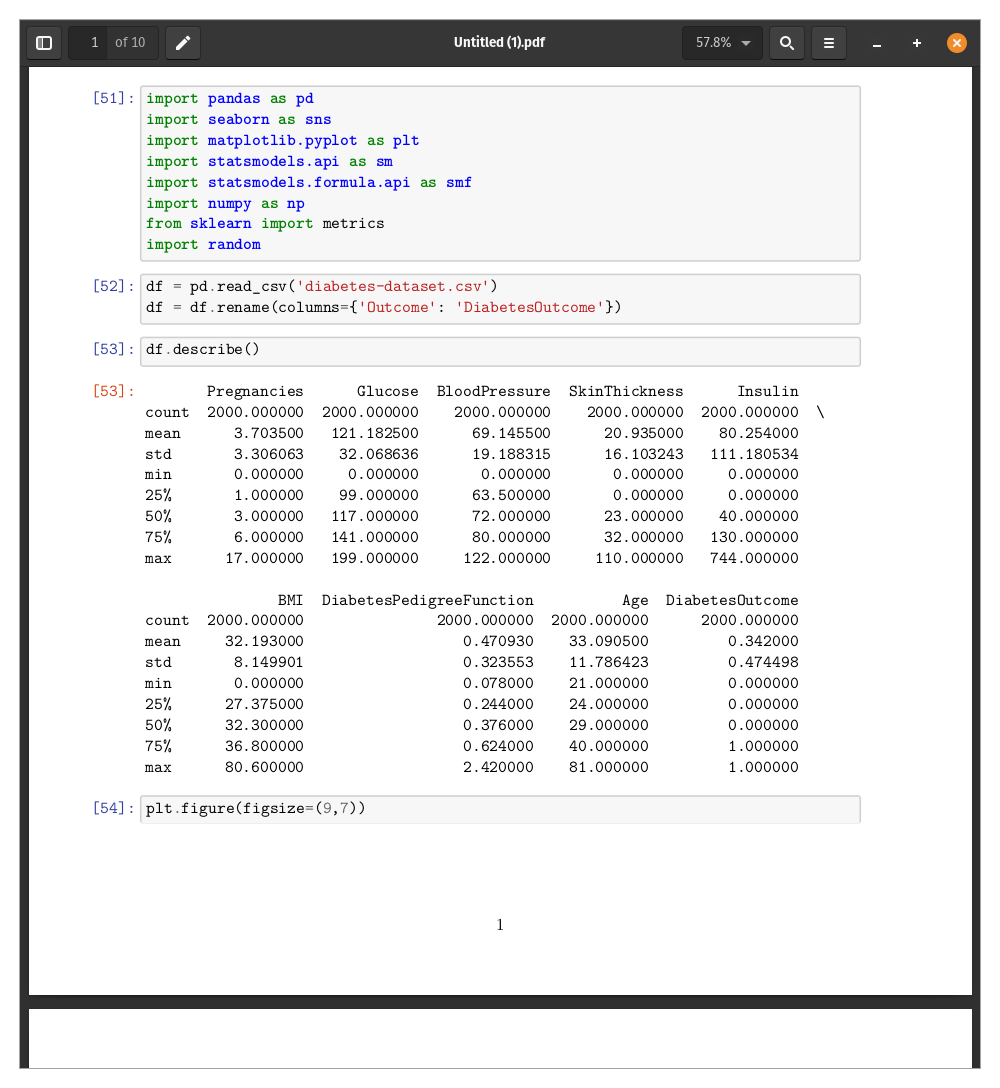
**Diabetes Dataset Analysis**

# Introduction

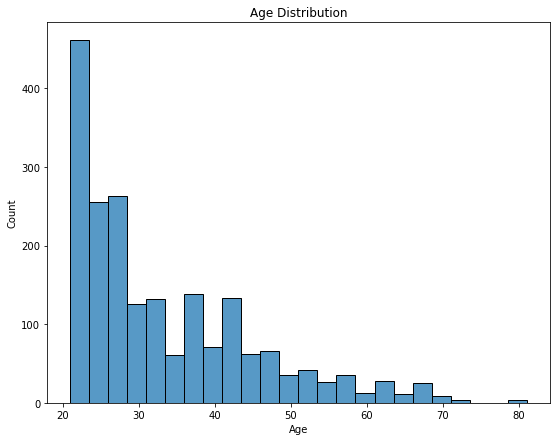
Diabetes is a condition that appears with extremely high levels of glucose in the blood system that lead to complications where the body cannot respond as needed to insulin. This disease is often chronic and often has dire consequences. The presence of a high amount of sugar in the blood also leads to other complications which may affect the nerves, the heart, and the eyes. Diabetes often occurs in two major types, type 1 and type 2. The latter is the most common and has led to a worldwide problem.



