Team 2 Project 2

Aaliyah, Aaron, James, Lucas, Ryan

Project Overview

› **Goal:** Create a model that *predicts whether an individual is at risk of developing Type II Diabetes based on the presence or absence of certain symptoms, prompting the individual to meet with their doctor if the model shows they are at risk.*

› *Type II Diabetes increased 20% between 2012 & 2022, costing around $412 billion, which includes medical costs and indirect costs like loss of productivity*.*

› *The creation of a model that can accurately predict whether Type II Diabetes may be present based on certain symptoms can help in mitigating the health and financial costs associated with this disease.*

› **Thompson, D. (2024, August 22). HealthDay News. Retrieved from HealthDay News: https://www.healthday.com/health-news/diabetes/number-of-americans-with-type-2-diabetes-jumped-by-20-in-a-decade*

Data Collection

RABIE EL KHAROUA · UPDATED 3 MONTHS AGO

47

🩸 Diabetes Health Dataset Analysis 🩸

Detailed Health Information for Diabetes Research

About Dataset

This dataset contains comprehensive health data for 1,879 patients, uniquely identified with IDs ranging from 6000 to 7878. The data includes demographic details, lifestyle factors, medical history, clinical measurements, medication usage, symptoms, quality of life scores, environmental exposures, and health behaviors. Each patient is associated with a confidential doctor in charge, ensuring privacy and confidentiality. This dataset is ideal for researchers and data scientists looking to explore factors associated with diabetes, develop predictive models, and conduct statistical analyses.
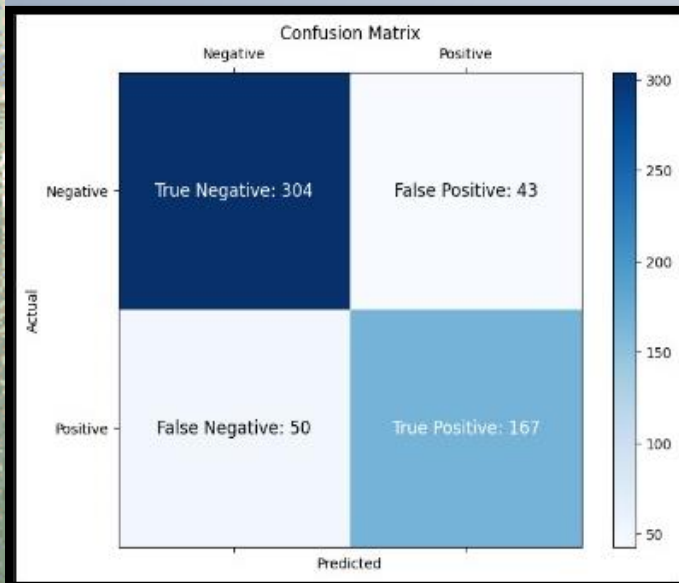
Data Exploration

› *Separation of Numerical and Categorical Data*
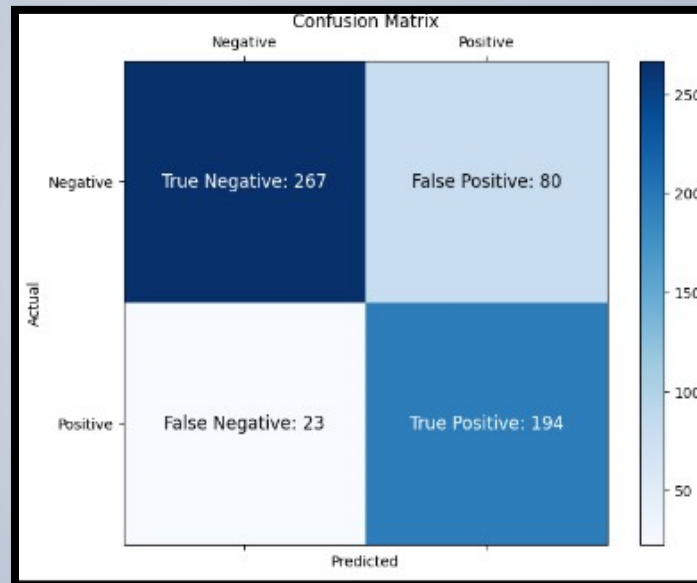
› *Feature Importance Analysis*

