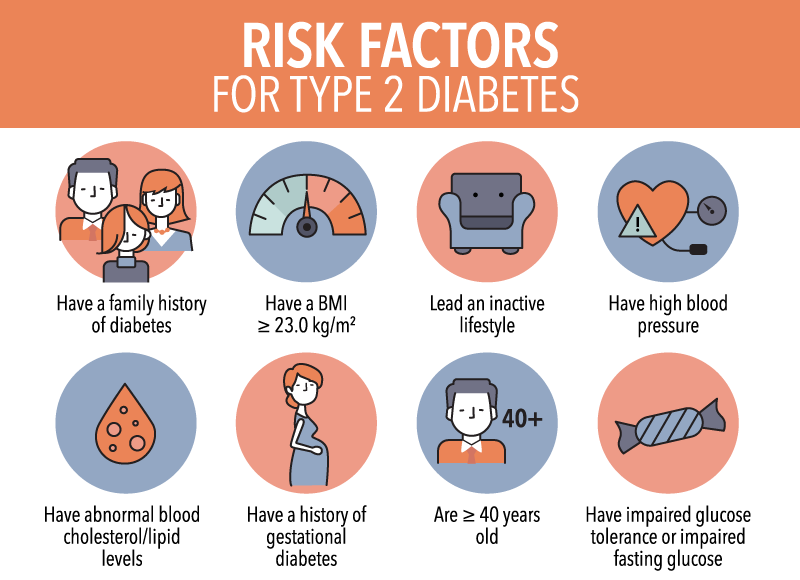
**Diabetes Health Campaign**

Diabetes is a costly, chronic condition in which the body does not produce or properly use insulin. An estimated 29.1 million people in the United States—9 percent of the population—have some form of diabetes. Of those, 25 percent do not yet know that they have the condition. In addition, an estimated 86 million U.S. adults have pre-diabetes—an elevated blood sugar level that is not high enough to be classified as diabetes, but which greatly raises their risk of developing type 2 diabetes and its complications. 1

Insulin is necessary to covert sugar, starches and other foods into energy for living. Major types of diabetes include:

* **Type 1 diabetes** is usually diagnosed in children and young adults and was previously known as juvenile diabetes. Insulin shots are the only way to keep blood glucose levels down in Type 1 diabetes sufferers. [Not usually preventable, genetic determinants.] 1
* **Type 2 diabetes**, which is the most common form of diabetes, causes the body to not produce enough insulin or causes the cells to ignore the insulin. [Usually develops over time as a person’s blood cells gradually become resistant to insulin. As the cells become more resistant, the pancreas must produce more insulin to overcome the resistance. If the condition is left untreated, the pancreas will be unable to produce sufficient insulin to allow blood cells to access sugars and other nutrients.] 1



**Treatment**

Diabetes is treatable.  Although there is no "cure" various treatments allow most diabetics to live relatively stable, normal lives.  Early screening, diagnosis and treatment also prevent or reduce the more serious consequences of the disease — emergency room visits, hospitalization, loss of sight, loss of limbs. Once diagnosed, diabetes requires self-management, including testing and monitoring blood glucose levels. Because treatment requires patient education, special equipment and supplies, it can become costly especially when it is not covered by health insurance.

**Costs of Diabetes**

Diabetes is the seventh leading cause of death in the U.S. Its complications, including heart disease, stroke, amputations, blindness and kidney disease, are both serious and expensive. The cost to treat an individual with diabetes is more than 200 percent higher than the cost to treat a patient without diabetes. According to the 2015 study by the Health Care Cost Institute, medical costs and productivity loss attributable to diabetes were estimated to be $245 billion in 2012. By comparison, the estimated total cost of diabetes in 2007 was $174 billion.

By 2012 the total cost of diabetes increased to $245 billion, meaning that the disease’s toll on the economy has increased by more than 40 percent since 2007, according to a [report](http://care.diabetesjournals.org/content/suppl/2013/03/05/dc12-2625.DC1/DC122625SupplementaryData.pdf) from the American Diabetes Association.

