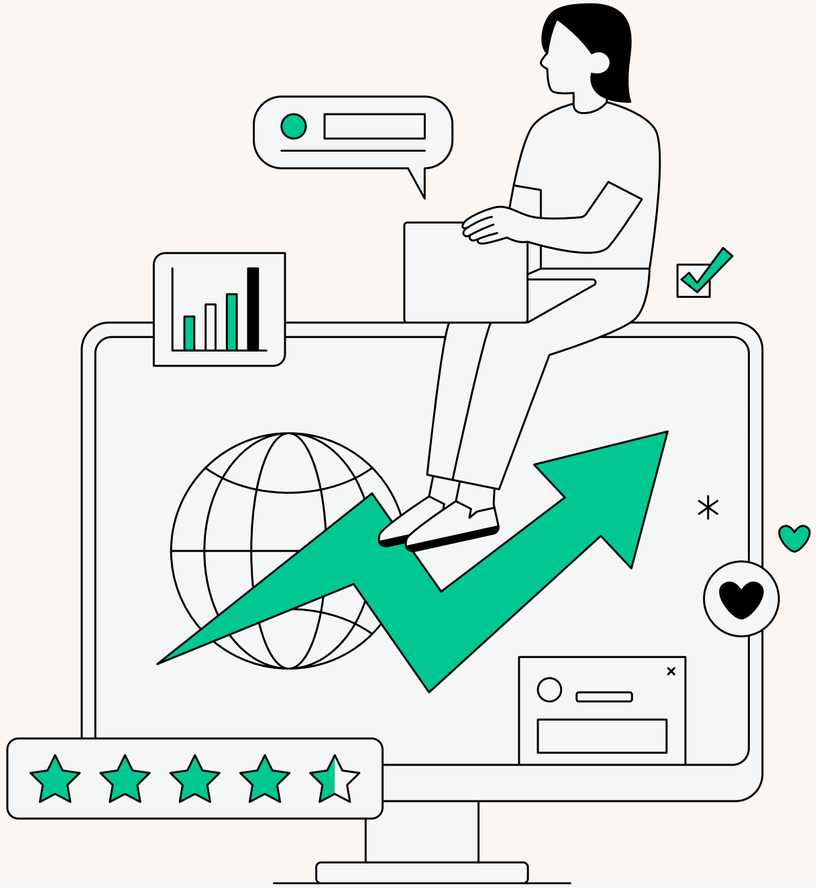
Alissa Crist
BU MET CS 677 - Fall 2023

# Credit Card Fraud

## A classification analysis



# Overview

This project analyzes anonymized, fictionalized data from Nidula Elgiriyewithana on Kaggle*.

**From Kaggle:**

**Dataset Description**: This dataset contains credit card transactions made by European cardholders in the year 2023. It comprises over 550,000 records, and the data has been anonymized to protect the cardholders' identities.

