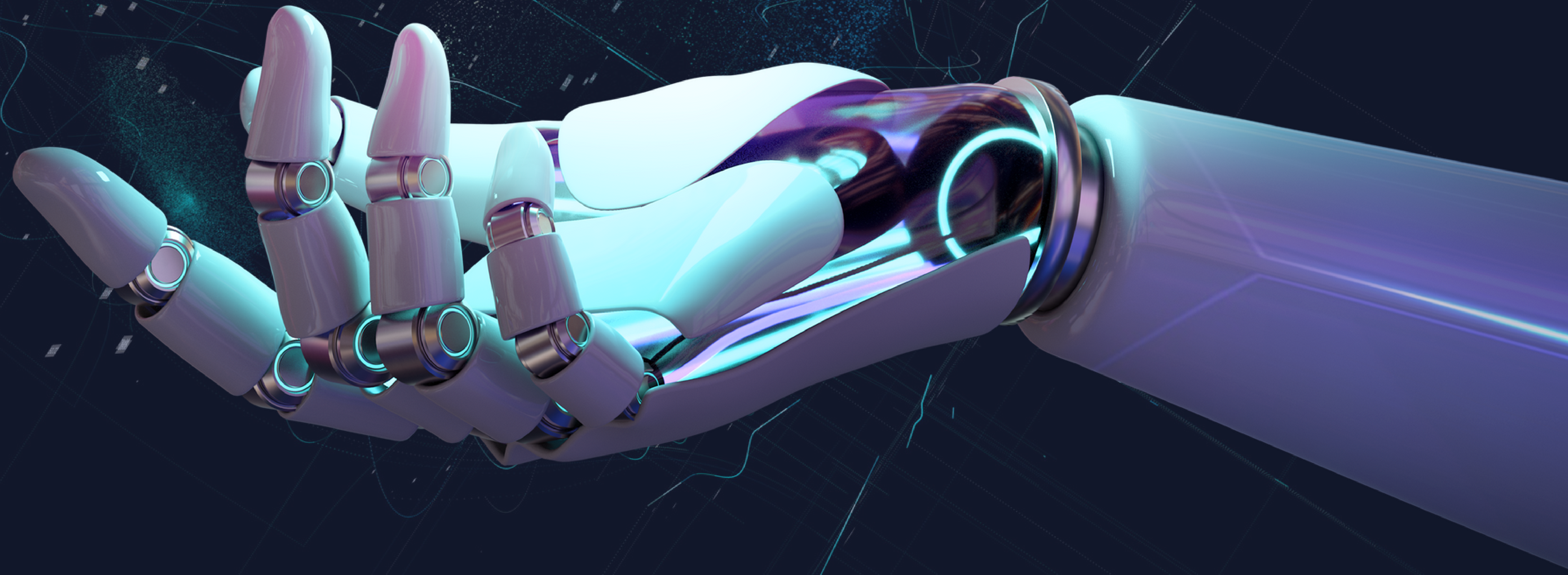
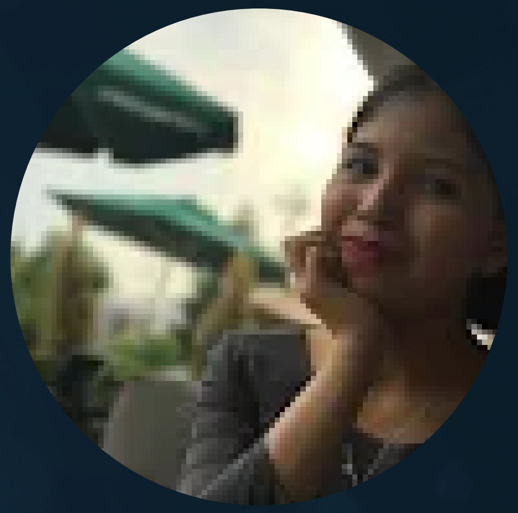
Behavior Analysis on Twitter for Bot Account Identification

A MACHINE LEARNING APROACH



Our Great Team



Adriana Leticia Martinez Estrada

David Arturo Hernandez Gomez

Contents

01/ Introduction
02/ Data Analysis
03/ Feature engineering
04/ Models
05/ Solving the problem!
06/ Future work

Social Media in our lives

Now a days the social media plays a fundamental role in our lives.

Food

Sports

Movies

...