Prevalence and Statistics:

The National Diabetes Statistics Report, a periodic publication of the Centers for Disease Control and Prevention (CDC), provides information on the prevalence and incidence of diabetes and prediabetes, risk factors for complications, acute and long-term complications, deaths, and costs. These data can help focus efforts to prevent and control diabetes across the United States. **This document is an update of the 2017 National Diabetes Statistics Report and is attached on Blackboard. Please read through the first 12 pages and locate the key facts and statistics related to 1). Prevalence of Diabetes (Diagnosed and Undiagnosed), 2) Incidence of Diagnosed Diabetes (Newly Diagnosed Diabetes), 3) Risk Factors for Diabetes-Related Complications, 4) Preventing Diabetes-Related Complications.**

* Take note that the rate of undiagnosed diabetes increases from 2.8% (2013-2016) to 7.8% (2018). This indicates that predicting diabetes in populations is very important and could be aided by the use of AI and machine learning.

Resources:

1. <https://www.ncsl.org/research/health/diabetes-overview-page.aspx>
2. <https://www.gov.sg/article/can-you-develop-diabetes>
3. <https://www.diabetesresearch.org/file/national-diabetes-statistics-report-2020.pdf>

# Health Analysis of Diabetes among Pima Indians

## Objective

In this notebook, we will try to build a machine learning model to accurately predict whether the patients in the dataset have diabetes. We will then use these models to better understand what contributes to our population of interest developing diabetes.

Your domain expertise is needed to help understand how to combat diabetes among Pima Indians. Ultimately, your goal is to learn about the population from historical data and apply your knowledge in the context of a public health plan.