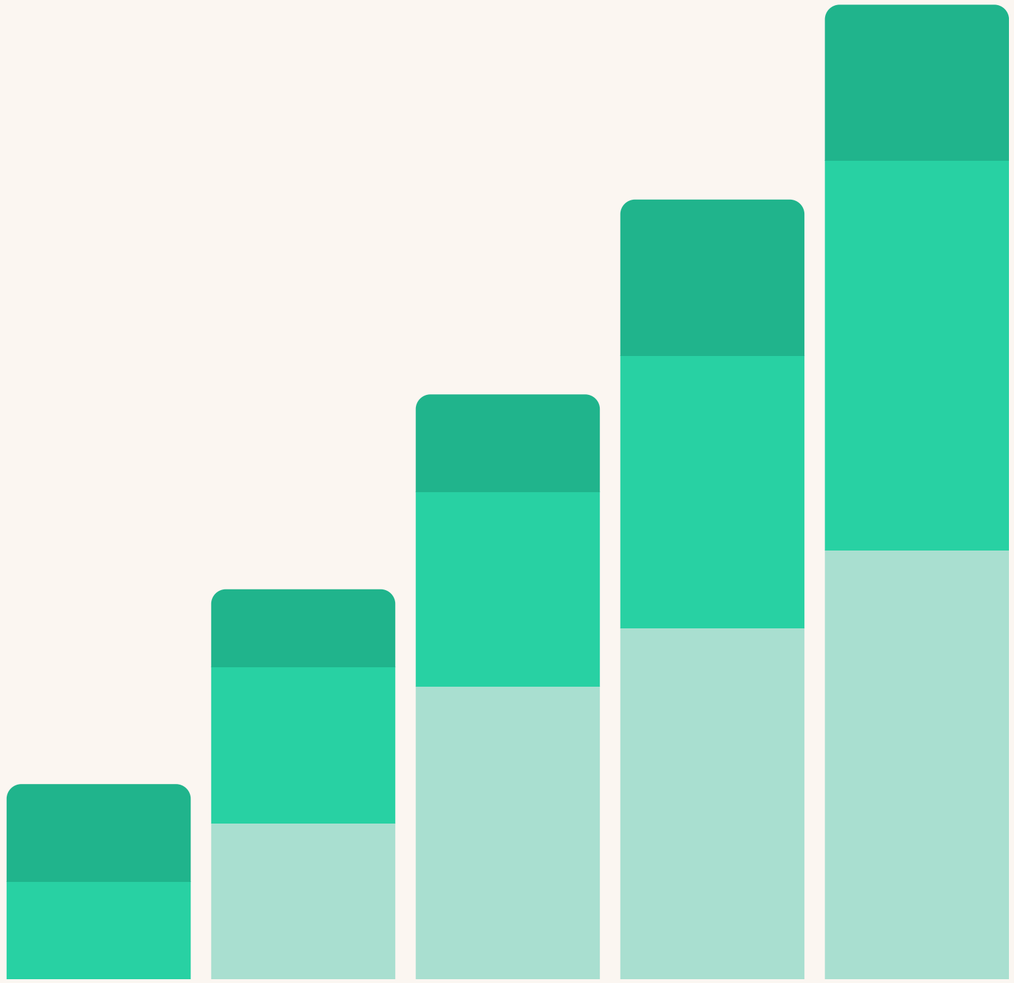
**Key Features of this Dataset:**

- id: Unique identifier for each transaction
- V1-V28: Anonymized features representing various transaction attributes (e.g., time, location, etc.)
- Amount: The transaction amount
- Class: Binary label indicating whether the transaction is fraudulent (1) or not (0)

**Dataset Source: https://www.kaggle.com/datasets/nelgiriyewithana/credit-card-fraud-detection-dataset-2023*

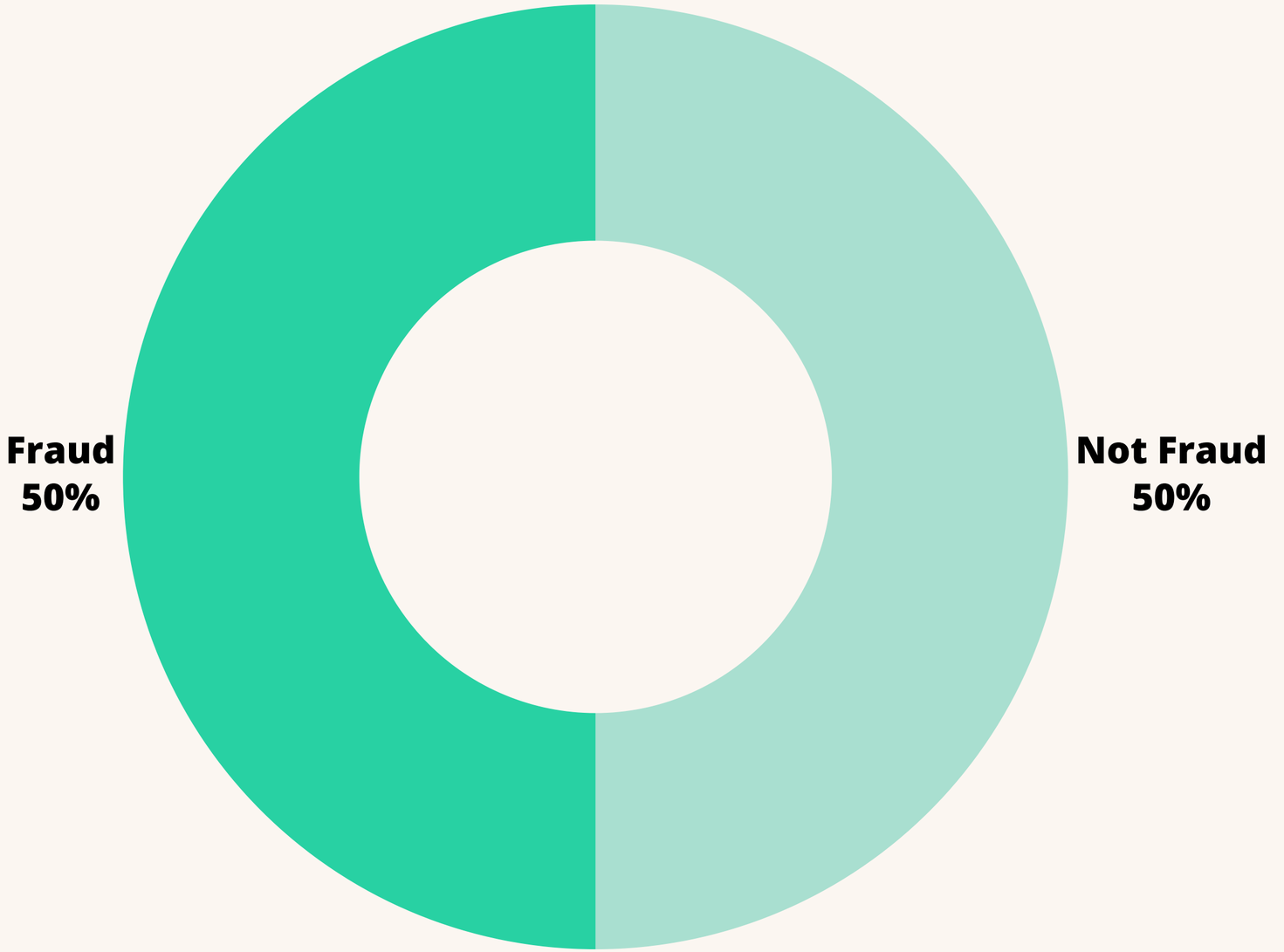
# Methodology used in the analysis

**This project utilizes classification algorithms such as Logistic Regression, k-NN, Naive Bayes', and Decision Tree. k-Means clustering is also used.**