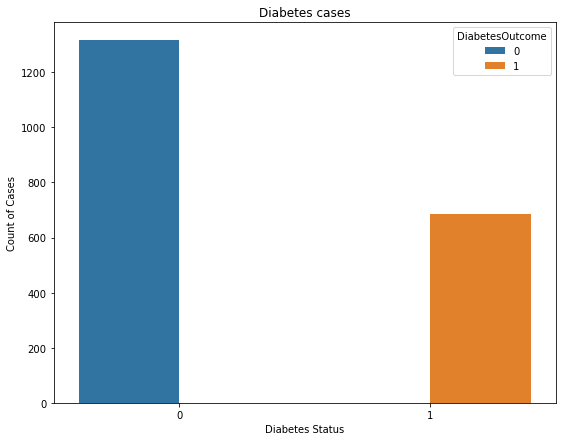
To handle the problem, it is necessary to investigate the disease and determine how it can be diagnosed and handled. The dataset represents a collection of data on 2000 people with and without diabetes. There are also a number of variables, such as the pregnancies one has had, glucose levels, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, and age. The dataset also captures the outcome of the diabetes test for the people, represented in either a 0 for no diabetes or a 1 for diabetes cases.

# Questions

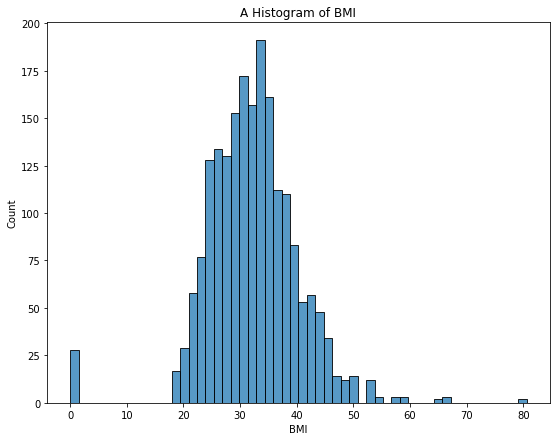
## What kind of patients are represented by the dataset?



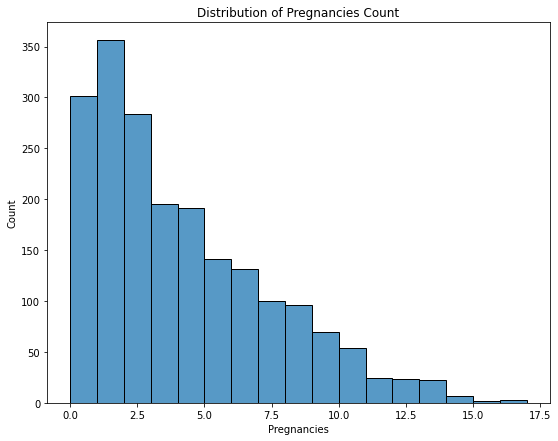
Most of the people captured in the dataset are young people, as indicated by the graph. The count decreases as the age increases, following the general trend in the visualization. The gap between people in their early twenties and the rest of the ages is also wide, with a difference of more than a hundred.



Looking at the people with diabetes versus those without, it is clear that most of those represented in the dataset do not have diabetes. The non-diabetes people are about twice the diabetes patients, but still a high count by the hundreds.

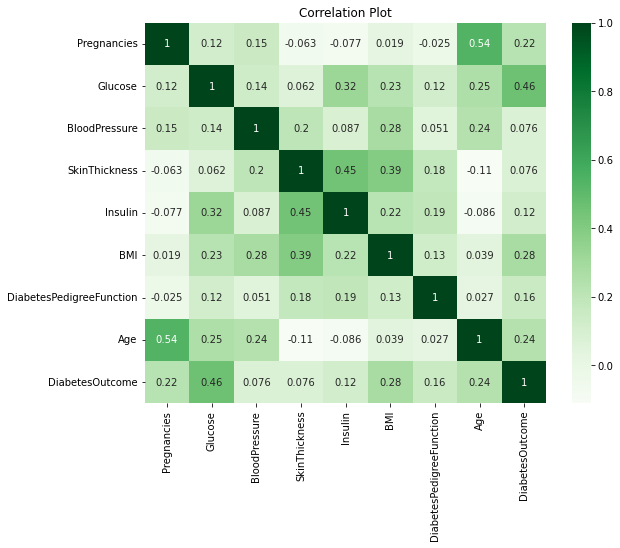


The BMI distribution shows a normal distribution curve. A majority of people have a BMI from 25 to 35. Some outlying cases have BMI above 50, which might indicate overweight and obese cases. There are also cases where the BMI is zero, which can indicate an error in the recording of the information. The dominance in the BMI from 25-35 is still on the overweight side. This could be an interesting indicator of the bad status of the population represented since the accumulation of fat in the body has been associated with type 2 diabetes.

