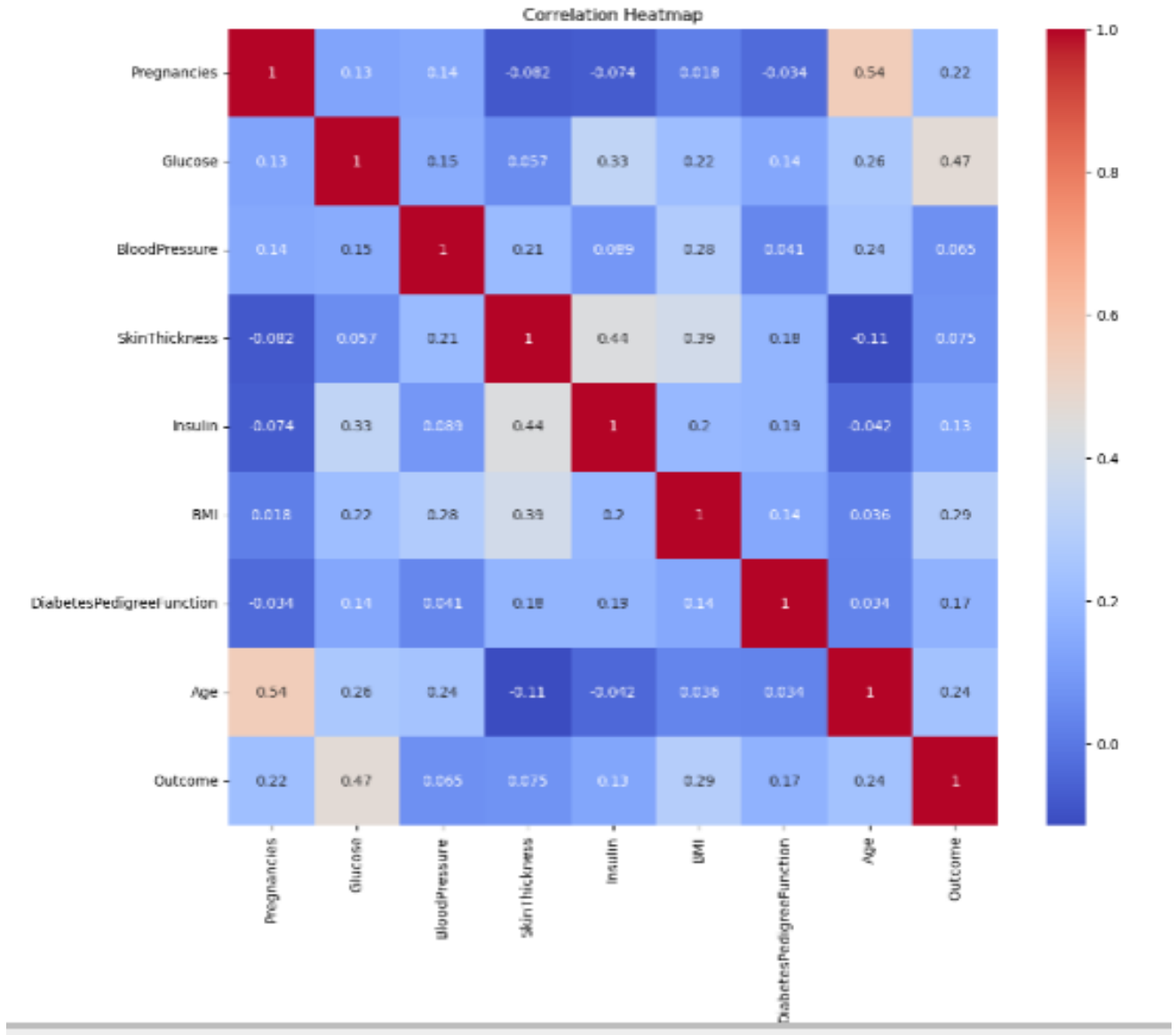
Project-Supervised Learning

TES BEYENE

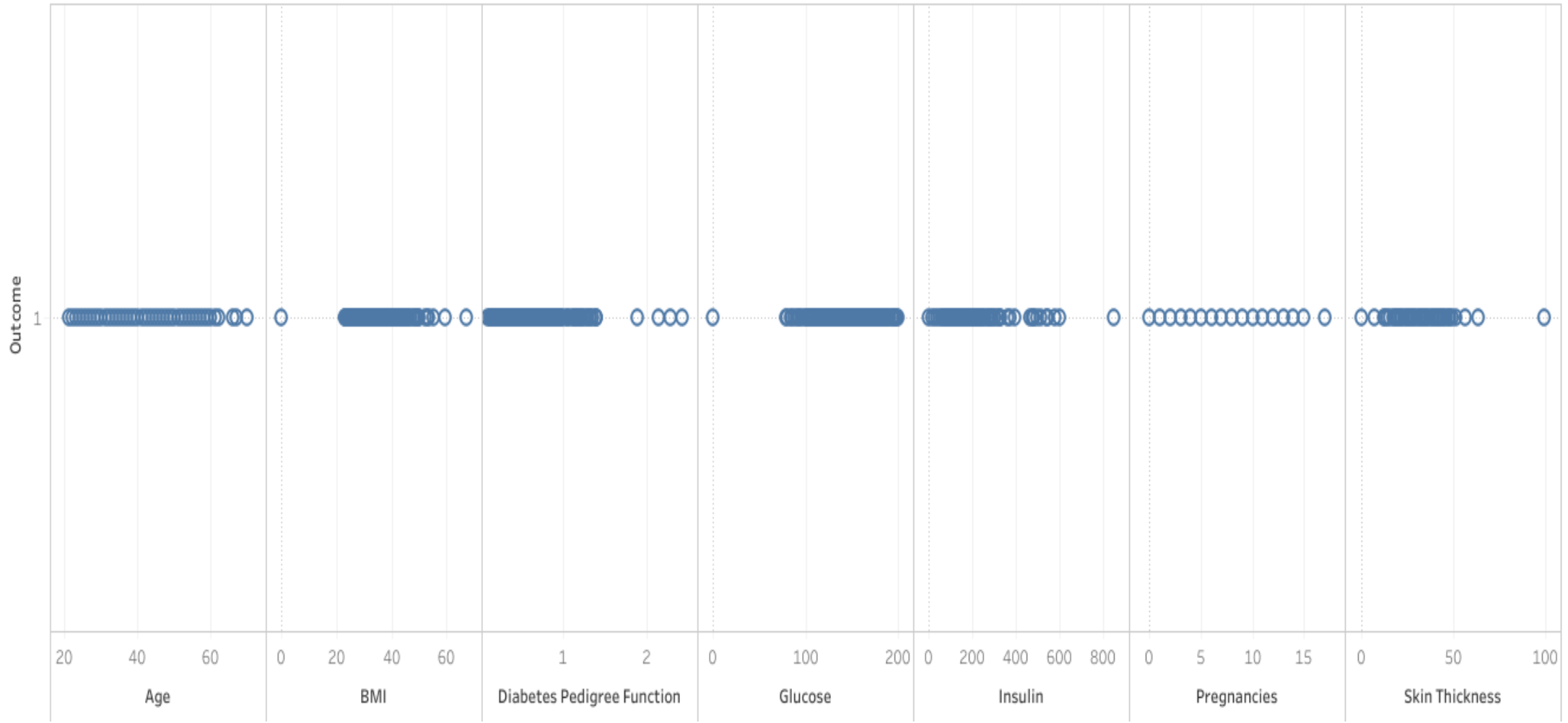
Project Goal

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

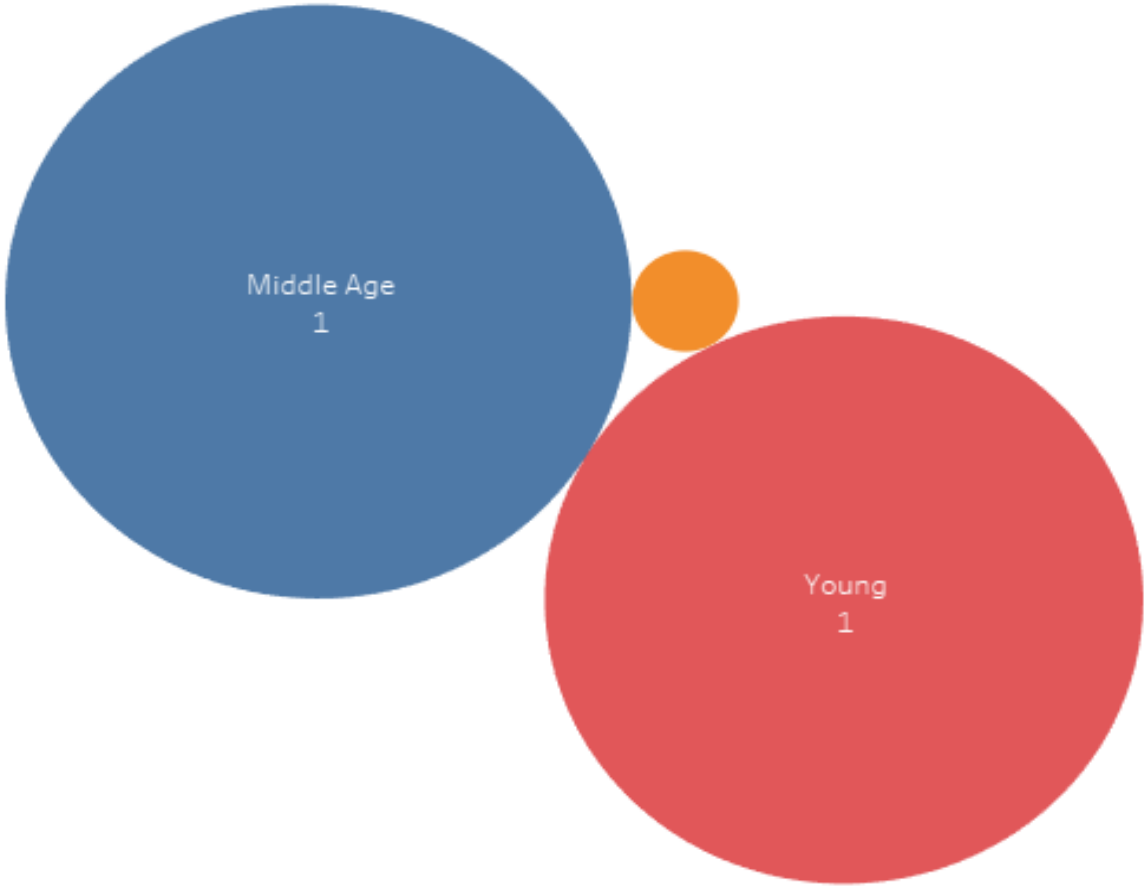
EDA & Visualization



Predictors Vs Diabtes Outcome



Age Group with Diabetes