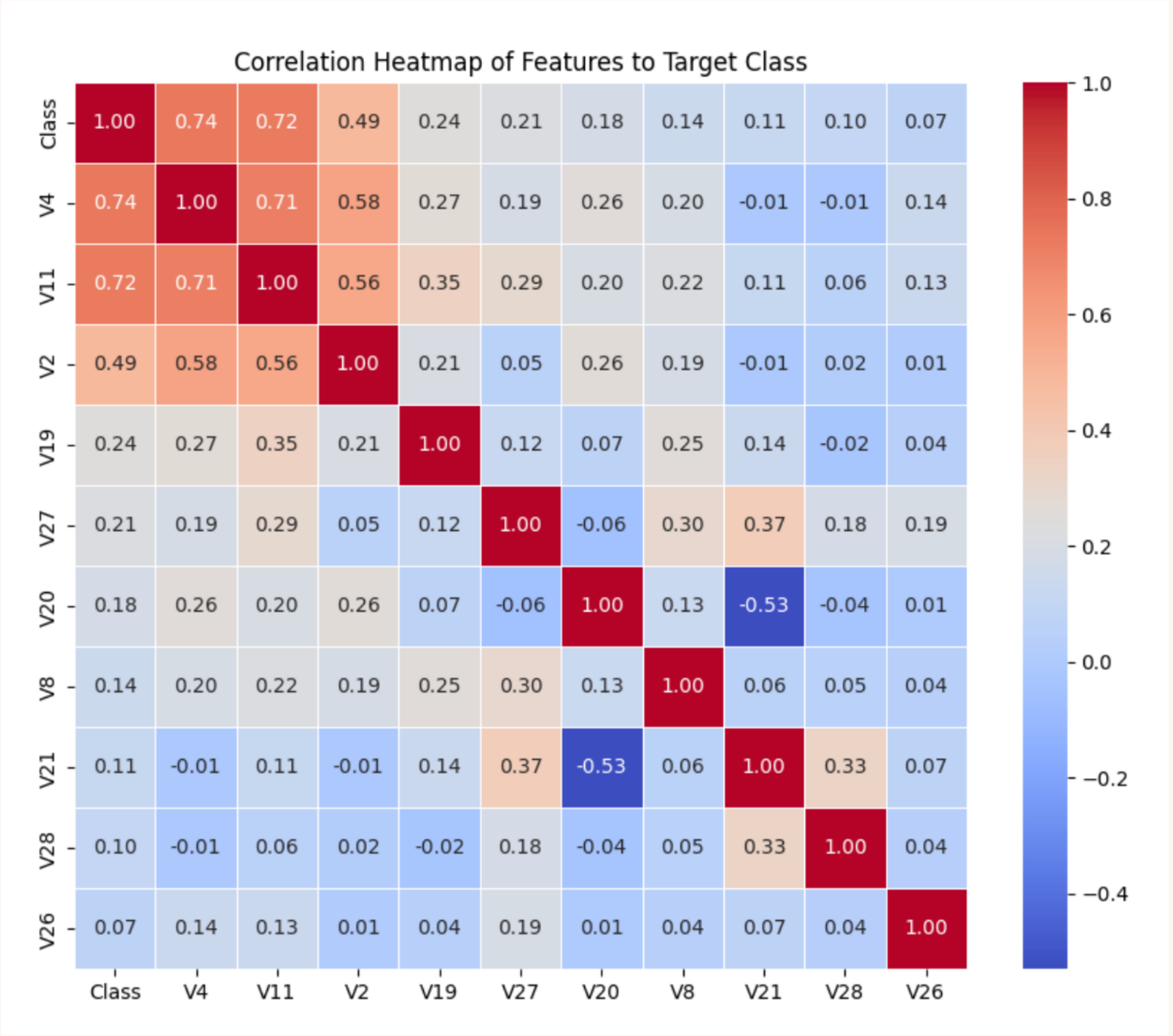
# Class Distribution

The dataset is evenly split between Fraud and Non-Fraud transaction classes



# Highest Correlated Features

| Feature | Correlation Score with Target Class |
|---|---|
| **V4** | 0.74 |
| **V11** | 0.72 |
| **V2** | 0.49 |
| **V19** | 0.24 |
| **V27** | 0.21 |



# Mean & SD for Key Features

| V4 | V11 | V2 | V19 | Amount |
|---|---|---|---|---|
| **Mean** | **Mean** | **Mean** | **Mean** | **Mean** |
| -2.88e-17 | -1.18e-16 | -1.32e-16 | 2.48e-17 | 1.20e+04 |
| **SD** | **SD** | **SD** | **SD** | **SD** |
| 1.0 | 1.0 | 1.0 | 1.0 | 6919.64 |