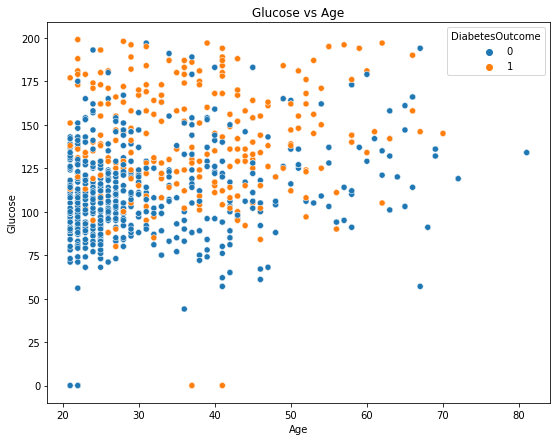


The dominating number of pregnancies is between 1 and 3. The number of pregnancies then decreases as the count increases. It is interesting to see that there are cases with more than 15 pregnancies.

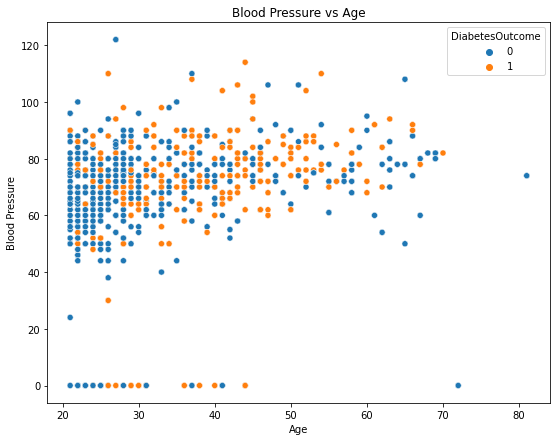
## What knowledge of the disease is represented in the dataset?



A correlation plot of the variables helps outline which variables might be related when it comes to diabetes cases. The plot shows that there is a high correlation between age and pregnancy. However, this information does not show how these relations relate to diabetes. Another highly correlated variable is insulin and skin thickness. The important correlation in the dataset is between glucose and diabetes outcomes. This goes to show that glucose can be a major determining factor with regard to diabetes. Other slightly correlated factors include BMI, age, and pregnancies.



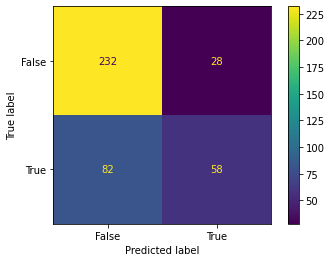
High glucose is also associated with more diabetes cases than not. This can be a good factor to consider when working towards reducing the condition in a population.