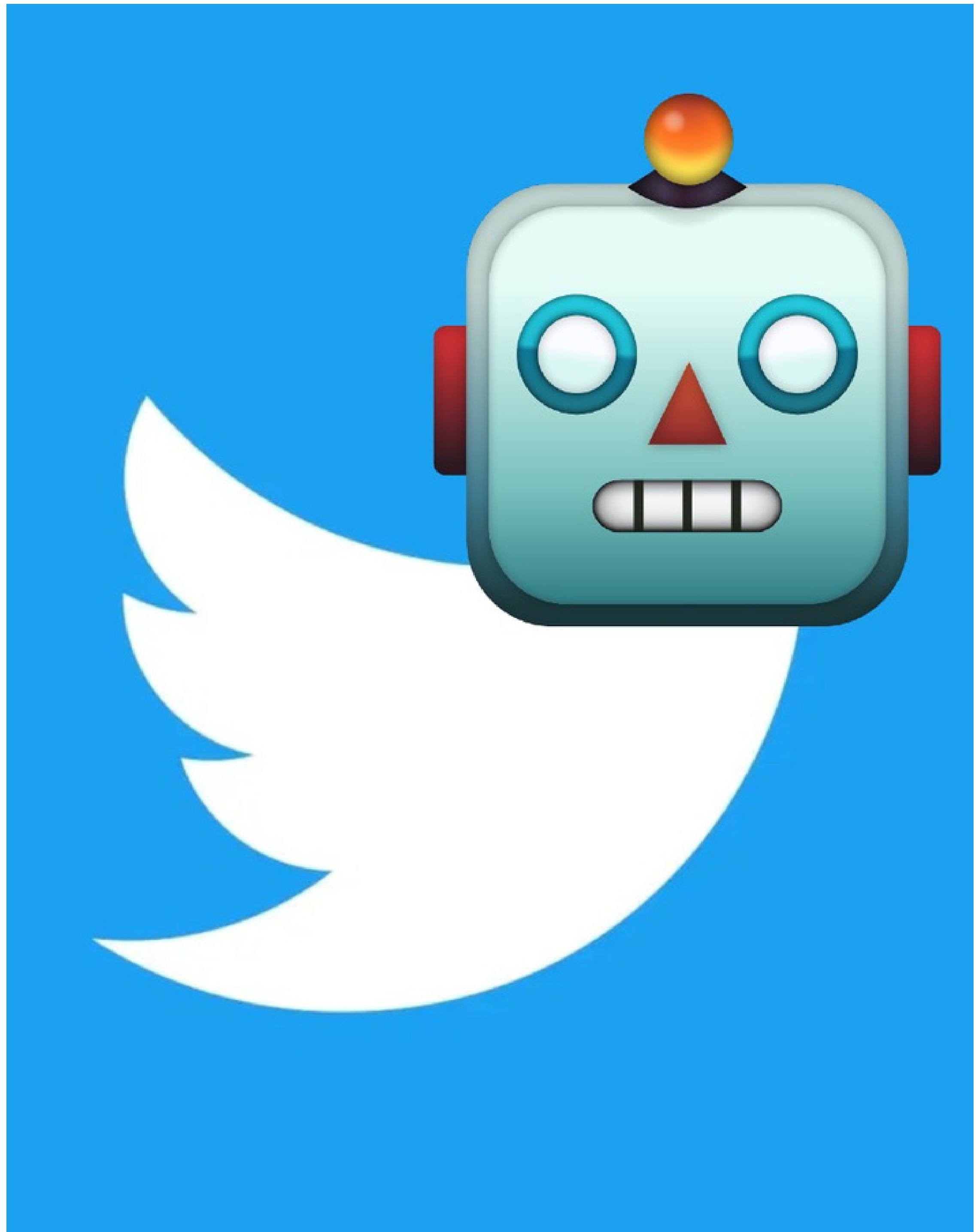
Science

Politics

Twitter and bots

Not everything we see is always true

- Influencers may not have that count of followers
- Not every post we see may be true
- Some accounts can automate various actions like posting tweets, retweeting, liking, following, and messaging



Twiplomacy #DigitalDiplomacy
@Twiplomacy · Follow

World Leaders and their Fake followers

Some of the most followed world leaders and their share of bot followers as determined by twitteraudit.com. Graphics prepared by @Saosasha @gzeromedia
#DigitalDiplomacy

8:23 AM · Feb 21, 2018 from Geneva, Switzerland

Lets see some numbers!