# 95%

## Average accuracy of classification models

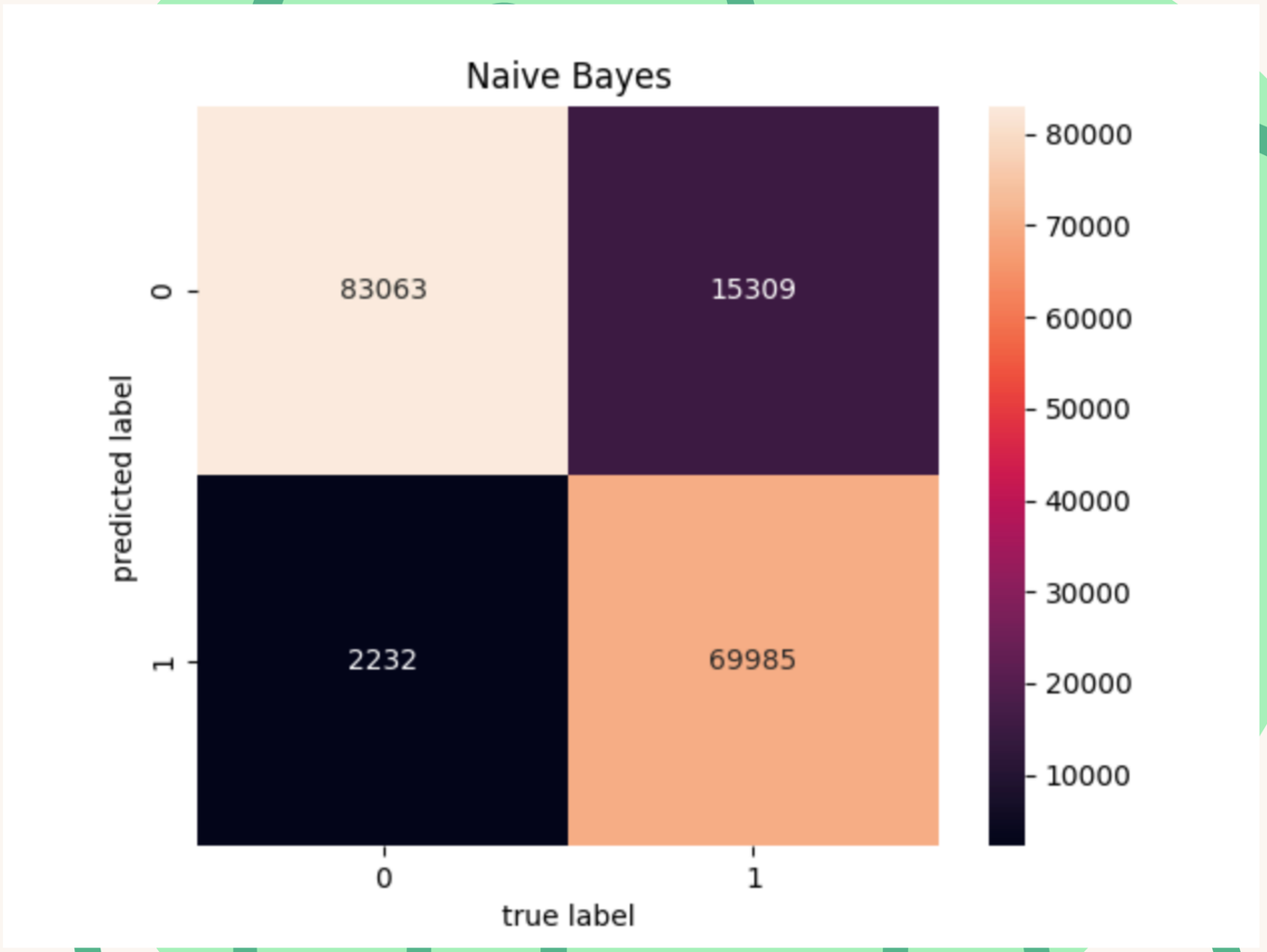
This includes Logistic Regression, k-NN,
Gaussian Naive Bayes', and Decision Tree