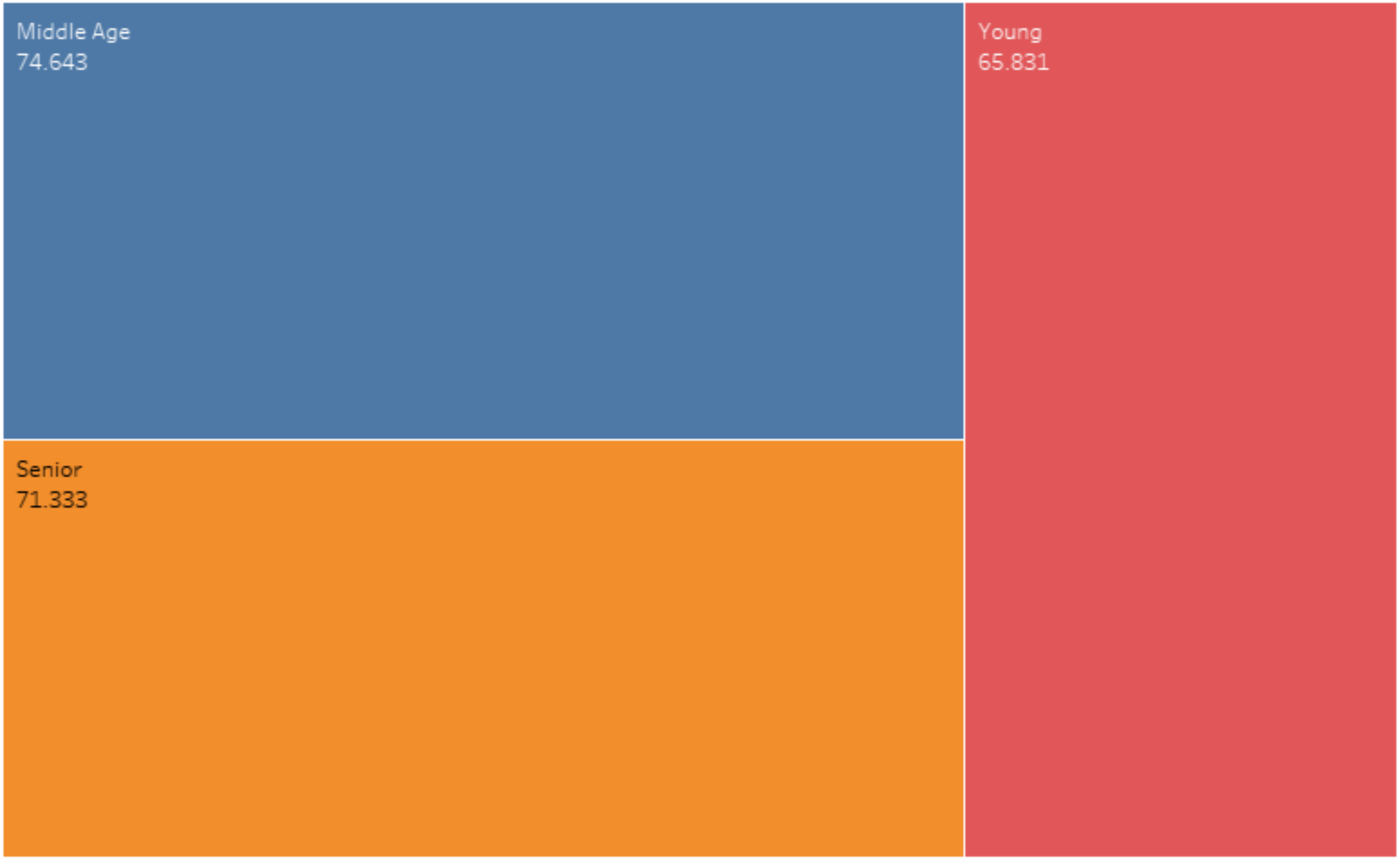
Outcome

1 1

Age Group

- Middle Age
- Senior
- Young

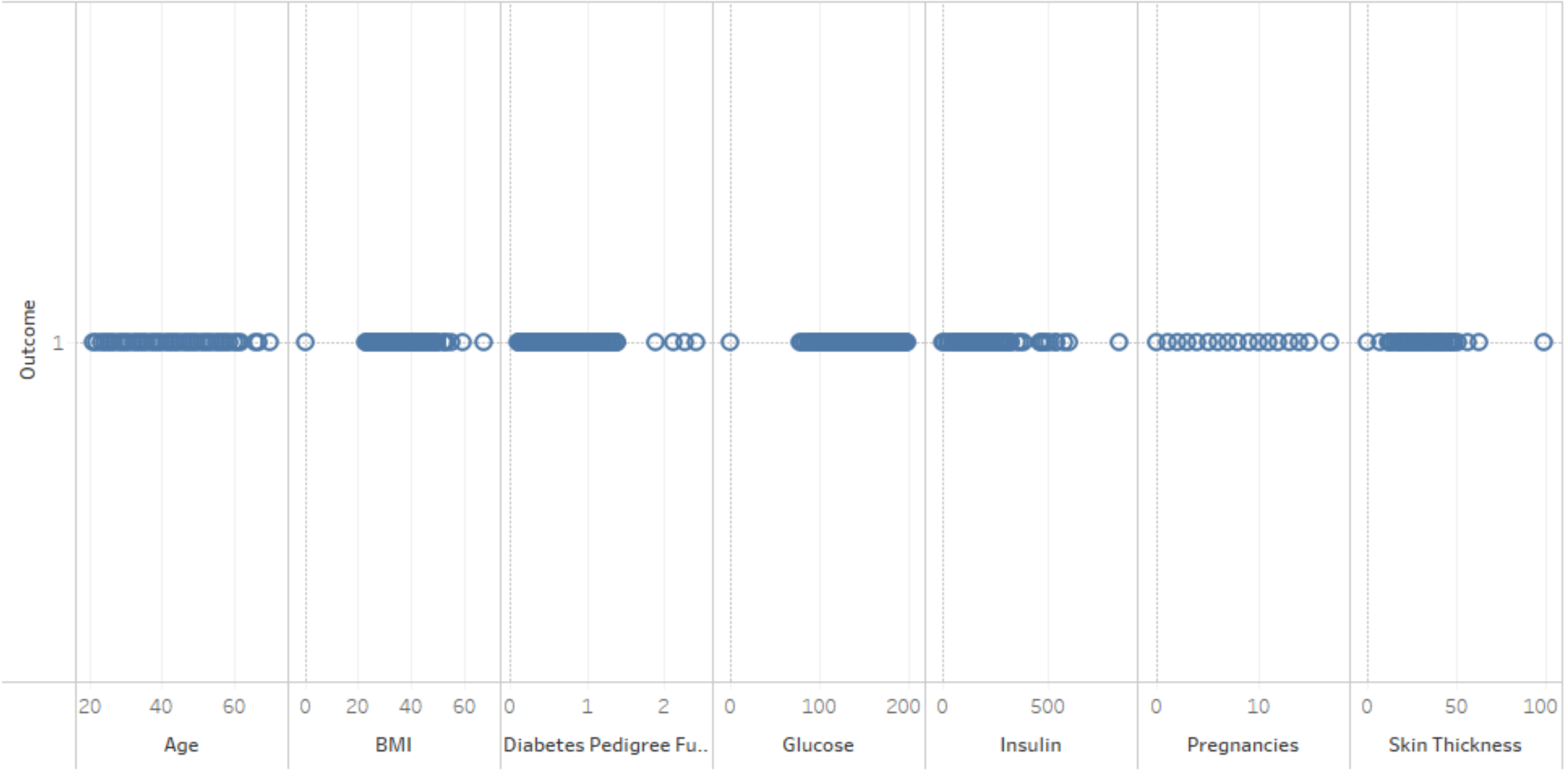
Average Blood Pressure by Age group

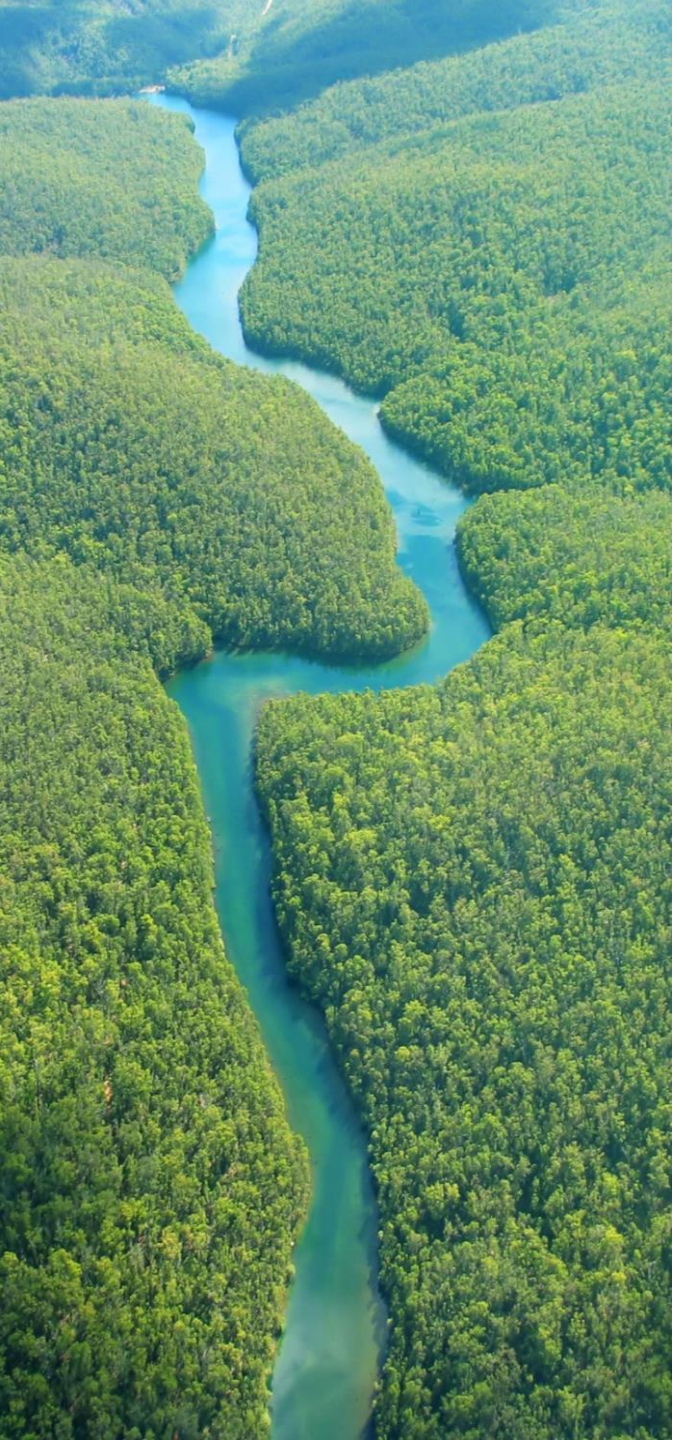


Age Group

- Middle Age
- Senior
- Young

Predictors Vs Diabtes Outcome





Findings

- The distribution for Diabetes Pedigree Function and Age were skewed. Took action to improve the distribution to avoid misleading our model training.
- There was significant imbalance in our outcome variable`(**binary classification of ratio of approximately 35% to 65%**)` which can potentially lead our model to favor into negative diabetes outcome). This is very dangerous outcome in this type of scenario as this will give false negative(meaning will give an outcome of negative diabetes to patients that have diabetes).
- Noticed outliers in **Insulin & Skin thickness**.
- Random Forest model has slightly better performance in the accuracy and precision metrics.
- The model Generally **performed okay** with **model tunning** it can perform better.
- Training, a model to determine whether a patient is positive or negative is very sensitive and delicate matter if a hospital or medical center is to depend on ML model. This requires extremely quality data, thorough preprocessing and feature engineering. The prediction of the model should surpass the normal six sigma quality.

Challenge & Future Plans

- Time was the measure constraint. Because of the need to cover the pending materials was not easy to play around on creating more visualizations.

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