For my final project in MAT 311: Introduction to Data Science, I set out to develop and compare multiple machine learning models to predict the likelihood of diabetes using the Pima Indians Diabetes dataset. The dataset, sourced from Kaggle, contains 768 medical records for female patients, each with eight diagnostic features such as glucose level, BMI, blood pressure, and age. Our goal was to build accurate, interpretable binary classification models to determine whether a patient was likely to have diabetes (Outcome = 1) or not (Outcome = 0).

The project followed a complete data science workflow: I began with exploratory data analysis (EDA) and used a decision tree to evaluate feature importance. I then reduced the dataset to the top five most predictive features to simplify the model and improve generalization. The data was split into training, validation, and test sets using a 70/20/10 ratio, and I implemented three models—K-Nearest Neighbors (KNN), Gaussian Naive Bayes, and Decision Tree—using Scikit-learn.

One of the biggest challenges in this project was structuring the code into modular Python scripts and ensuring that the entire machine learning pipeline ran smoothly from start to finish. Debugging data splits, managing random seeds, and producing reproducible results a lot of effort. It was also initially difficult to get clean metrics and graphs for all three models in a way that could be interpreted clearly. I ended up having to learn scikitlearn by myself from scratch, as I was having a very hard time understanding how to do everything However, through enough trial and error as well as incremental testing, I were able to build a well-structured repository with functional model training, evaluation, and ROC analysis.

Seeing the final models run with decent test accuracy and AUC scores was incredibly rewarding. It validated my design decisions—especially feature selection and data handling—and provided a sense of accomplishment in translating conceptual data science knowledge into a working real-world application. Overall, this was a rewarding project and I know have a lot of knowledge I can bring into the real world, as well as a project to add to my portfolio.