# Gaussain Naive Bayes'

**Accuracy:** 90%

**TPR**: 97%

**TNR**: 84%

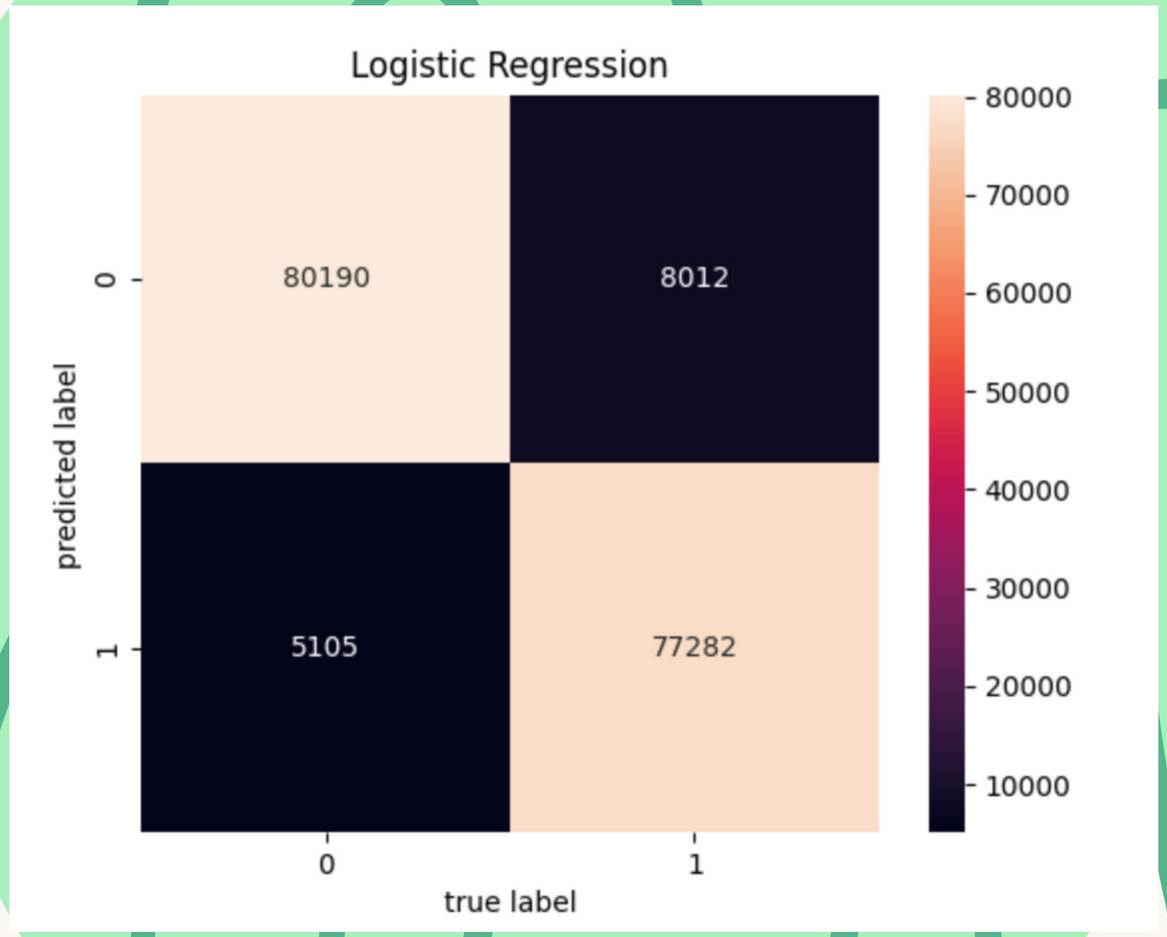


# Logistic Regression

**Accuracy:** 92%

**TPR**: 94%

**TNR**: 91%

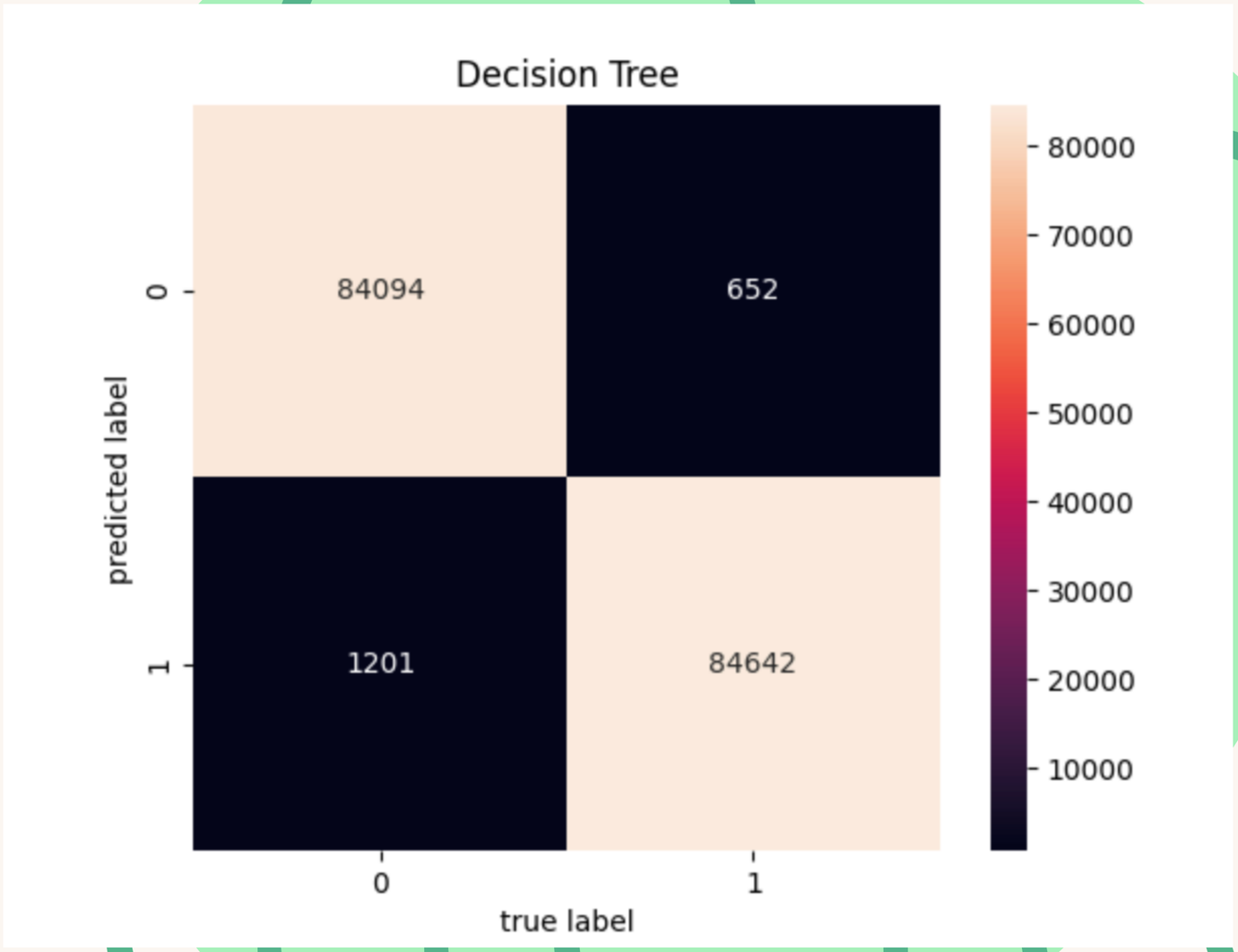