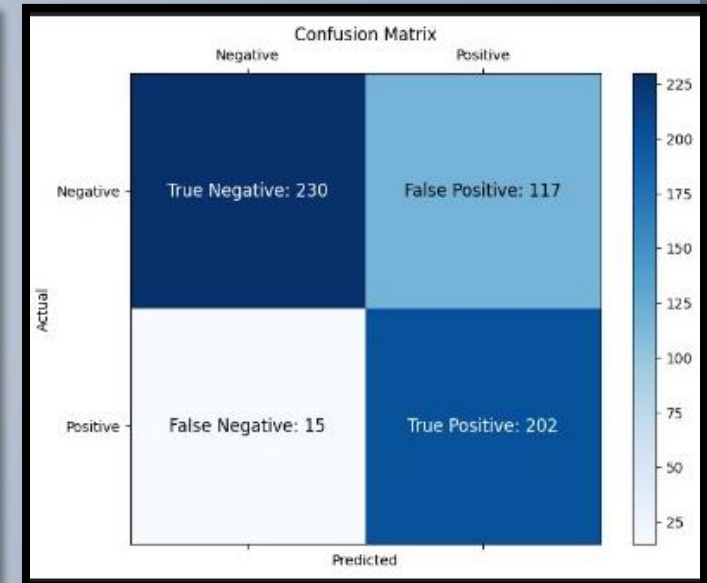› *Scaling the Data*

Experiment in Logistic Regression