



# CREDIT CARD FRAUD DETECTION

---

RUBY FUNG

# AGENDA

---

PROBLEM STATEMENT

METHODOLOGY

EXPLORATORY DATA ANALYSIS

SUPERVISED LEARNING

UNSUPERVISED LEARNING

CONCLUSION AND RECOMMENDATIONS



# PROBLEM STATEMENT

---

66

+ 104%

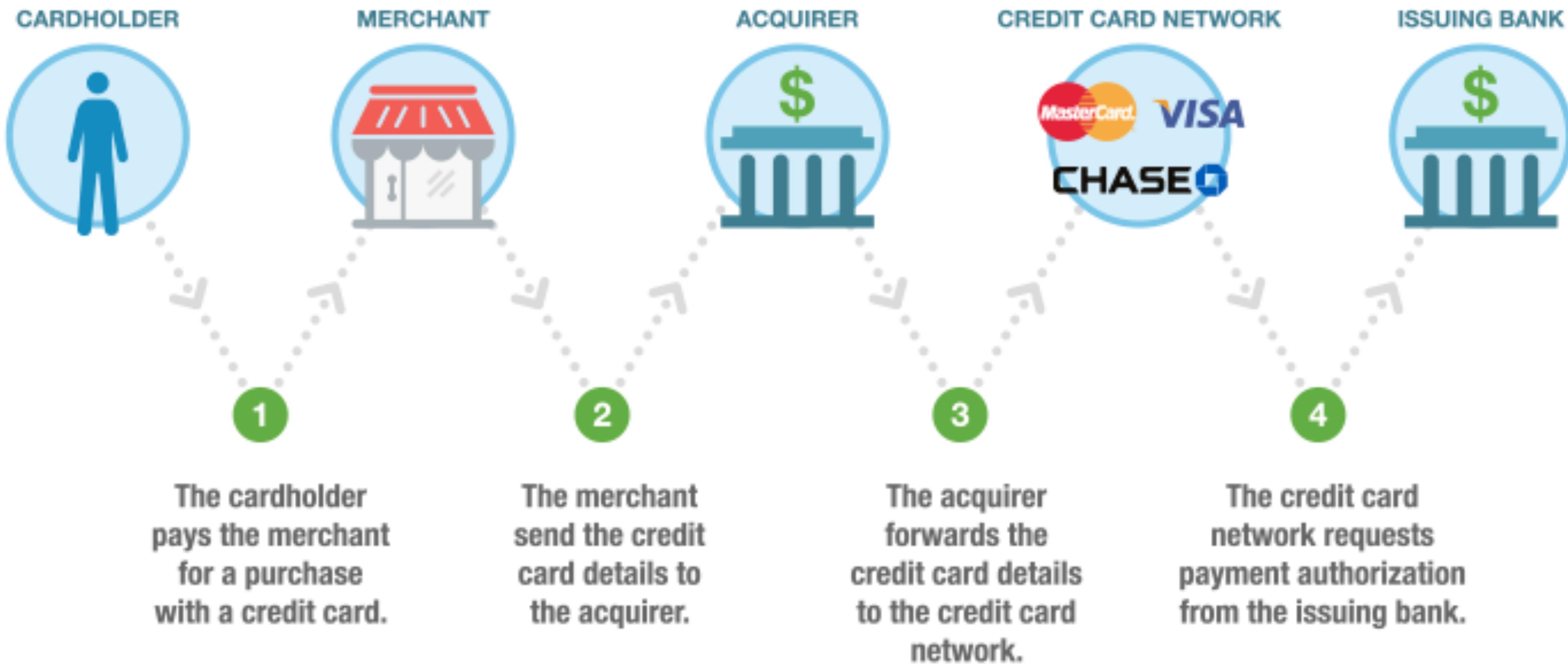
- *Federal Trade Commission*

> \$35 billion

Global card loss by 2020

- *Nilson report*

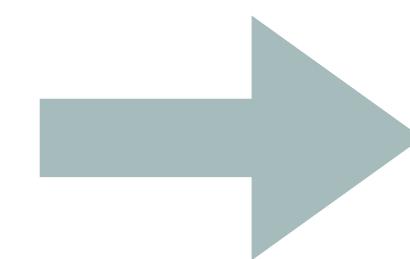
# STAKEHOLDERS



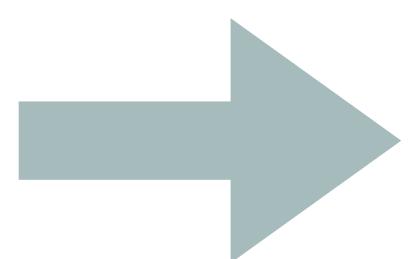
# PROBLEM STATEMENT

---

Credit card fraud  
detection system



Credit card holders:  
Prevent loss



Credit card issuers:  
Reduce cost



# METHODOLOGY

---

# SUPERVISED V.S. UNSUPERVISED

---

	Supervised	Unsupervised
Input Data	Labeled data	Unlabelled data
Number of Classes	Known	Unknown
Types	Classification	Clustering



# EXPLORATORY DATA ANALYSIS

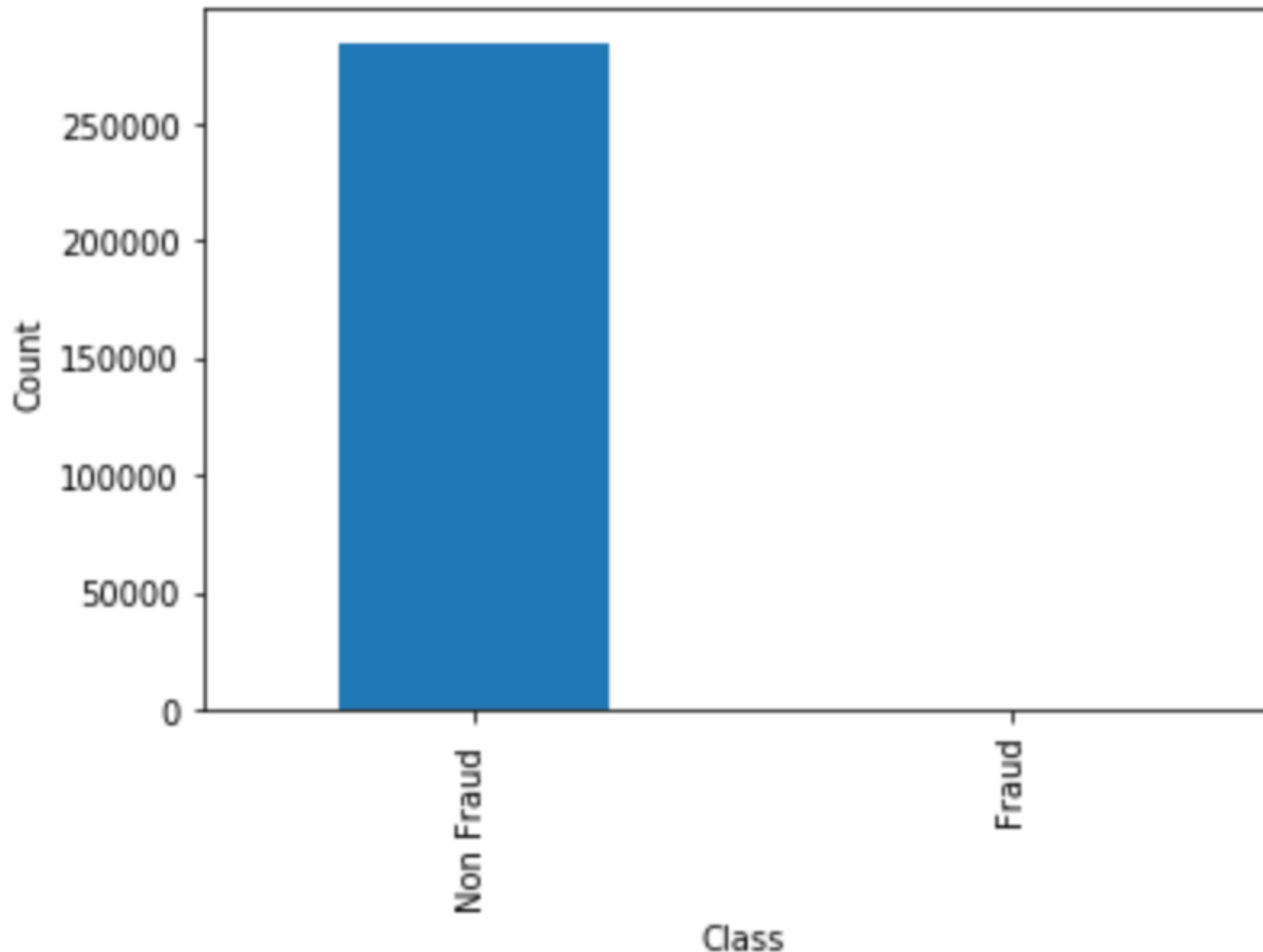
---

# CLASS DISTRIBUTION

---

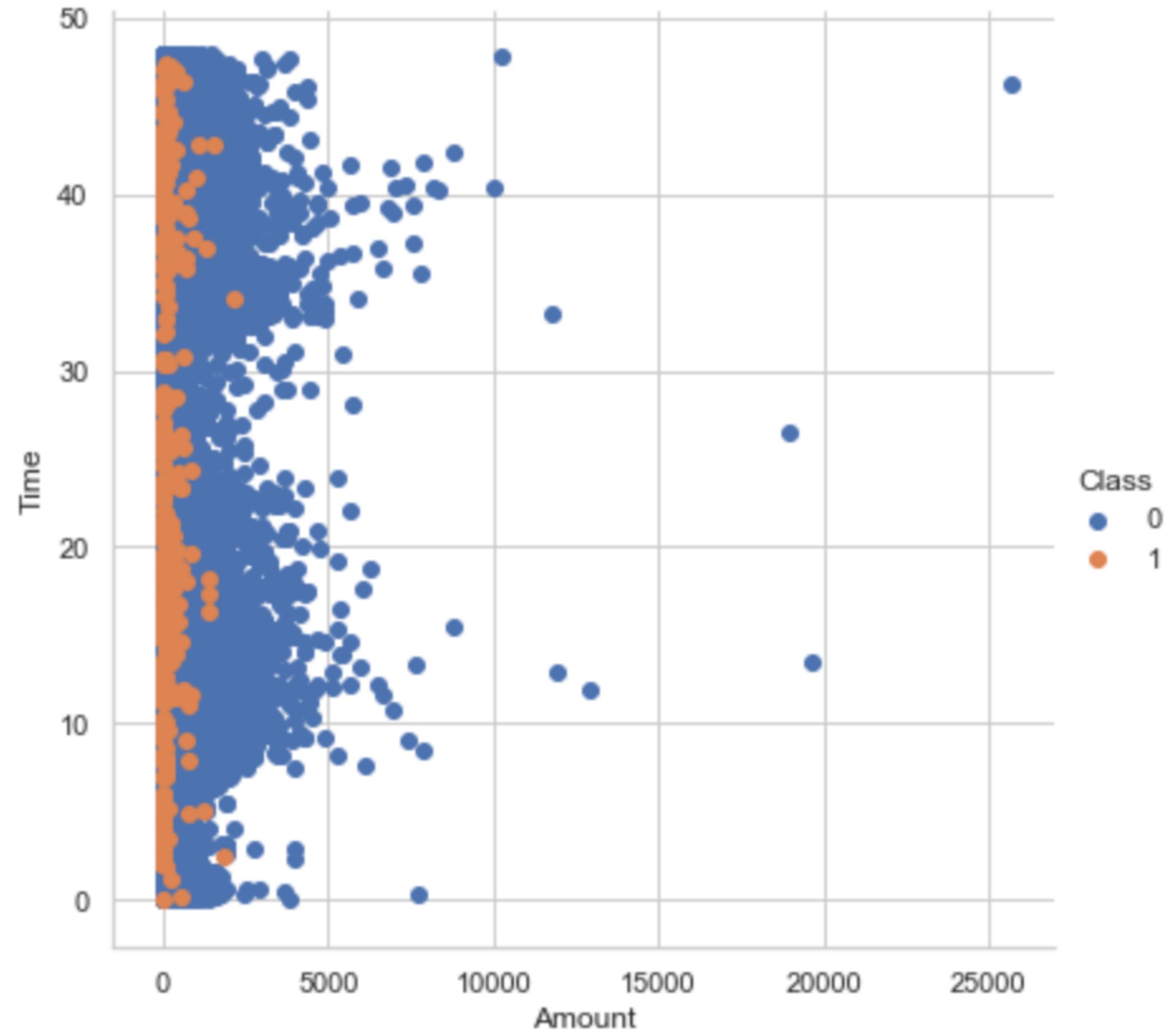
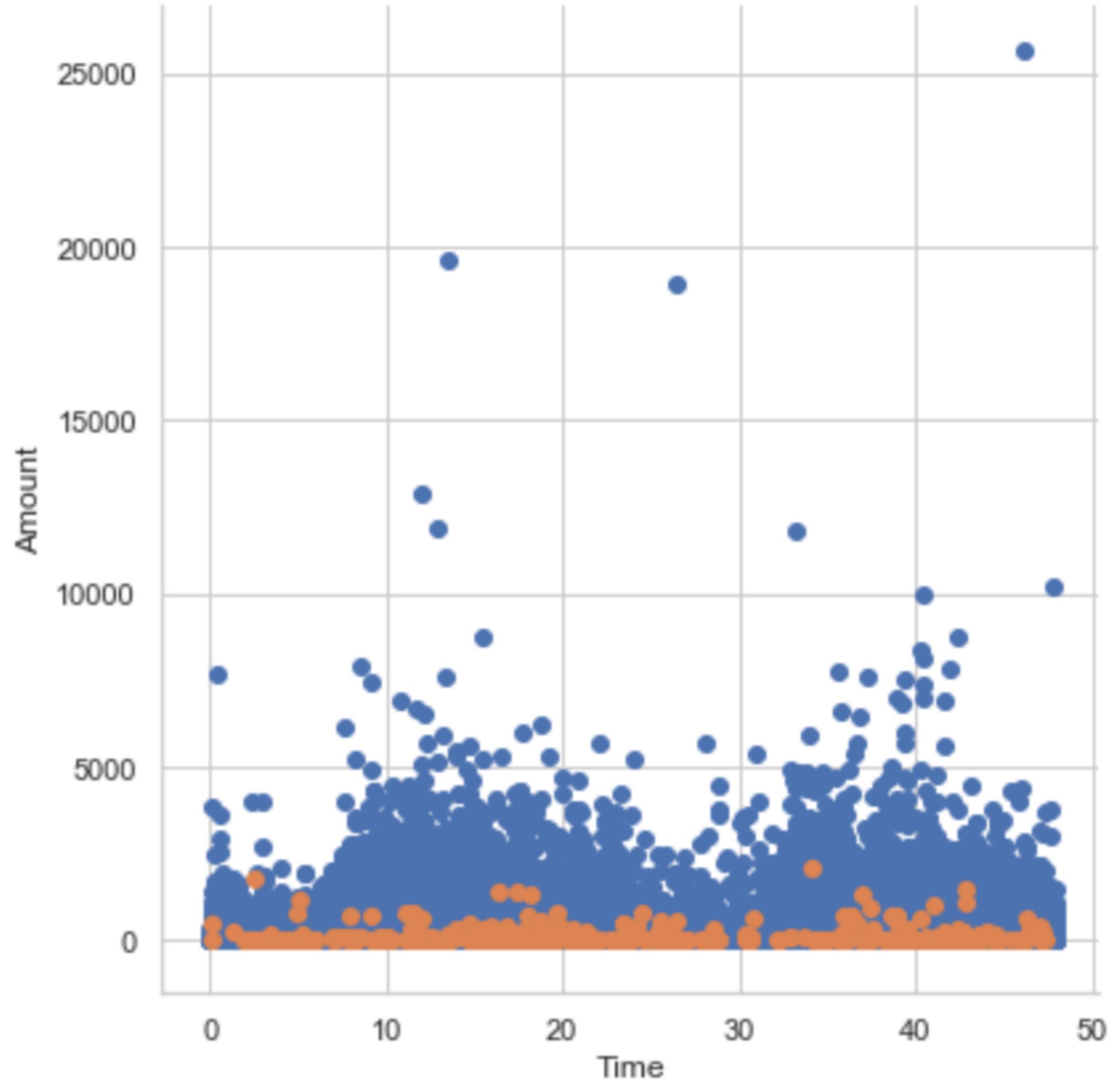
**99.83%**

