

A stack of several credit cards is shown, fanned out slightly. The top card is dark grey or black with a silver chip and a silver signature strip. Below it, a red card is visible, followed by a gold card, a blue card, and another gold card. The bottom-most card is partially obscured but shows a silver chip. The cards are resting on a dark, textured surface.

# Credit card Fraud Detection

# Why this project?

---

- Credit card fraud can cause financial loss and identity theft to both financial institutions and their clients.
- A fraud detection system is required to prevent such fraudulent activities and protect the customers' identities.
- According to a report published by the Nilson Report in December 2021, the global losses due to credit card fraud was \$28.58 billion in 2020.

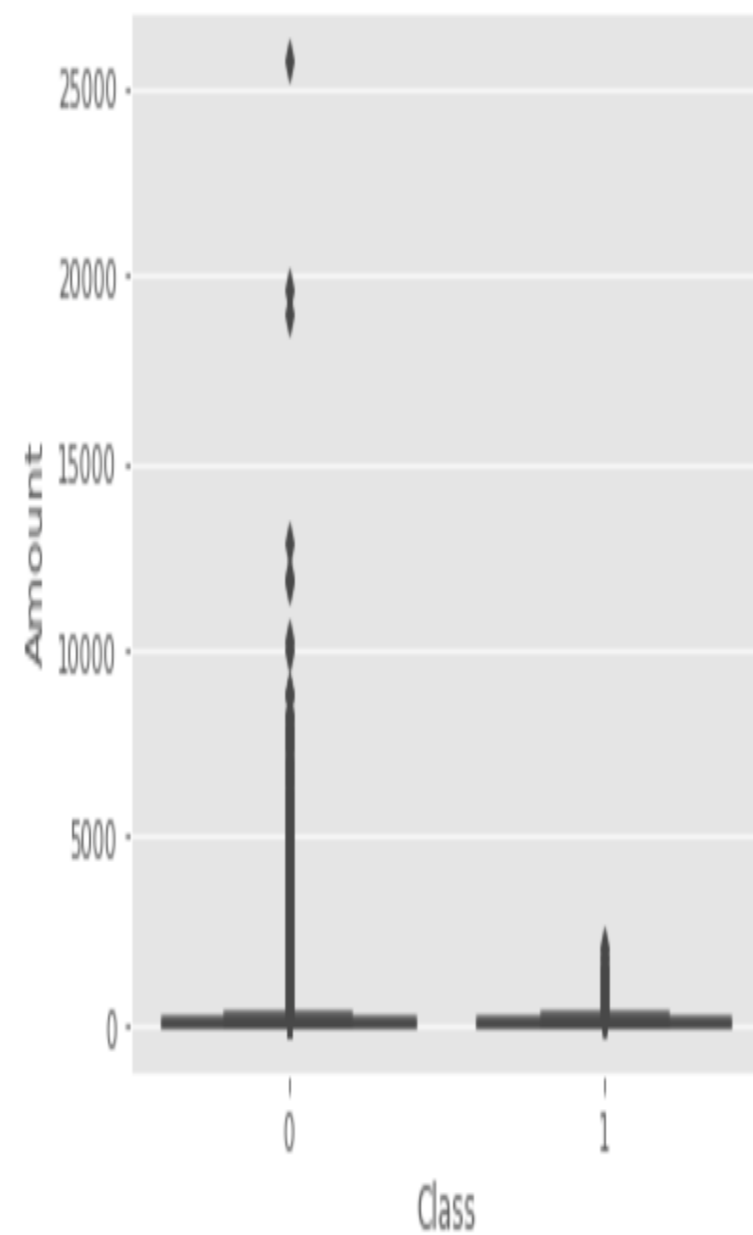
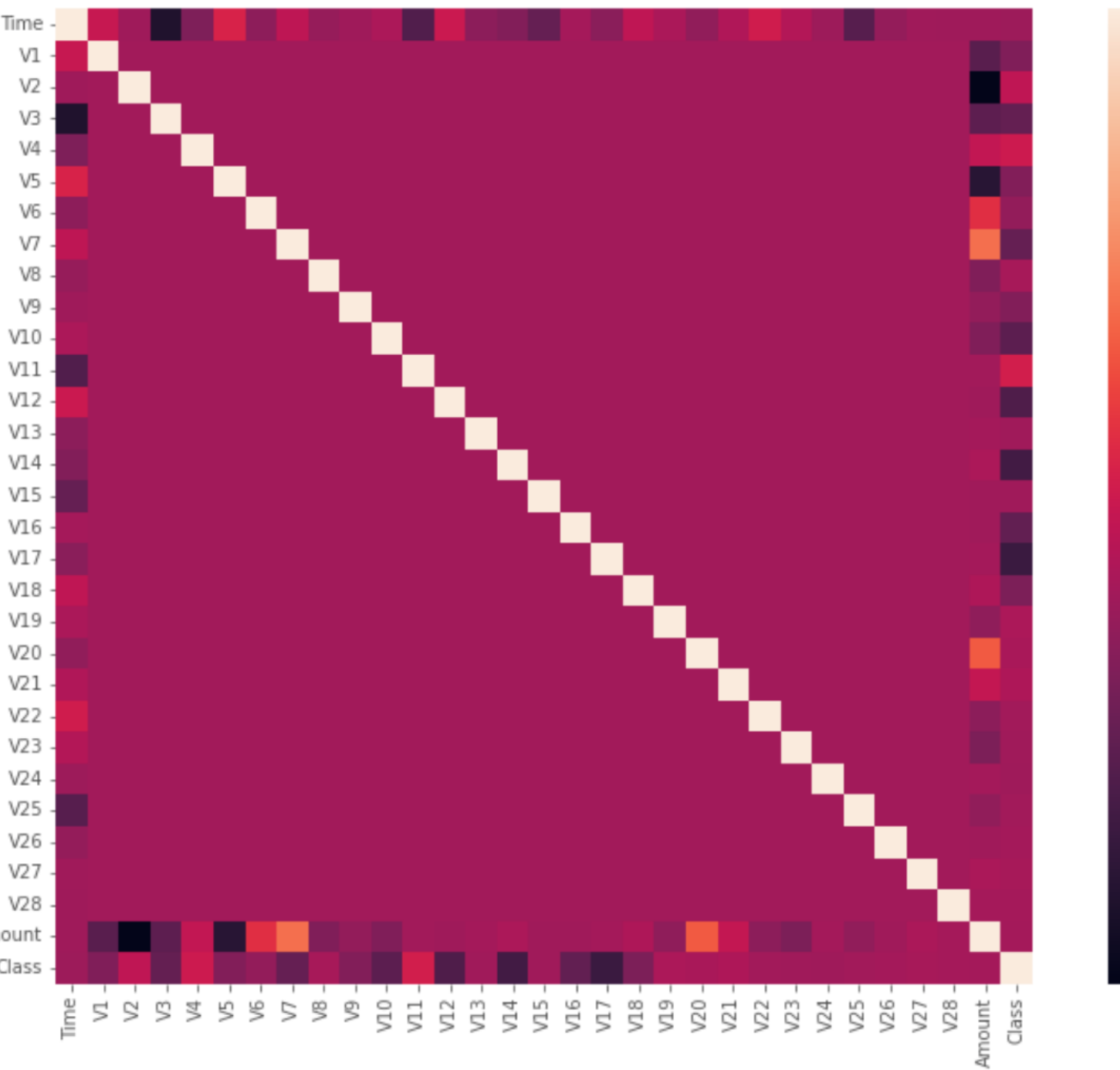
# Dataset

- 
- Source : Kaggle
  - <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>
  - The dataset contains credit card transactions that took place in September 2013 in Europe, involving a total of 284,807 transactions.
  - 492 Fraudulent Transactions
  - The dataset is imbalanced as the proportion of fraudulent transactions is very low.