Model Tuning

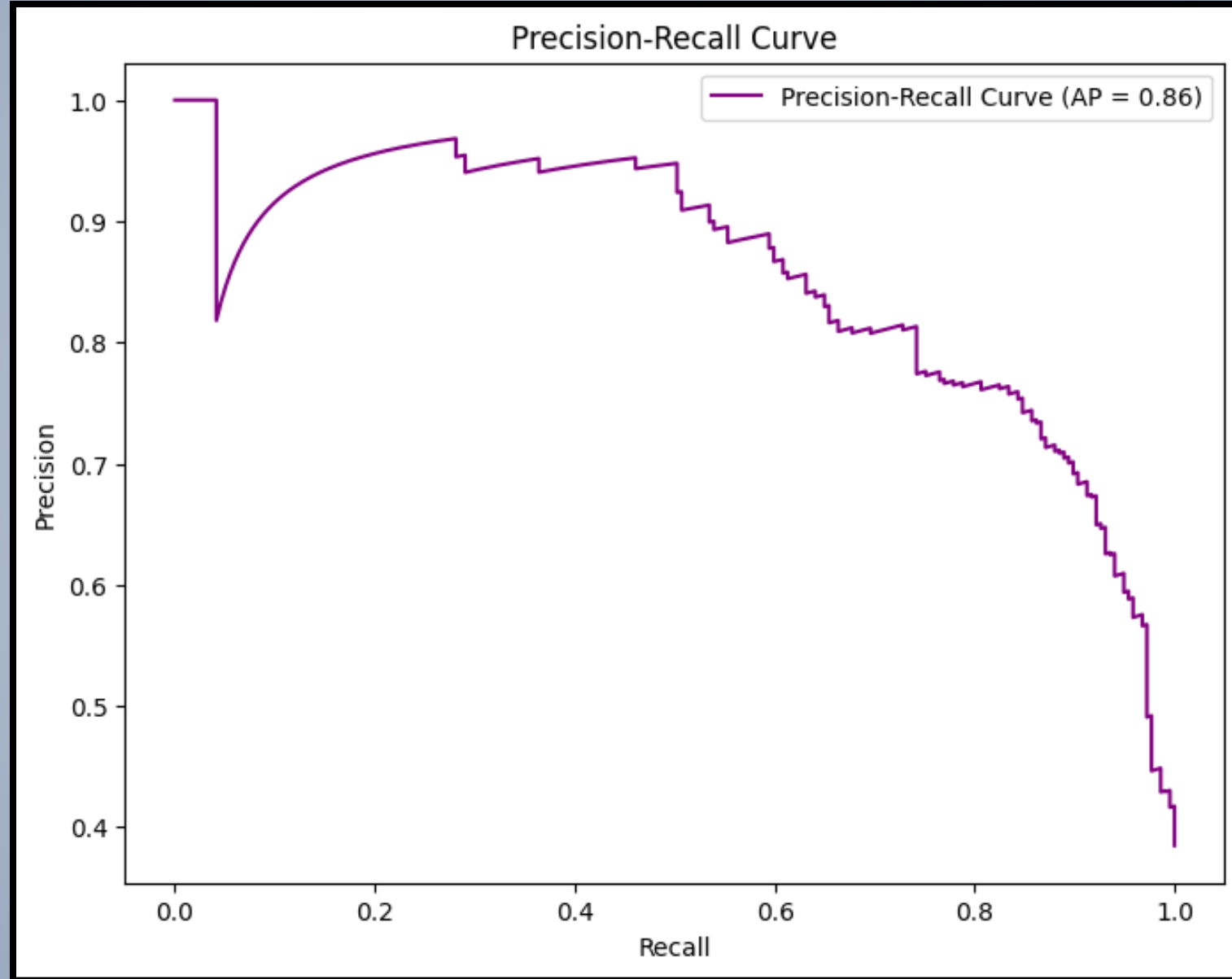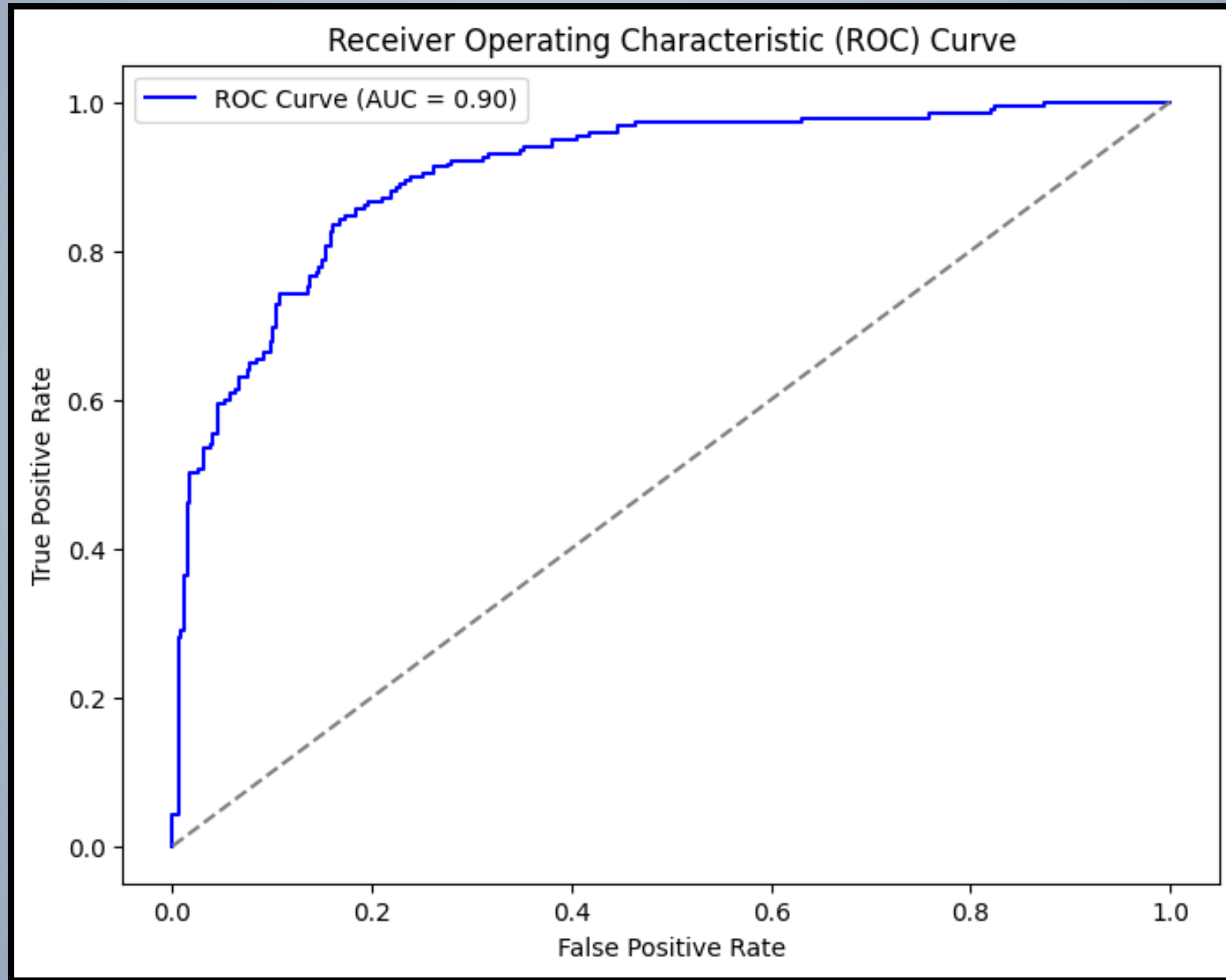