**0.17%**



# RELATIONSHIP BETWEEN AMOUNT & TIME

---





# SUPERVISED LEARNING

---

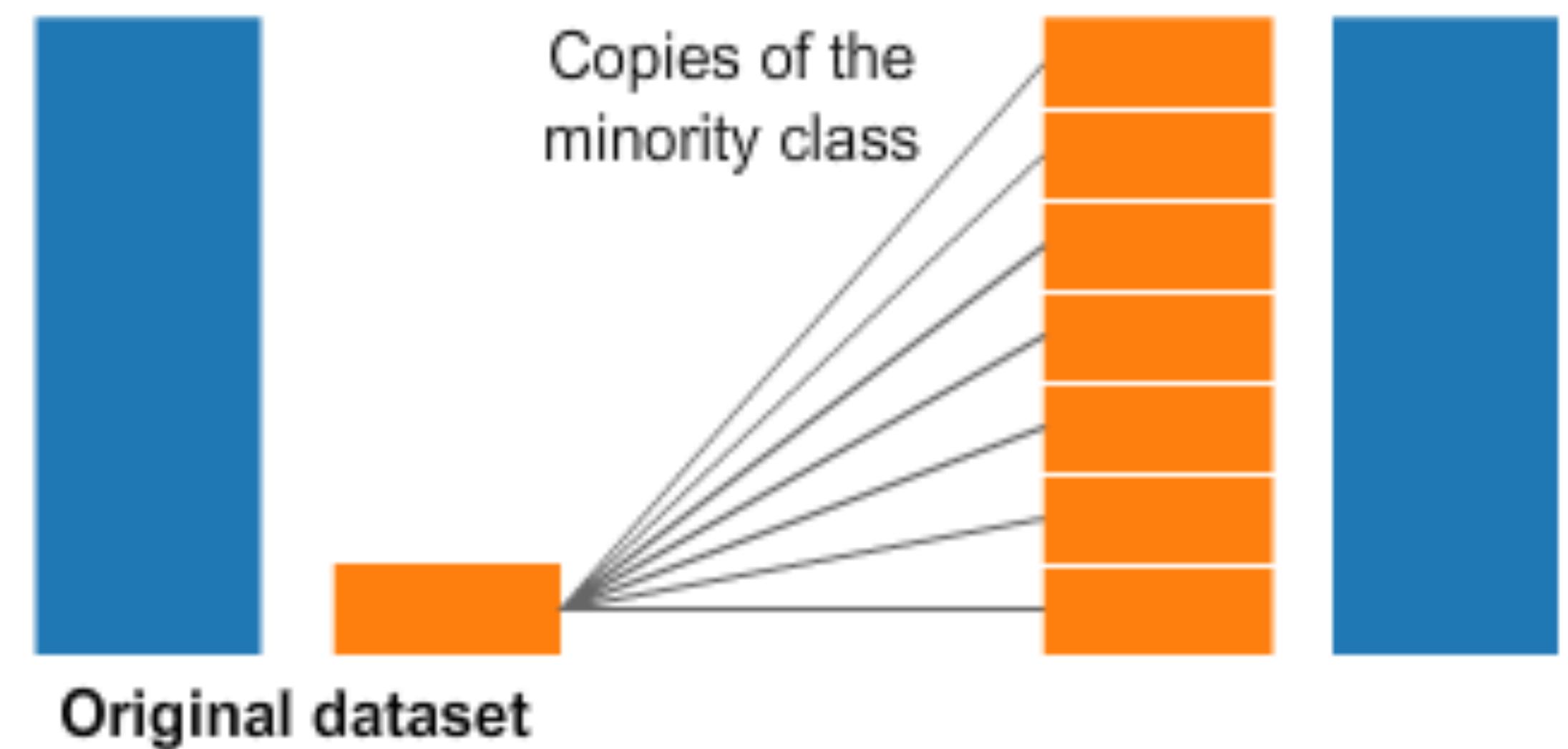
# RESAMPLING METHODS

---

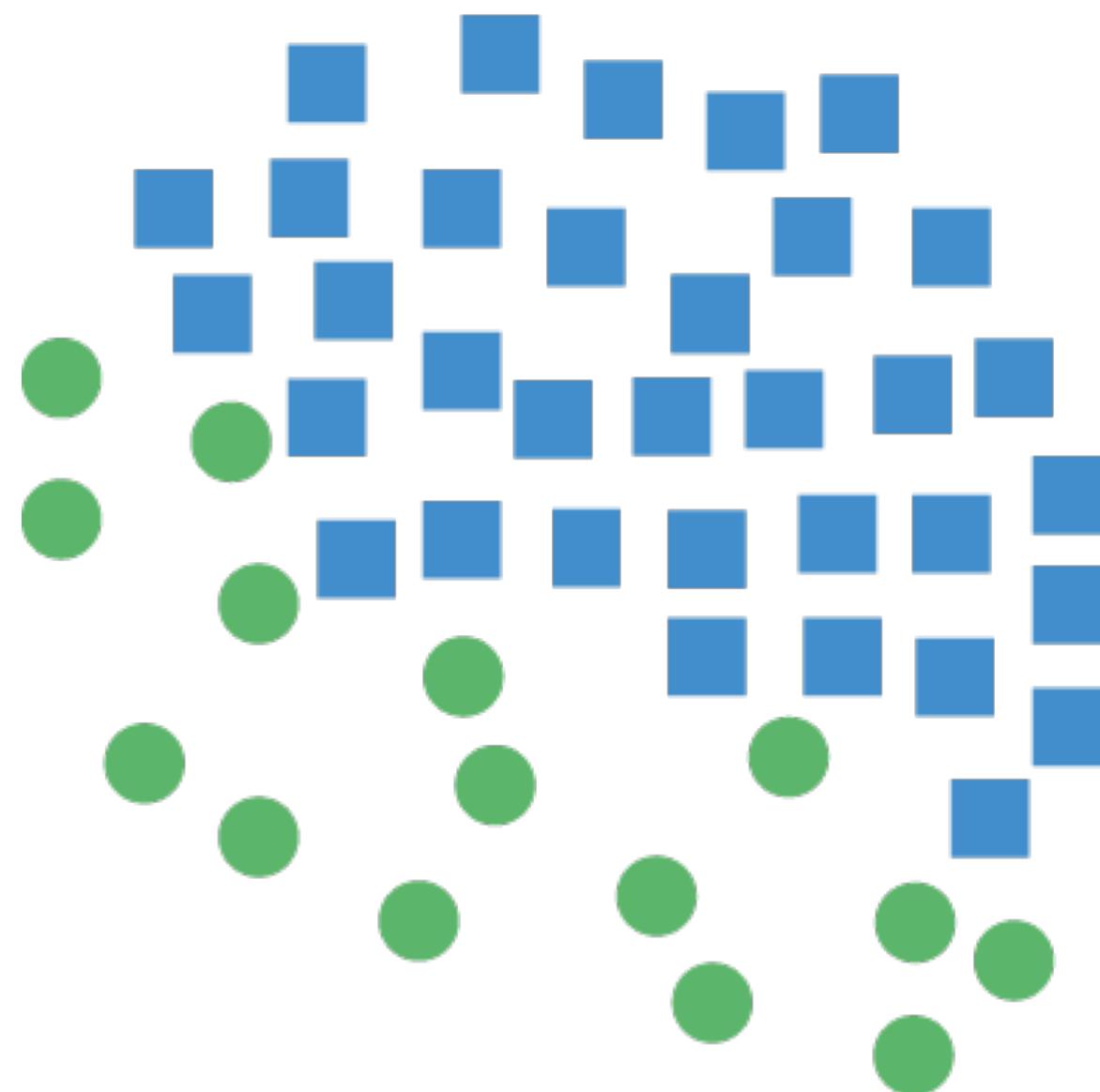
## Undersampling



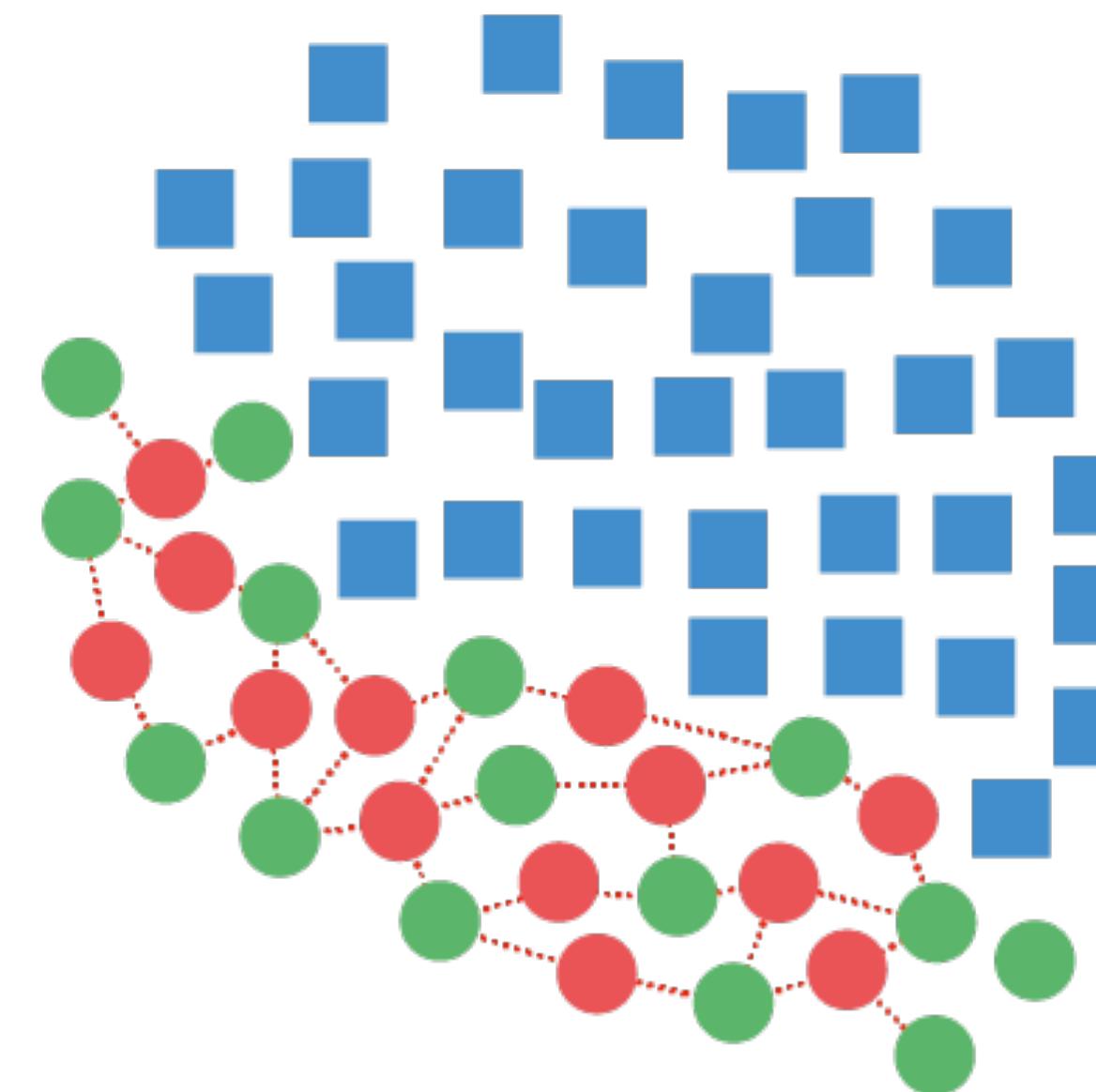
## Oversampling



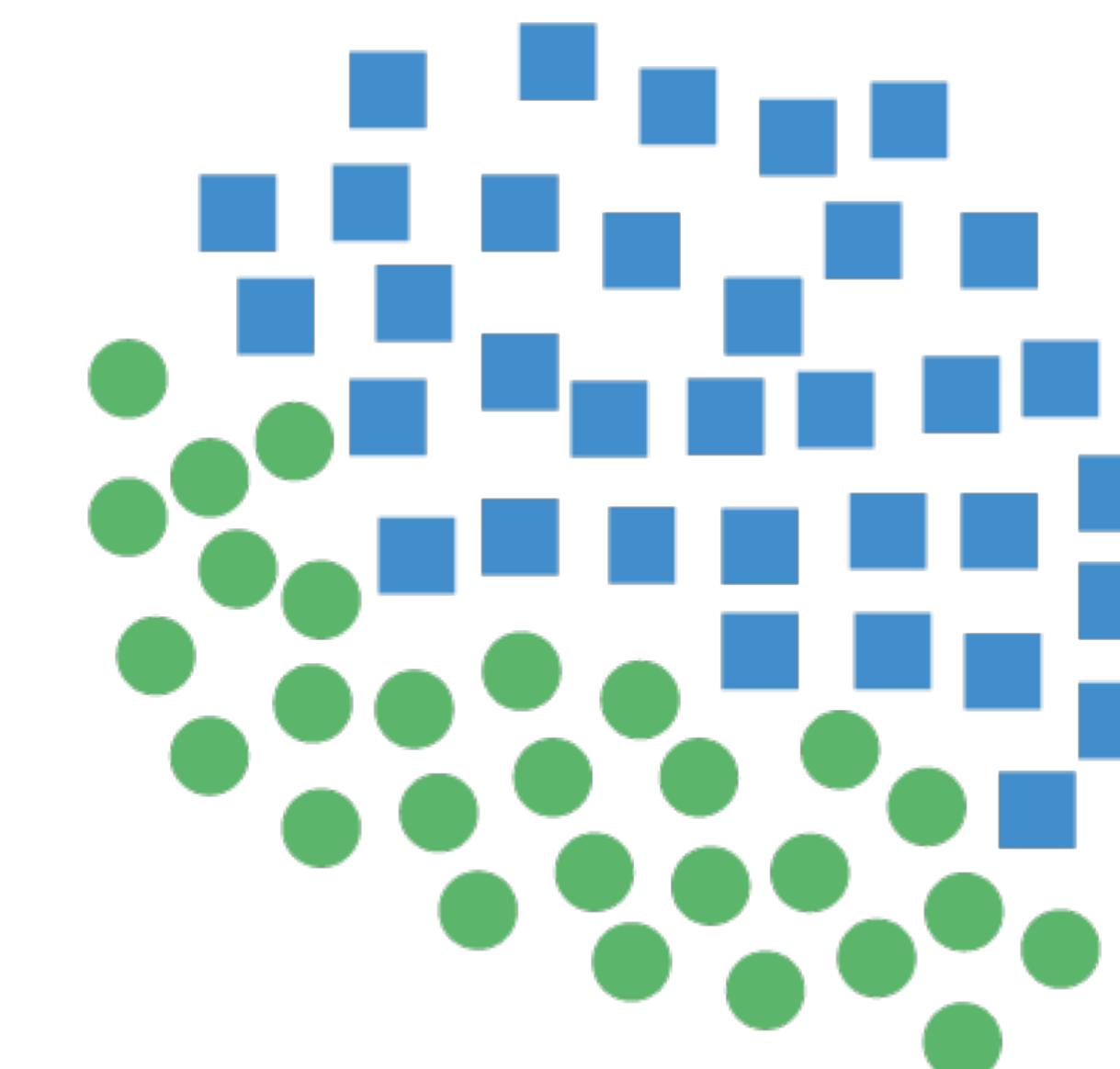
## Synthetic Minority Oversampling Technique



