DIABETES DIAGNOSTIC MEASURES

By Aman Negassi

If anyone wanted to get an idea how significant Diabetes is to our health and a community's, it was estimated to be the 7th leading cause of death in 2016 according to the World Health Organization. Among adults over the age of 18, it has almost doubled going from 4.7% in 1980 to 8.5% in 2014 on a global scale. It's prevalence has been increasing rapidly among low-and middle-income countries than in high-income countries although there was a 5% increase in premature mortality attributed to diabetes in high-income countries from 2010-2016. The chronic disease is a major cause of blindness, kidney failure, heart attacks, stroke, and lower limb amputation. In a step-by-step fashion, most of the food you eat is broken down into glucose (sugar) and released into your bloodstream. When your blood sugar goes up, it signals the pancreas to release insulin. The insulin allows the blood sugar into your body's cells for use as energy. If you have diabetes, your body either does not make enough insulin or cannot utilize the insulin as well as it should. Either way, too much blood sugar stays in your bloodstream and can lead to serious complications such as heart disease, vision loss, and kidney disease. A lot of people may be familiar with Type 1 and Type 2 but are not familiar with gestational diabetes. Gestational diabetes develops in pregnant women who have never had diabetes which puts the baby at risk for health problems.While it goes away after the baby is born, the mother and the baby is likelihood of contracting Type 2 diabetes increases as well as the baby likely to becoming obese in their adoloscence. It's estimated in the United States that 88 million adults, more than 1 in 3 have prediabetes and worse, more than 84% do not even realize they have it which makes the data analysis all the more important.

DATA

This is the dataset looking at female patients of Pima Indian heritage with the column variables being the diagnostic measures. Important to note, Diabetes Pedigree Function means the likelihood of diabetes based on family history whether hereditary or not and the possibilities.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

DATA CLEANING/WRANGLING

As noted above, the values for Outcome, whether a patient has diabetes or not is 1 and 0. That's why a replace method will be used where 1 is YES and 0 is NO.

```
diabetes['Outcome'] = diabetes['Outcome'].replace(1, 'YES')
diabetes['Outcome'] = diabetes['Outcome'].replace(0, 'NO')
```

Before

```
diabetes['Outcome'].unique()
```

```
array([1, 0])
```

After

```
diabetes['Outcome'].unique()
```

```
array(['YES', 'NO'], dtype=object)
```

They become int (64) to object data types.

EXPLORATORY DATA ANALYSIS

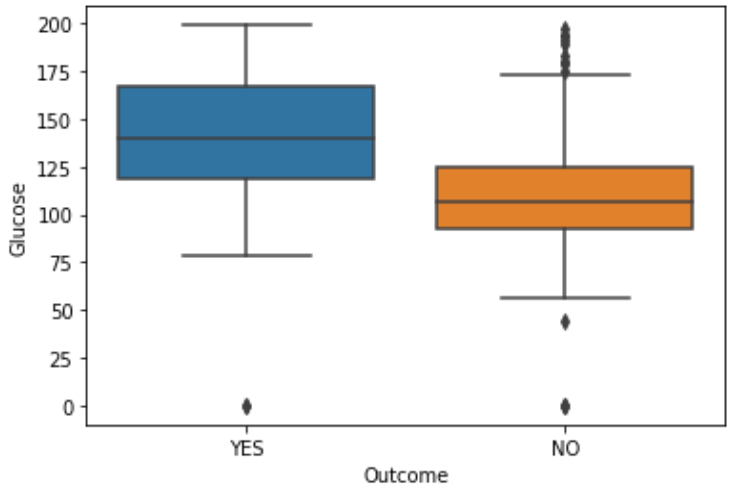
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000

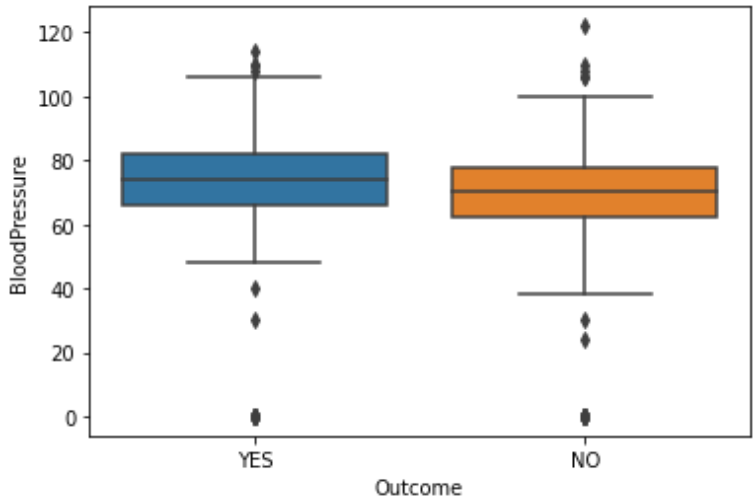
	min	max
Pregnancies	0	17
Glucose	0	199
BloodPressure	0	122
SkinThickness	0	99
Insulin	0	846
BMI	0	67.1
DiabetesPedigreeFunction	0.078	2.42
Age	21	81
Outcome	NO	YES