The solution

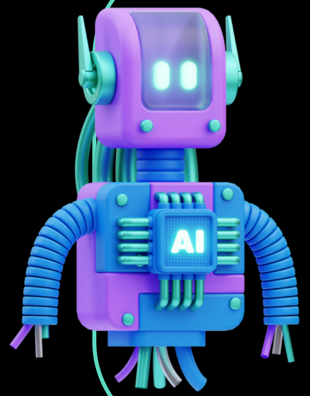
ENVIRONMENT

Colab

All the solution is build around colab tool for ML

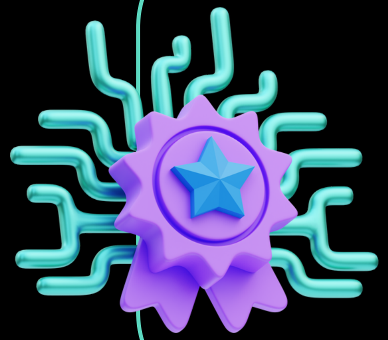
libraries

The library versions are listed on the requirements file to ensure the correct functionality.



Python

The code is develop in Python 3



Repository

All the code are hosted on a GitHub repository using git for change control.

Data Analysis

The data was extracted from the following repository.

Botometer

The data consist of two datasets

- Self-identified bots from https://botwiki.org. Labels and user objects.
- Verified human accounts. Labels and user objects.

Data Analysis

1. Understand data

- Look for fields with information not relevant to the model.
- Drop columns with same value in all registers.

2. Pipelines

- The numerical data was scaled using **StandardScaler** technique.
- Categorical data was encod to multiple categories using **OHE.**
- **Nulls** was inputed using mean and most frecuent value respectively.

3. Split

The data was split into a 70-30 train-test split in a stratified manner.

4. Reproducibility

All the methods have a seed to ensure the same result between environments..

