Predicting Diabetes in Women

Rebeca Rodriguez, Emma Beebe
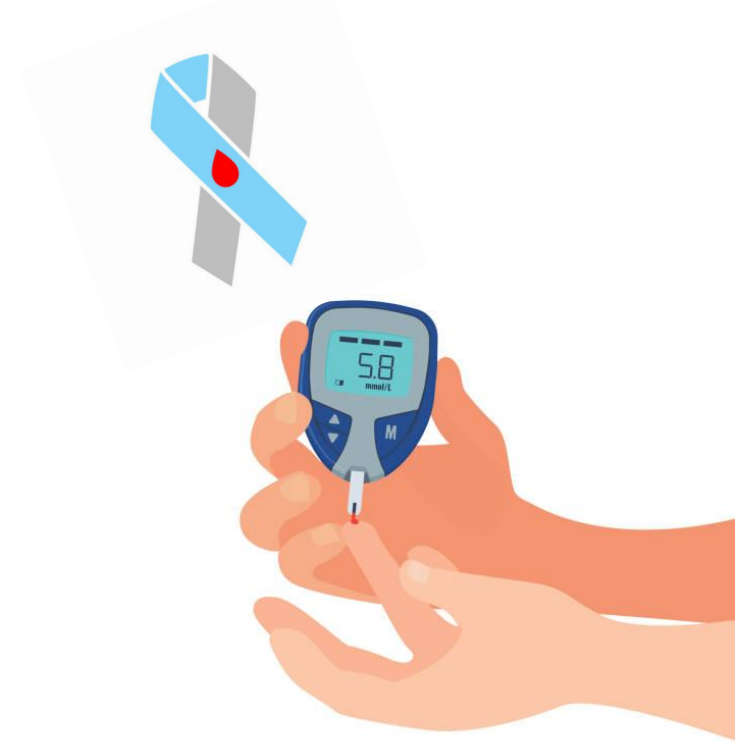
Meet the Team

Rebeca Rodriguez

MEng in CS

Emma Beebe

MEng in EE

Table of Contents

Background

❖ What is diabetes?

❖ Why diabetes in women?