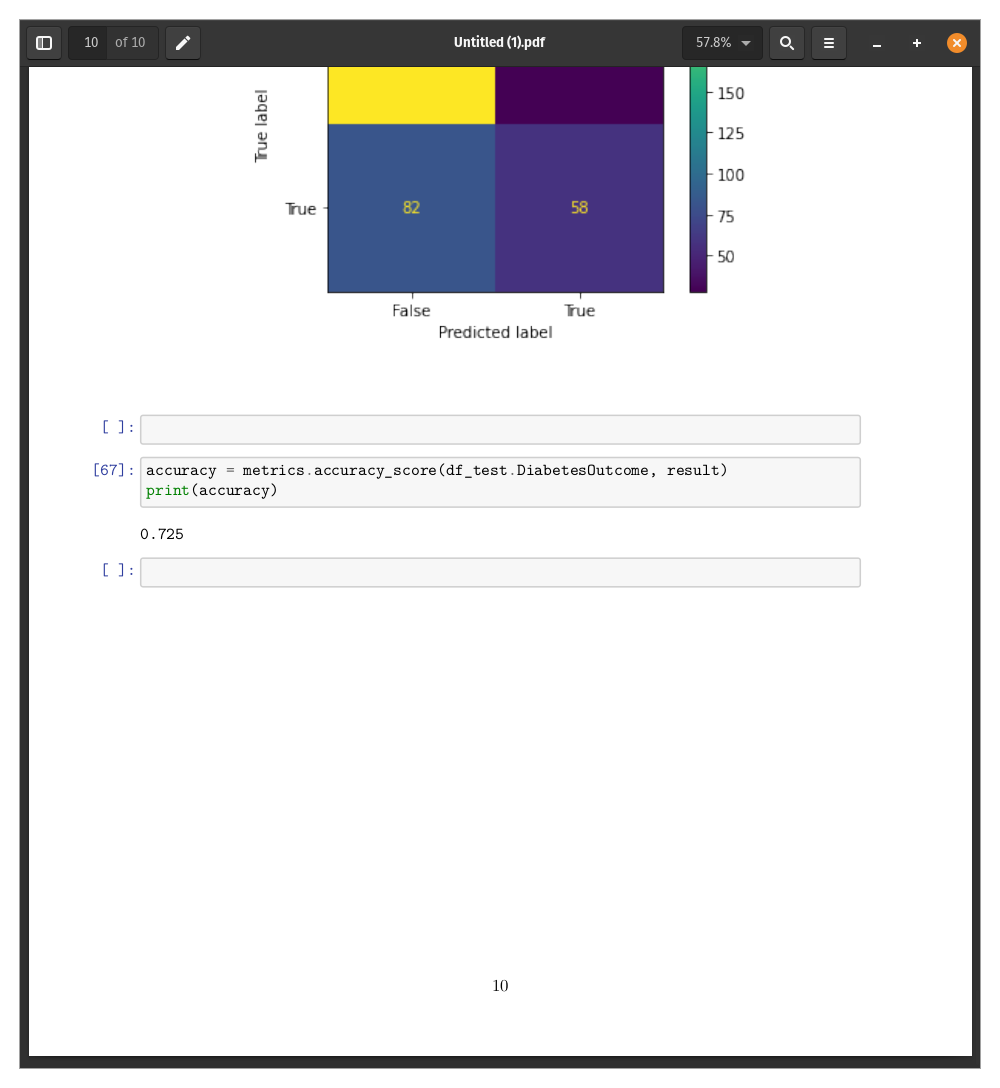
Younger people have a higher range in blood pressure than older people. The non-diabetes cases are also more prevalent in the younger population than the older. The opposite is also true, where diabetes cases are more pronounced in the older generation. Another interesting observation is that there are also higher cases of diabetes in the younger population than it would be comfortable to have. This is indicative that age is no longer a major determining factor when considering diabetes cases.

## Is it possible to predict the diabetes cases?

Predicting the diabetes outcome can provide insight into whether it is possible to determine the disease given the other factors. This can help in further elaborating whether there is a connection between the variables and the diabetes outcome. The approach uses a generalized linear model (GLM) to create the model and predict the outcome. The results are predicted in probabilities prompting the need to pick a threshold value to translate the two outcomes of the diabetes cases.



This confusion matrix shows good prediction results, with the accuracy going to more than 70%.



This is a good prediction model and can be useful to use the factors to predict whether a person is likely to have diabetes or not. Improvements can also be made to increase the prediction rate of the model.

## How do we improve the safety of patients by using the knowledge to prevent diabetes?

It is clear from the previous analyses and plots that diabetes results from some key factors. For instance, insulin appears to be among the highest predictors of the disease. Its occurrence in high levels seems to correlate highly with diabetes. Other factors, such as glucose, can also predict the cases of diabetes. To keep the population safe, there is a need to keep the levels of glucose to levels that will promote health. Factors such as age cannot be controlled, but having a healthy life can result in less risk of getting the disease. The concern about the BMI defaulting to higher levels than considered normal means that there is a need to sensitize the change of lifestyle to avoid overweight problems. Despite the connection between BMI to diabetes being weak, being overweight or obese has other problems that might bring problems to one's life. It is important to know that the case being investigated is type 2 diabetes which can be caused by the said factors.

# Conclusion

The analysis still shows that there is more to diabetes than the highlighted factors. Although the prediction level for the cases is 70%, there is still room for improvement. This is possible by introducing more factors and data points to improve the model and probably discover new knowledge on the disease. However, the analysis has indicated strong connections with a few factors, such as age, insulin, glucose and BMI. Input from the medical community can also help improve the results by providing insights on how to navigate through such results. Diabetes cases are still increasing as time goes on, which makes it a significant problem across the world.