Original Dataset



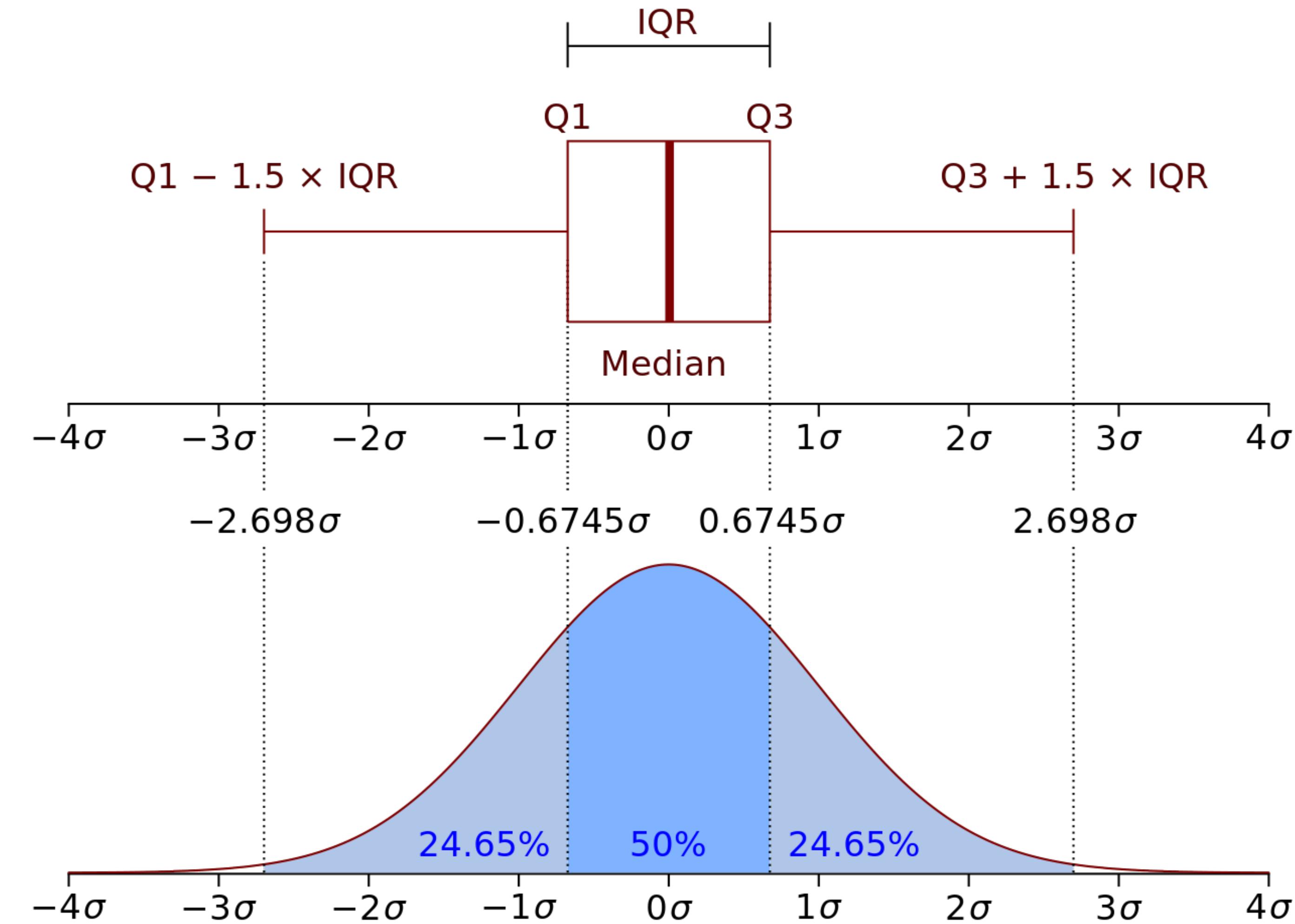
Generating Samples



Resampled Dataset

# INTERQUARTILE RANGE (IQR)

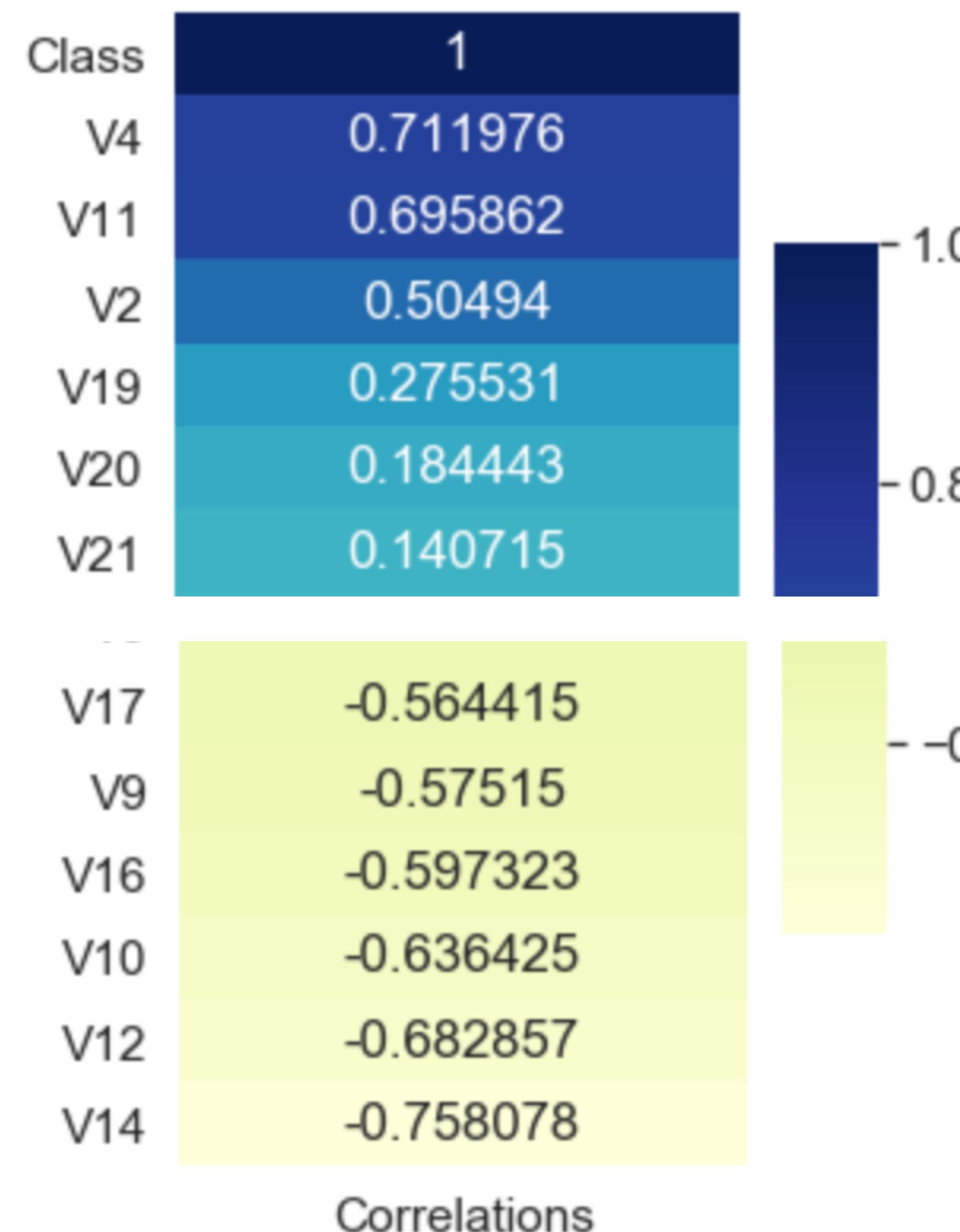
---

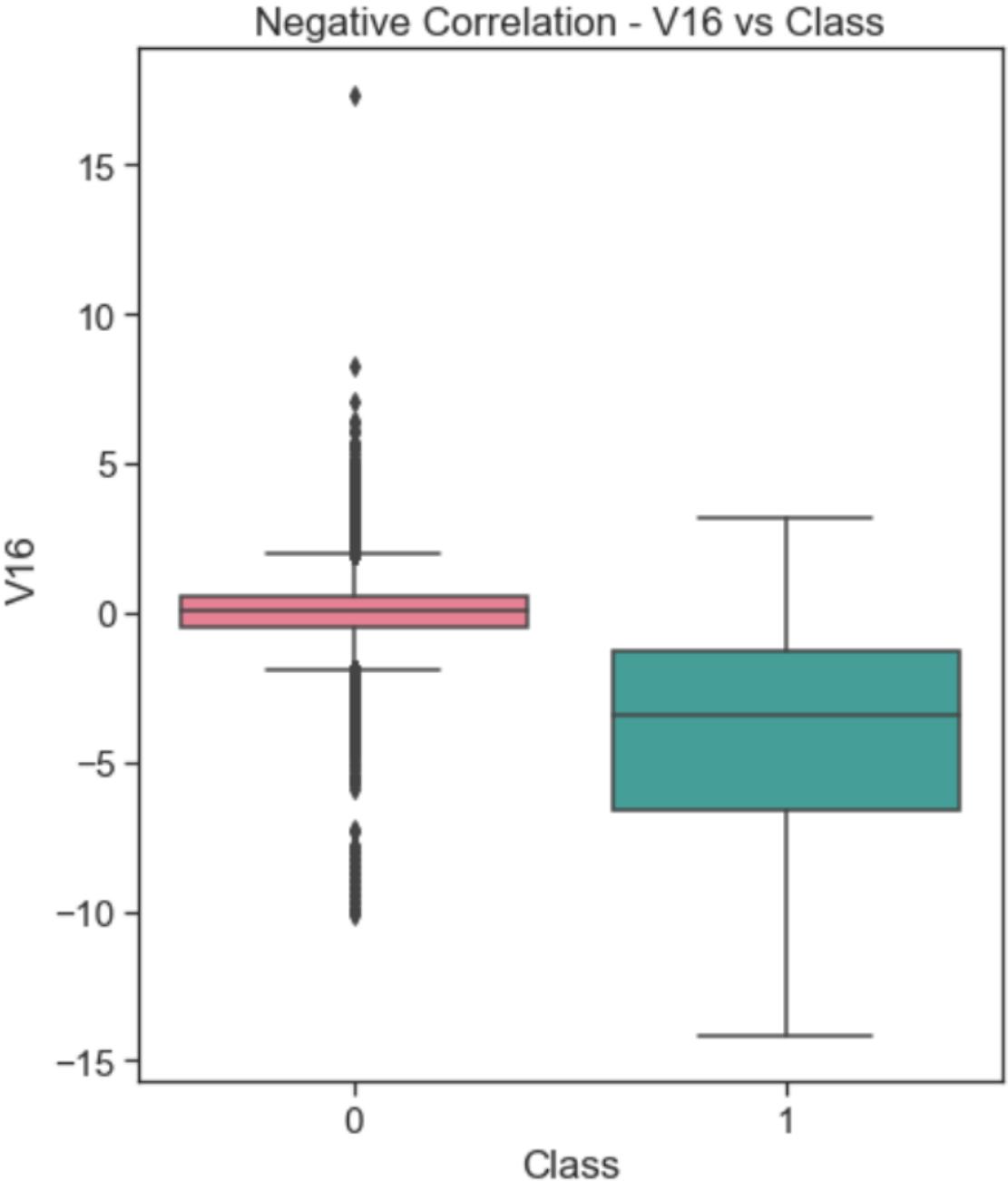
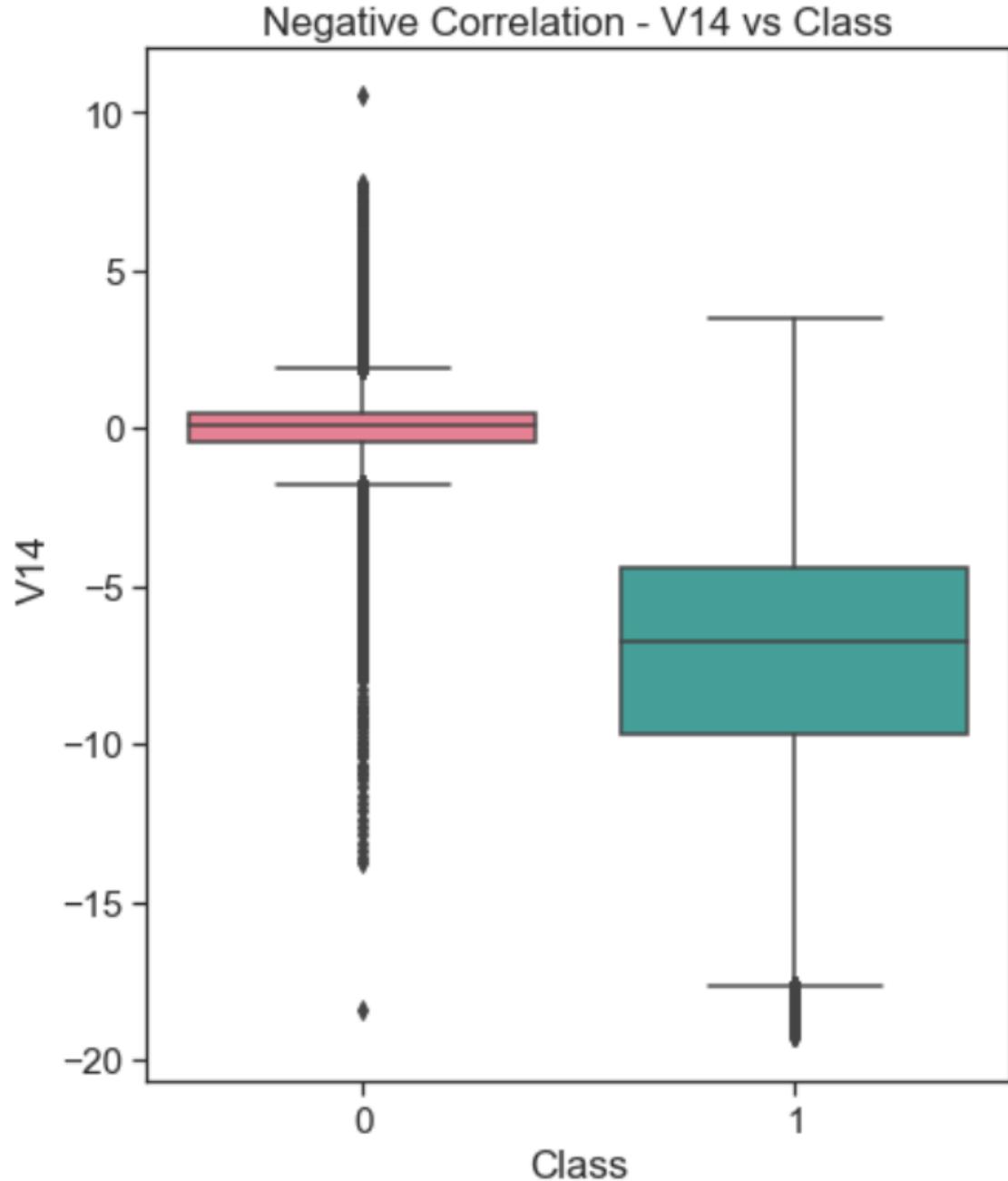
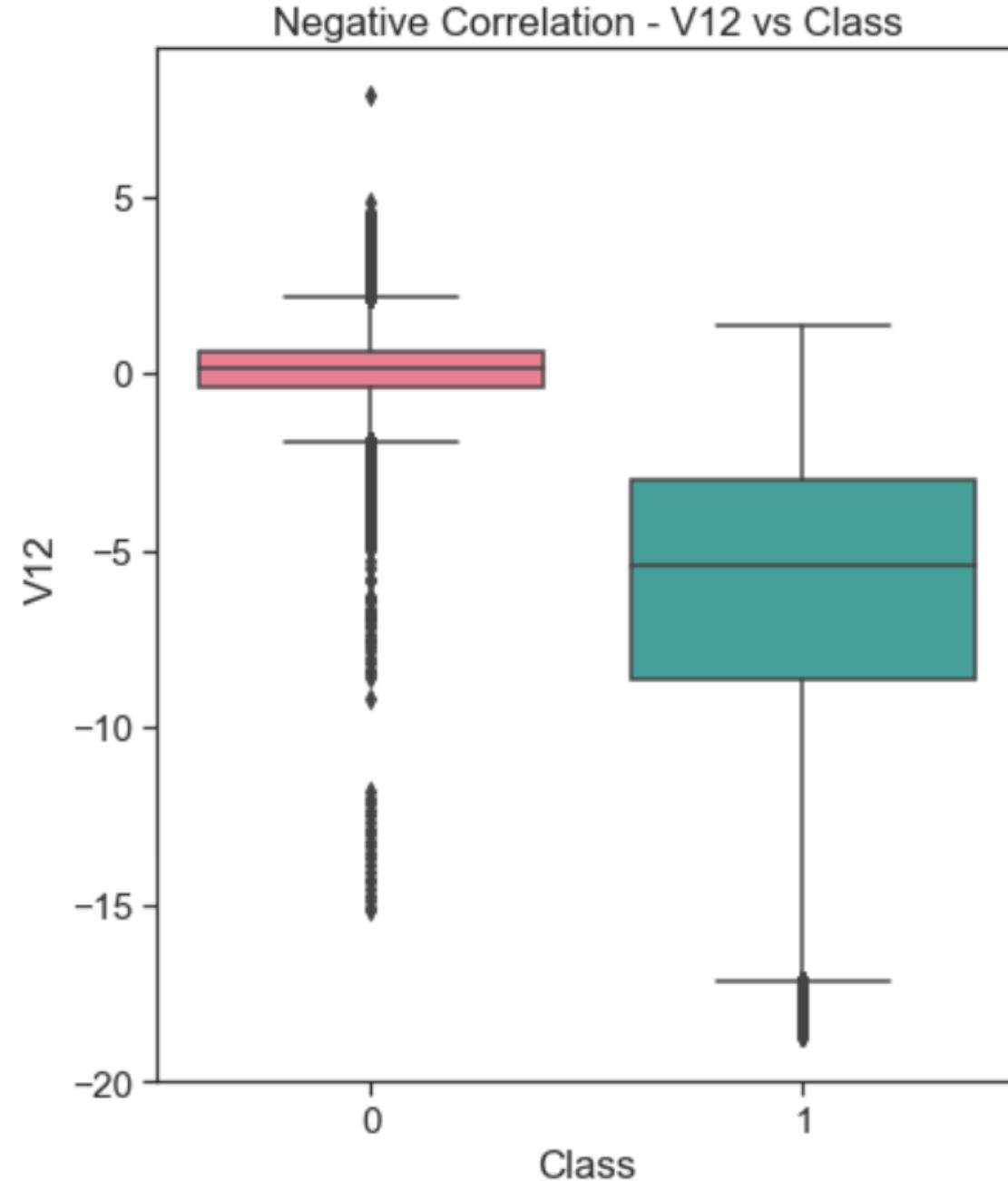
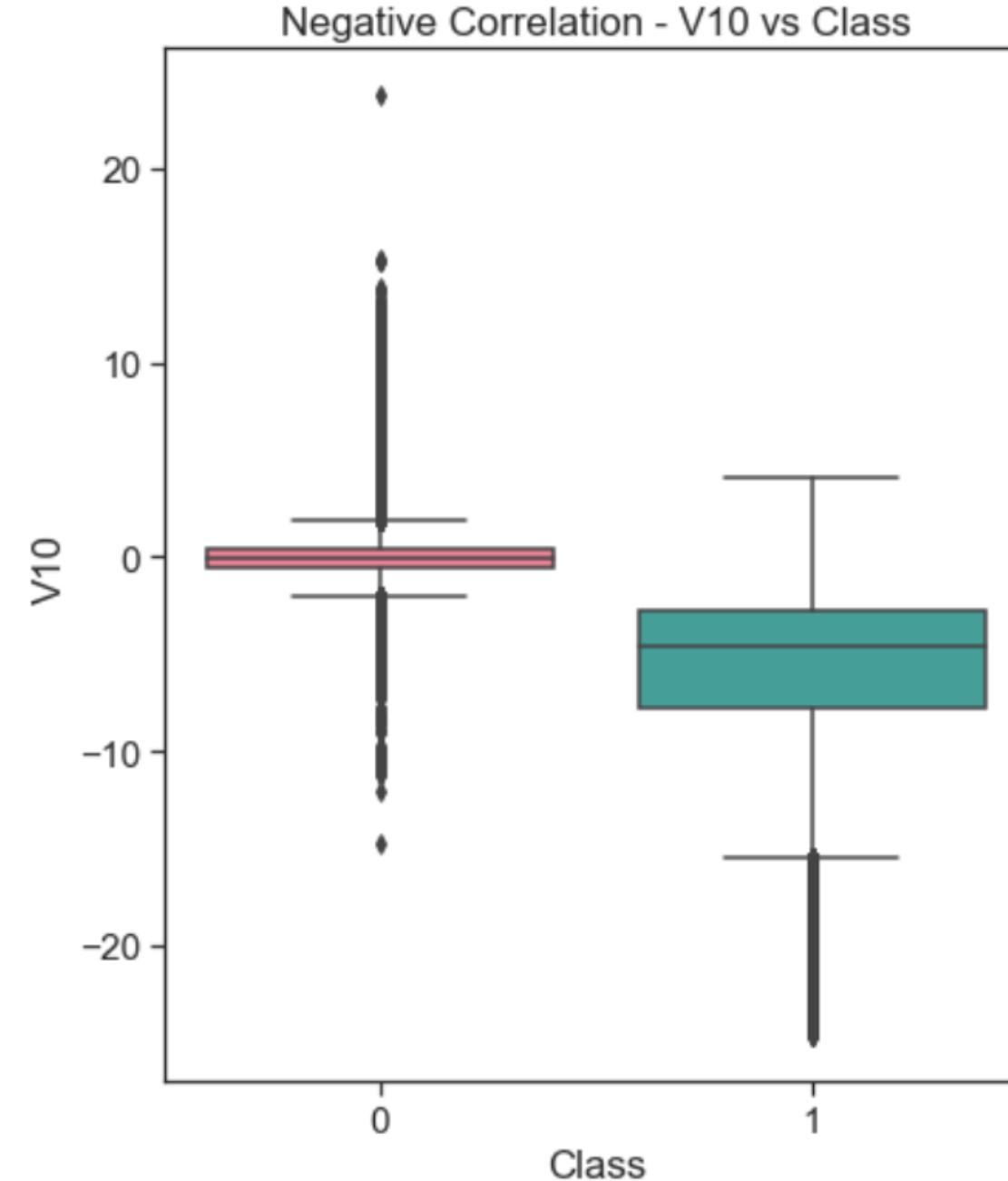
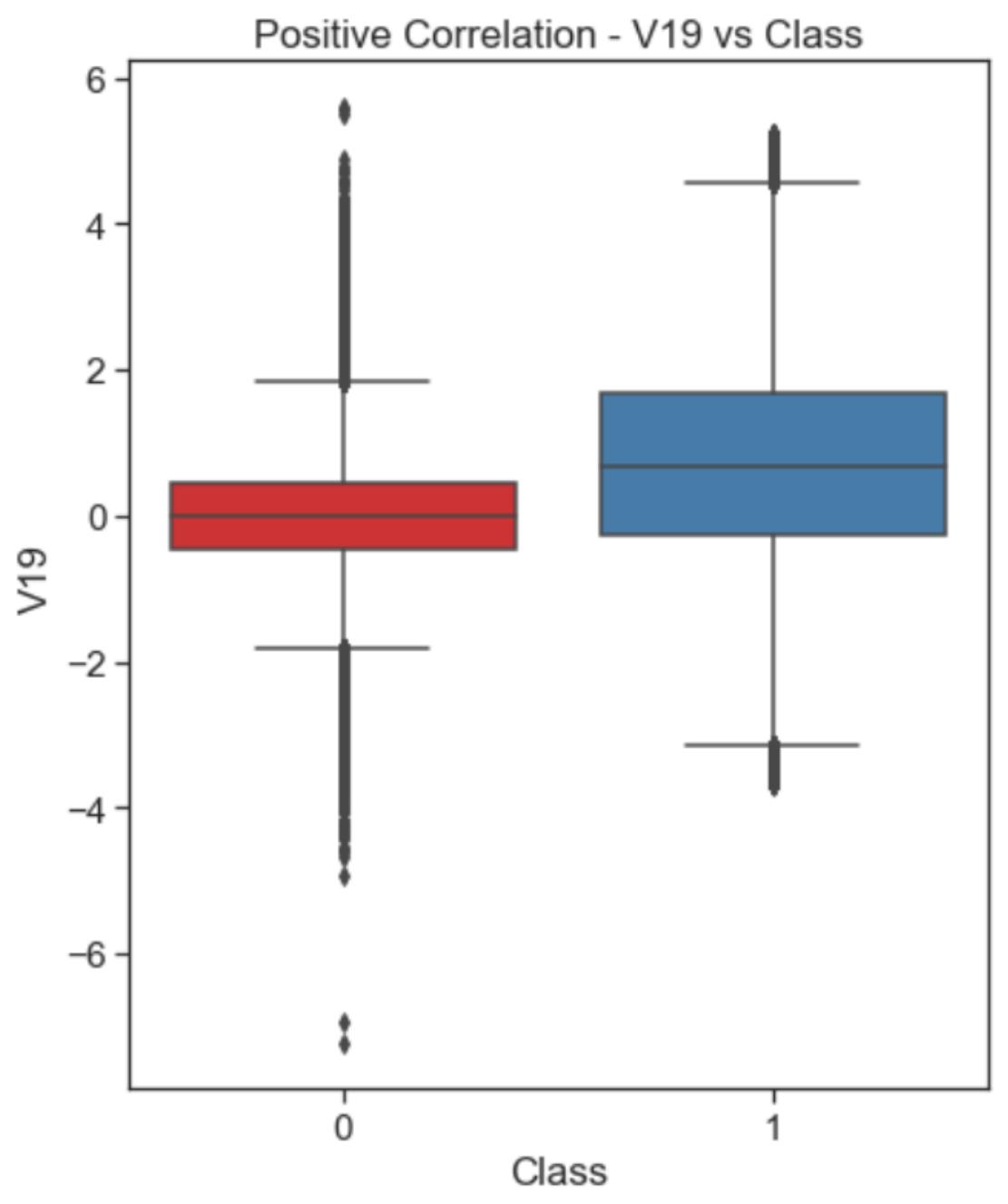
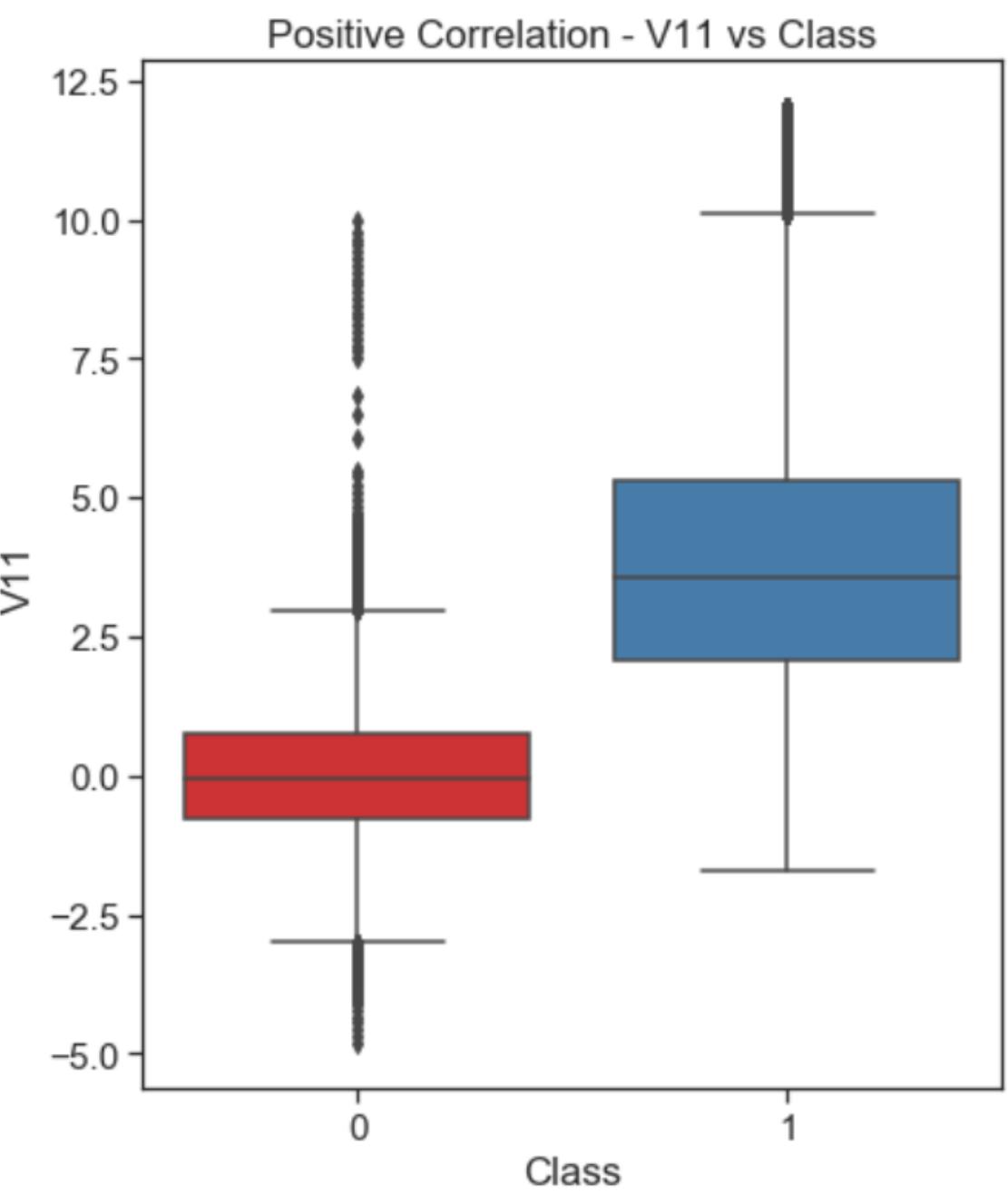
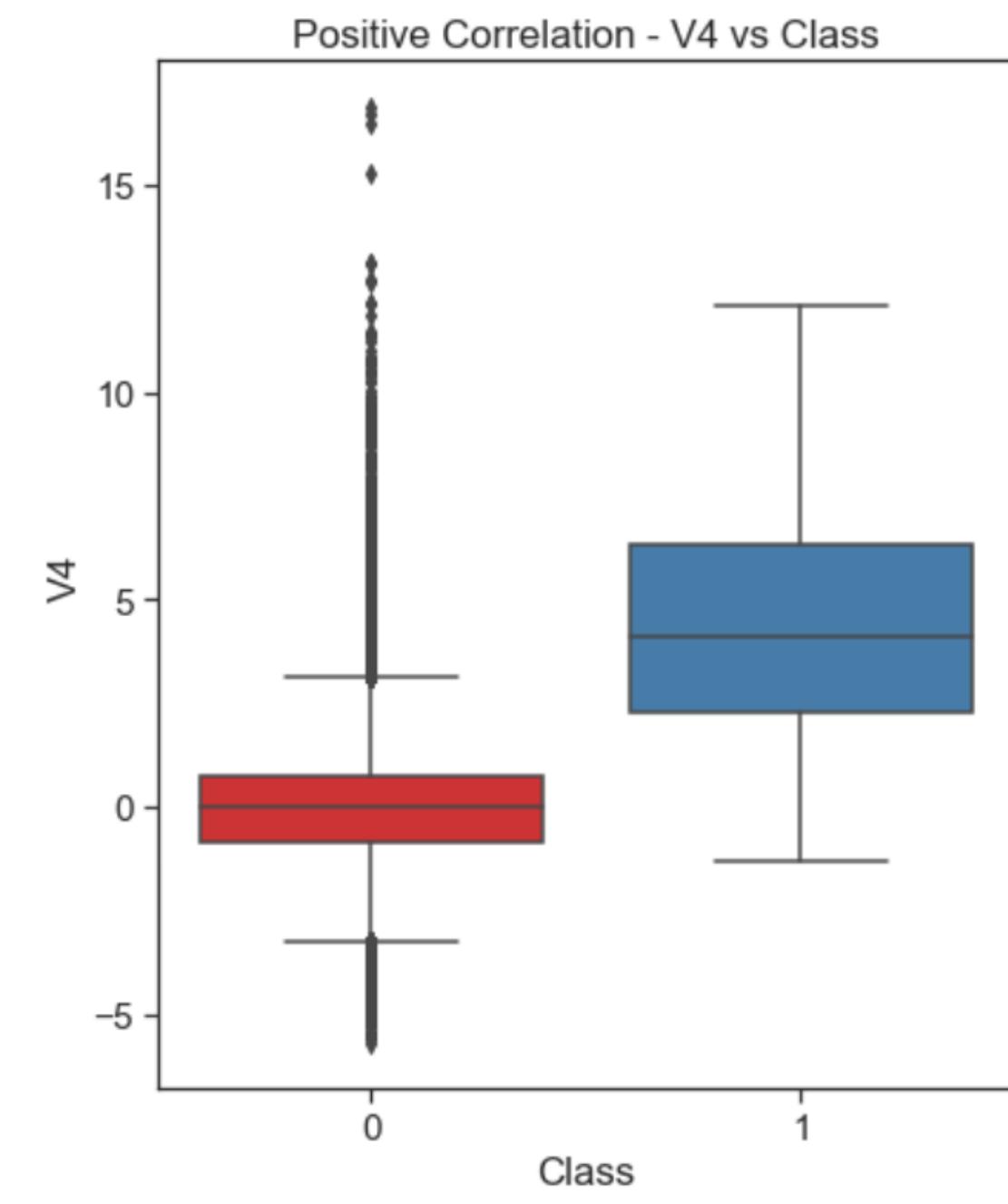
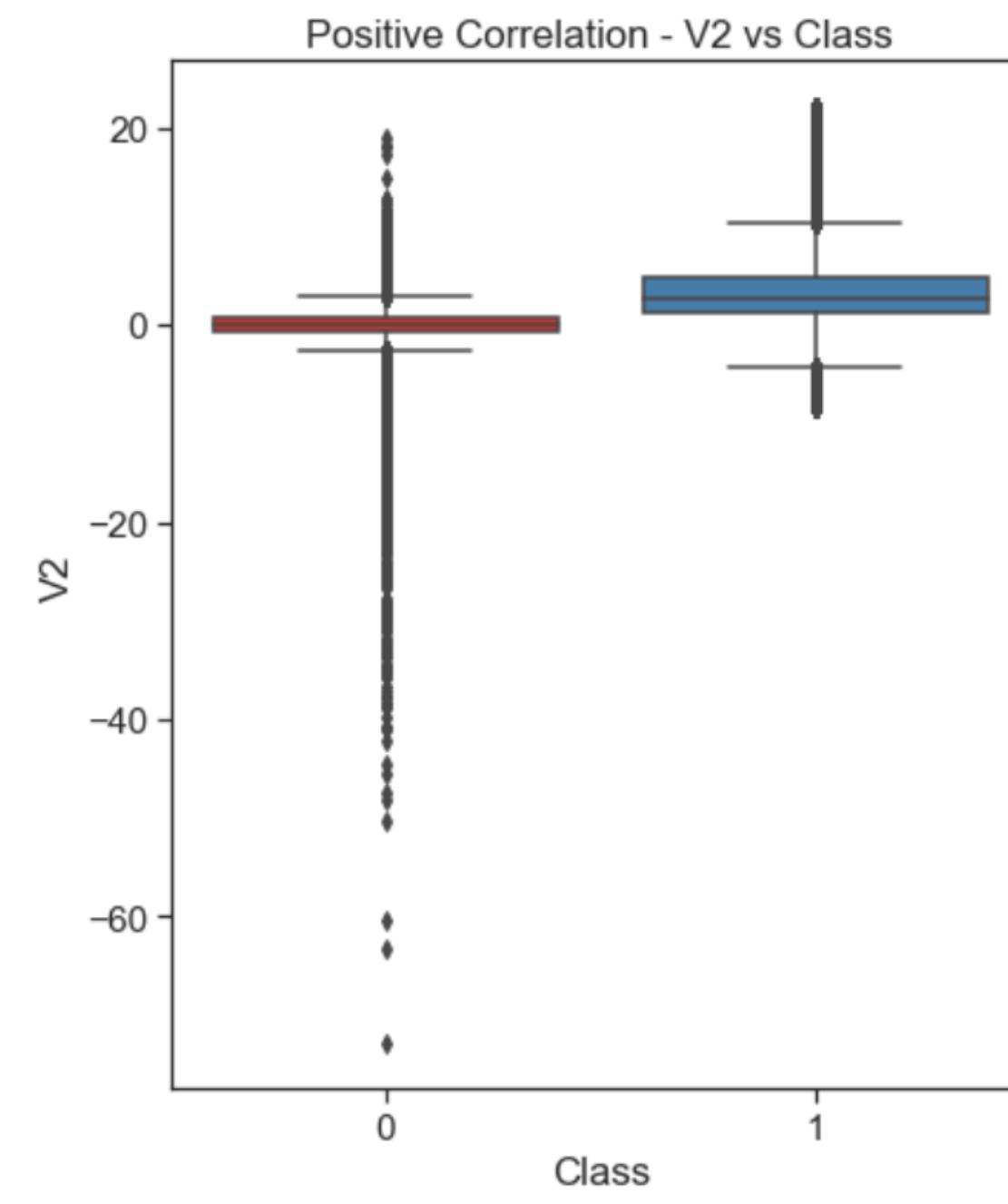