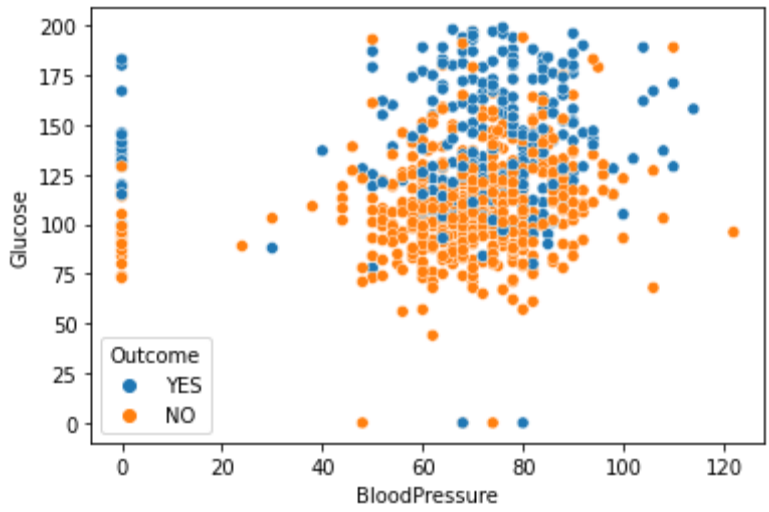
```
diabetes.shape
```

```
(768, 9)
```

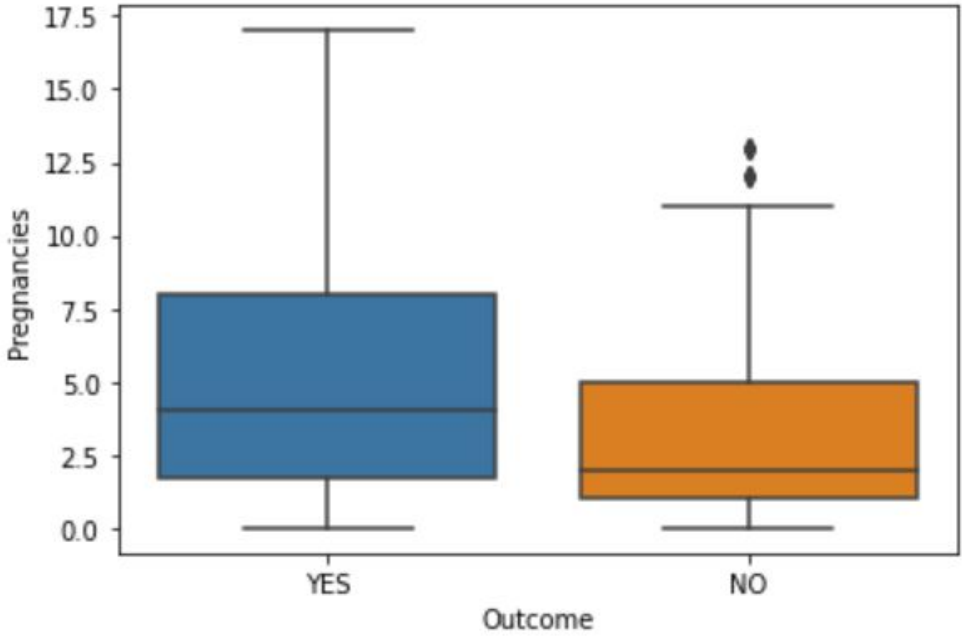
There are 768 observations meaning 768 patients assessed by the following variables listed above.