Why Accuracy

- Easy interpretation

Models

- L2 (RidgeClassifier)
- DecisionTreeClassifier
- Logistic_regression
- SVM

74%

RidgeClassifier

Accuracy

92%

SVM

Accuracy

97%

Decision Tree

Accuracy

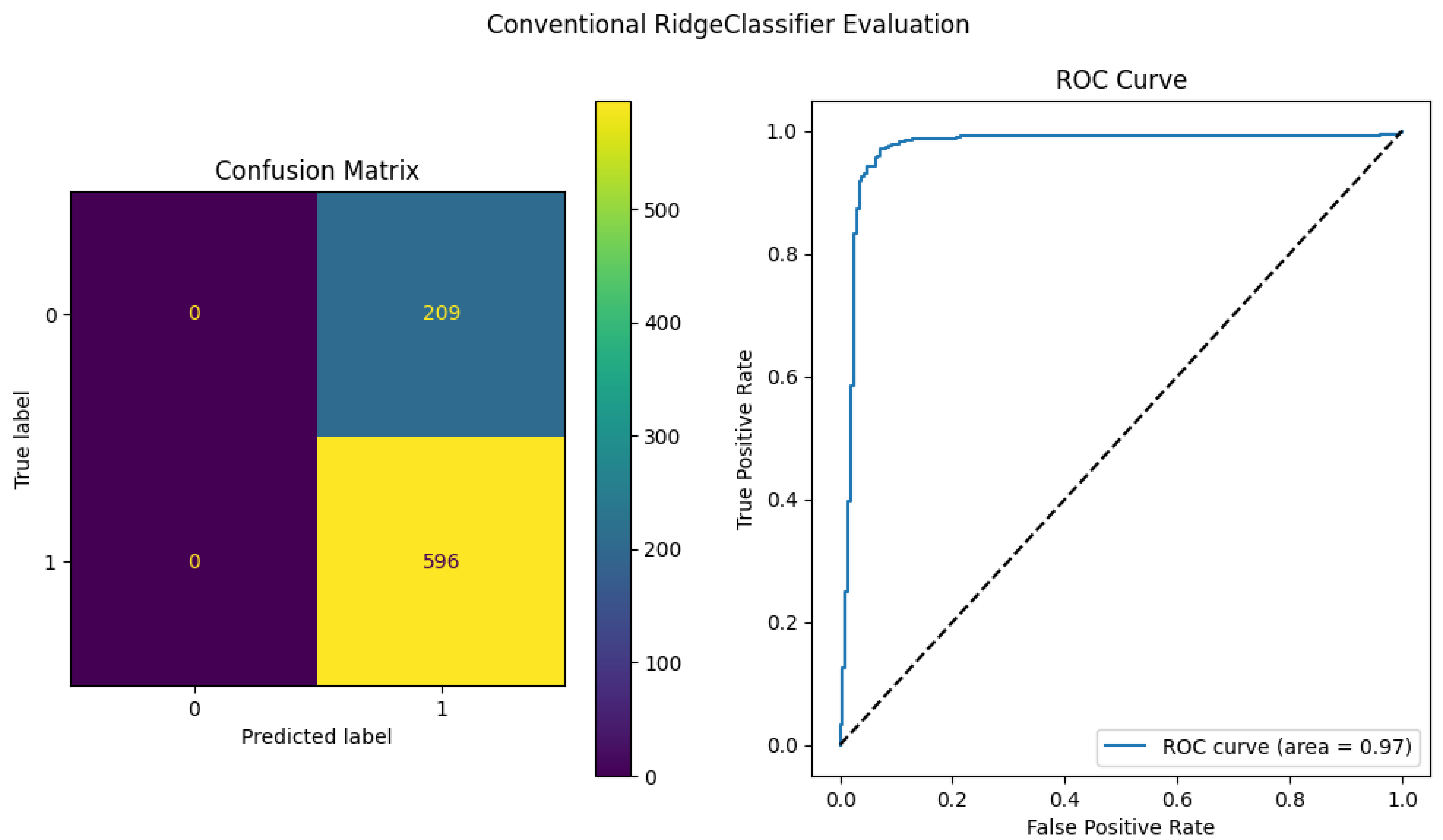
94%

Logistic Regression