# Decision Tree

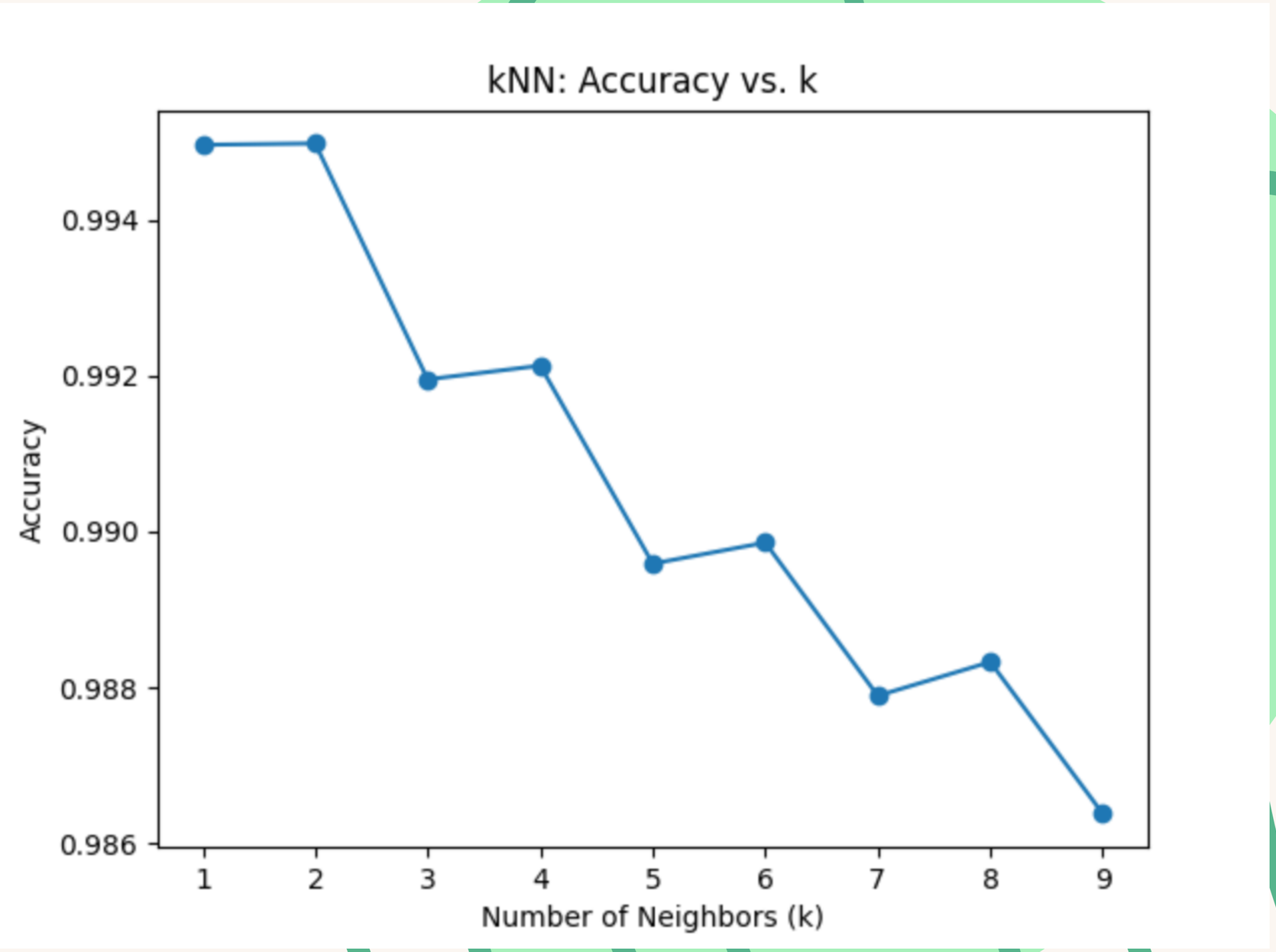
**Accuracy:** 99%

**TPR**: 99%

**TNR**: 99%



# k-NN
# Finding k

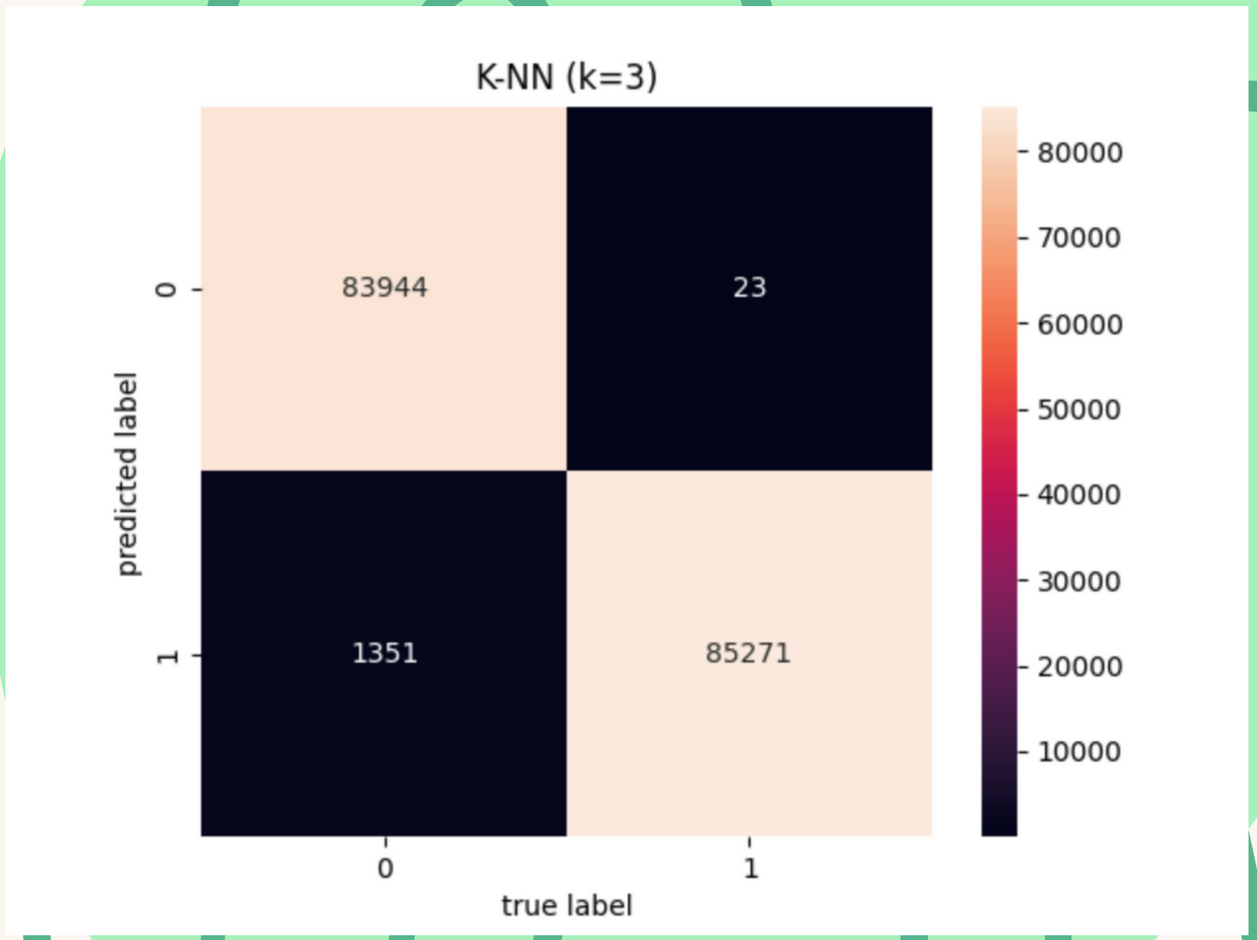


# k-NN (3)

**Accuracy:** 99%

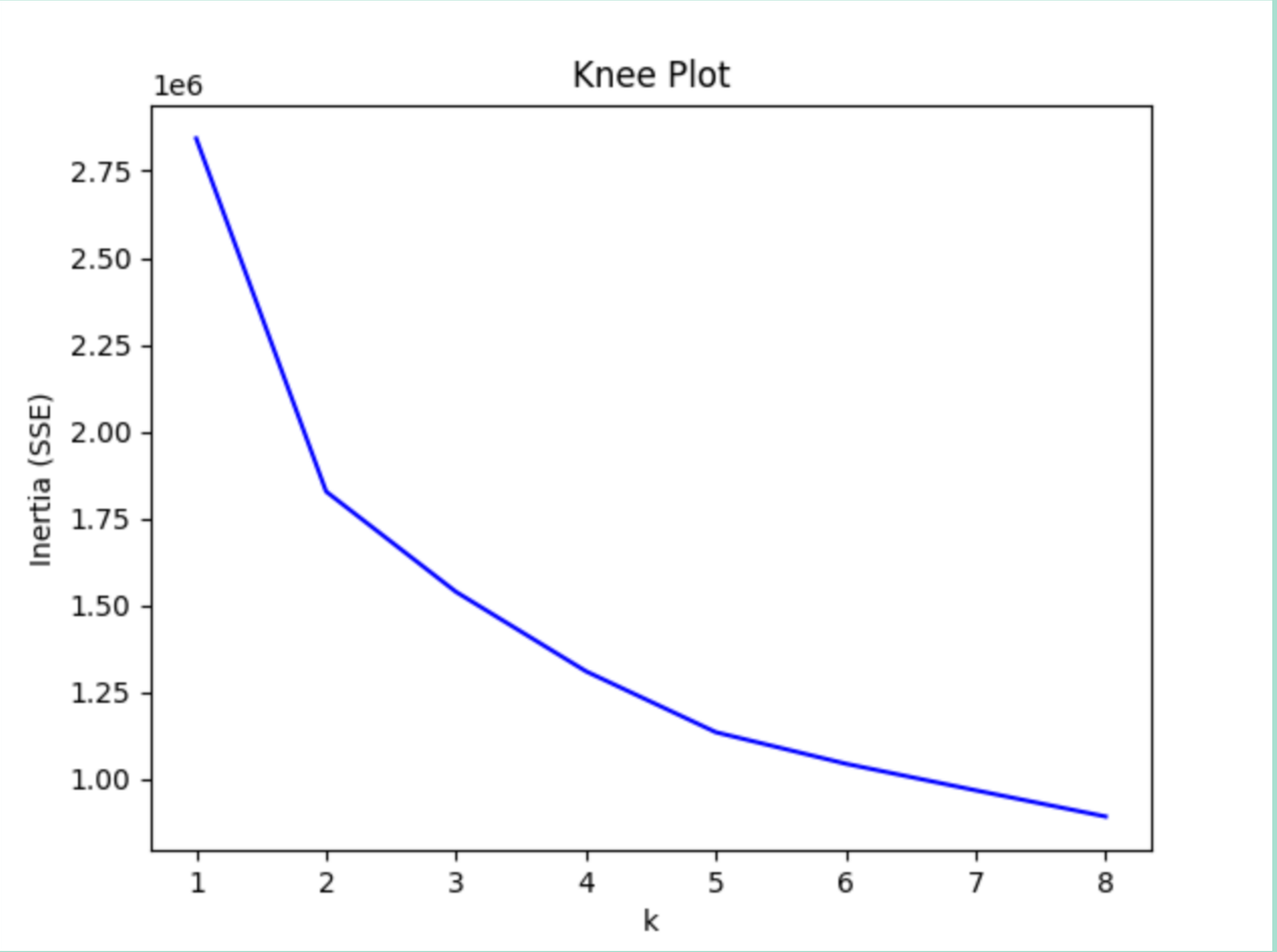