Source: [https://en.wikipedia.org/wiki/Interquartile\\_range](https://en.wikipedia.org/wiki/Interquartile_range)

# HIGHEST POSITIVE AND NEGATIVE CORRELATIONS

Correlation of Variables with Class

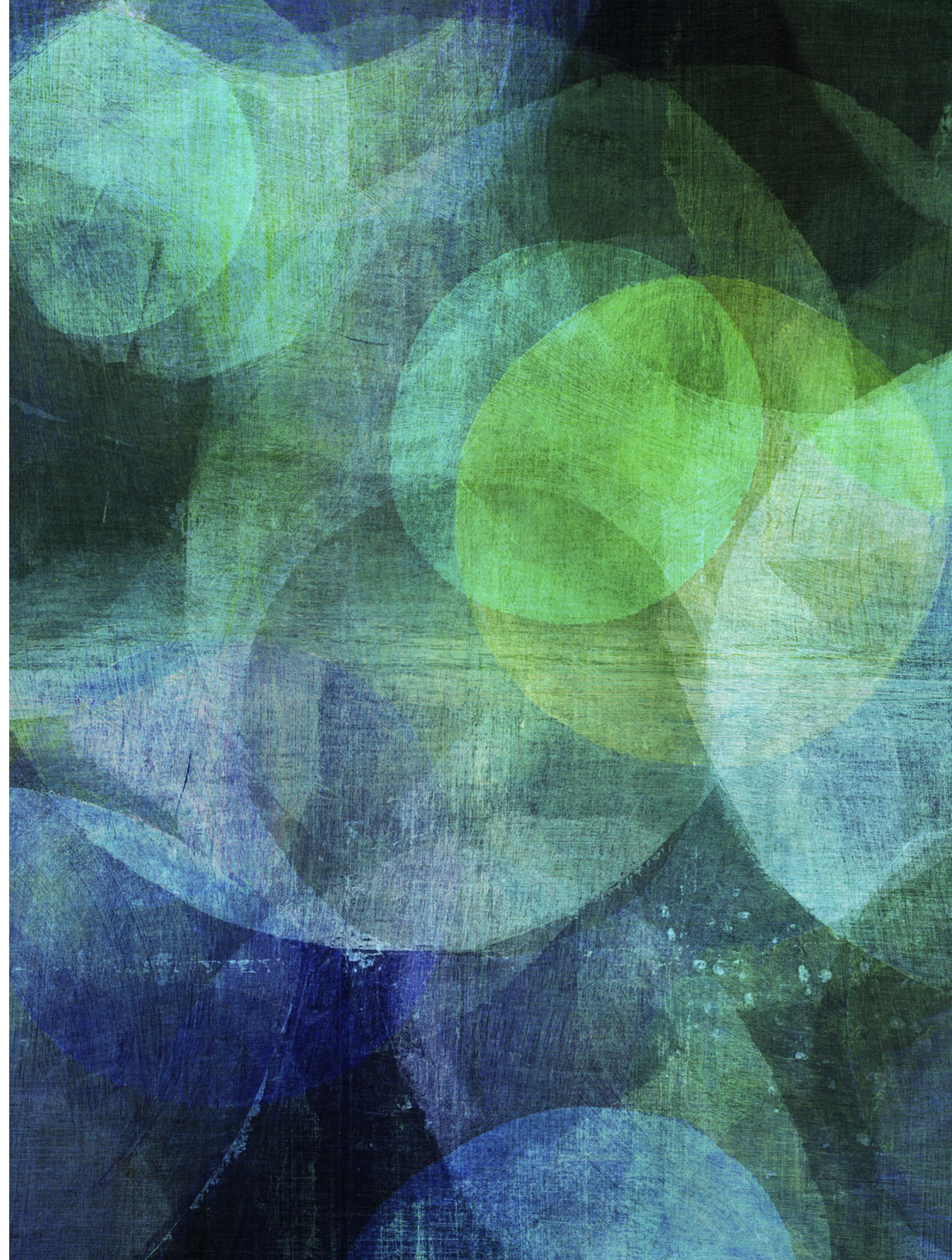




# REMoval of extreme outliers

From top 4 features which are positively & negatively correlated with class

An outlier is then a data point that lies outside the interquartile range.