# Confusion Matrix [Logistic Regression]
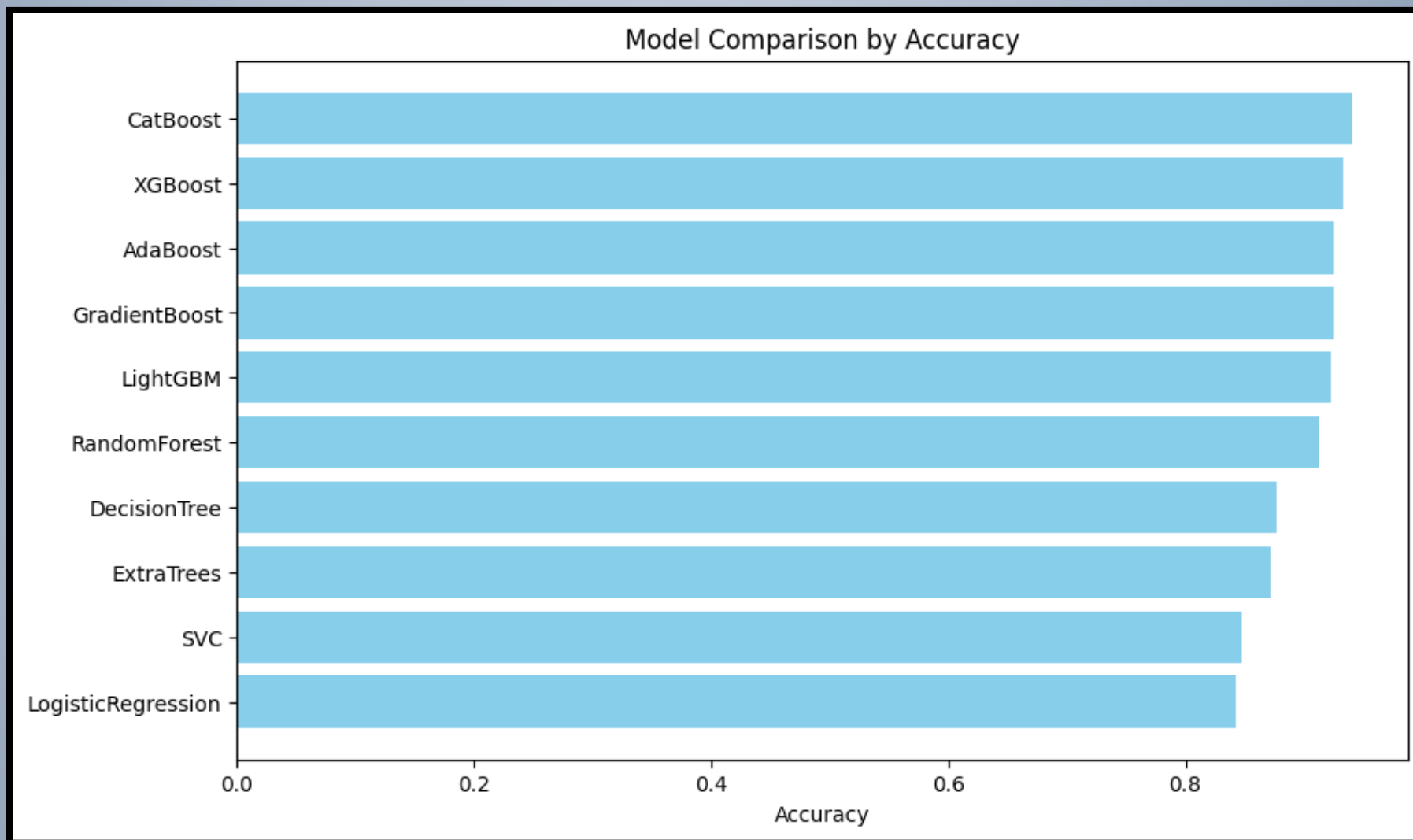