❖ How can it be detected?
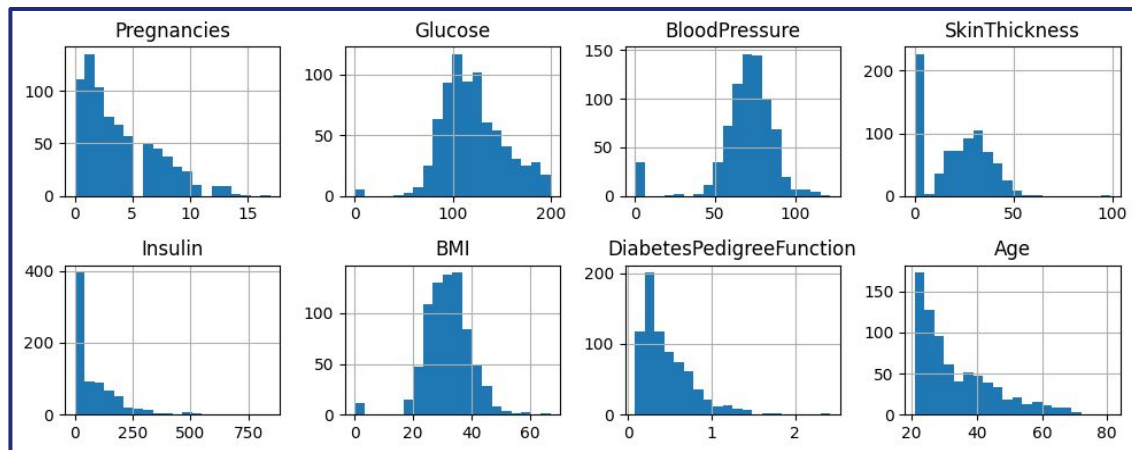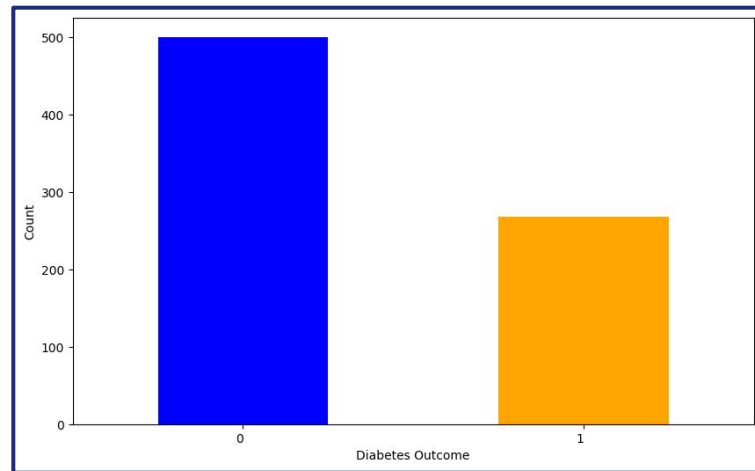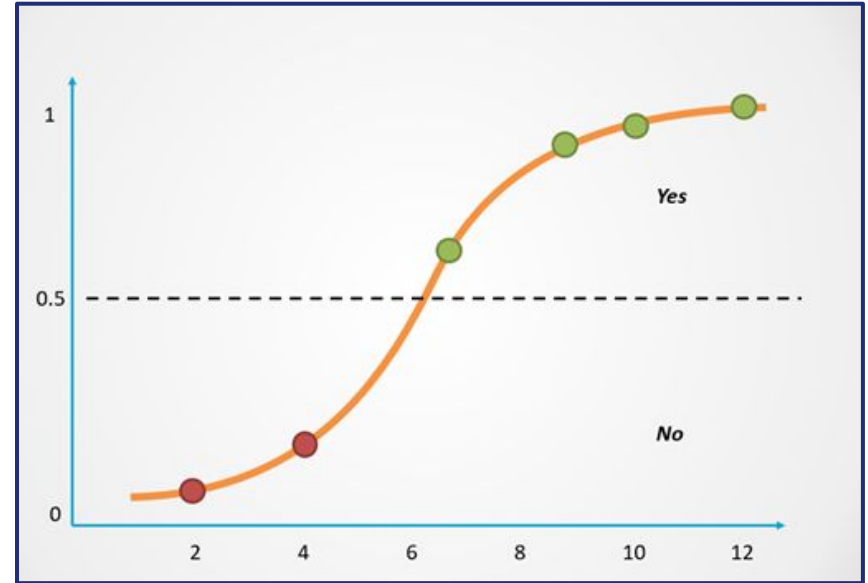
Dataset

