Data Preprocessing

- Detected no nan values, however 0 at glucose, insulin etc. implies missing data.
- Filled zeroes with median of that column grouped by the outcome. E.g.: Glucose zeroes filled with median of glucose medians of diabetic patients if an individual is diabetic, median of healthy people if not.
- This was applied to “glucose”, “blood pressure”, “skin thickness”, “insulin” and “BMI” columns.
- Zeroes at “pregnancies” do not imply missing values, because it is possible and sensible to not have any pregnancies.

Feature Engineering

- Created a new feature “is Woman” using pregnancies and age where if 1 implies woman, 0 is not definitive about the sex.
- Detected similar pregnancy groups and grouped them together as a categorical value, 1 pregnancy, 1-6 pregnancies and 7-9 pregnancies. Increasing pregnancy implies a higher chance of diabetes.
- Similar grouping with age and BMI.

Model Building

- Using visualizations and maps from data analysis, first tried to use all features, however since my dataset is extremely small, it caused overfitting.
- Then observed feature importance using random forest and decided to use the first 4 features in my models. And when I this my models improved, logistic regression gave 0.75 accuracy previously, now 0.82; random forest gave 0.82 accuracy previously, now 0.89. It is clear that random forests perform better than logistic regression in traditional models.
- Then moved onto the neural networks, I tried two networks one with dropout and one without. Tried to keep my model shallow as possible because, again, this is a very small set, and it is too easy to overfit.
- However, this didn’t work either. I get similar accuracies as model without dropout. So, I thought maybe my model was too shallow, made it deeper and added more dropout layers.
- After getting 0.89 again, I thought maybe it is more useful to improve random forest rather than neural networks, NN might be a bit overkill maybe.

- This got 0.86 accuracy in validation set, 0.89 in test set. Was not very helpful either.
- Finally in a for loop tried different iterations and depths. Got the best accuracy in est:12, depth:3. However, they were all between 0.85-0.89, so this also was not a satisfactory improvement.
- Finally tried with 5 features instead of 4 to see if this will change accuracy. Got 0.83, 0.87 so not much improvement.

Result

- It is easy to see that logistic regression should not be used. However, I couldn't improve my model in a way that neither helps deciding between RF or NN, nor reaching the accuracy of at least 0.90.
- From that I realized that I should improve feature engineering.
- Even though I feel like my model failed here are some conclusions about data:
    - Insulin, glucose, skin thickness and diabetes pedigree function correlates largely with diabetes, with bigger than 0.1 or close.
    - Blood pressure, age, BMI, pregnancies relatively less, and my new feature isWoman implies nothing.