## 5 CLASSIFIERS

---

1. Baseline Logistic Regression
2. Optimised Logistic Regression
3. Naive Bayes
4. Decision Tree
5. Random Forest

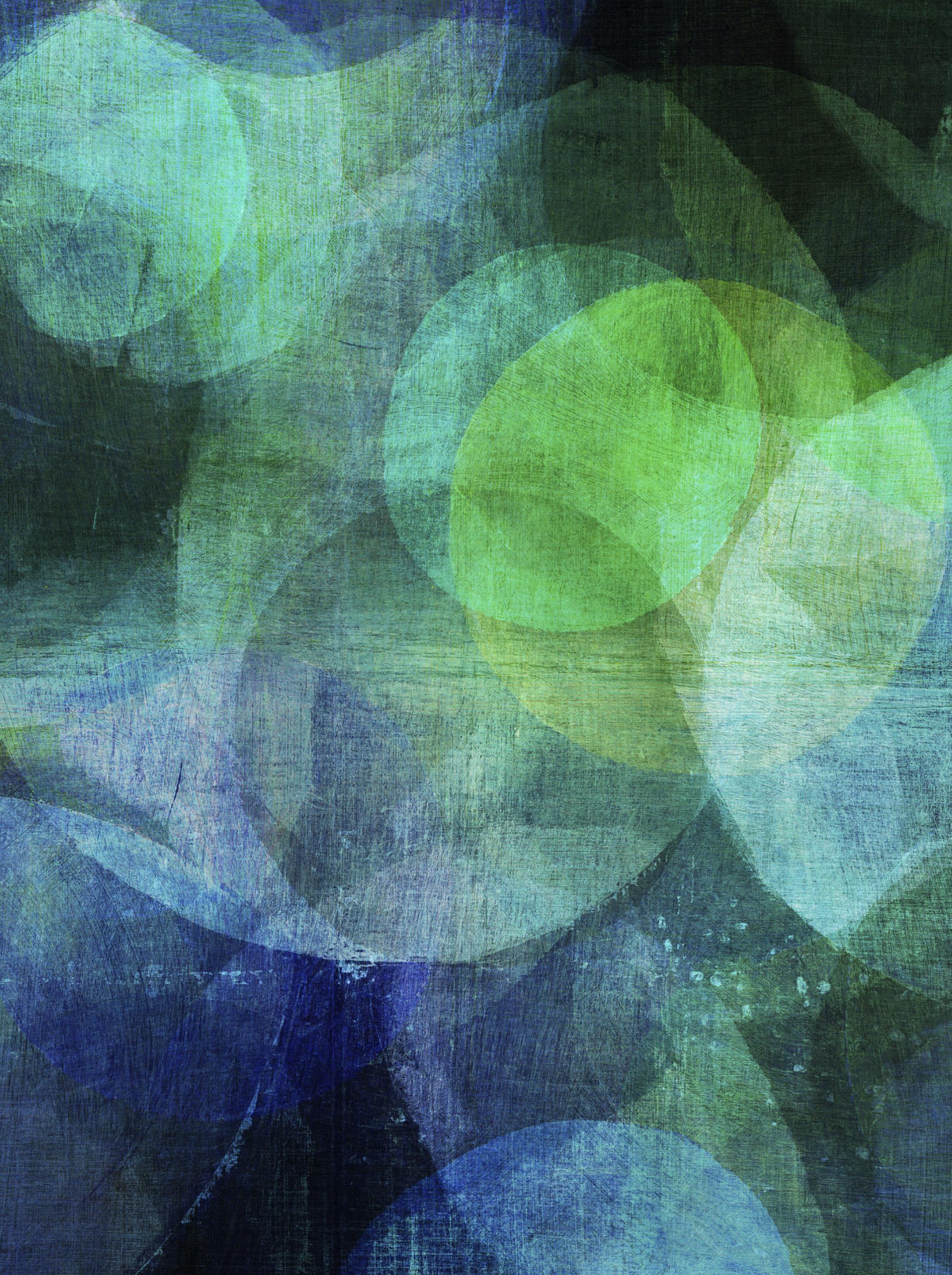
**1. Classification report**

**2. Confusion matrix**

**3. Precision-Recall Curve**

**4. ROC Curve and AUC**

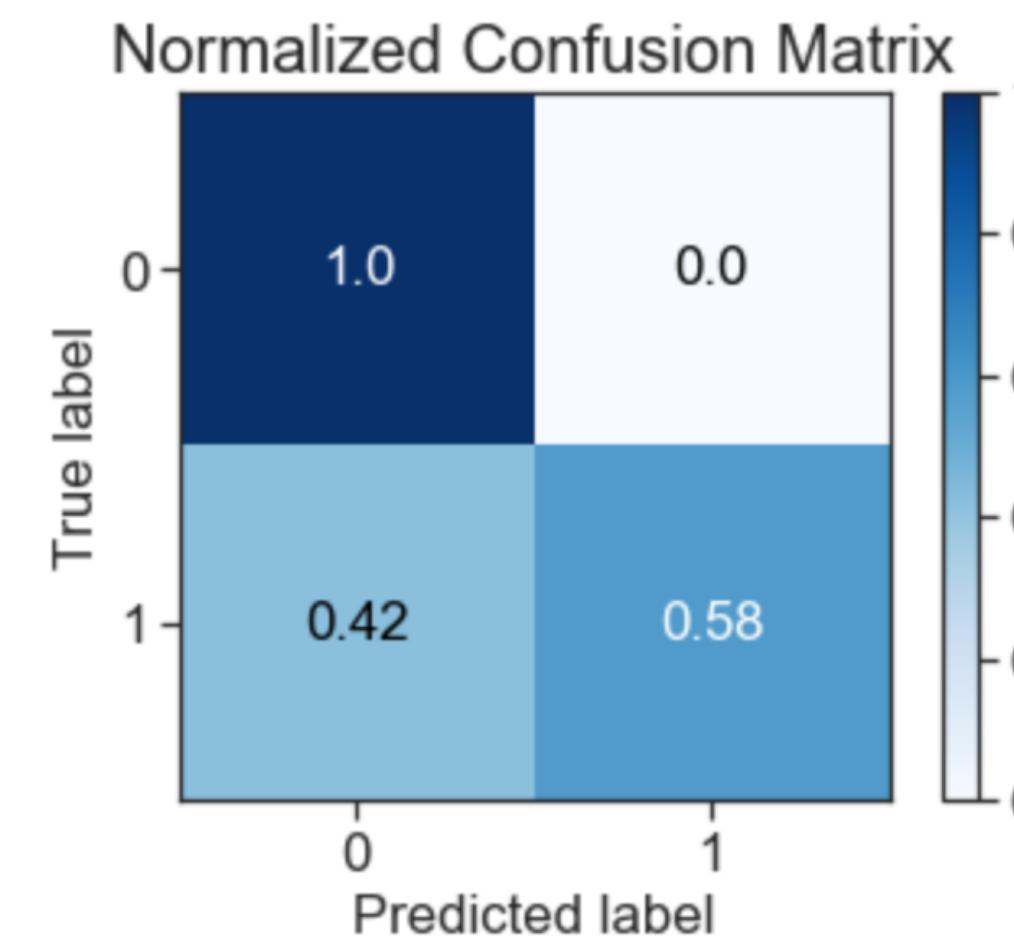
**PERFORMANCE EVALUATION**



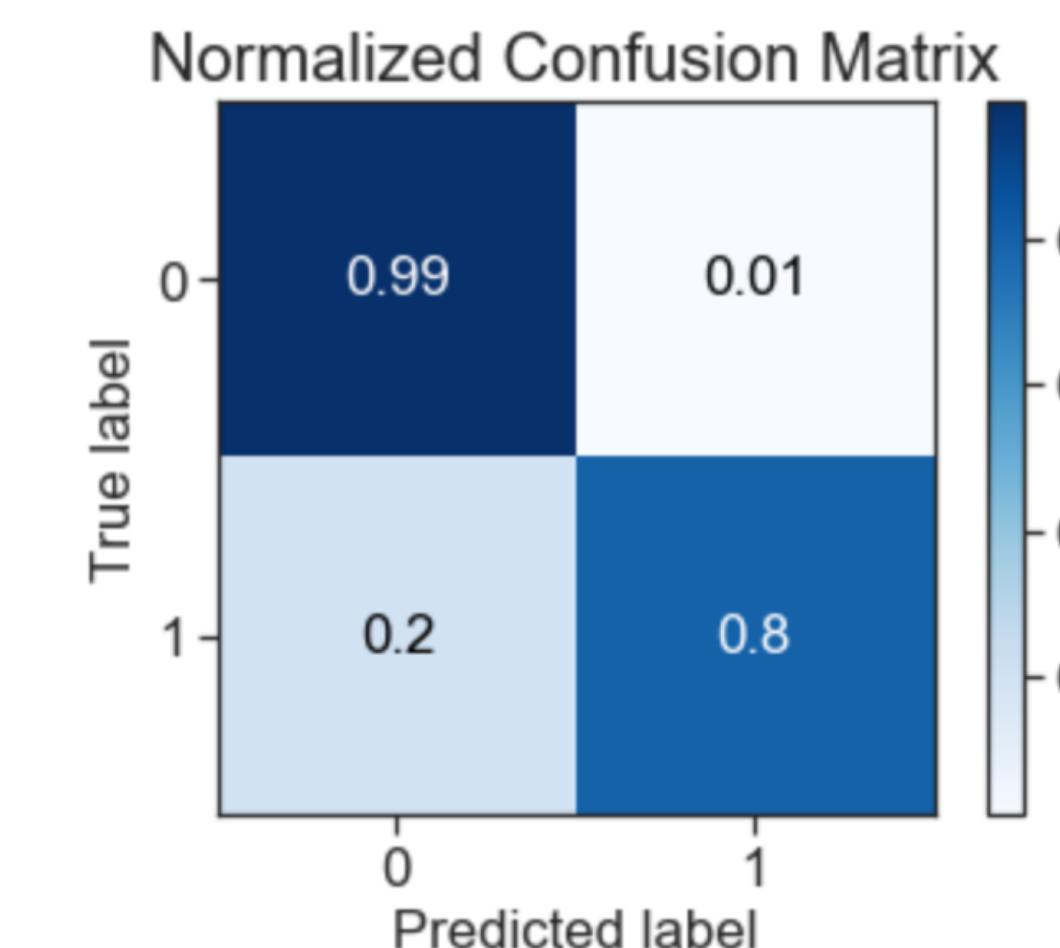
# CONFUSION MATRIX

---

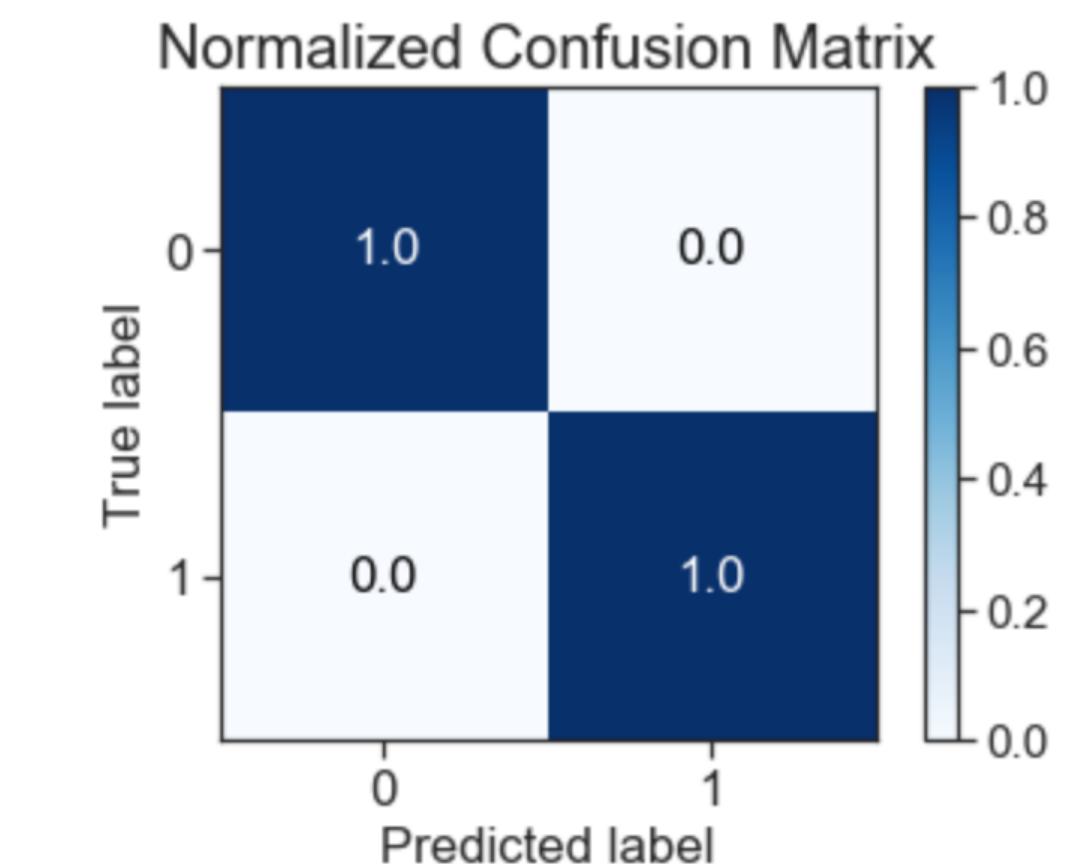
Confusion Matrix for ['Baseline Logistic Regression']



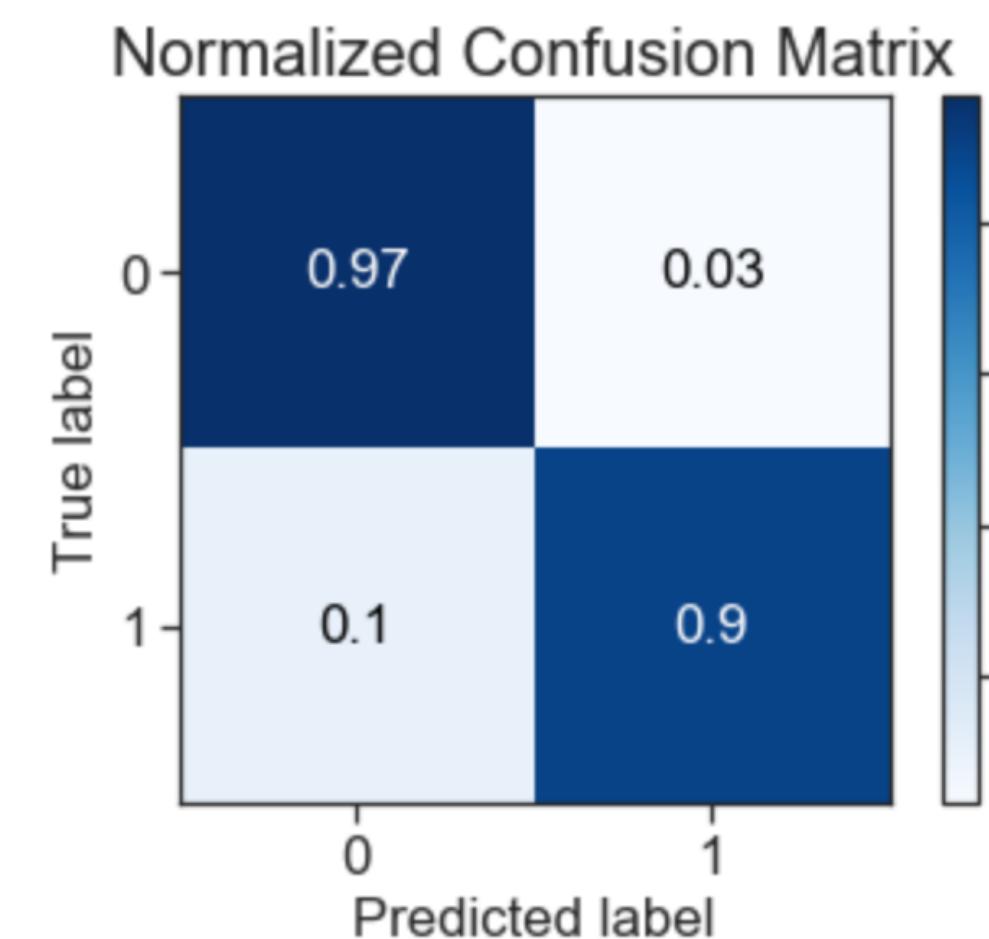
Confusion Matrix for ['Naive Bayes']



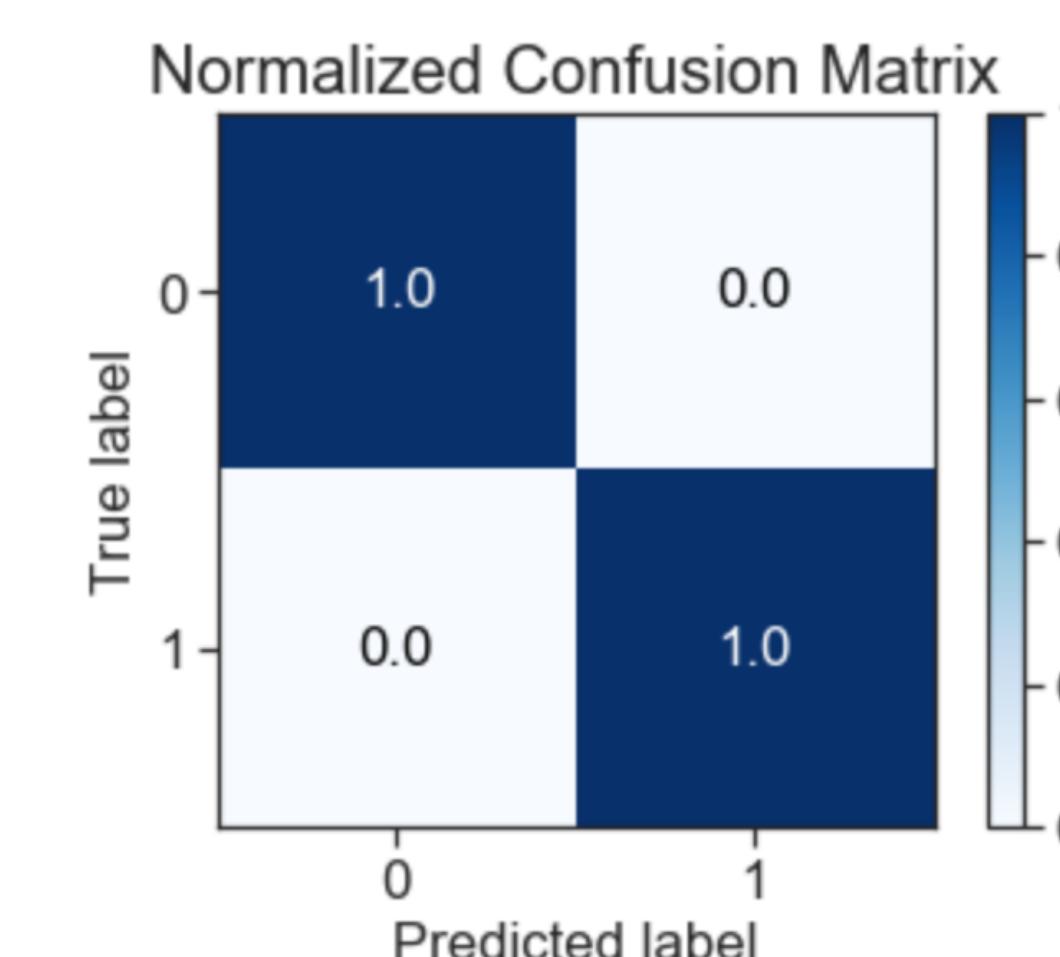
Confusion Matrix for RandomForestClassifier()