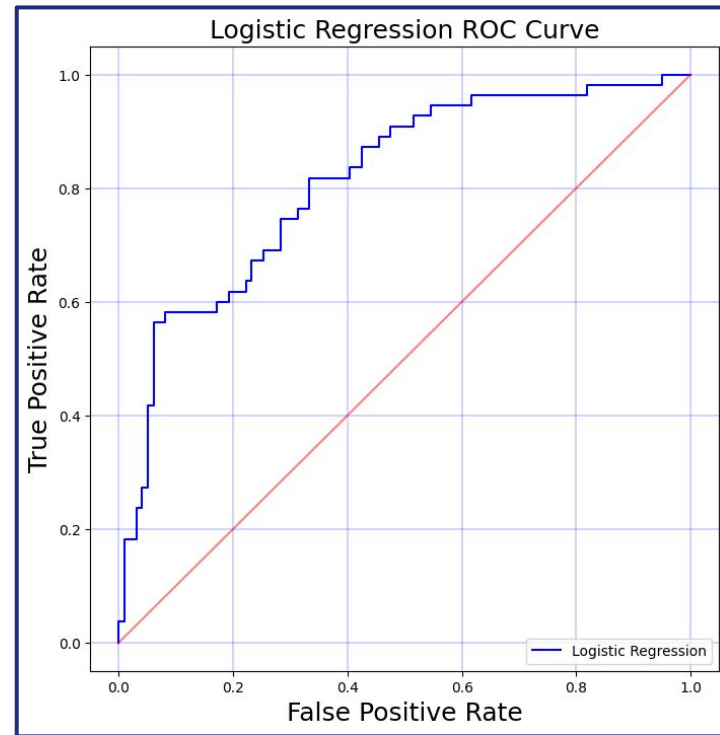❖ Data Ingestion

    ➢ 768x8

❖ Train-Test Split

    ➢ 80/20
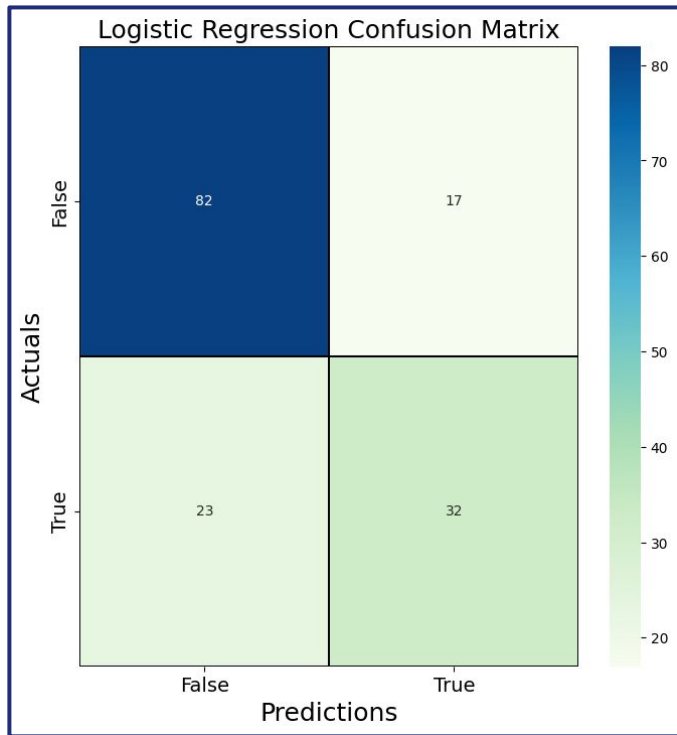
Logistic Regression

❖ Estimates probabilities on a logistic regression curve
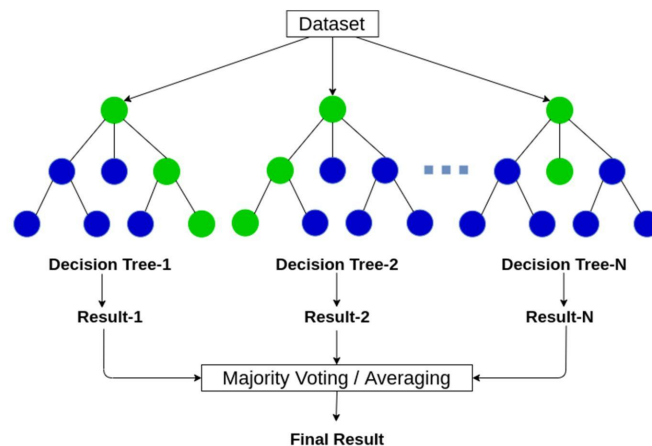
