**Fraudulent Transactions Prediction Dataset Explanation**

**Team Members:**

* Andrii Prykhodko 51773650
* Artur Okseniuk 20258565
* Maksym Kashevarov 37544938
* Mark Shafarenko 83967814
* Serhii Shymko 52264213
* Karim Saliev 84082025
* Tae Yoon ...

**Summary of the Dataset**

Credit card fraud is a significant issue for financial institutions worldwide, leading to substantial monetary losses annually. Detecting fraudulent transactions effectively is crucial to mitigate these losses. This project aims to identify fraudulent transactions using machine learning.

**Dataset containment:**

* Number of entries: 284,807
* Number of columns: 31
* Column types:
  + 30 columns of type float64
  + 1 column of type int64 (Class)

**Key columns**

Class: Indicates whether a transaction is fraudulent (1) or legitimate (0).

**Type of problem: classification**

The problem at hand is a classification problem. The goal is to classify whether a credit card transaction is fraudulent based on the features provided. Classification problems involve predicting categorical outcomes and, in this case, there are two categories: fraudulent (1) and legitimate (0).

**Dependent and Independent Variables**

**Dependent Variable (Y Target): Class**

This is what we are trying to predict (whether the transaction is fraudulent or not). The dependent variable, also known as the target variable, is the outcome that we are trying to predict. In this dataset, the dependent variable is the 'Class' column. It indicates whether the transaction is fraudulent (1) or legitimate (0).

**Independent Variables (X Features)**

The independent variables, also known as features, are the inputs used to predict the dependent variable. In this dataset, there are 30 features that represent various transaction attributes transformed by PCA (Principal Component Analysis). These features include: V1, V2, V3, ..., V28, along with Amount (transaction amount) and Time (time elapsed from the first transaction).

Изображение выглядит как диаграмма, снимок экрана, текст, линия

Автоматически созданное описание  
After constructing a decision tree with a depth of 6, we achieved an accuracy of 99.94%. To better understand the resulting decision tree, we visualized it in a plot to examine its details. From this diagram we can manually simplify the algorithm:

* **If V17 <= -2.677**:
  + **False**: If V14 <= -4.219:
    - **False**: All subsequent nodes can be removed due to a consistent result of false.
  + **True**: If V12 <= -1.925:
    - **True**: If V14 <= -3.465:
* **True**: If V26 <= -0.255:
  + **False**: All subsequent nodes can be removed due to a consistent result of true.
  + **True**: If V27 <= 1.208:
    - **True**: All subsequent nodes can be removed due to a consistent result of true.
* **False**: If V28 <= 0.113:
  + **False**: All subsequent nodes can be removed due to a consistent result of false.