Confusion Matrix for ['Optimized Logistic Regression']

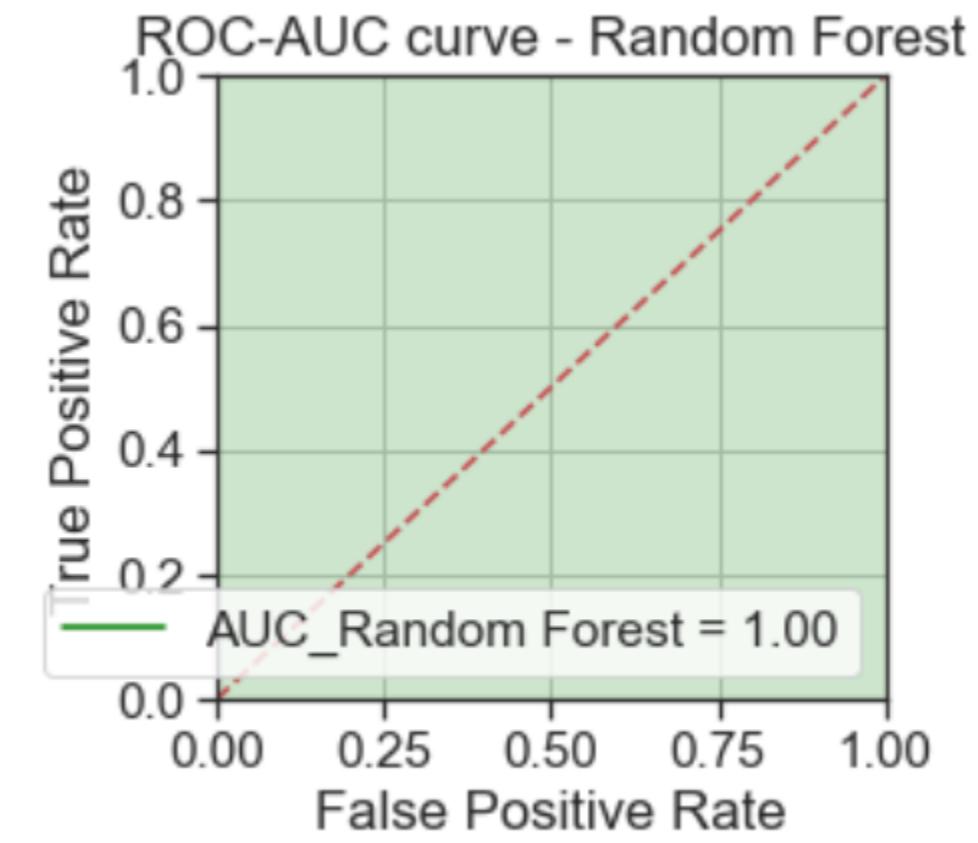
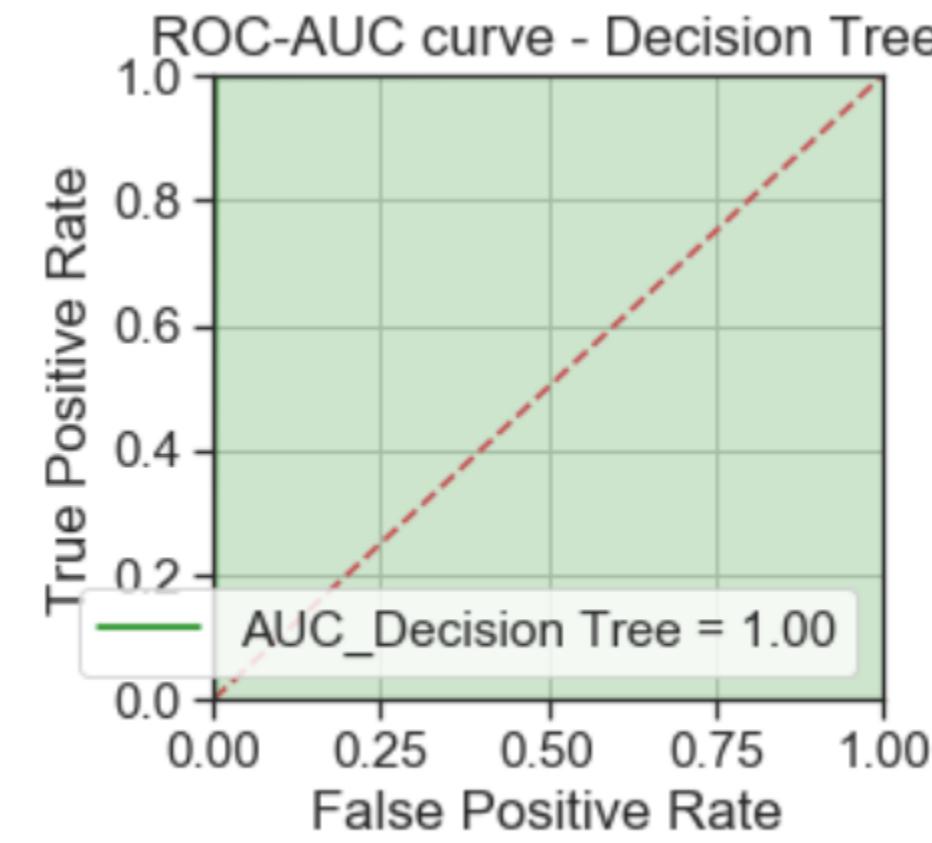
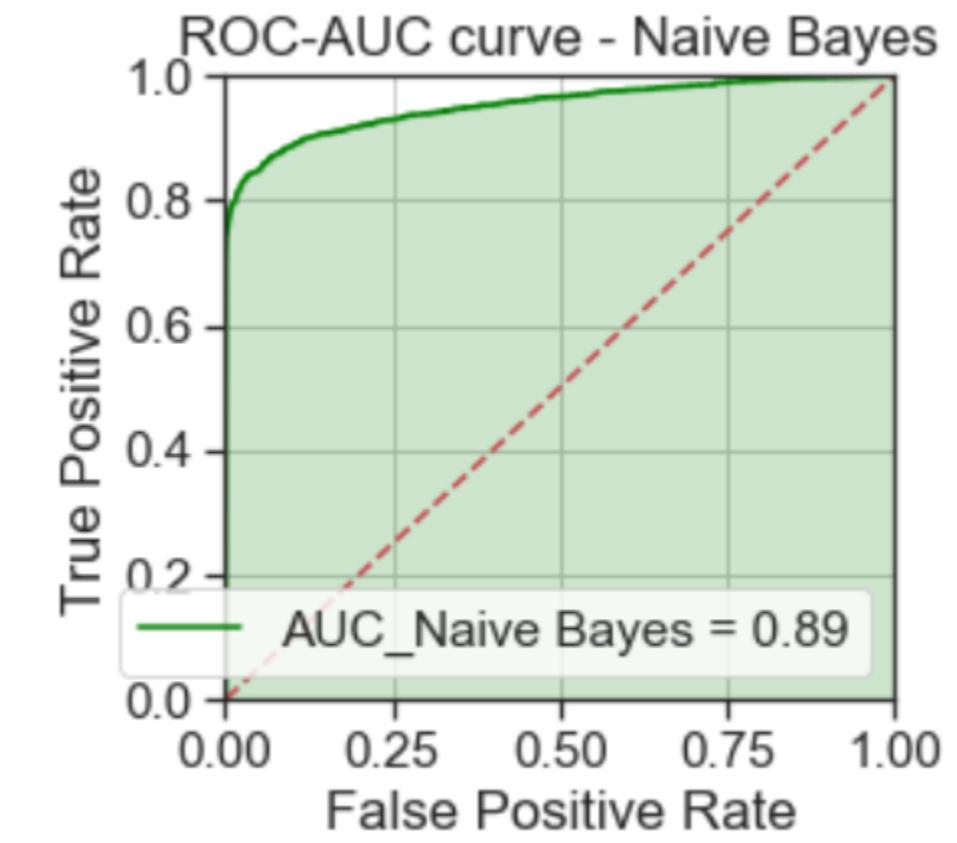
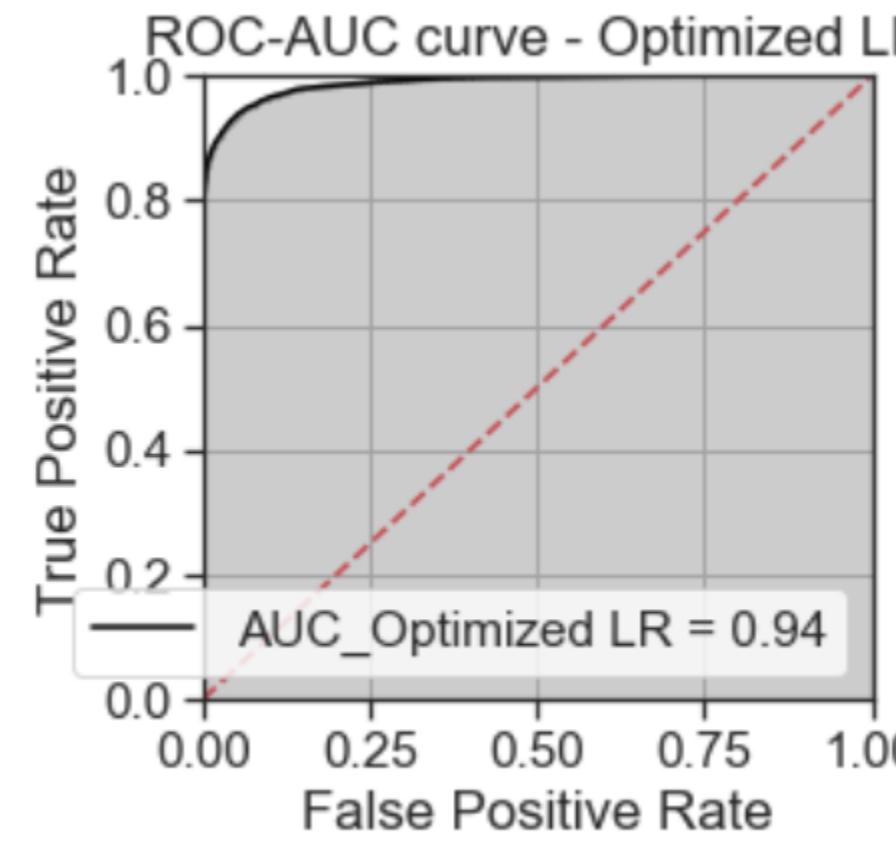
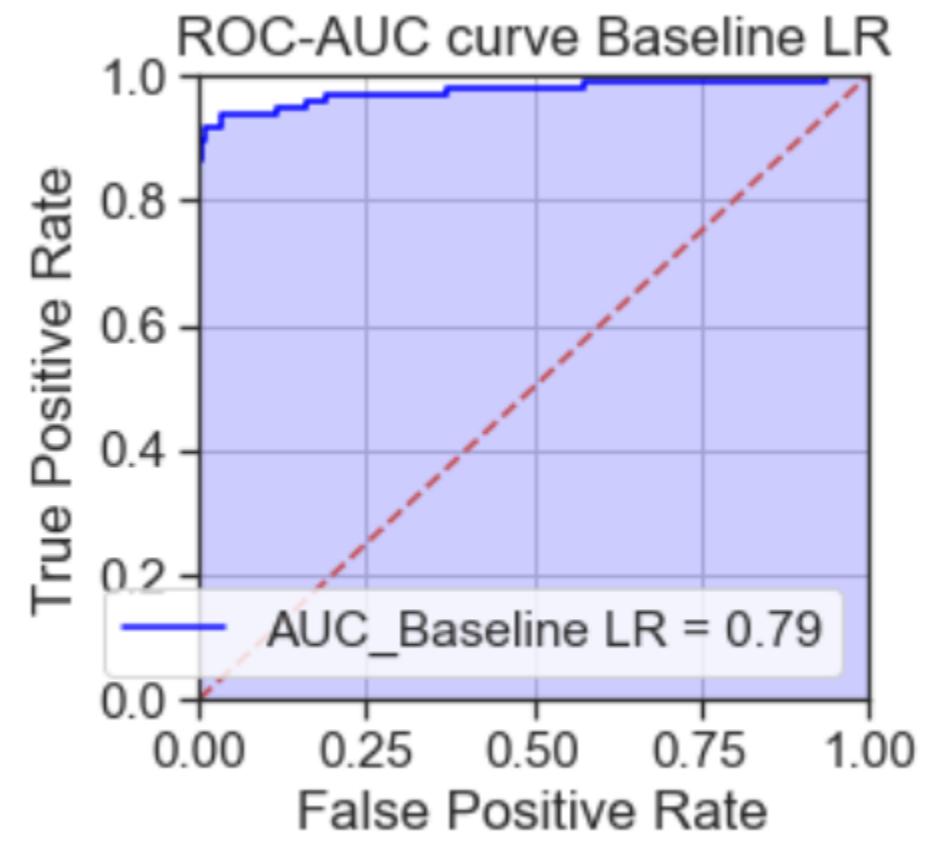
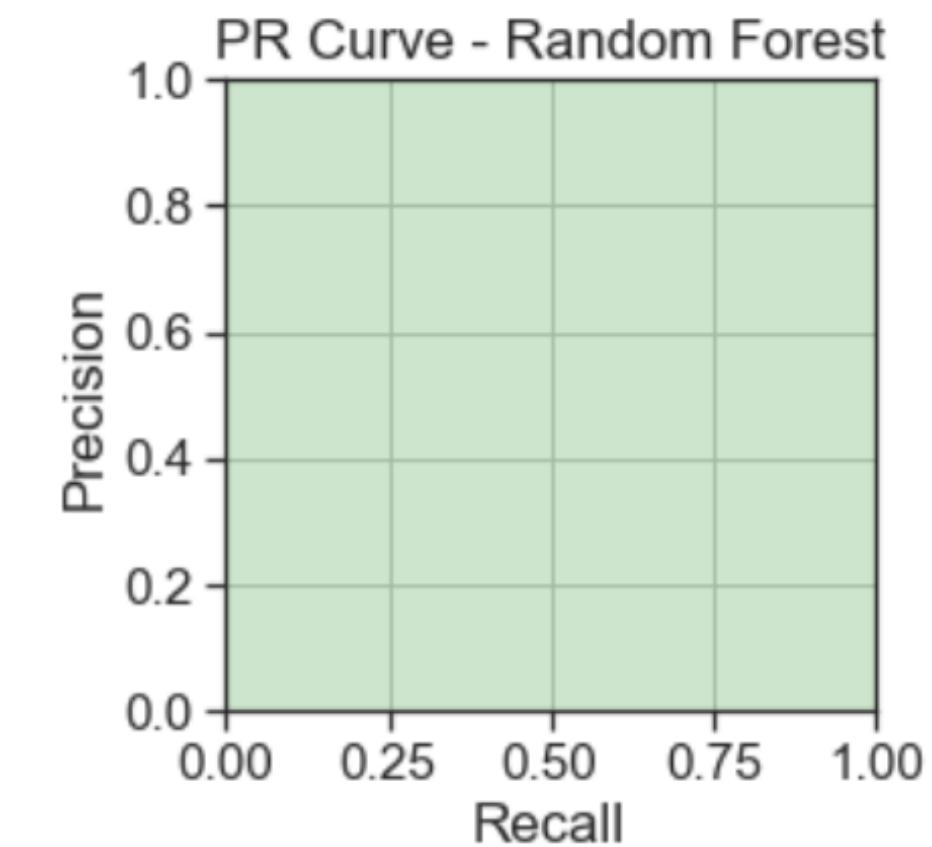
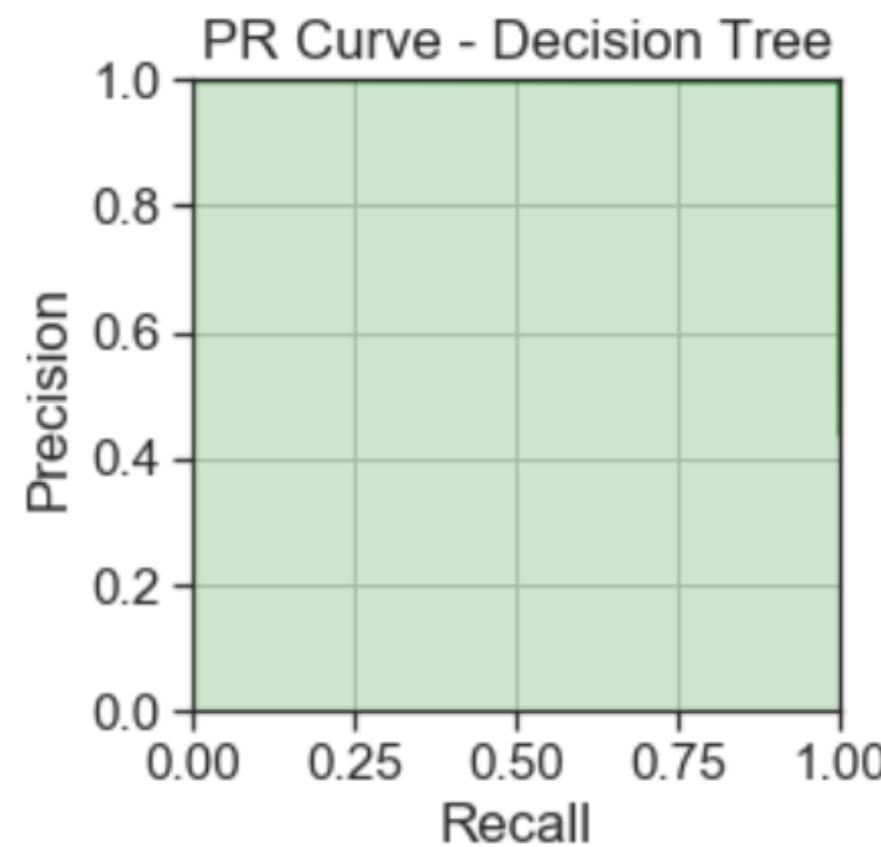
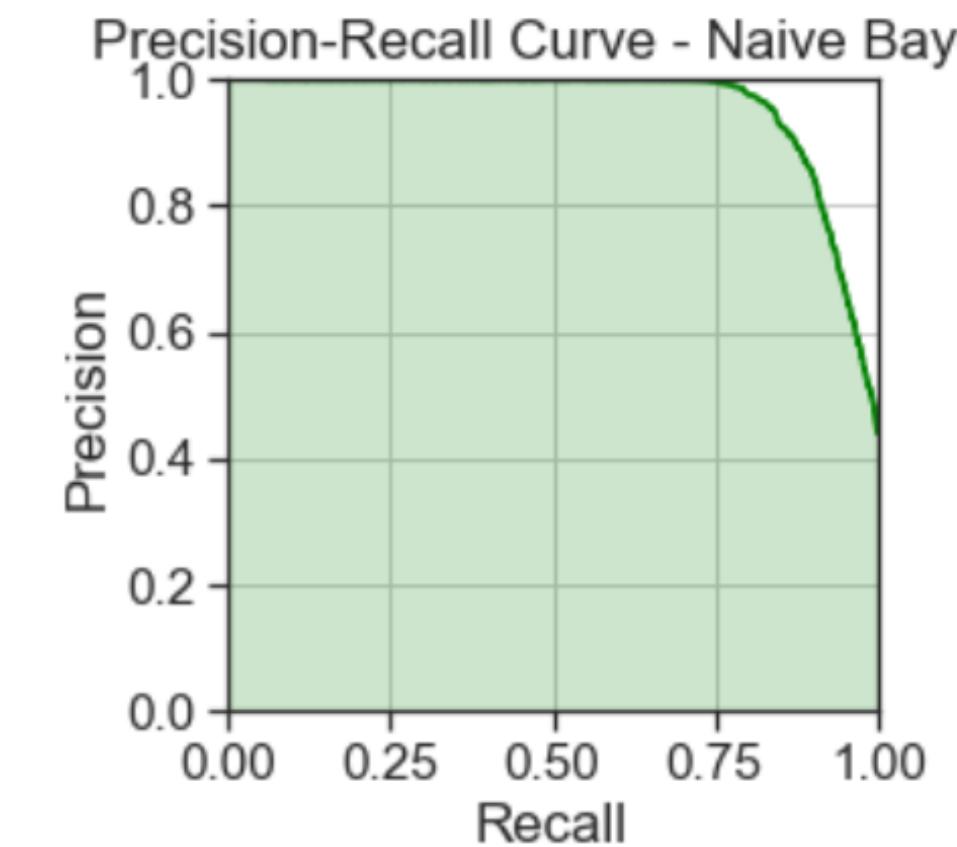
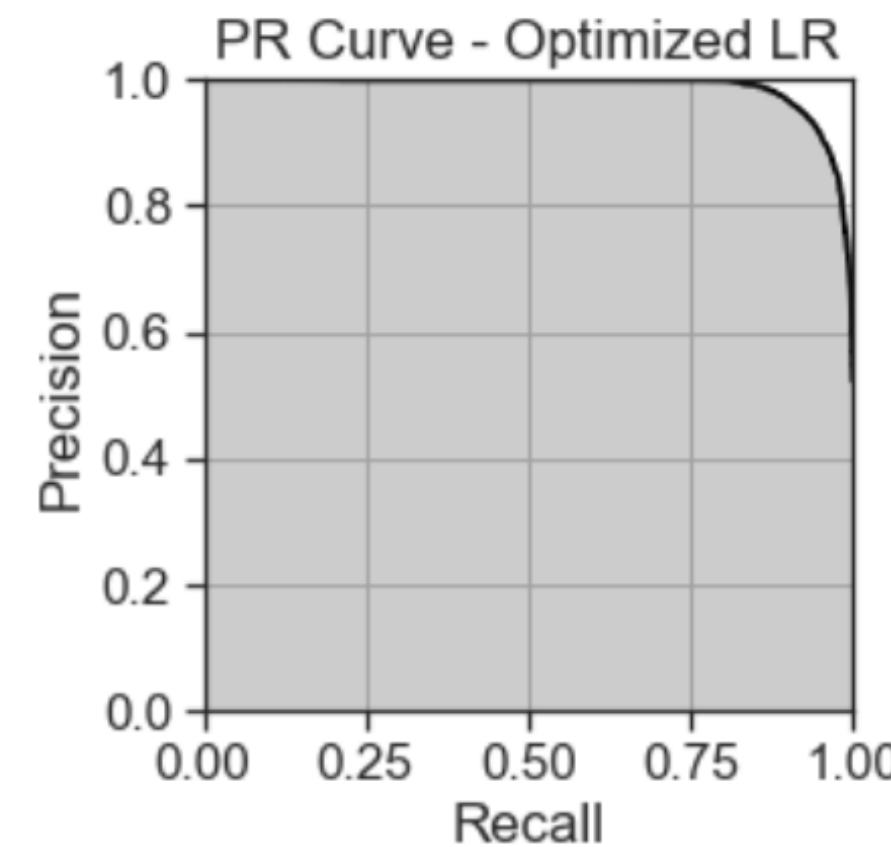
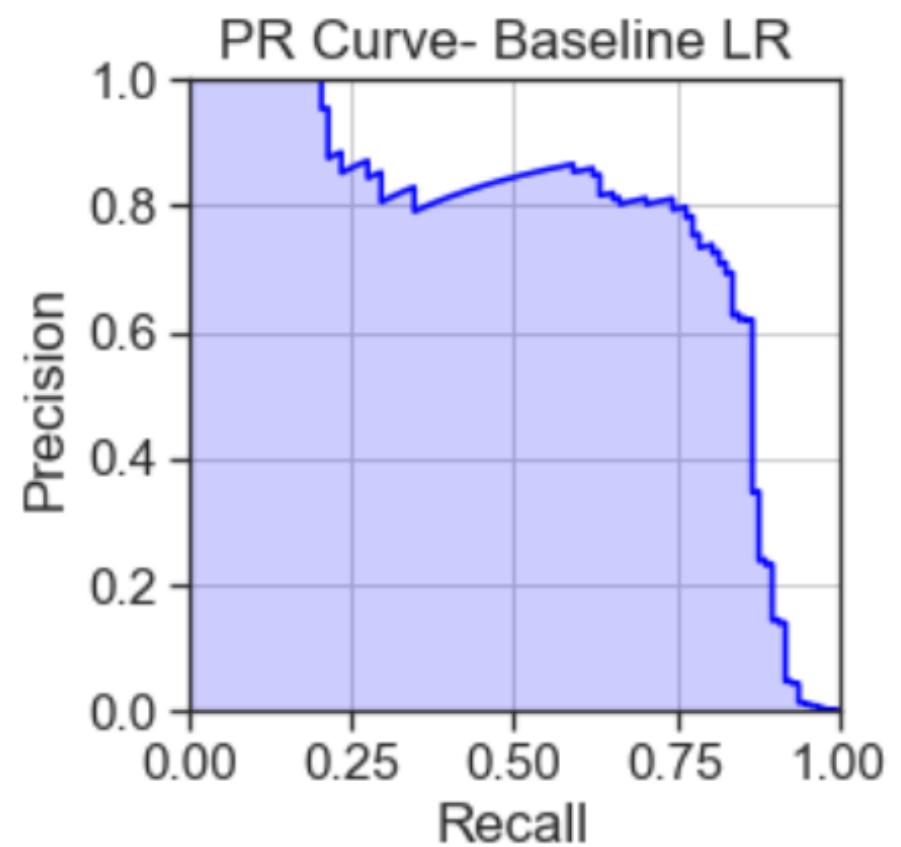