❖ Used for classification and predictive analysis
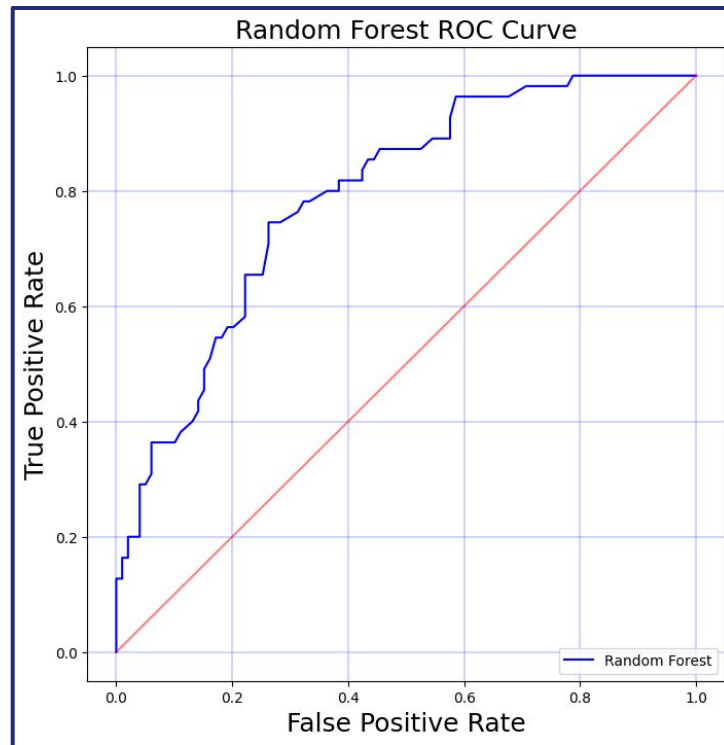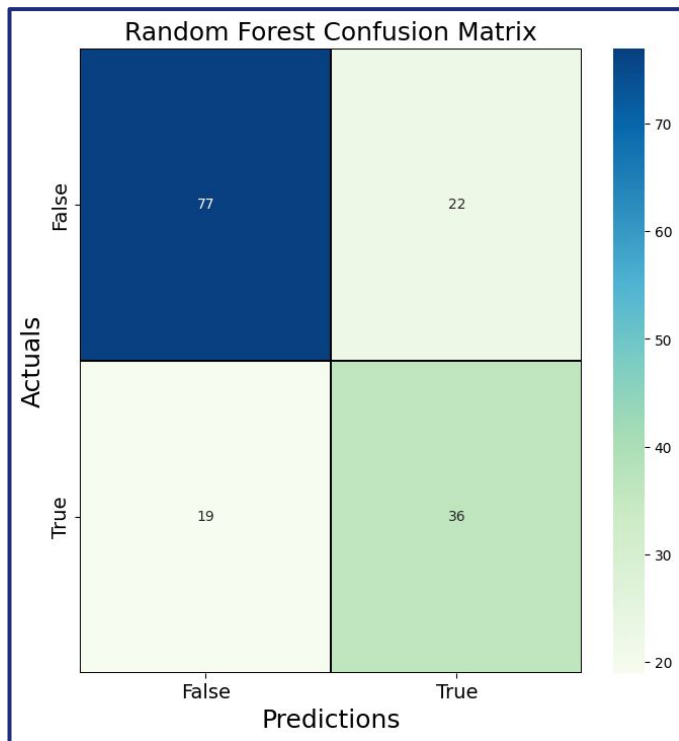
Logistic Regression Results

Random Forest

- ❖ Multiple decision trees
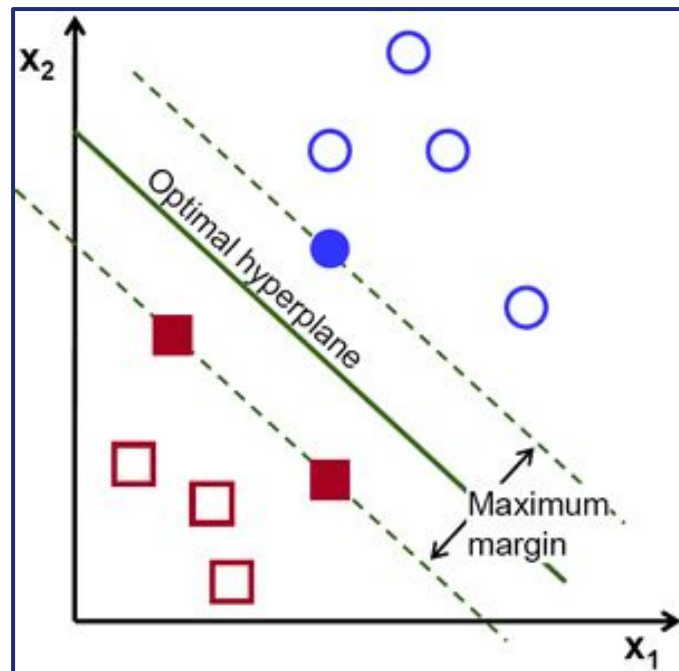- ❖ Used for classification and regression
- ❖ Ensemble method

