knowing and understanding your assumption may not be reliable or valid and it may even cause bias.That's why I used this particular approach to make the data fair and most reliable for all readers.

Explanation of the results :

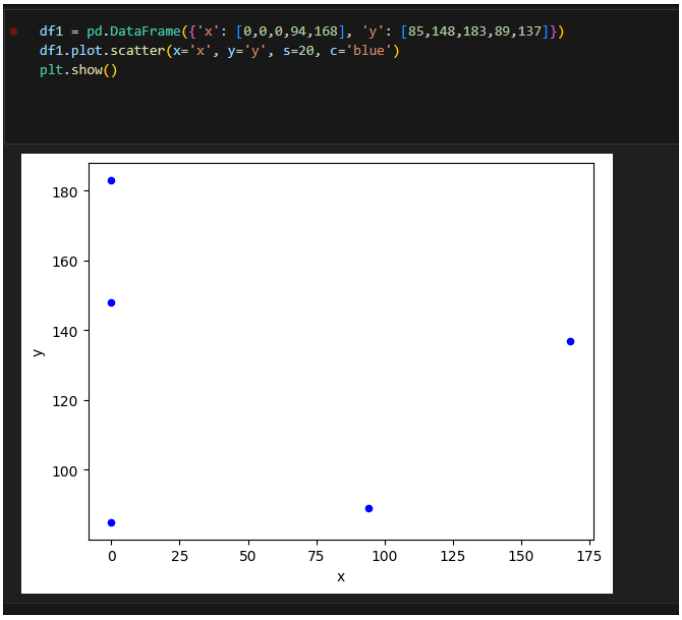

Fig-1 = Scatter Plot that I had constructed on Vs code (Refer to the Final Conda Environment also uploaded on github to see the data more clearly)

As said in the csv file. Most of the results already show the initial glucose level of the people as they have not taken any insulin shots. However, for the people who have taken, really did in fact helped in controlling and reducing the glucose to normal glucose levels. For glucose level that still seem to be high even after taking the insulin shots conclude that their initial glucose level may even be higher than expected , or it can be affected by other factors like age or bmi, as mentioned in the data set, where an older person would be in need of more insulin shots to reduce the glucose level compared to a person whose still young.

How I went through the project ? I firstly tried to break down the data and analyze the data that was provided in the data set in WSL, this included all possible correlation between all the columns present, seeing all the different columns with the help of df.head() or df.tail() from which we can make the most related question that can affect both the variables directly or indirectly, after I was done analysing and seeing a correlation, I tried to see which will most connect with the topic of diabetes or not. Once I found out a question that was both correlated in someway as well as connecting with the topic of diabetes , I made a scatter plot of the first 5 data points of the data set because it helps connect and show the relationship between the 2 variables of the question. I even made the scatter plot to answer the question that I had constructed based on these data points given.I even researched some commands of plotting a proper scatter plot.Once I was done with the scatter plot, I analysed the results and answered the question on behalf of what I thought as well as what the data was showing with additional research done on google.(Links of those research links are present above)

Link I researched for the commands to make a scatter plot in VS code : https://saturncloud.io/blog/how-to-create-a-python-scatter-plot-from-a-pandas-dataframe-with-many-columns/#:~:text=To%20create%20a%20scatter%20plot%20from%20a%20Pandas%20DataFrame%2C%20we,scatter()%20method.&text=In%20this%20example%2C%20we%20have,each%20containing%20100%20random%20values.&text=This%20will%20create%20a%20scatter,plotted%20along%20the%20y%2Daxis.

How I created Conda Environment ?

The teacher had already provided us the csv file of the data set called diabetes. Once the file was downloaded, I inputted it into the "Data Science Fundamental Folder" that was present in the folder " linux" so that I can start working on WSL in Vs Code. Once the file was loaded, I opened my vs code , then connected with the WSL and opened the file. There I made the final conda environment called data-science-final.ipynb with the help of add file option in the Vs Code so that I can start reading the data of the file that I had loaded with the help of python version 3.9.To install the

python , I simply went to the data-science-final.ipynb file , where there was add kernel option and simply added the python version 3.9 from there. Once the python was added I installed pandas,numpy, matplotlib,seaborn,etc with the help of pip install. (You can even refer to the data-science-final file as well if you want to check up the commands I have used that I have uploaded already on github). Once these were installed, I started importing basic commands to read the file properly such as , df.columns, df.head(), df=pd.read_csv('Diabetes_Csv'). This is how I created the Conda Environment.