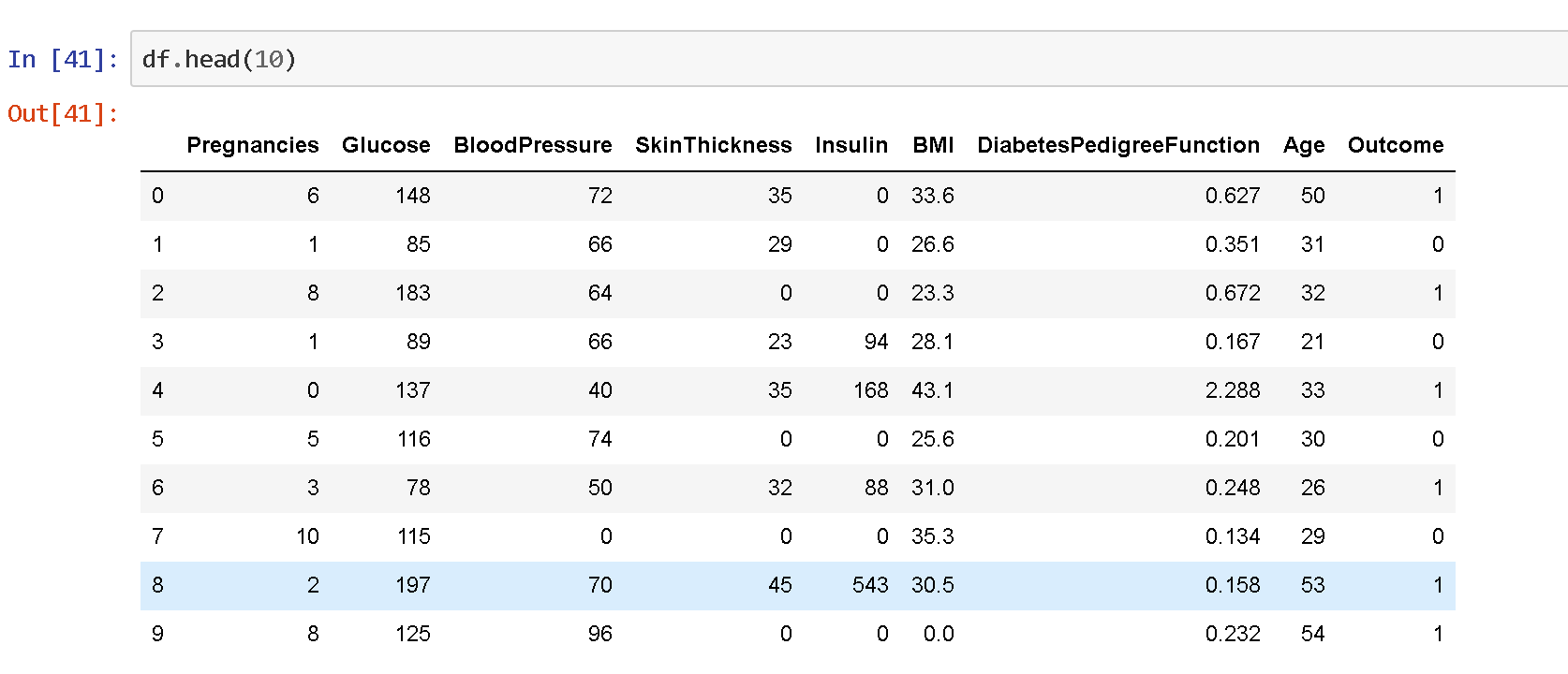
## Context

As we consider the relationships between variables, keep in mind how the information could be applied to a public health campaign. For instance, which of the explanatory variables might cause diabetes and which might be an effect of diabetes?

# **Exploratory Data Analysis**

It's important to explore the data set before diving in. It's common that data is imperfect, and you are likely to find issues that need to be addressed before you can move forward with your analysis.

We'll start by looking at a sample of the data and some attributes of each feature in the data set.



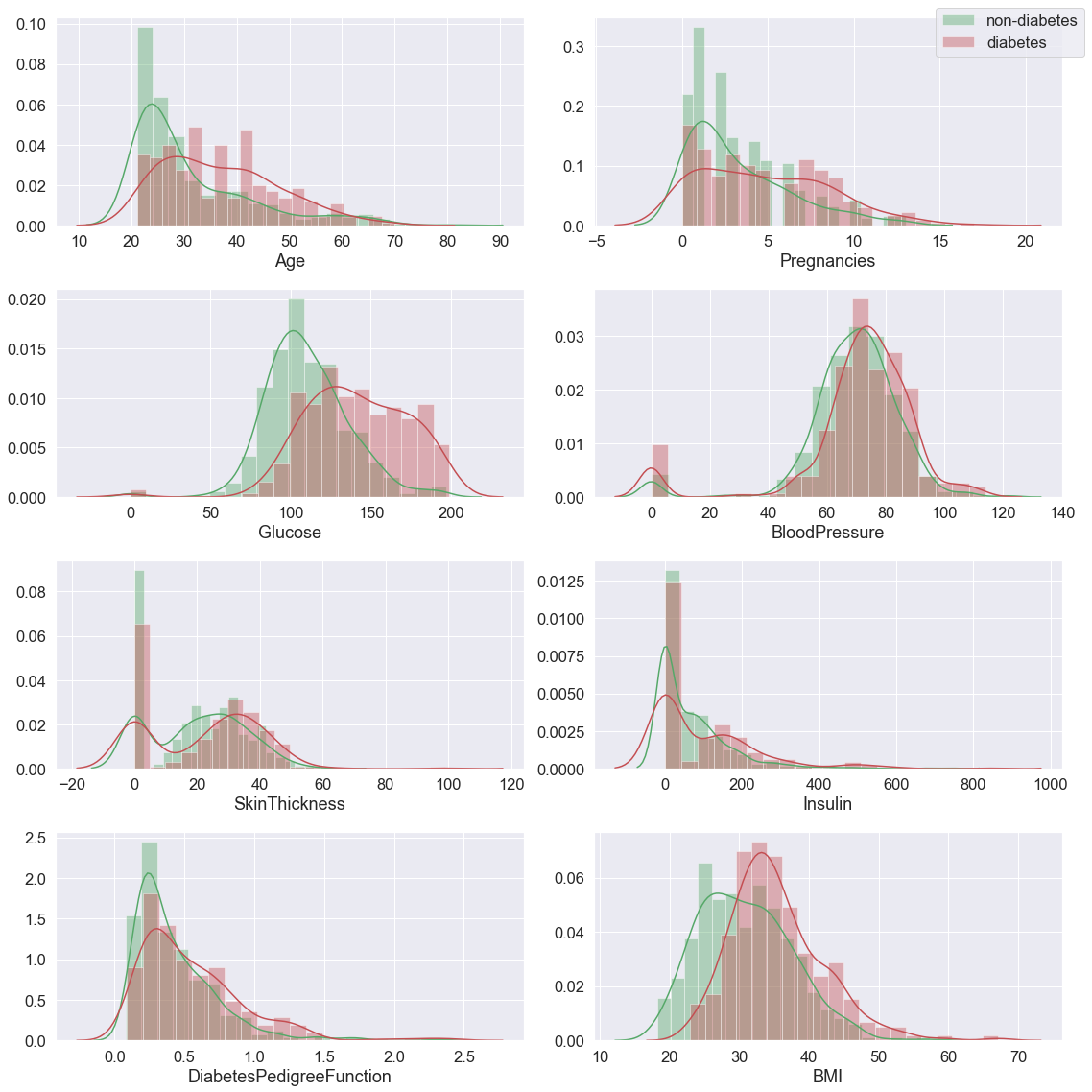
Do you see any data points that seem erroneous for BMI?\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Why is it important to identify and remove this erroneous data from the analysis? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## Data Visualization