Random Forest Results
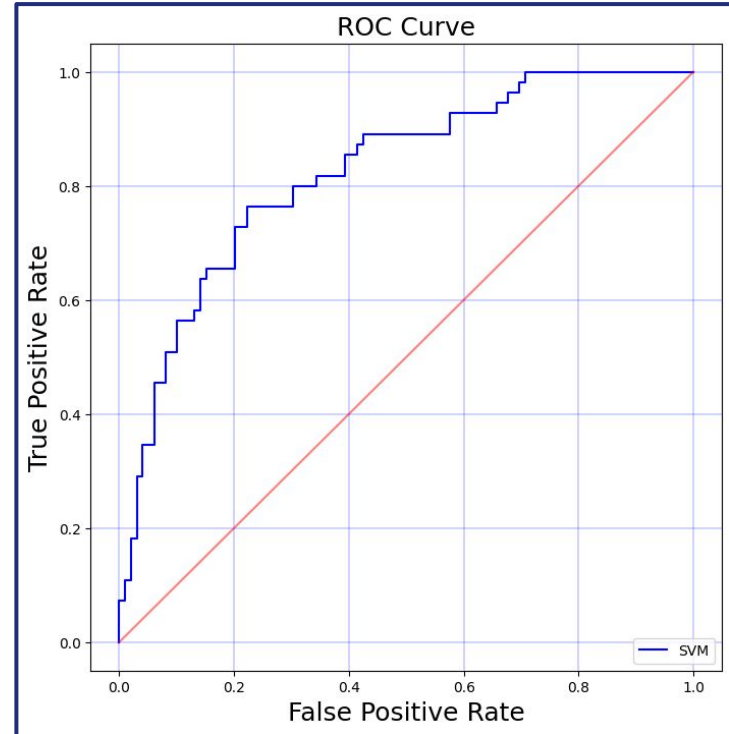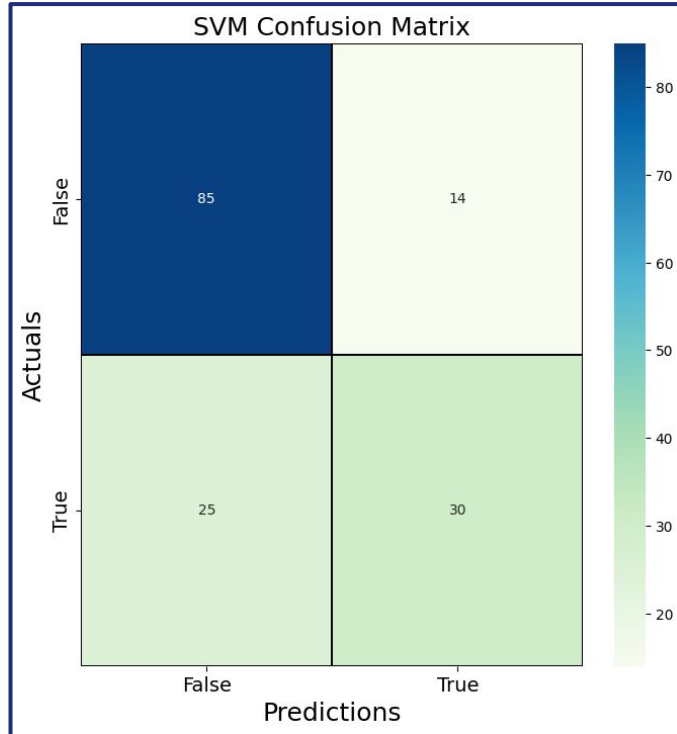