Glucose
200
175
150
125
100
75
50
25
0
YES
NO
Outcome

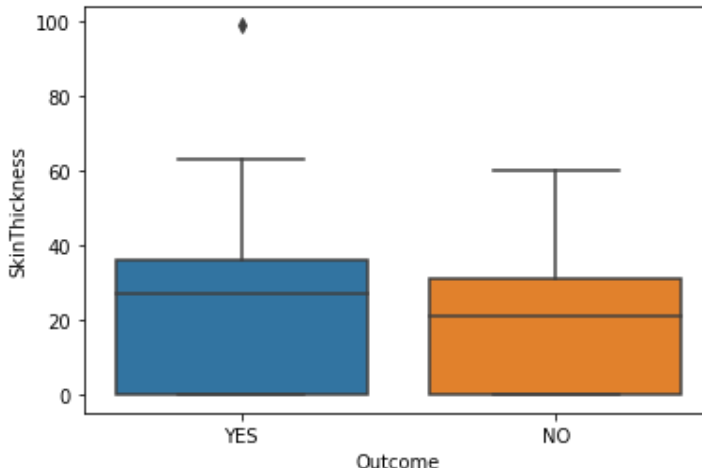

BloodPressure
120
100
80
60
40
20
0
YES
NO
Outcome

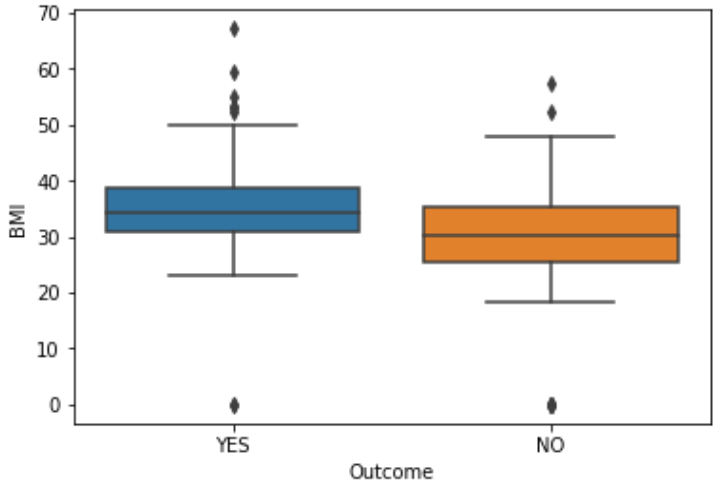

Glucose
200
175
150
125
100
75
50
25
0
Outcome
YES
NO
0
20
40
60
80
100
120
BloodPressure

Patients who contracted diabetes on average had higher glucose levels where it seems blood pressure levels on average were no different between those who have diabetes versus those who do not. Despite that, there is a direct relationship between blood pressure and glucose levels.



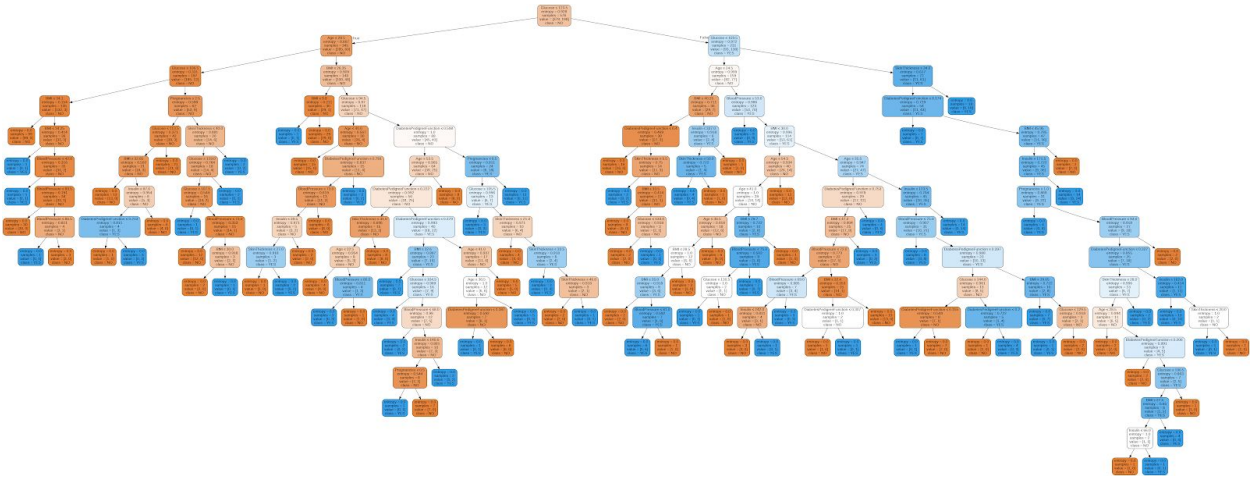
Patients who contracted diabetes on average had more pregnancies which is not surprising considering that gestational diabetes is more prevalent in diabetes.