### Distributions

Now we'll explore the distribution of each variable, looking for big differences between those who had diabetes and those who did not have diabetes.

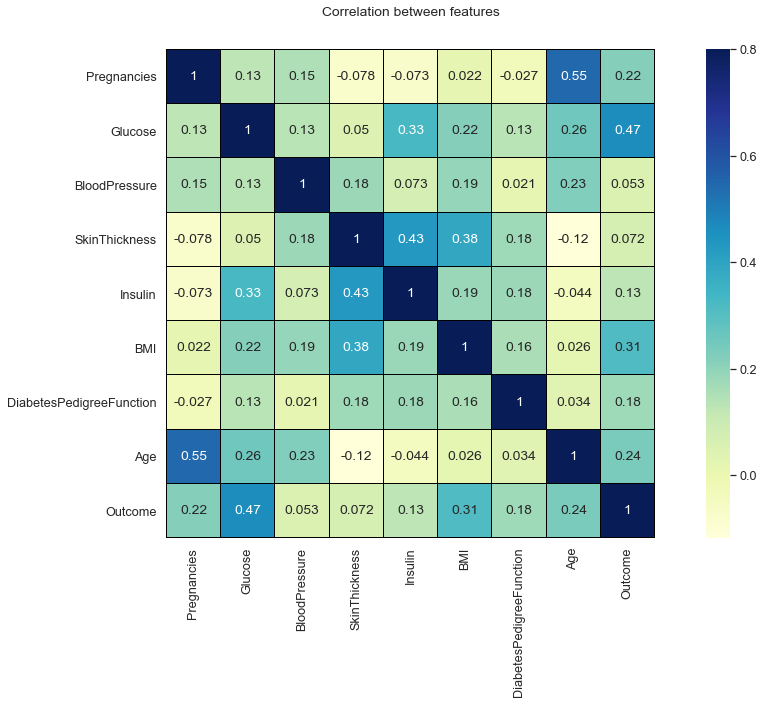


Which of these variables seem to have high separation between classes? That is - do you see any big differences between red (diabetes) and green (no diabetes) distributions for any of these variables? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## Correlation between features

Variables within a dataset can be related for lots of reasons. It can be useful in data analysis and modeling to better understand the relationships between variables. The statistical relationship between two variables is referred to as their correlation.

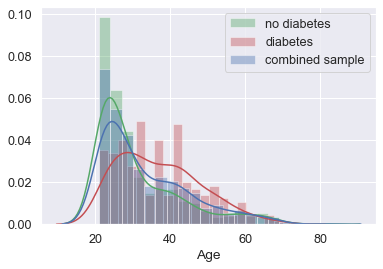
A correlation could be positive, meaning both variables move in the same direction, or negative, meaning that when one variable’s value increases, the other variables’ values decrease. Correlation can also be neural or zero, meaning that the variables are unrelated.