Confusion Matrix for ['Decision Tree']



# PRECISION-RECALL CURVE, ROC CURVE

---



Classifier	SMOTE - Accuracy rate(%)	RUS - Accuracy rate(%)
Baseline Logistic Regression	Train 99.92, Test 99.91	Train 99.92, Test 99.91
Optimized Logistic Regression	Train 98.75, Test 98.78	Train 98.17, Test 97.65
Naive Bayes	Train 90.28, Test 90.37	Train 91.38, Test 91.11
Decision Tree	Train 100, Test 99.85	Train 100, Test 93.88
Random Forest	Train 100, Test 99.98	Train 100, Test 96.11

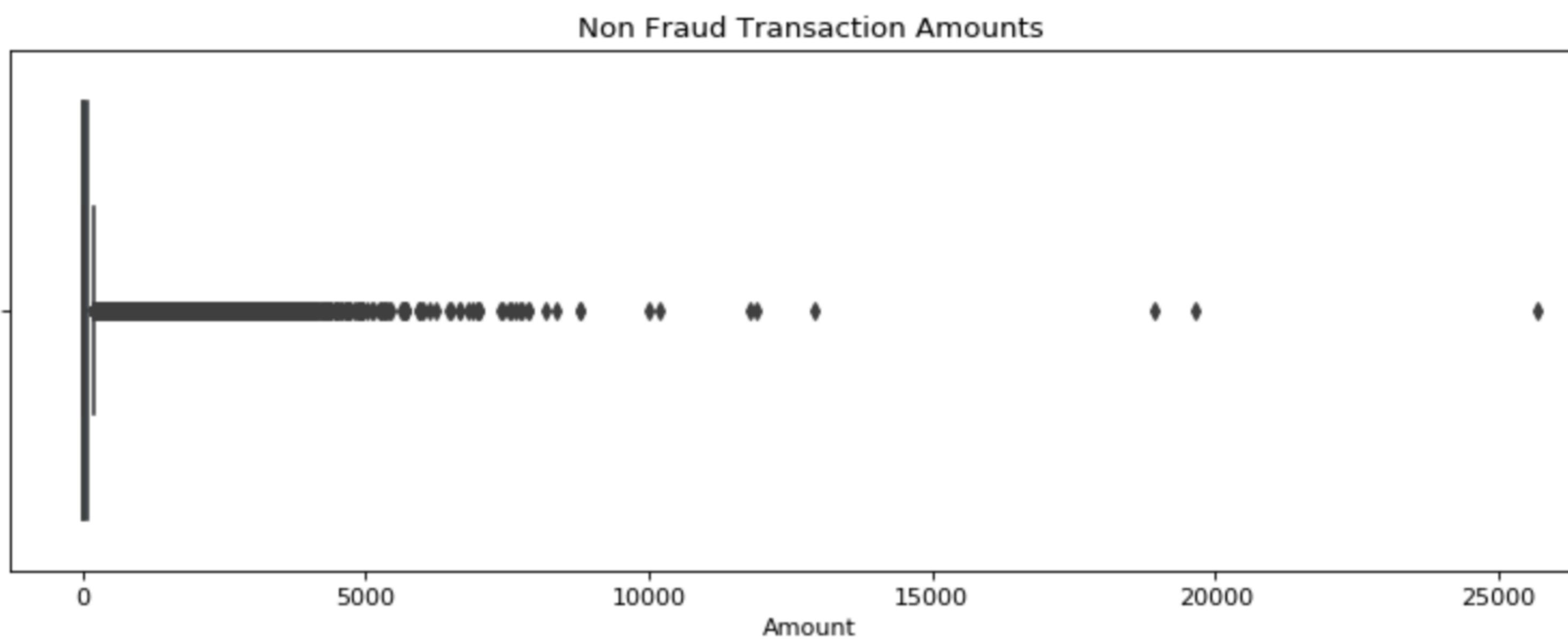
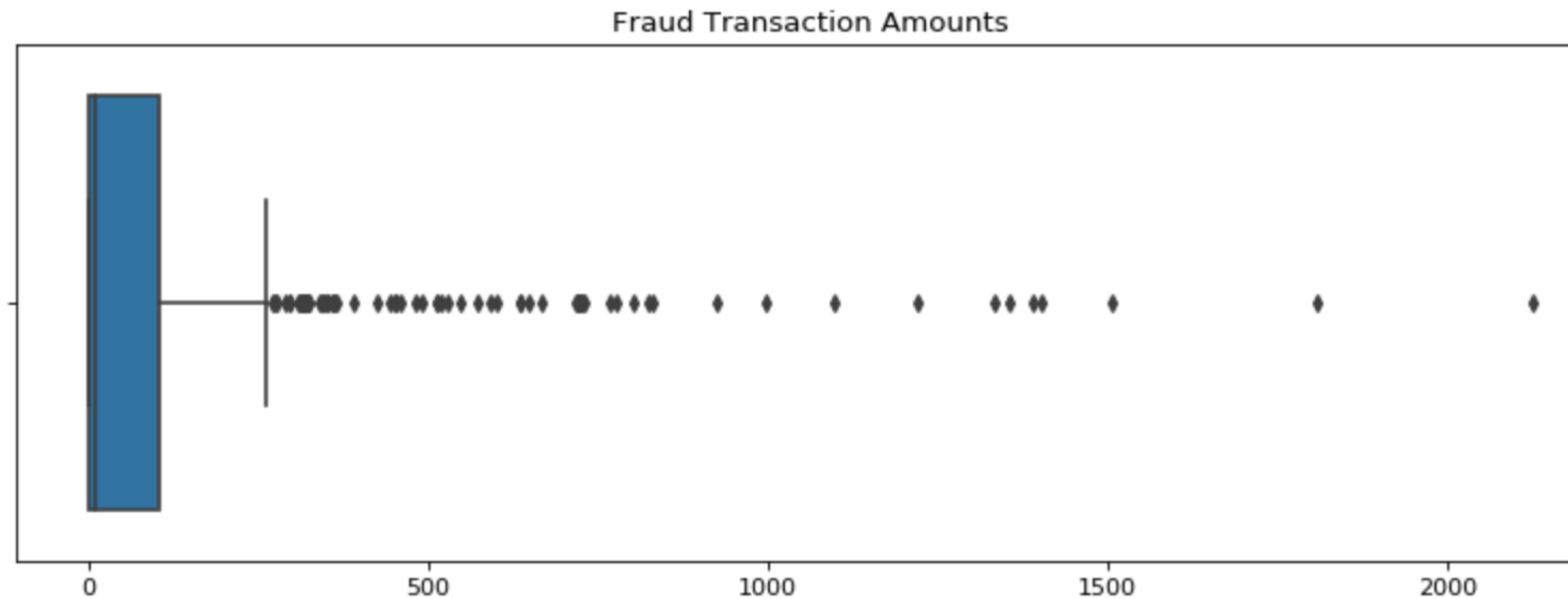