The patients who contracted diabetes had a higher BMI and thicker skin than their counterparts on average although it's not very noticeable. It is quite telling because it shows diabetes goes beyond being physically active and eating well.

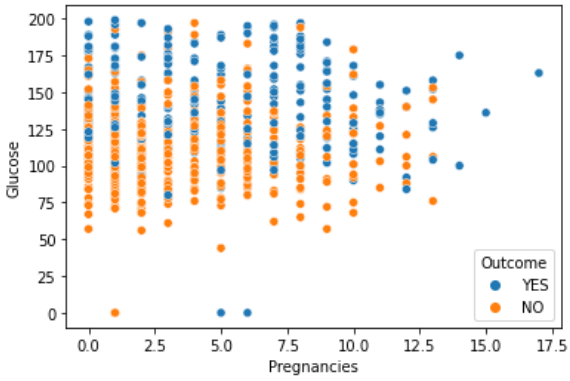
MODELS

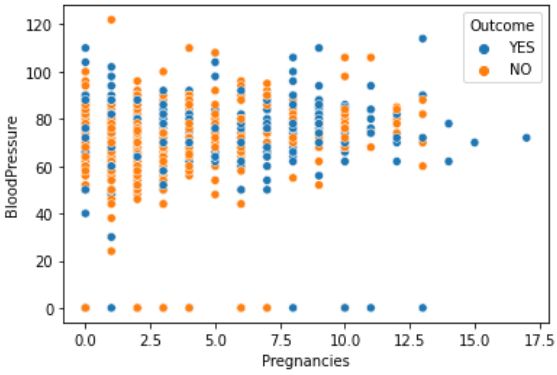


```
Gini impurity  model - max depth 3
Accuracy: 0.6979166666666666
Balanced accuracy: 0.692271662763466
Precision score 0.573170731707317
Recall score 0.7131147540983607
```

These are the findings from the Decision Tree model.

INSIGHTS





This was an interesting discovery as we would expect a direct relationship considering stress of a pregnancy and giving birth to a human being. Gestational diabetes may be common in pregnancies as noted before but it is not always going to lead to higher glucose and blood

pressure levels. It's because of this that the current response is worth being re-evaluated. It has been noted by the American Diabetes Association that Diabetes may be underreported as a cause of death. Only 35 to 40% of people with diabetes who died had diabetes listed anywhere on the death certificate and about 10 to 15% had it listed as the underlying cause of death.

FUTURE IMPROVEMENTS AND RECOMMENDATIONS

```
Xlr, Xtestlr, ylr, ytestlr = train_test_split(diabetes[['Insulin','BloodPressure','Glucose', 'BMI']].values,
                                              (diabetes.Outcome == 'YES').values,random_state=5)
```

```
clf.fit(Xlr, ylr)
print(accuracy_score(clf.predict(Xtestlr), ytestlr))
```

```
0.796875
```

I insist on relying on Insulin, Blood Pressure, Glucose, and BMI as variables for modeling such as Logistic Regression. The accuracy score for determining whether a patient has diabetes or not was higher than all of the other models not to mention that it is also recommended to use 4 for the # of variables instead of any other given amount. I will say it is possible that could change if provided a larger sample size or a different sample size. It might be worth exploring other diagnostic measures as a means of improving the score. Those variables give a strong indicator specifically as I have used different variables but with the same amount for the train/test split.

CREDITS

A special thanks to my mentor Nate Sutton for helping me plan the course that got me to this point. You have helped me expand my horizon and I cannot thank you enough. I also owe a special thanks Kenneth Gil-Pasquel for your constant support on the technical side. You have challenged me while helping me learn that has pushed me to become more self-sufficient which has changed my approach as a Data Scientist. An honorable mention to Blaine Bateman for his help leading me to this Capstone when the other Capstone seemed so ambiguous.

REFERENCES

https://www.cdc.gov/diabetes/basics/diabetes.html#:~:text=Diabetes%20is%20the%20seventh%20leading,diabetes%20has%20more%20than%20doubled.

https://www.diabetes.org/resources/statistics/statistics-about-diabetes

https://www.kaggle.com/uciml/pima-indians-diabetes-database

https://www.who.int/news-room/fact-sheets/detail/diabetes