Accuracy

THE RESULTS

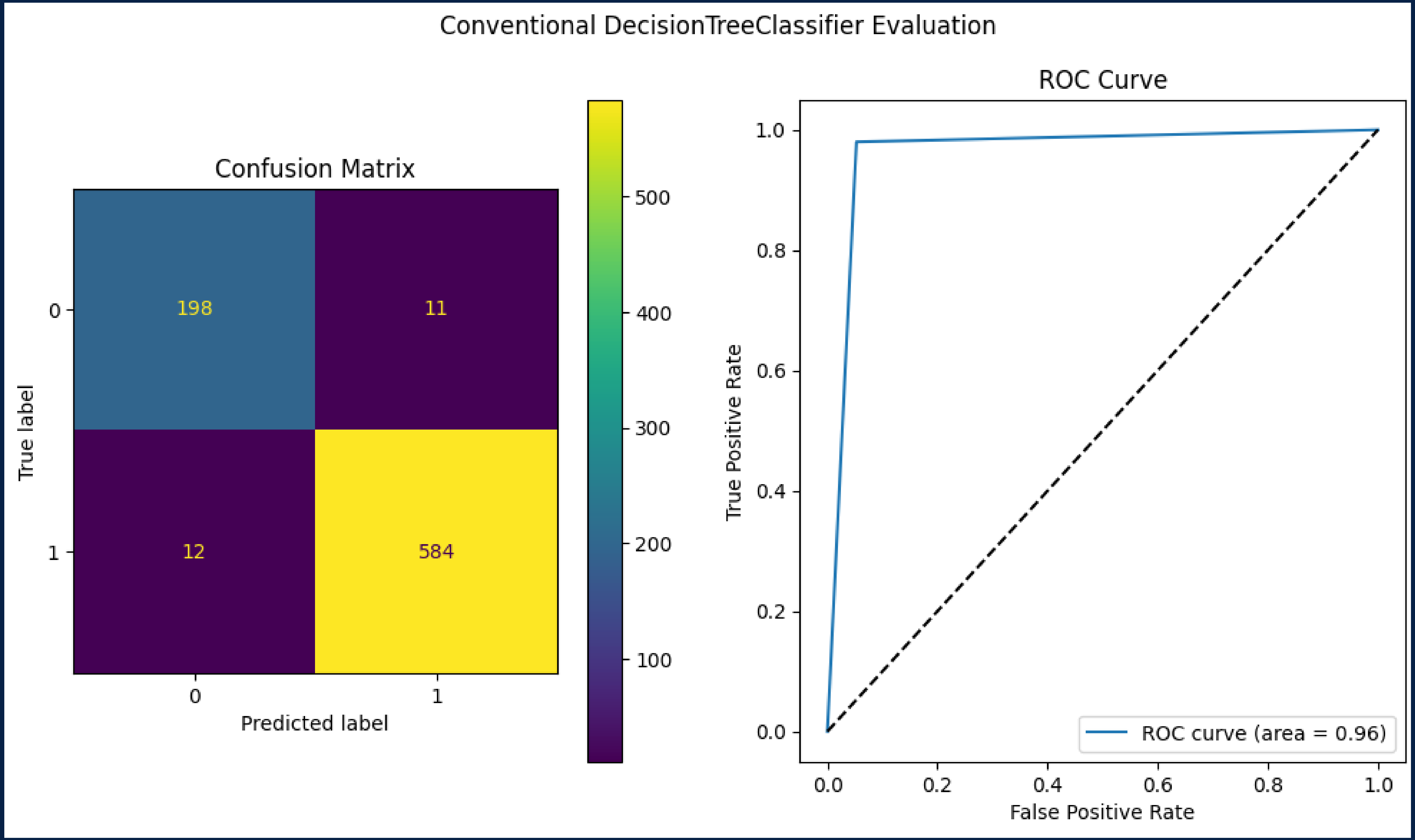
Ridge classifier



Confusion matrix Ridge classifier models

ROC Curve Ridge classifier

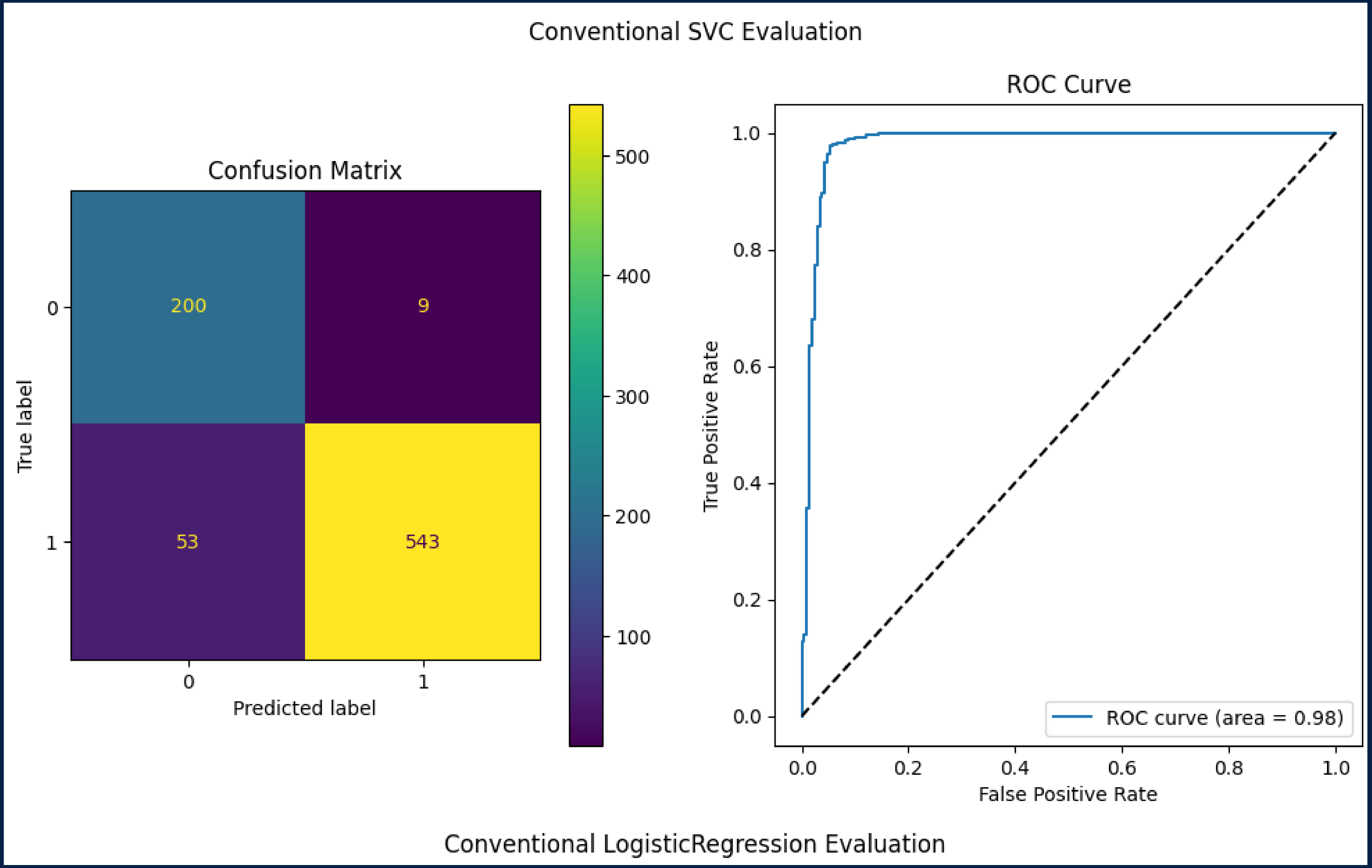
Decision Tree Classifier



Confusion matrix Decision tree classifier