Random Forest Confusion Matrix

Random Forest ROC Curve

Support Vector Machine

❖ Finds a hyperplane in N-dimensional space that classifies the data points
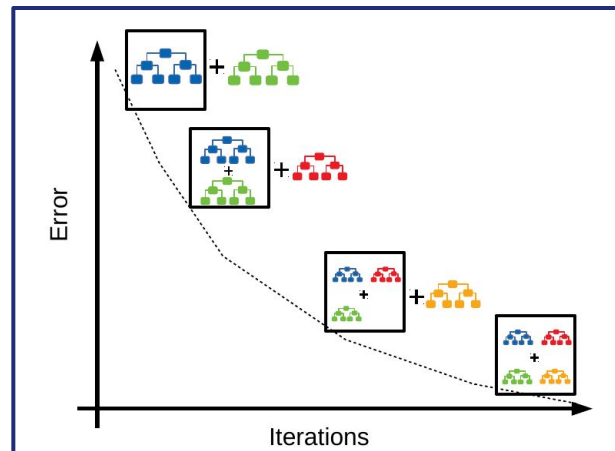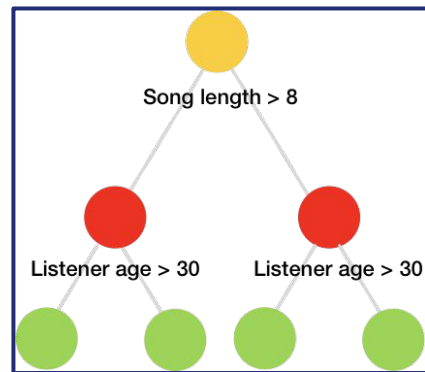
❖ Can be used for both regression and classification tasks
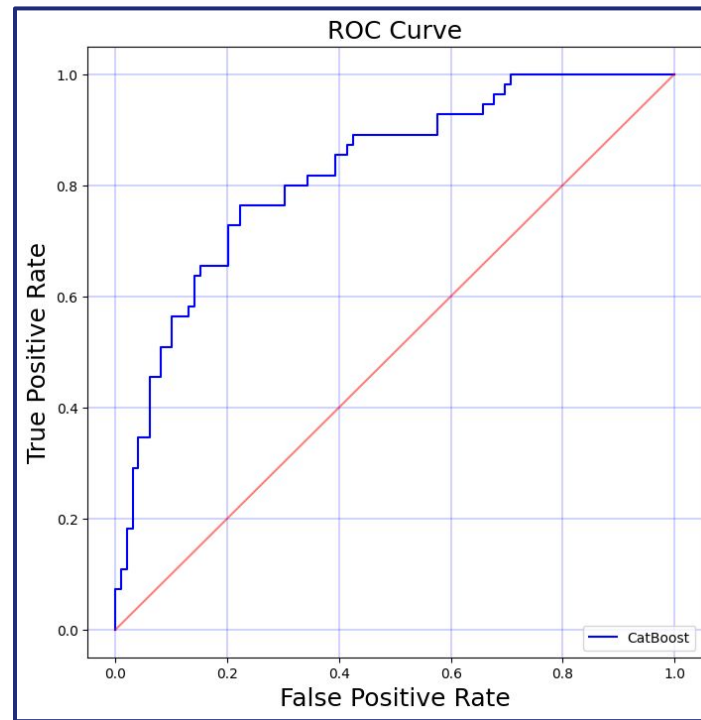
Support Vector Machine Results
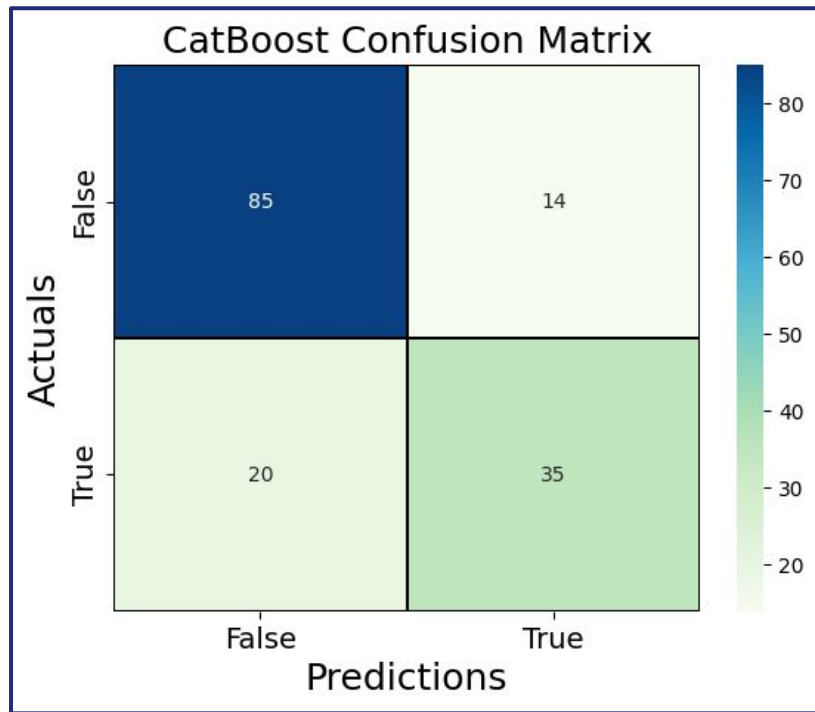
CatBoost