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V2
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...	-0.018307	0.277838	-0.11047
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	...	-0.225775	-0.638672	0.10128
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	...	0.247998	0.771679	0.90941
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	...	-0.108300	0.005274	-0.19032
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	...	-0.009431	0.798278	-0.13745
5	2.0	-0.425966	0.960523	1.141109	-0.168252	0.420987	-0.029728	0.476201	0.260314	-0.568671	...	-0.208254	-0.559825	-0.02639
6	4.0	1.229658	0.141004	0.045371	1.202613	0.191881	0.272708	-0.005159	0.081213	0.464960	...	-0.167716	-0.270710	-0.15410
7	7.0	-0.644269	1.417964	1.074380	-0.492199	0.948934	0.428118	1.120631	-3.807864	0.615375	...	1.943465	-1.015455	0.05750
8	7.0	-0.894286	0.286157	-0.113192	-0.271526	2.669599	3.721818	0.370145	0.851084	-0.392048	...	-0.073425	-0.268092	-0.20423
9	9.0	-0.338262	1.119593	1.044367	-0.222187	0.499361	-0.246761	0.651583	0.069539	-0.736727	...	-0.246914	-0.633753	-0.12079

# Exploratory data analysis

- The data was visualized using histograms, box plots, and scatter plots to gain insights.
- Investigated the correlation between different features to identify potential predictors of fraud.