ROC Curve Decision tree classifier

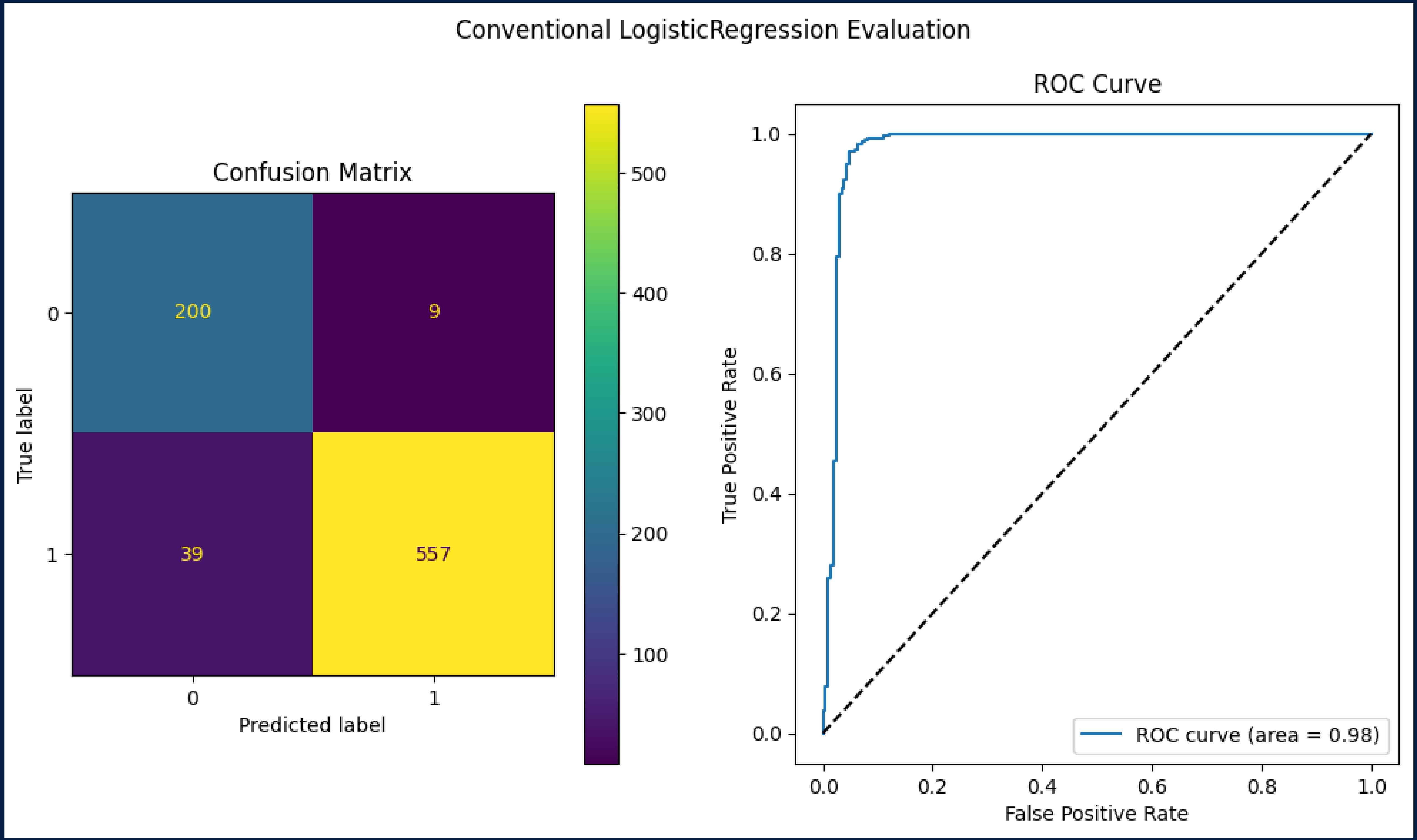
Support Vector Machine



Confusion matrix SVM

ROC Curve SVM

Logistic Regression



Confusion matrix Logistic Regression

ROC Curve Logistic Regression

Tryit Yourself

GRACIAS!