❖ Member of the boosting family

❖ Reduces the need for much of
   hyperparameter tuning

❖ Robust to overfitting

CatBoost Results
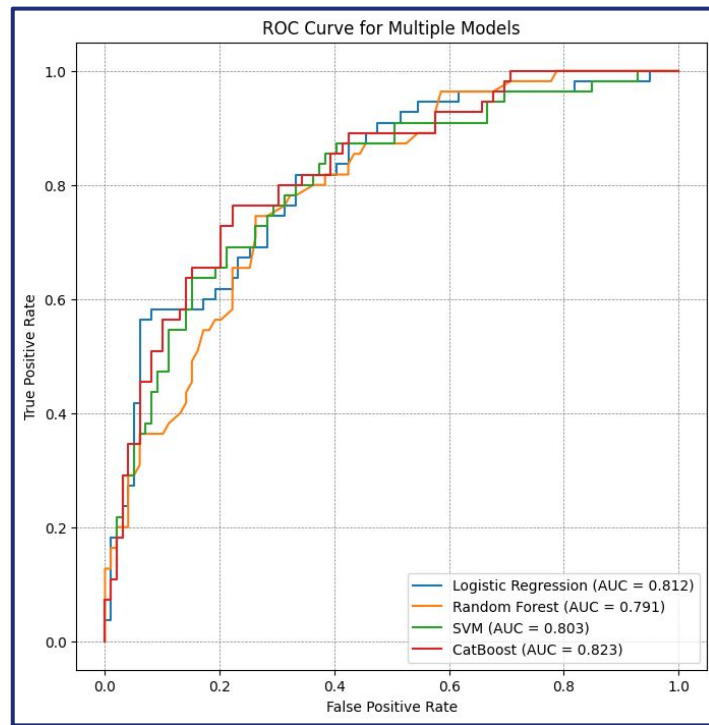
Results Comparison

Model	Area Under Curve (AUC)	Accuracy
Logistic Regression	0.812	0.740
Random Forest	0.791	0.734
SVM	0.803	0.747
CatBoost	0.823	0.778

ROC Curve for Multiple Models

Lessons Learned

❖ Machine learning models require extremely large amounts of data to be effective

❖ Women have many unique cases that could lead to diabetes
  ➢ Polycystic Ovarian Syndrome (PCOS)
  ➢ Gestational Diabetes
  ➢ Menopause

Future Work

❖ Diabetes prediction system implemented in hospitals

❖ More data categories

Final Comments

Questions?

———