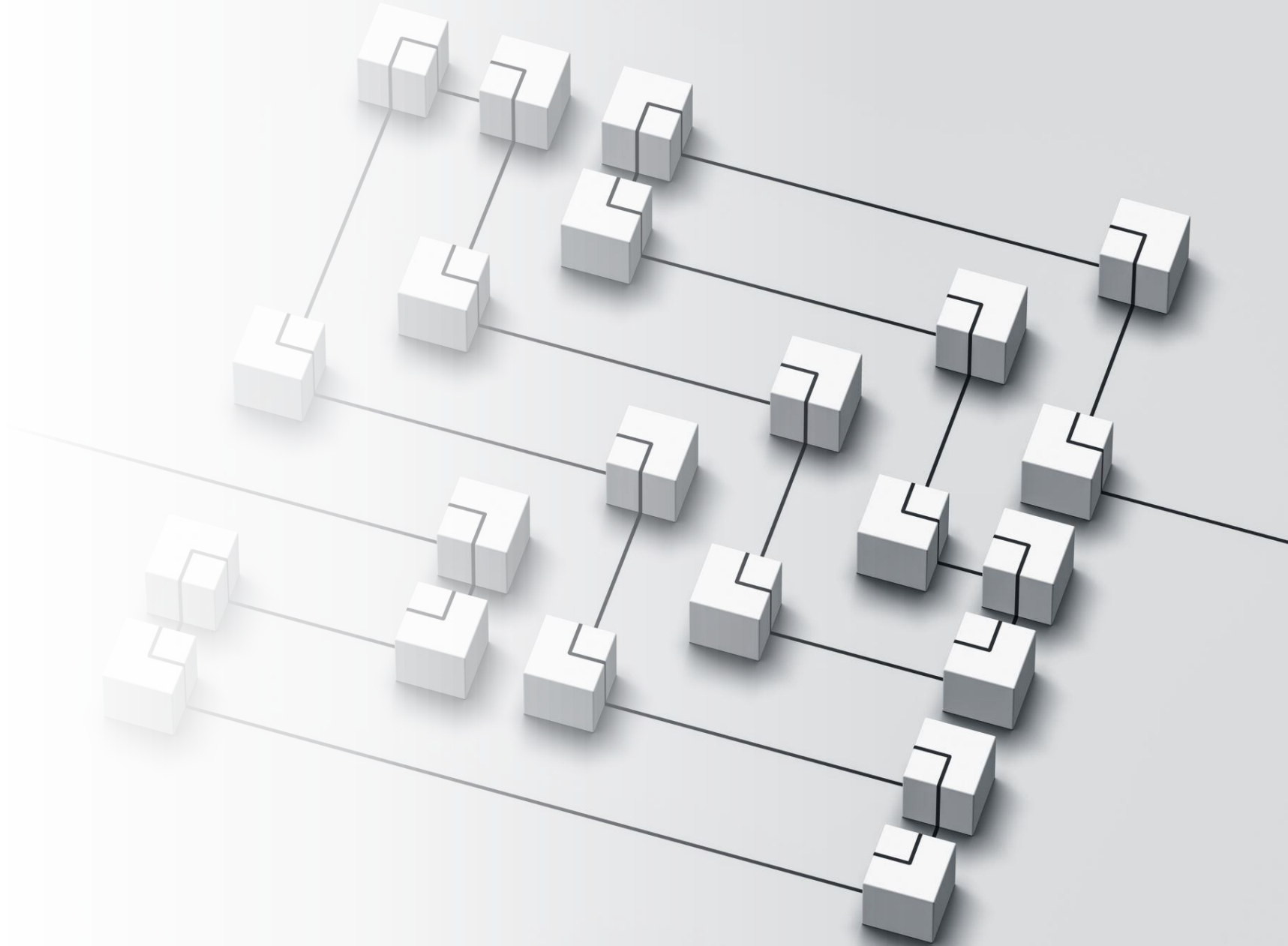
# Data preprocessing

- The time and amount variables were scaled to have similar ranges using a standard scaling method.
- The dataset was down sampled using random under-sampling to balance the class distribution between fraudulent and non-fraudulent transactions.


```
0      1040  
1       393  
Name: Class, dtype: int64
```

## Model

- Used several machine learning models, including XGBoost, logistic regression, decision trees, random forests, and support vector machines.
- Evaluated the models using different metrics, such as accuracy, precision, recall, and F1-score.







Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.9991	0.8636	0.5816	0.6951
Decision Tree	0.9991	0.7115	0.7908	0.7488
Random Forest	0.9996	0.9730	0.7714	0.8605
Gradient Boosting	0.9989	0.7375	0.6020	0.6629
AdaBoost	0.9993	0.8554	0.7245	0.7845
K-Nearest Neighbors	0.9995	0.9383	0.7755	0.8492
Support Vector Machine	0.9993	0.9531	0.6224	0.7531

# Challenges

- The dataset was imbalanced, which made it difficult to train models that accurately capture the patterns in the data.
- The absence of information regarding the data collection process and the features created challenges in interpreting the results obtained from the models.

# Future of the project


- Finding a more balanced and informative dataset to improve the efficacy and reliability of the analysis.
- Using advanced machine learning techniques, such as deep learning, to enhance the precision and effectiveness of the models.
- Building a real-time fraud detection system capable of monitoring transactions in real-time and sending alerts to customers and financial institutions to detect and prevent fraudulent activities.