1

2

3

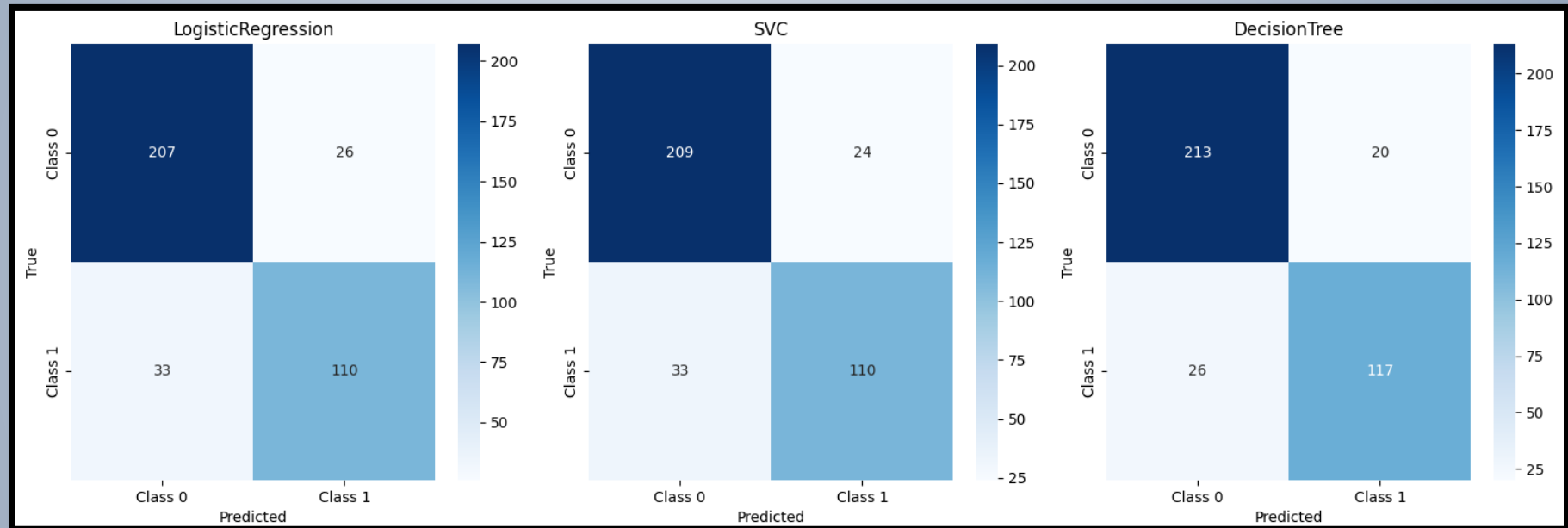# Precision Recall [Logistic Regression]

ROC Curve [Logistic Regression]