Which variables are highly correlated? Does this make sense to you? Discuss.\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

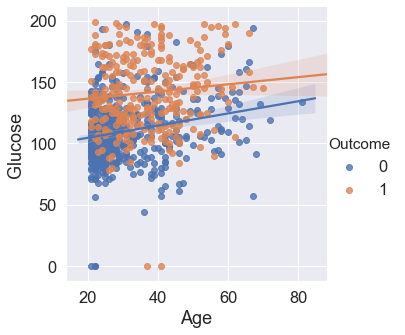
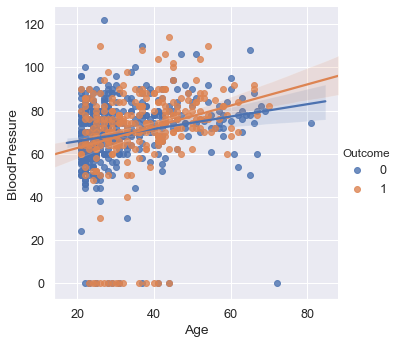
### Taking a further look:

Some of these variables look like they may be correlated. While correlation does not equal causation, plots like this can help us understand how our features interact.



Does this population seem to develop diabetes later in life, or in earlier years?\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Do you think a 40 year old seems much more likely to have diabetes than a 50 year old in this population?\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_



Do you think an 80 year old is more or less likely than a 45 year old to have high blood pressure? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Given that a 60 year old in this population has high blood pressure, would you guess that they do or do not have diabetes?\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

What can you tell about the relationship between glucose and diabetes?\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Predictive Modeling

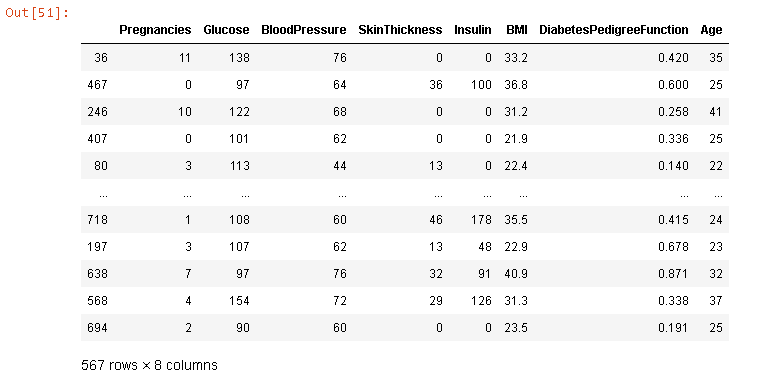
### Split into testing and training data

To make sure our model will perform well on new observations, we will hold out a few observations, hiding them from the model. Once we think we've built a good model, we will reintroduce these observations and ask our model to make predictions about them as a test to see how well it performs.

**Vocabulary:**

* **Training Data**: data we use to teach our model
* **Testing Data**: data we hold out until our model is built to test how it performs

Now we will split our predictors and our outcome variables into training and testing observations and take a look at our training data!



## Logistic Regression

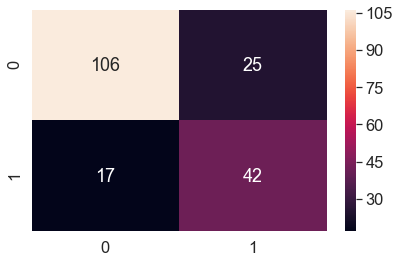
Logistic regression is the appropriate regression analysis to conduct when the dependent variable is binary. Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

## Visualizing Logistic Regression Performance with a Confusion Matrix

Now let's use a "Confusion Matrix" to visualize how our classifier is doing at classifying our test set.

A confusion matrix is a table that visualizes the performance of a classification algorithm. Each row corresponds to the true class label, here either diabetes or no-diabetes, while each column corresponds to the predicted class. The entry in each cell tells us the number of records of an actual class that were predicted as a given class. The table makes it easy to see how the algorithm is making mistakes. If it was perfect then we'd see zeros everywhere except on the diagonal (the predicted class was always the actual class).