**TPR**: 98%

**TNR**: ~100%

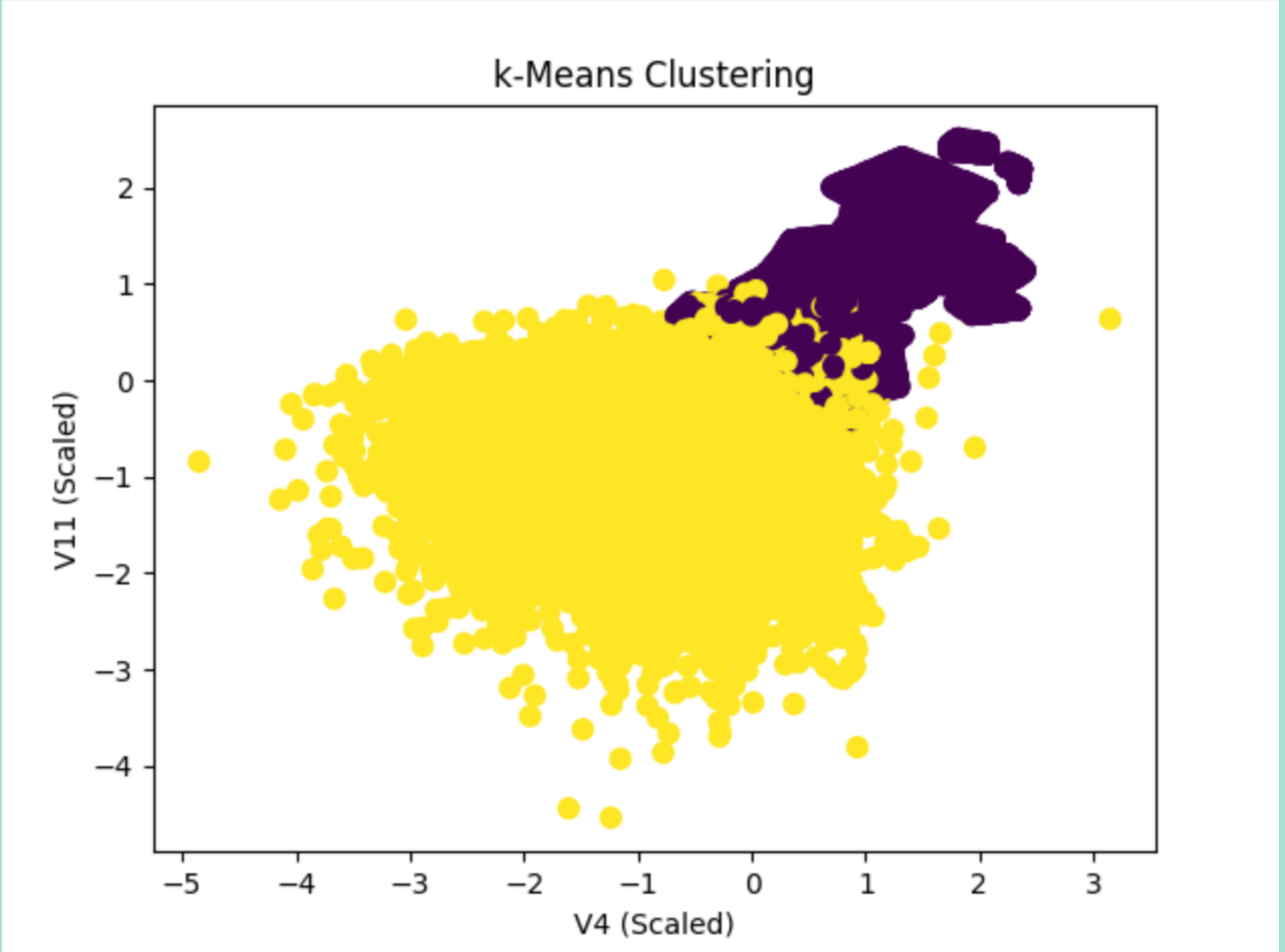


# k-Means Clustering

"Elbow/Knee"Method to select optimal number of clusters (2)



# k-Means Clustering



# Summary

- Machine Learning is essential in classification problems to predict a result based on given features
- Can be better to approach data “blind” to avoid biased analysis
- On real-world data, other methods for categorical and continuous data analysis could be applied to each feature to uncover more trends