Model Comparisons
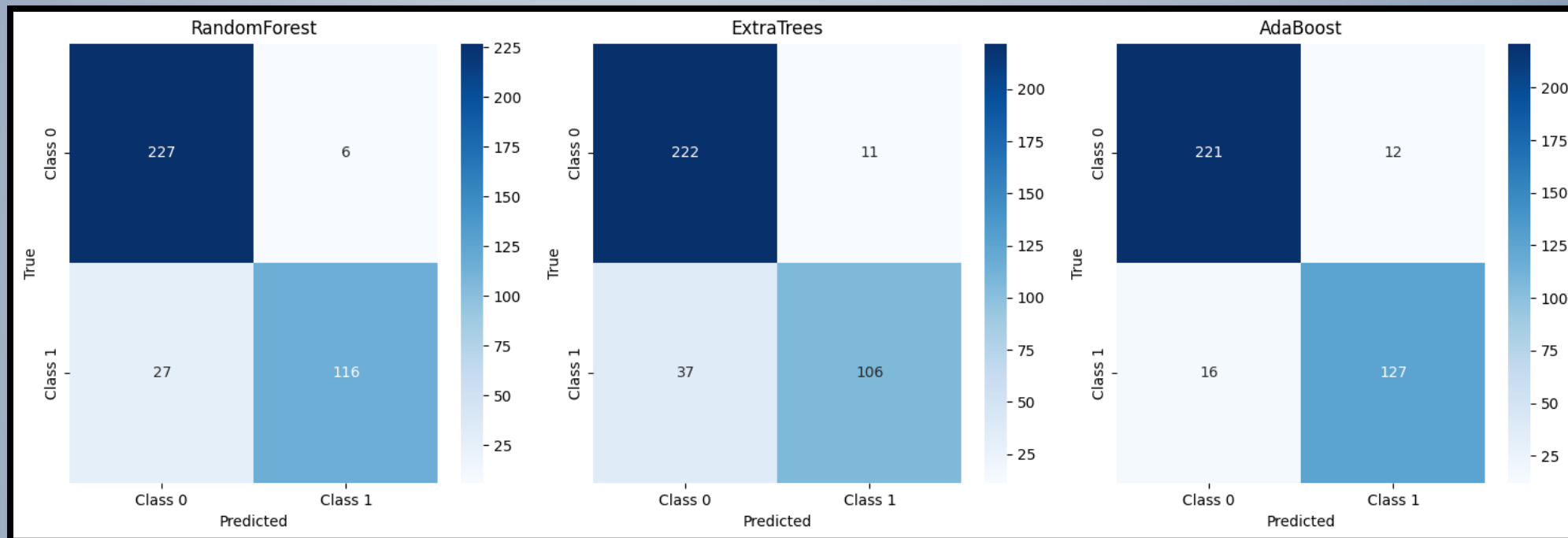