Accuracy 77.89473684210526

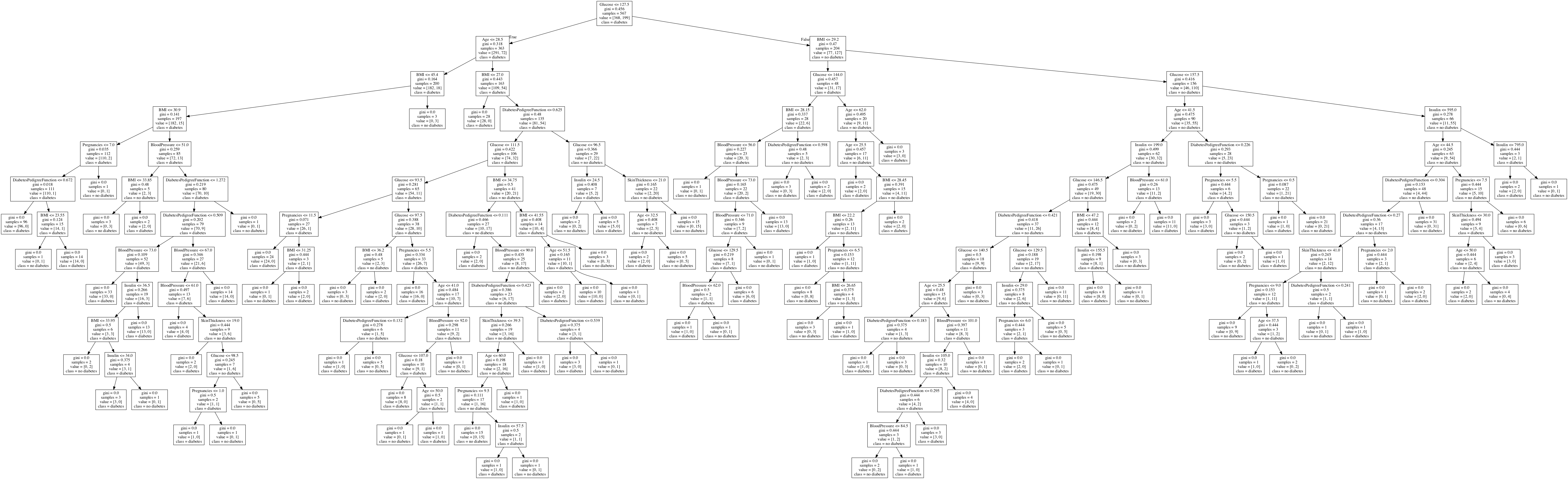


## Decision Tree

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy), each representing values for the attribute tested. Leaf node (e.g., Hours Played) represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

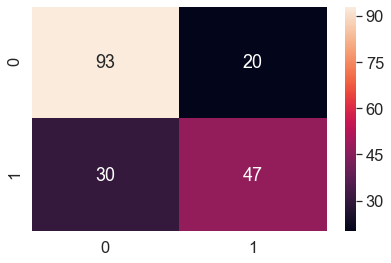
The accuracy of your decision tree on testing data is: 73.68421052631578

## Visualizing the Decision Tree Model



## Visualizing Decision Tree Performance with a Confusion Matrix

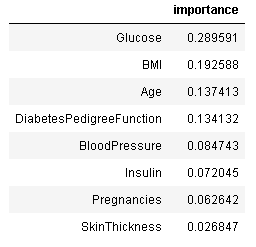
Accuracy 73.68421052631578



## Feature Importance

We can understand how our model makes decisions and which variables are important predictors. To take a closer look at the relative importance of each variable as a predictor of diabetes, we can investigate the variable importance from sklearn's feature importances.

To learn more about feature importance from the gini index, check out: <https://towardsdatascience.com/explaining-feature-importance-by-example-of-a-random-forest-d9166011959e>



If you were going to create educational material for this population, which of the explanatory variables might you discuss? How might you decide that?\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Discuss the limitations of this information without context.\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_