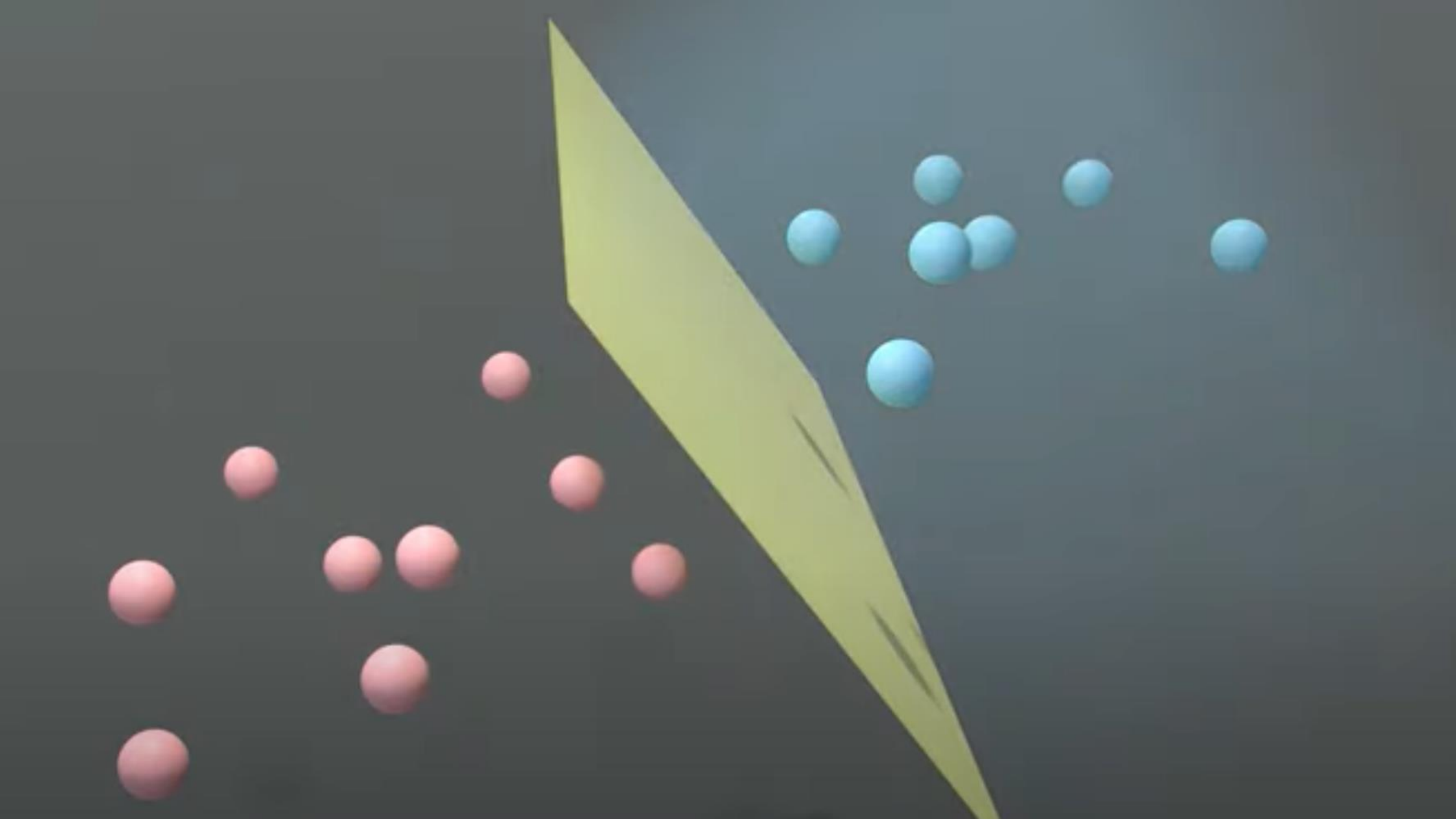


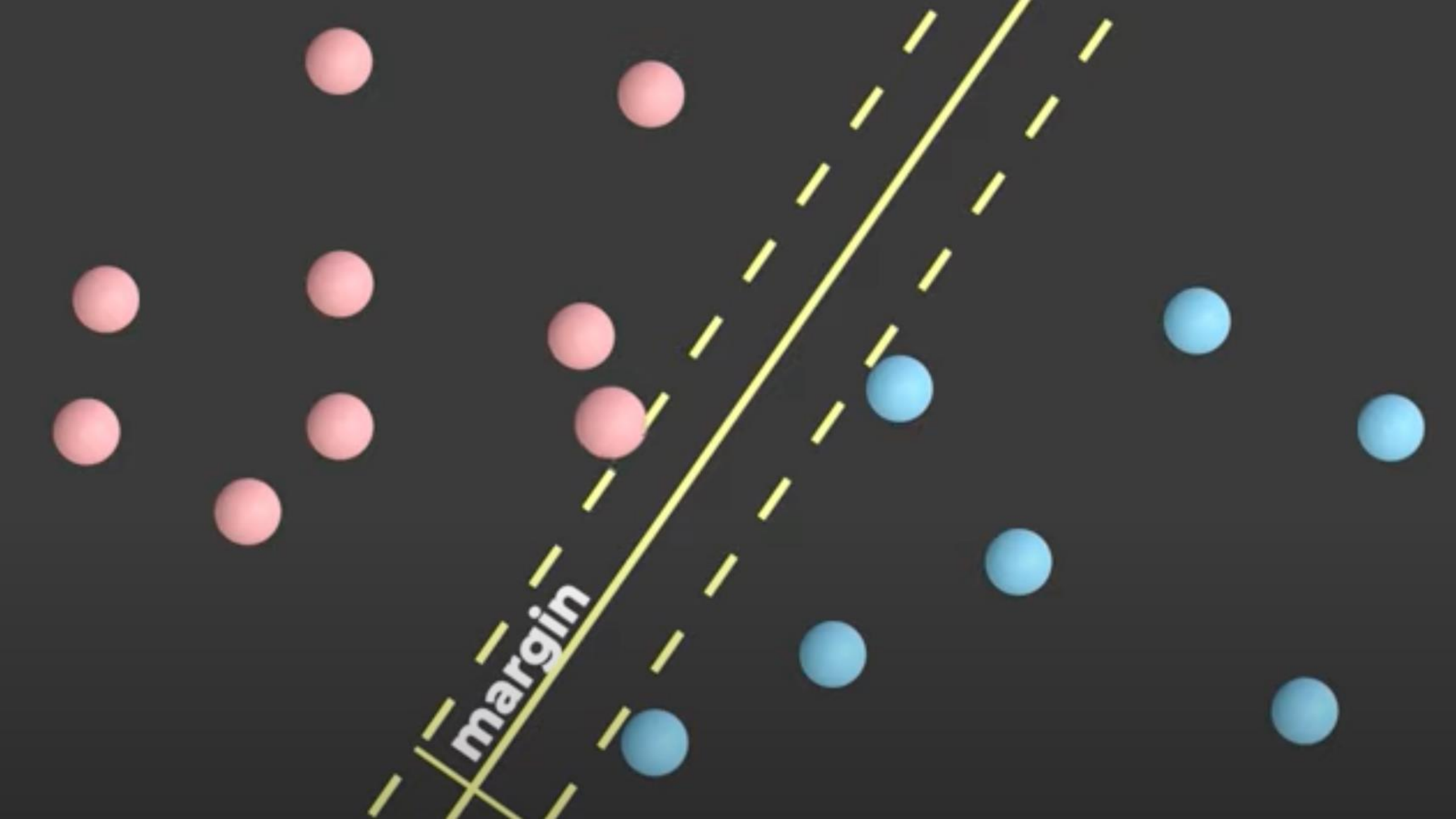
---

# SUPPOT VECTOR MACHINE



- 
- SVM is a supervised machine learning algorithm that can be used for both classification and regression problems.
  - SVM finds a boundary or hyperplane that separates different classes of data points.
  - The hyperplane is chosen to maximize the margin or distance between the hyperplane and the closest data points from each class.
  - This helps the algorithm better classify future data points based on their location in relation to the hyperplane.





# Types of SVM

- SVMs are divided into two main categories: linear and non-linear SVM.
- The linear SVM uses a linear hyperplane to separate the data.
- The non-linear SVM, on the other hand, utilizes kernel functions to transform data into a higher-dimensional space where it can be linearly separated.



# Advantages of SVM

- It can handle high-dimensional data with a small sample size.
- It is efficient.
- SVM can handle non-linearly separable data by using kernel functions to map the data into a higher-dimensional space.
- less prone to overfitting than other classification methods.
- It is flexible.
- It is effective in handling small to medium-sized datasets.

# Disadvantage of SVM

- SVM can be sensitive to the choice of kernel function and its parameters.
- It can be computationally expensive, especially for large datasets.
- It may not work well with noisy data or data with overlapping classes.

# Application of SVM

- SVM is widely used in many areas, including image classification, text classification, bioinformatics, and finance.
- It is used in face recognition systems, spam detection, and gene expression analysis.
- It can be used for predicting stock prices, credit default risk, and fraud detection.