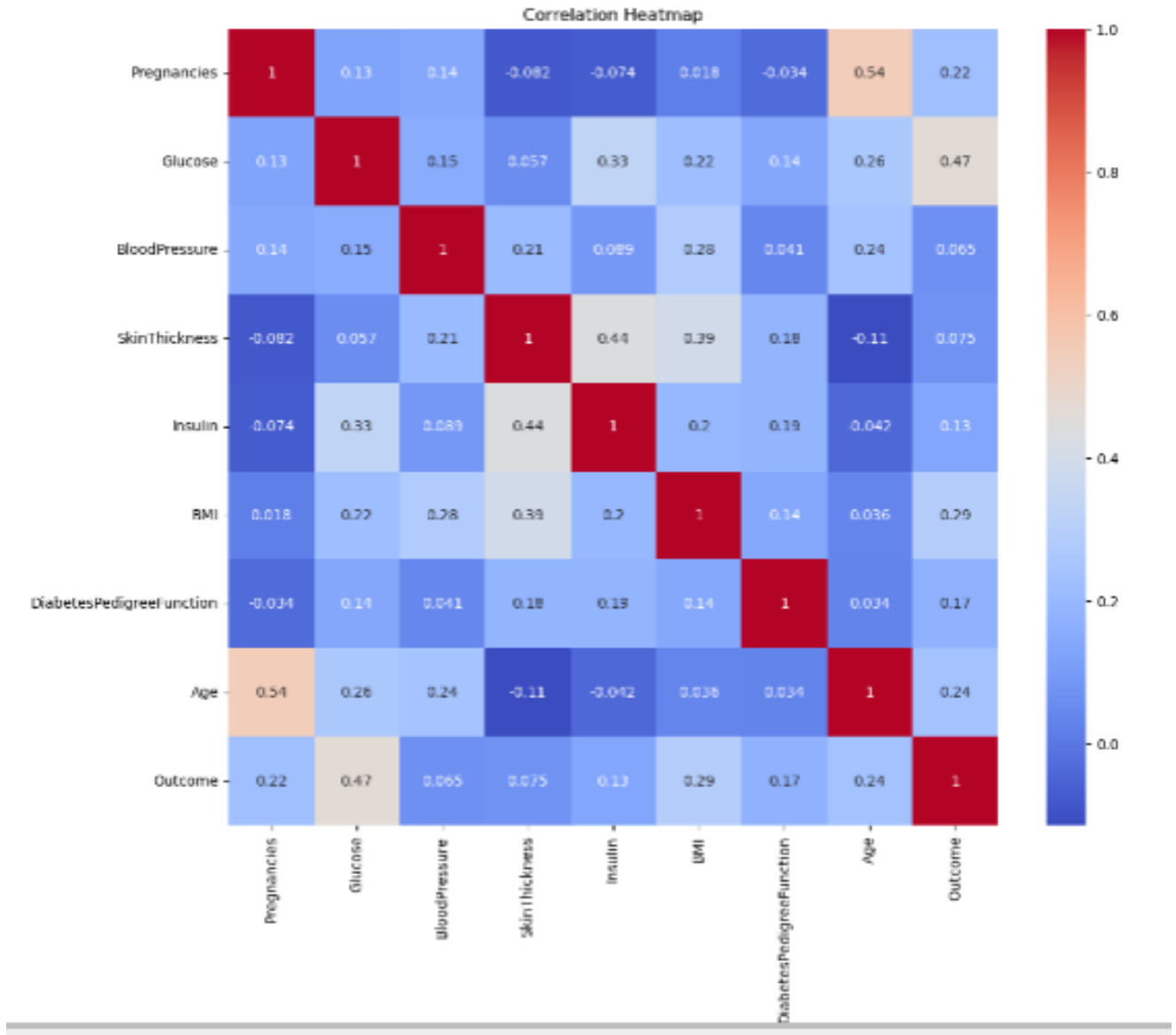
Project-Supervised Learning

TES BEYENE

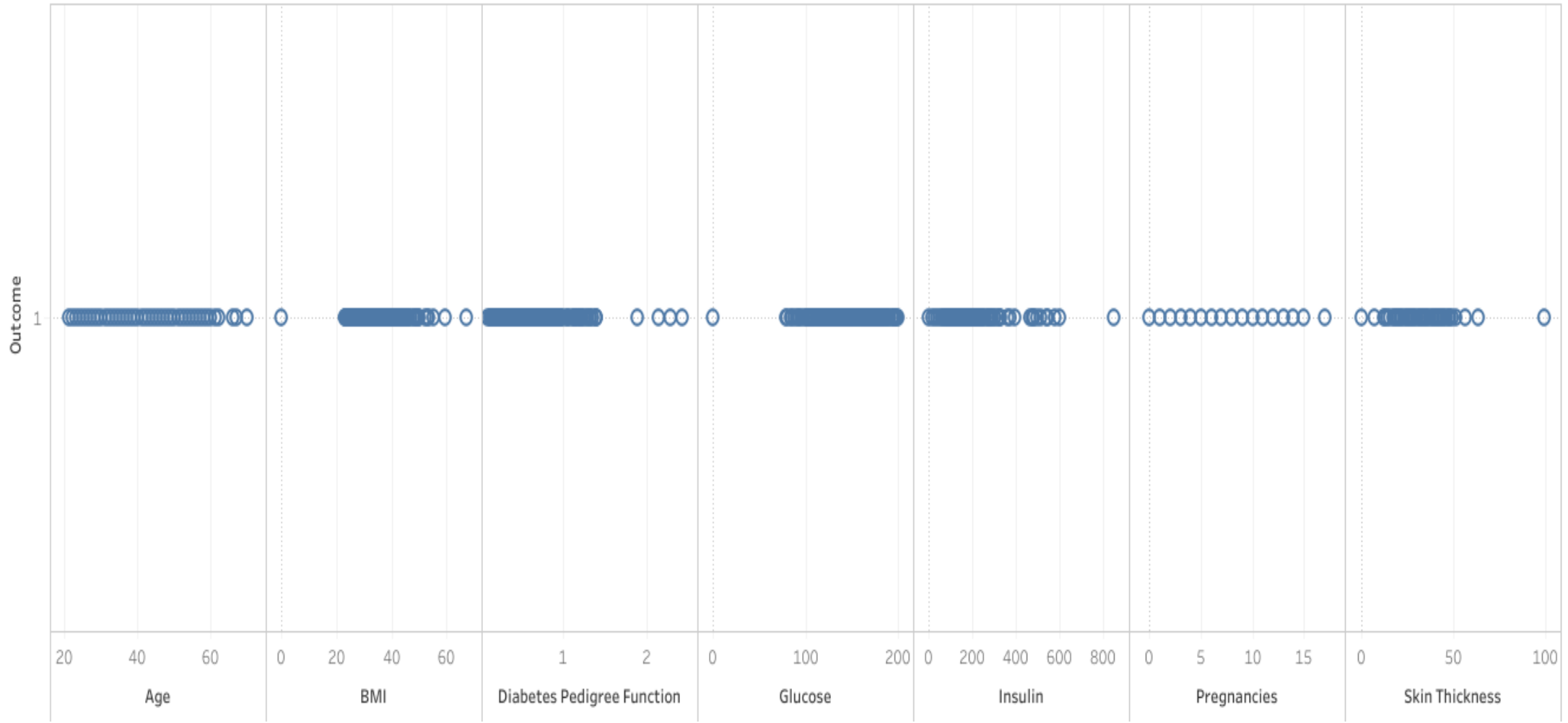
Project Goal

- Applying supervised learning techniques to a Diabetes data set and use data visualization tools to communicate the insights gained from the analysis.