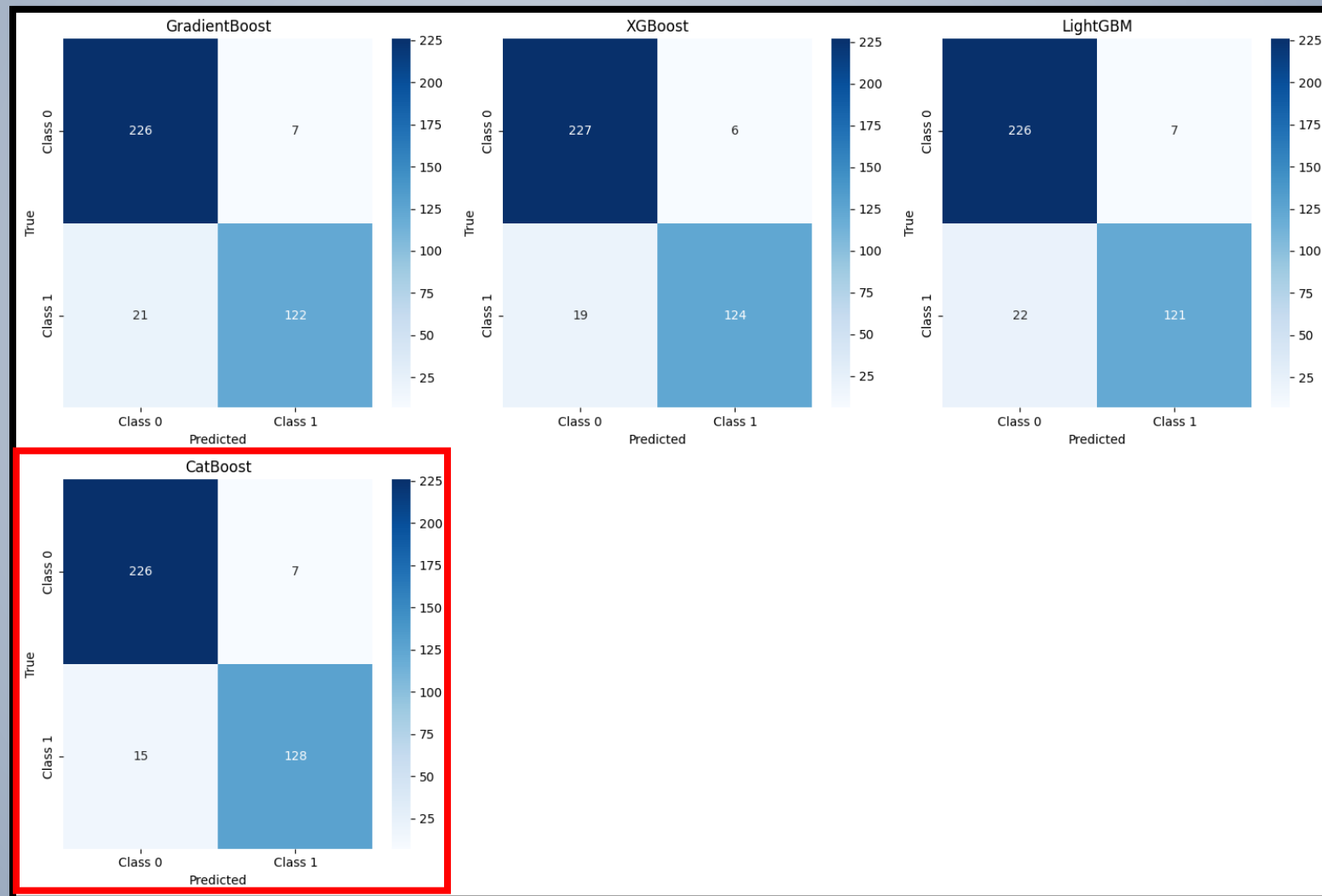# Accuracy Comparisons

