
Supervised Learning Project

Christopher Martinez Demmans

Goals of Project

Use diabetes dataset to create a machine learning model capable of taking a set of information about a patient and then using it to predict whether or not they have diabetes with a high accuracy.

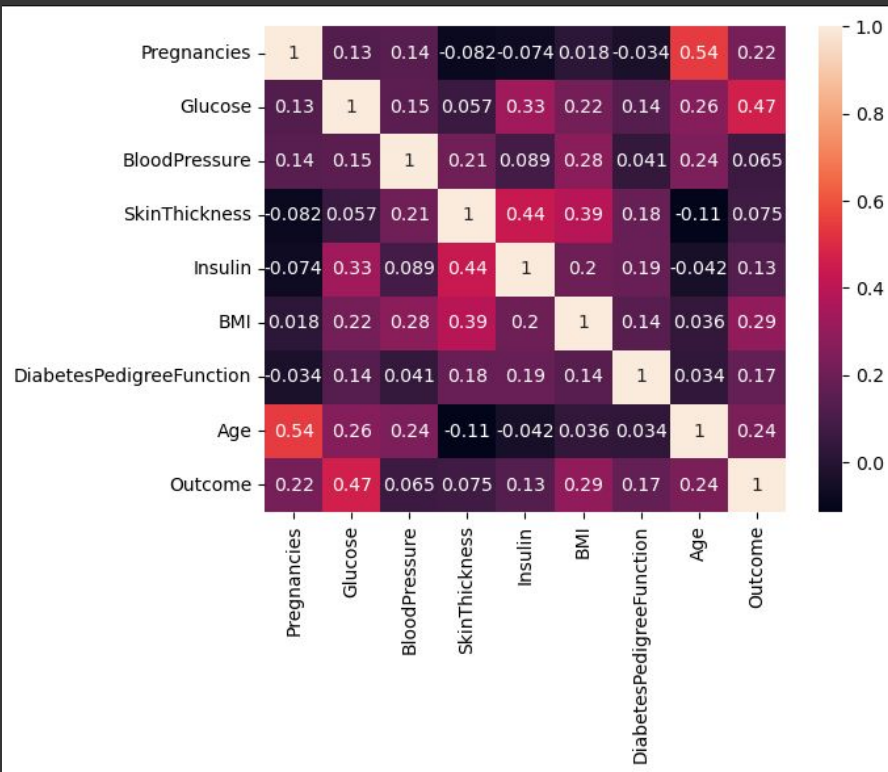


The Data

Our data set includes key indicators as to whether or not someone as diabetes.

- Glucose
- BMI
- Insulin
- Age
- Blood Pressure

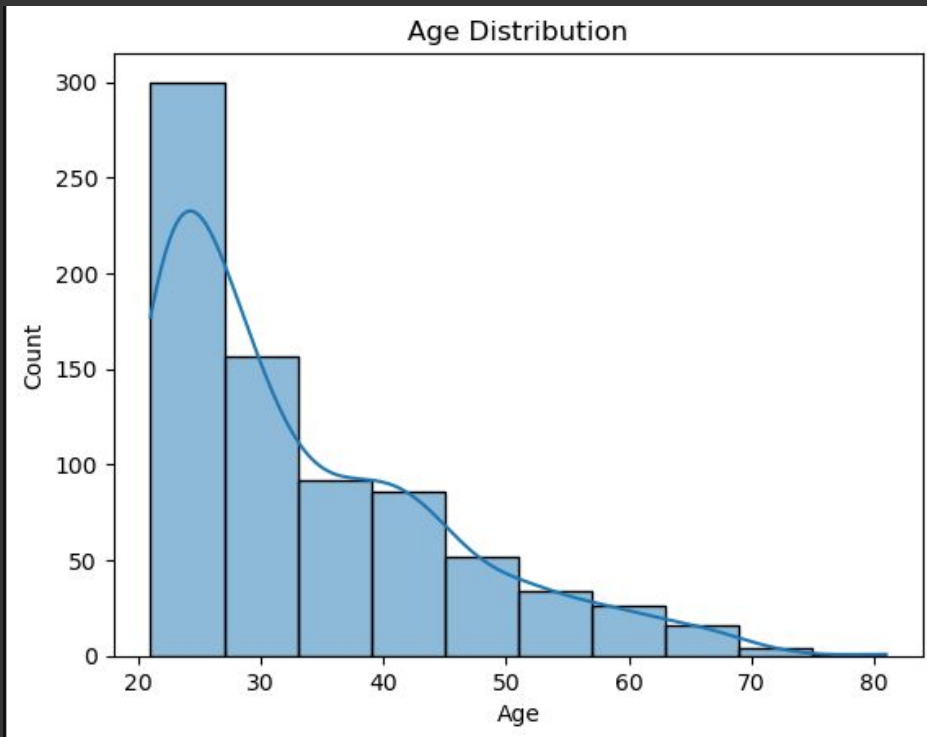
EDA:



Observations:

Glucose levels seem to have the strongest correlation with outcome, followed by BMI.

Who do these models apply to?



Age Range:

Our data mostly comes from people in the 20-30's age range.

Our model will be best for predicting diabetes in that group of people.

—

Machine Learning Models: Random Forest and SVM.



Random Forest

How good are the models we created?

→ **Accuracy :74%**

The higher the better.

→ **Recall: 69%**

More likely to capture all true positive cases.



SVM

How good are the models we created?

→ **Accuracy :76%**

The higher the better.

→ **Recall: 56%**

Model is more reserved when it comes to deciding whether someone has diabetes than the random forest model.

Which Model Should We Use?

The random forest model.

It is better to assume someone has diabetes and get them tested than to not know.

Has similar accuracy to SVM model.

Thank you for reading!