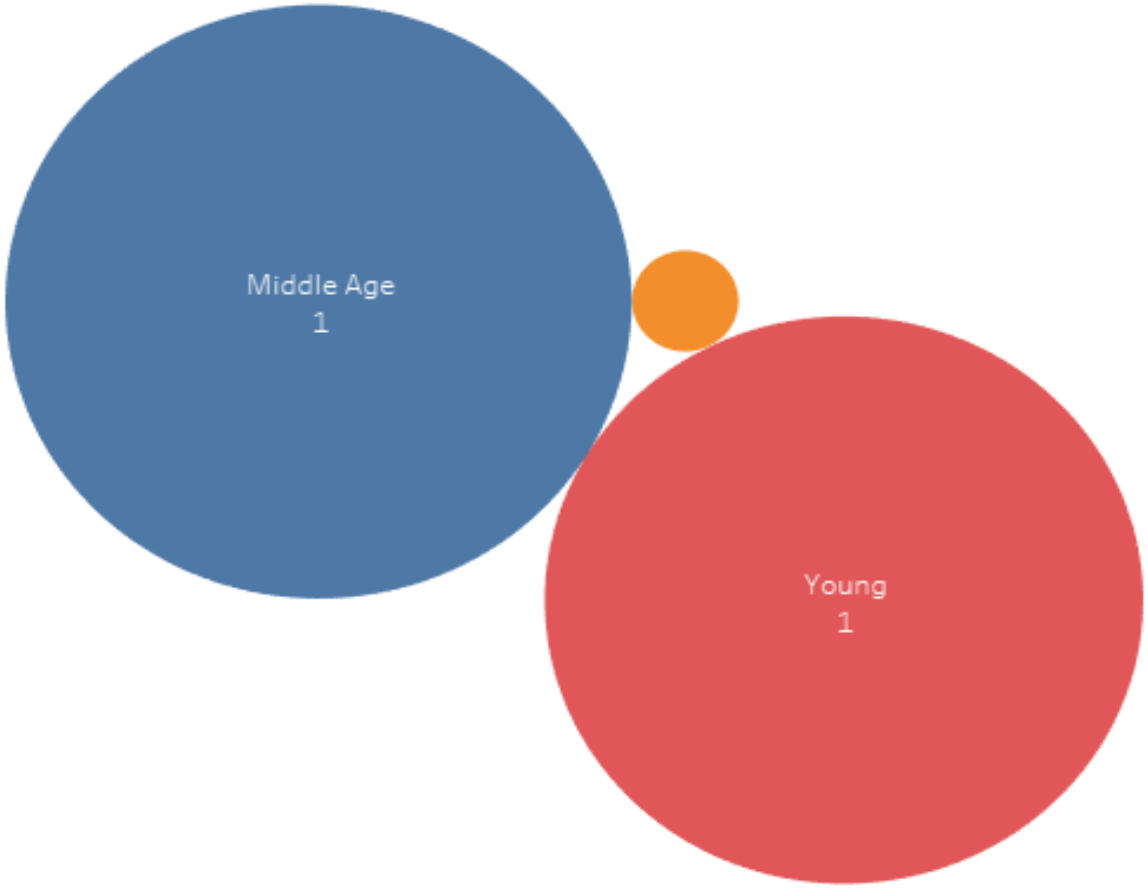
EDA & Visualization



Predictors Vs Diabtes Outcome



Age Group with Diabetes



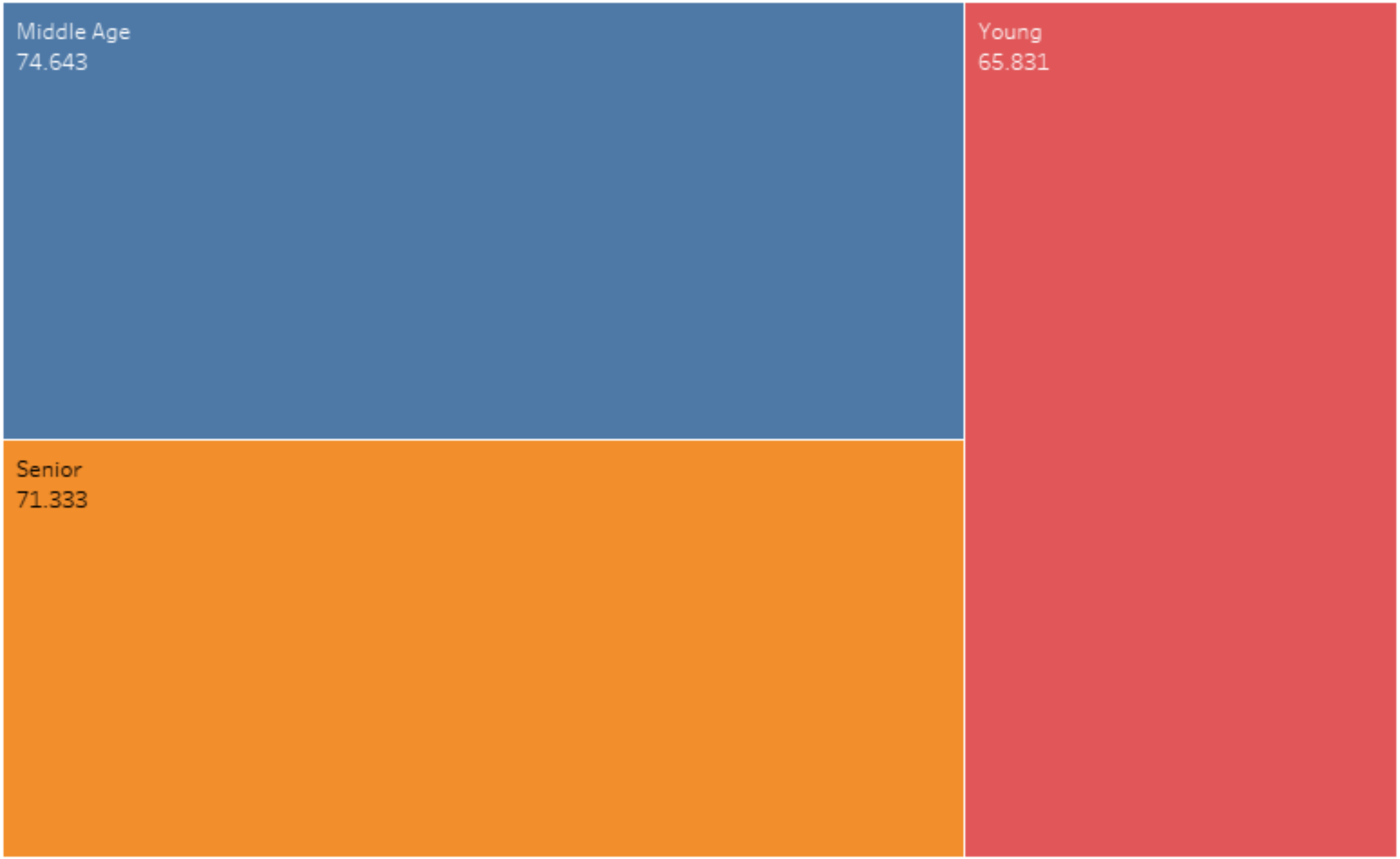
Outcome

1 1

Age Group

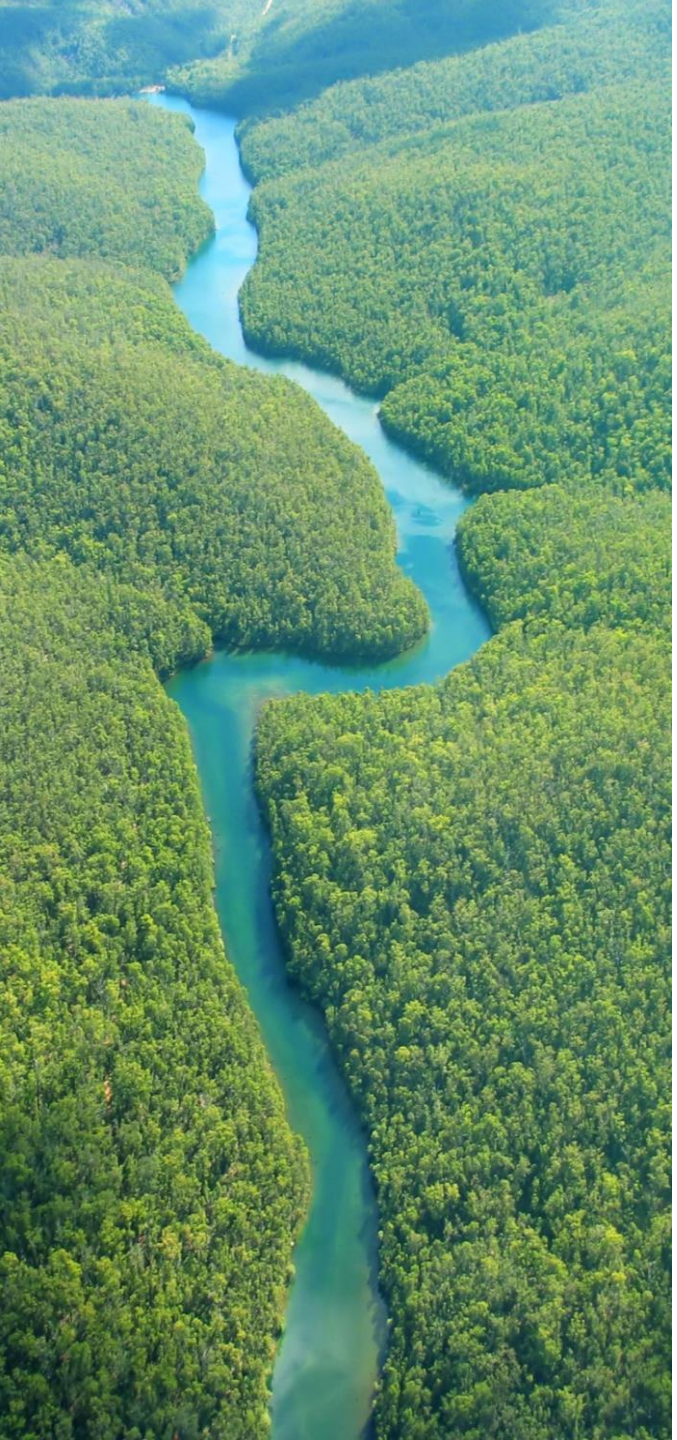
- Middle Age
- Senior
- Young

Average Blood Pressure by Age group



Age Group

- Middle Age
- Senior
- Young



Findings

- The predictors Pregnancies, Insulin, DiabetesPedigreeFunction and Age have skewed distribution.
- The predictors BMI, Blood Pressure & Glucose have some kind of outlier.
- Because of **the non-complex nature** of the dataset Logistic regression unexpectedly performed better than Random Forest model.
- The model Generally performed okay and with model tunning it can perform better.
- Looking the heatmap the correlations that we can consider are the correlations between outcome and Glucose and also the correlation between Age & pregnancies.
- Since the dataset was not large and relatively not complex the logistics regression unexpectedly performed better than the presumably Random Forest model.

Challenge & Future Plans

- Since the dataset had many zeros, I had challenges performing logarithmic transformation as part of feature engineering.

- If I had more time, I could have tried many other ML models to see how they perform in the dataset.