# UNSUPERVISED LEARNING

---

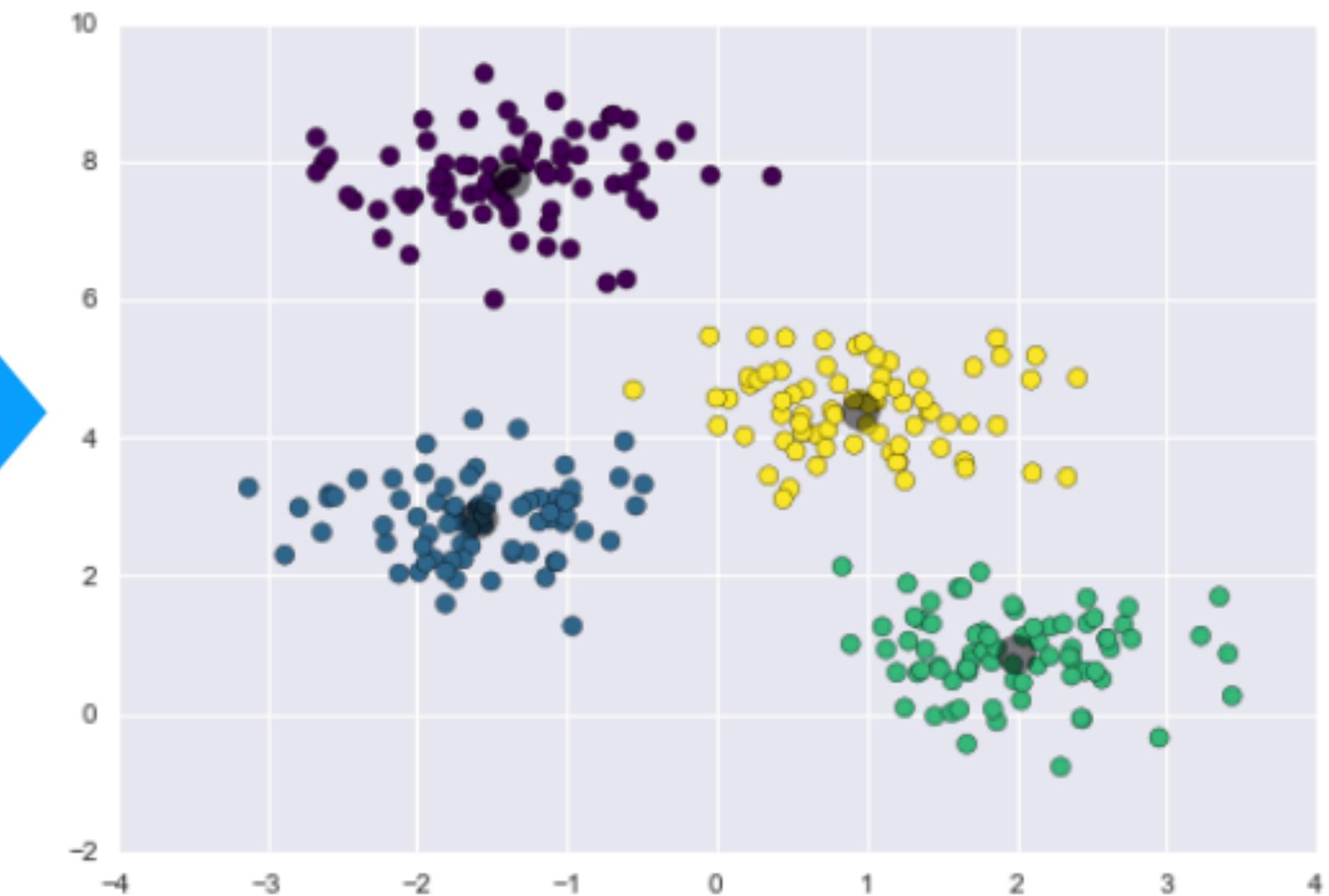
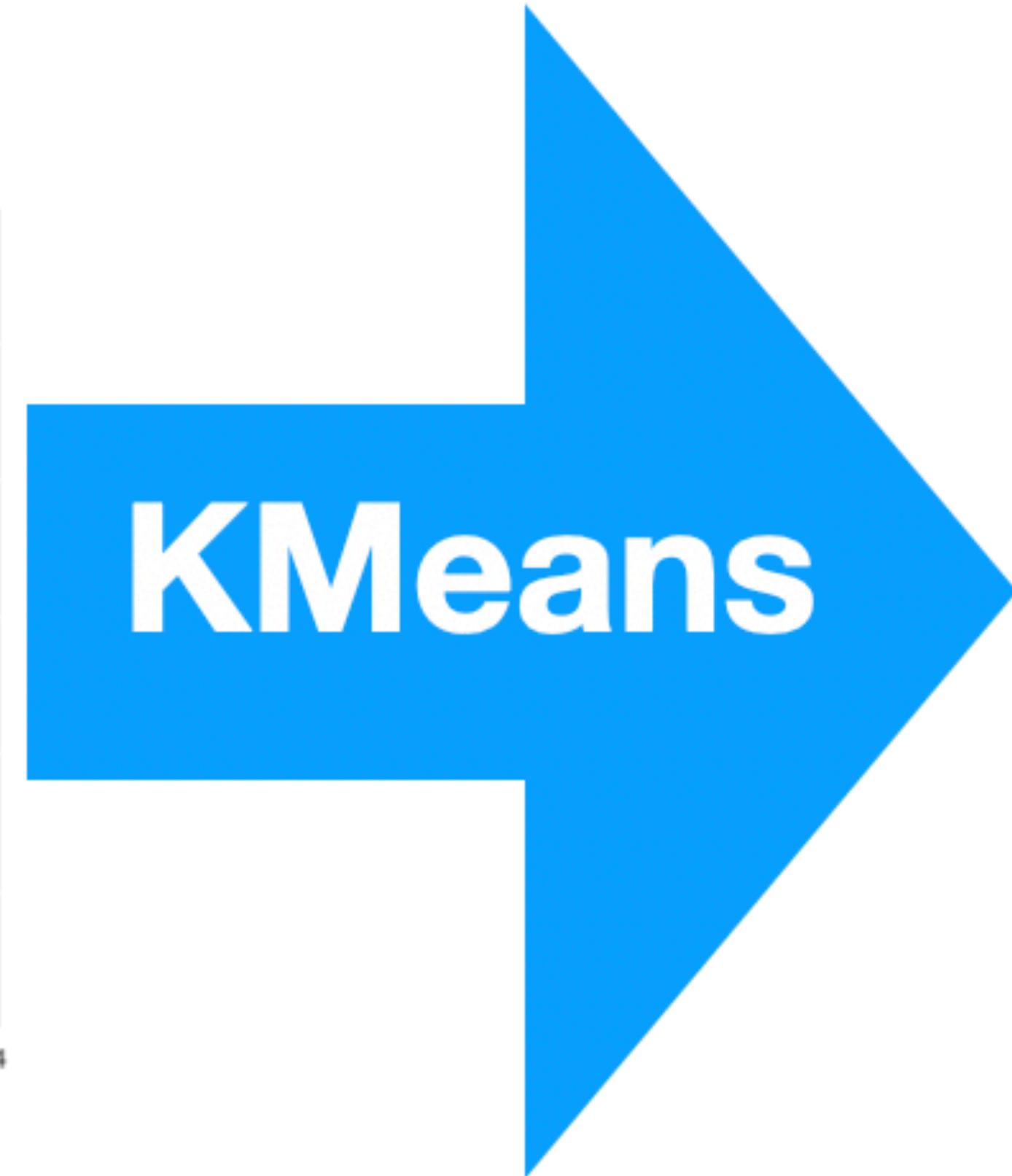
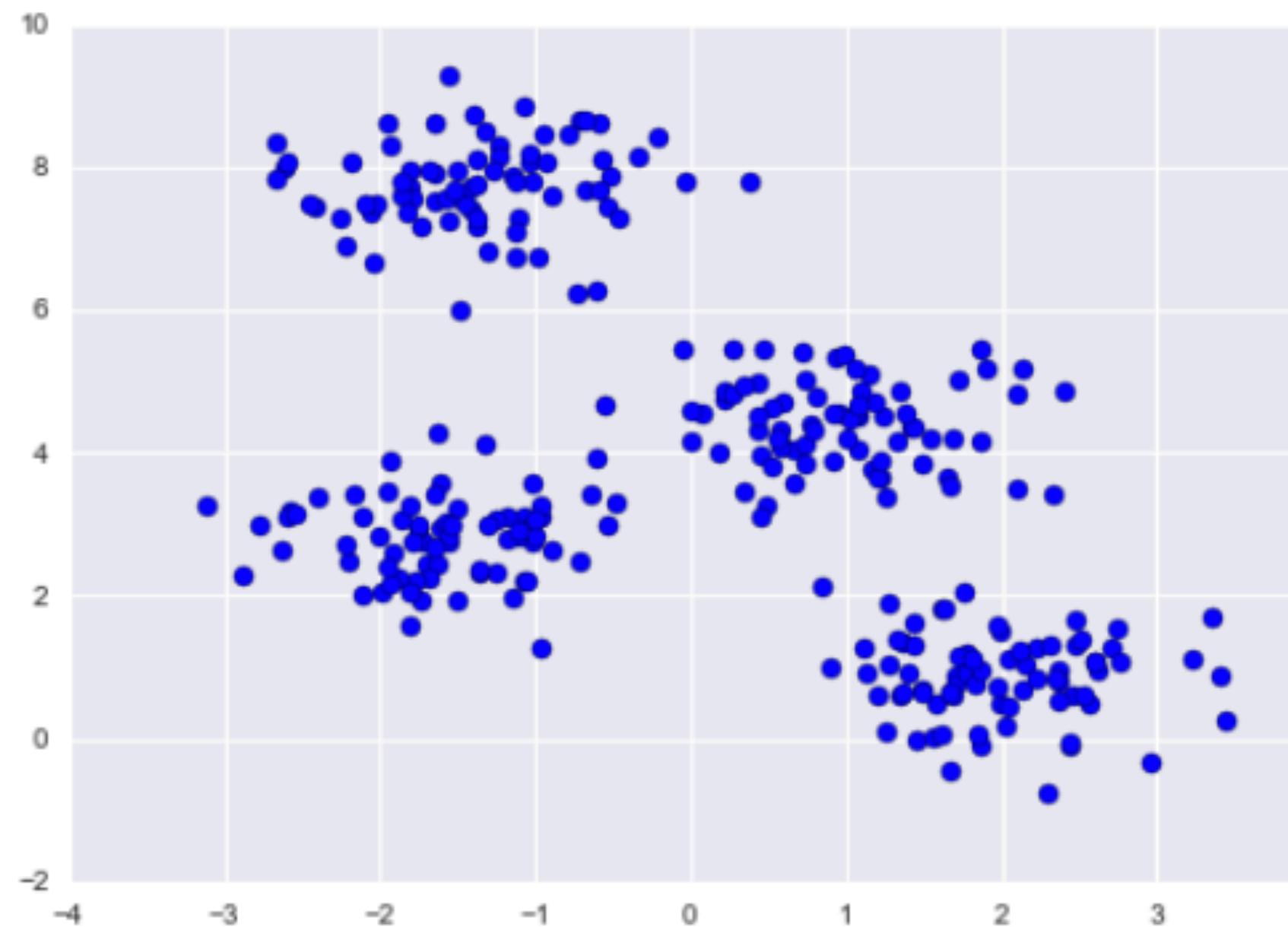
# ABNORMAL BEHAVIOURS

---



# K-MEANS CLUSTERING

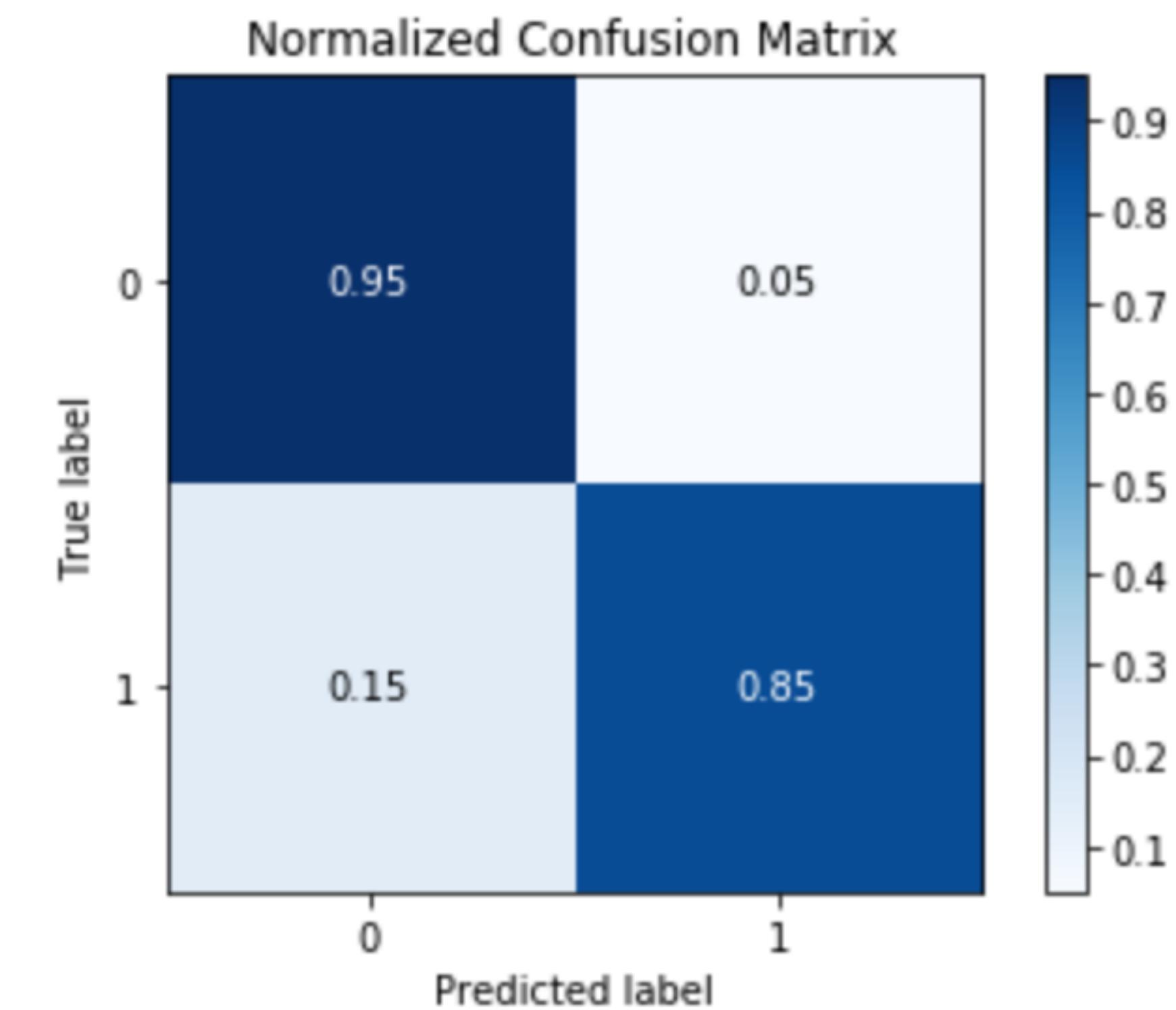
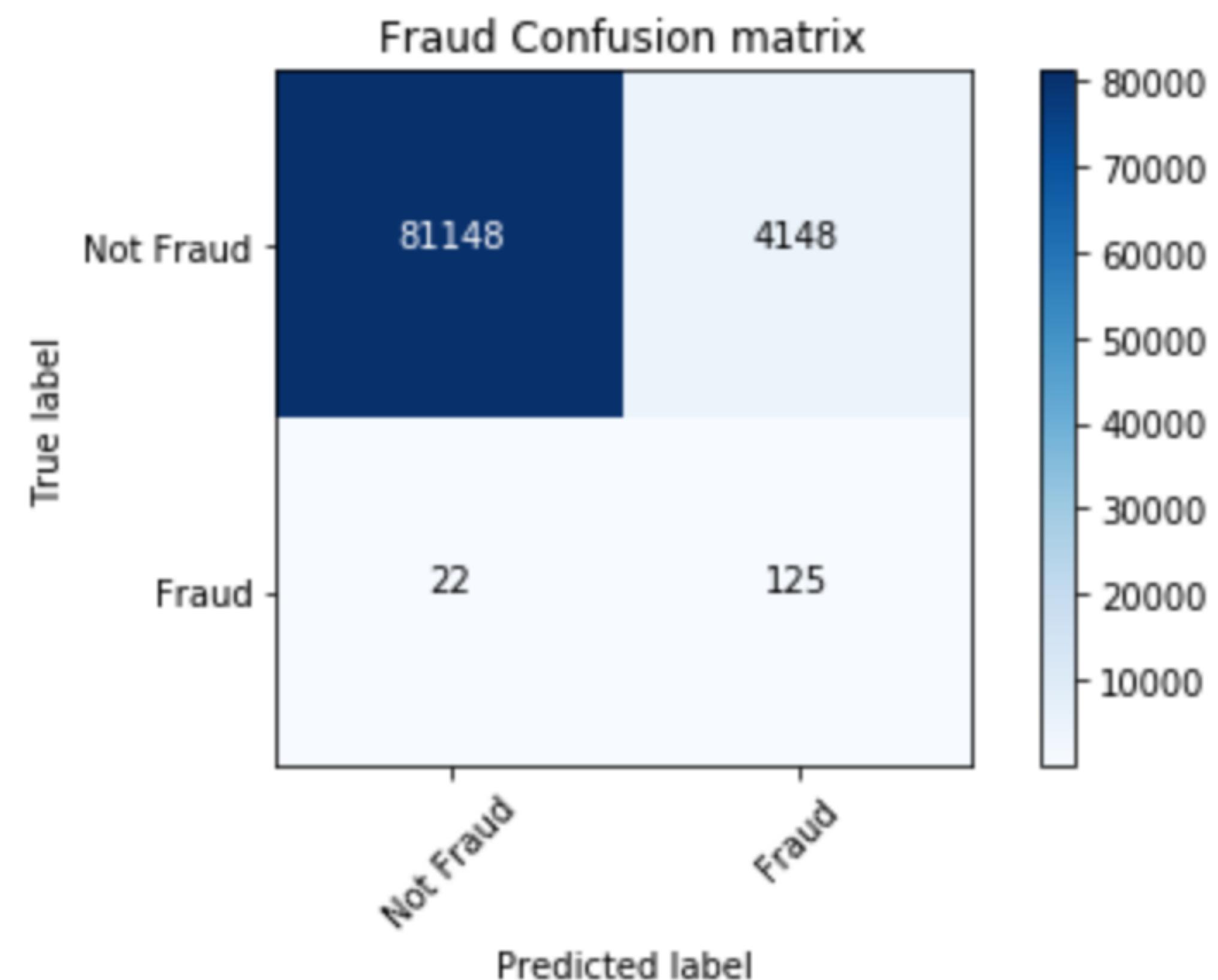
---



# K-MEANS CLUSTERING RESULT

---

Confusion matrix, without normalization



“

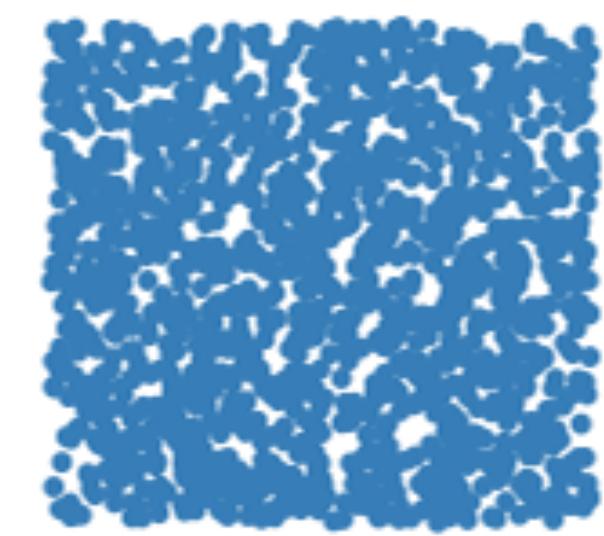
0.9

ROC score

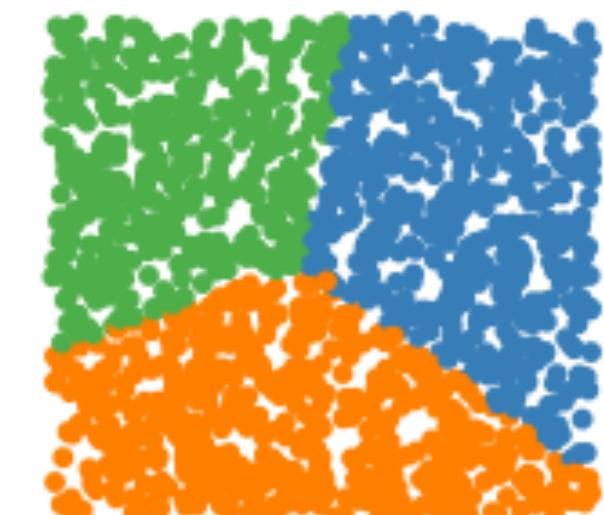
# K-MEANS V.S. DBSCAN

---

DBSCAN



k-means



“

100%

DBSCAN



# CONCLUSION

---

# SUPERVISED V.S. UNSUPERVISED

---

	Supervised	Unsupervised
Input Data	Labeled data	Unlabelled data
Number of Classes	Known	Unknown
Types	Classification	Clustering

# 1. GRID SEARCH CV

# 2. PCV-TRANSFORMED DATA

---

## RECOMMENDATIONS