Model Comparison by Accuracy

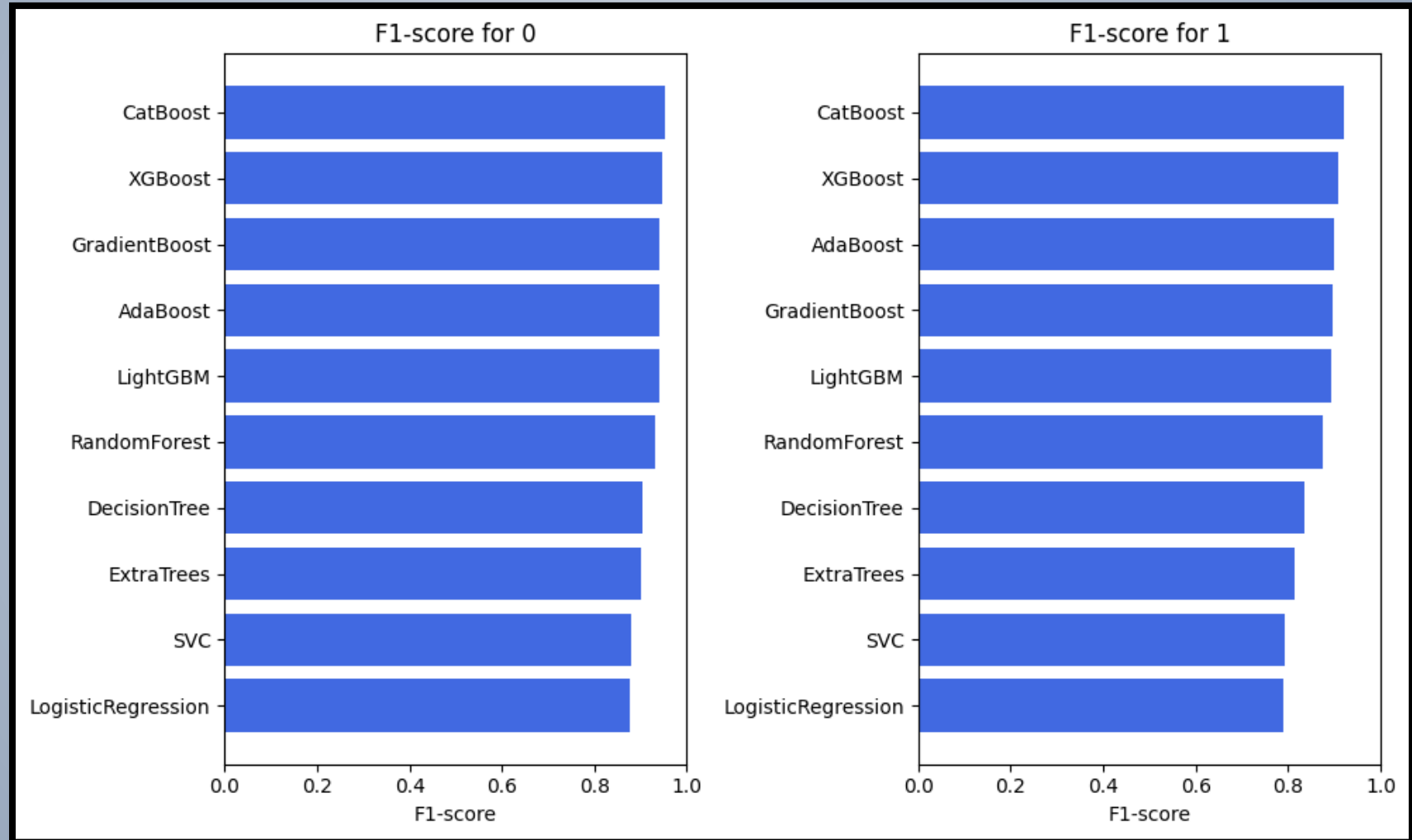# Confusion Matrix Comparisons

Confusion Matrix Comparison Continued

Confusion Matrix Comparison Continued

F1 Score Comparisons

Chosen Model: CatBoost

```
**Accuracy**:
 0.9414893617021277

**Confusion Matrix**:
 [[226   7]
 [ 15 128]]

**Classification Report**:
              precision    recall  f1-score   support

           0       0.94      0.97      0.95       233
           1       0.95      0.90      0.92       143

    accuracy                           0.94       376
   macro avg       0.94      0.93      0.94       376
weighted avg       0.94      0.94      0.94       376
```

Interactive demonstration

https://huggingface.co/spaces/kopeck/diabetespredictiontool

Unanticipated Insights

› Automate processes via a ML pipeline in order to avoid missing steps.

› More
positive diagnosis than negative in the dataset

› Improving on model using Neural Network Models and increased actual patient data

Real World Applications / Benefits

- *Targeted Care*

- *Healthcare practitioners could target specific contributors to Type II Diabetes based on the data that has been submitted by the patient. Based on that initial data, models can be run with potential future data to show that improvements in one or more contributors can lead to prevention – if not present – or remission – if the disease is present.*

- *Increase in Health Awareness*

- Viewing the body as an interconnected system, anybody at risk of developing or has already developed Type II Diabetes, based on certain metrics and symptoms, must adjust the way they live, which, if followed through, will benefit other aspects of their health outside of their relationship with prediabetes or Type II Diabetes. Such factors as diet management, exercise programs, sleep hygiene, relaxation techniques, and an overall focus on improving one's quality of life, have the power to move past the initial focus of these lifestyle changes – prevention or mitigation of Type II Diabetes – and directly impact in a positive manner other aspects of a patient's life.

Real World Applications / Benefits

- *Cost Minimization*

  - *The use of an AI model to detect the presence of prediabetes or Type II Diabetes allows individuals to obtain an accurate prediction of their risk factor without having to pay for what may be an unnecessary screening at the doctor's office or hospital.*

- *Remote Accessibility*

  - With the increase in telemedicine, access to online diagnostic applications such as one created here can be sent to a qualified physician for review, allowing the potential patient to speak with a doctor directly in the digital space, avoiding the hassles of scheduling and commuting to a physical location.