## Work Sample 1: Diabetes Risks Prediction

Objective

Build a predictive model to identify individuals at risk of diabetes using common health indicators. The goal is to deliver accurate predictions while also making the model easy to interpret.

Methods

This analysis uses patient health data from over 1,000 individuals to build a predictive model for diabetes risk. I evaluated eight different machine learning algorithms to determine the most effective approach. After comparing performance metrics such as accuracy and AUC, logistic regression emerged as the most robust and interpretable model, offering reliable predictions while maintaining transparency in how each feature contributes to the outcome.

Visual Summaries
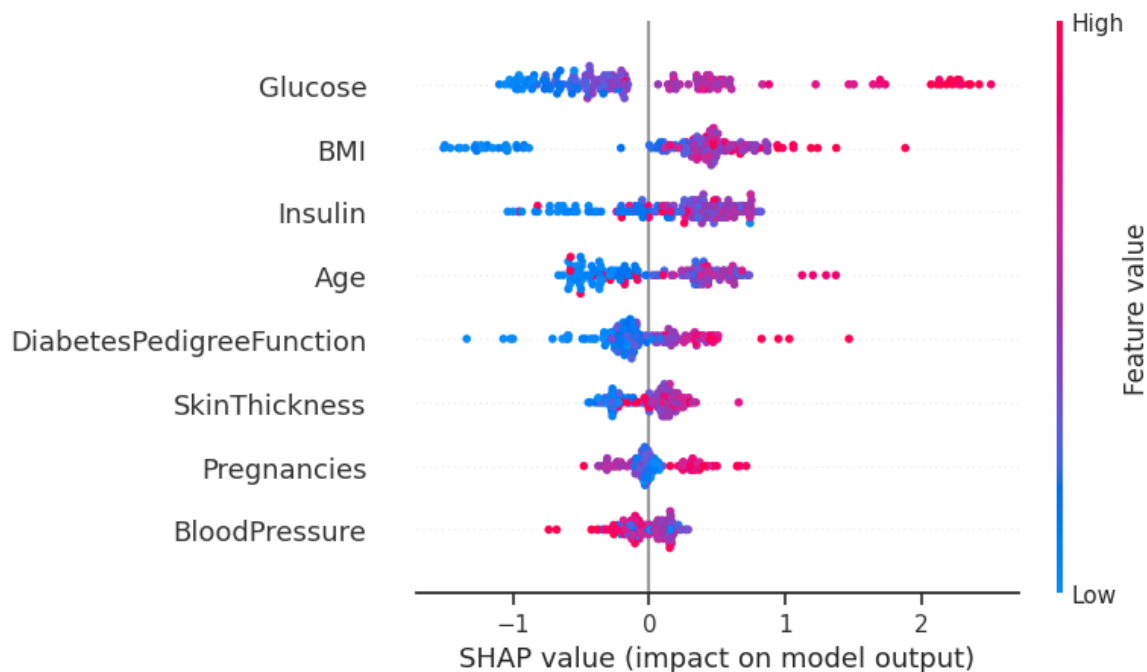
The SHAP (SHapley Additive exPlanations) summary plot below illustrates how different features influence the model's diabetes risk predictions. SHAP values indicate whether a feature increases or decreases an individual's predicted risk. Each dot represents one patient, with colors showing the feature's value — red for high and blue for low.

For instance, high glucose levels (red dots on the right) significantly raise the predicted risk, while low glucose levels (blue on the left) lower it. Features like BMI, insulin, and age exhibit similar patterns. In contrast, blood pressure values cluster tightly around zero, regardless of being high or low, suggesting that this feature has minimal impact on the model's predictions.

While SHAP helps us understand association, it doesn't imply causation. To explore the causal impact, I applied Average Treatment Effect (ATE) analysis. Among all features, the Diabetes Pedigree Function — a score representing family history of diabetes — showed the strongest causal relationship with risk. Its ATE score was significantly higher than all other variables. This

underscores the importance of genetic factors in diabetes risk and highlights the need for individuals with a family history to take proactive steps in monitoring and managing their health.

Causal Impact (ATE) of Features on Diabetes Risk