CREDIT CARD FRAUD DETECTION

RUBY FUNG

AGENDA

PROBLEM STATEMENT

METHODOLOGY

EXPLORATORY DATA ANALYSIS

SUPERVISED LEARNING

UNSUPERVISED LEARNING

CONCLUSION AND RECOMMENDATIONS

PROBLEM STATEMENT

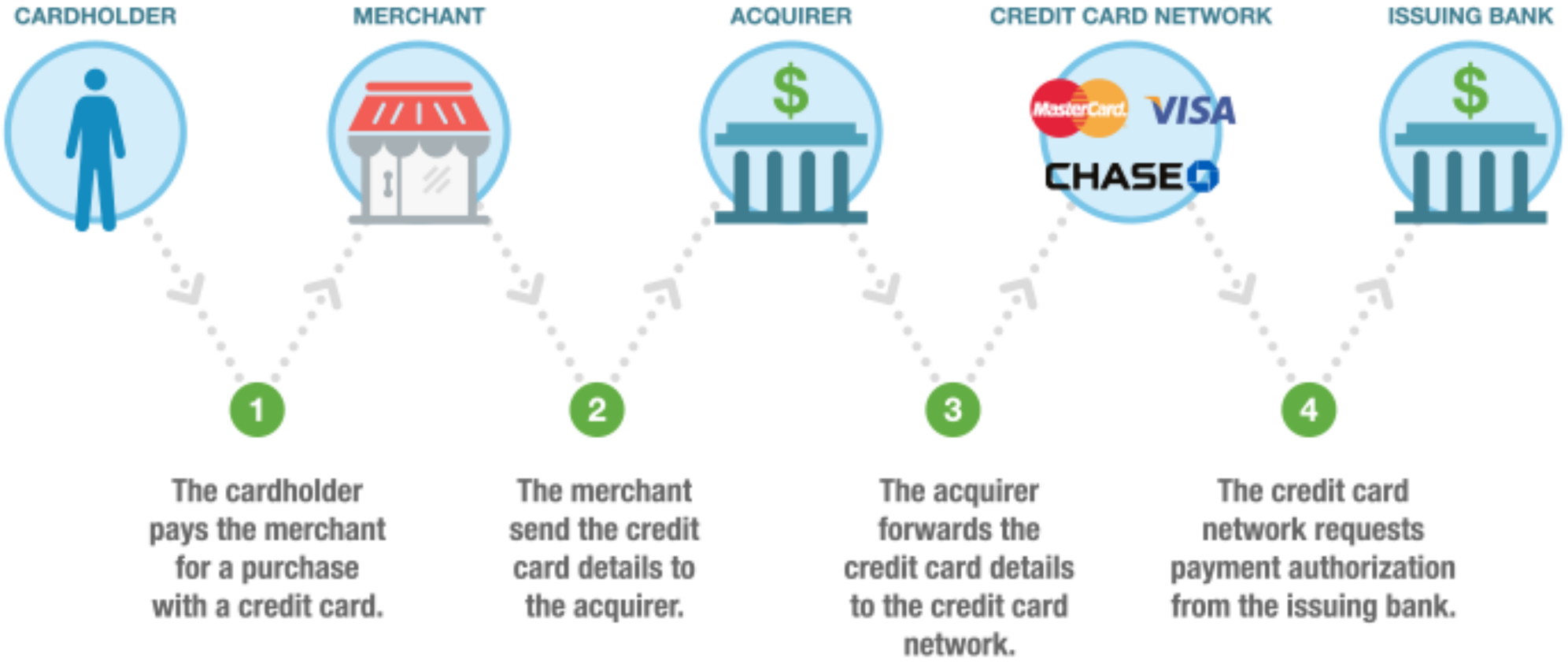
“

> $35 billion

Global card loss by 2020

-Nilson report

STAKEHOLDERS



Source: https://sites.google.com/site/martinpirjanloo23662366/

PROBLEM STATEMENT

Credit card fraud detection system

Credit card holders: Prevent loss

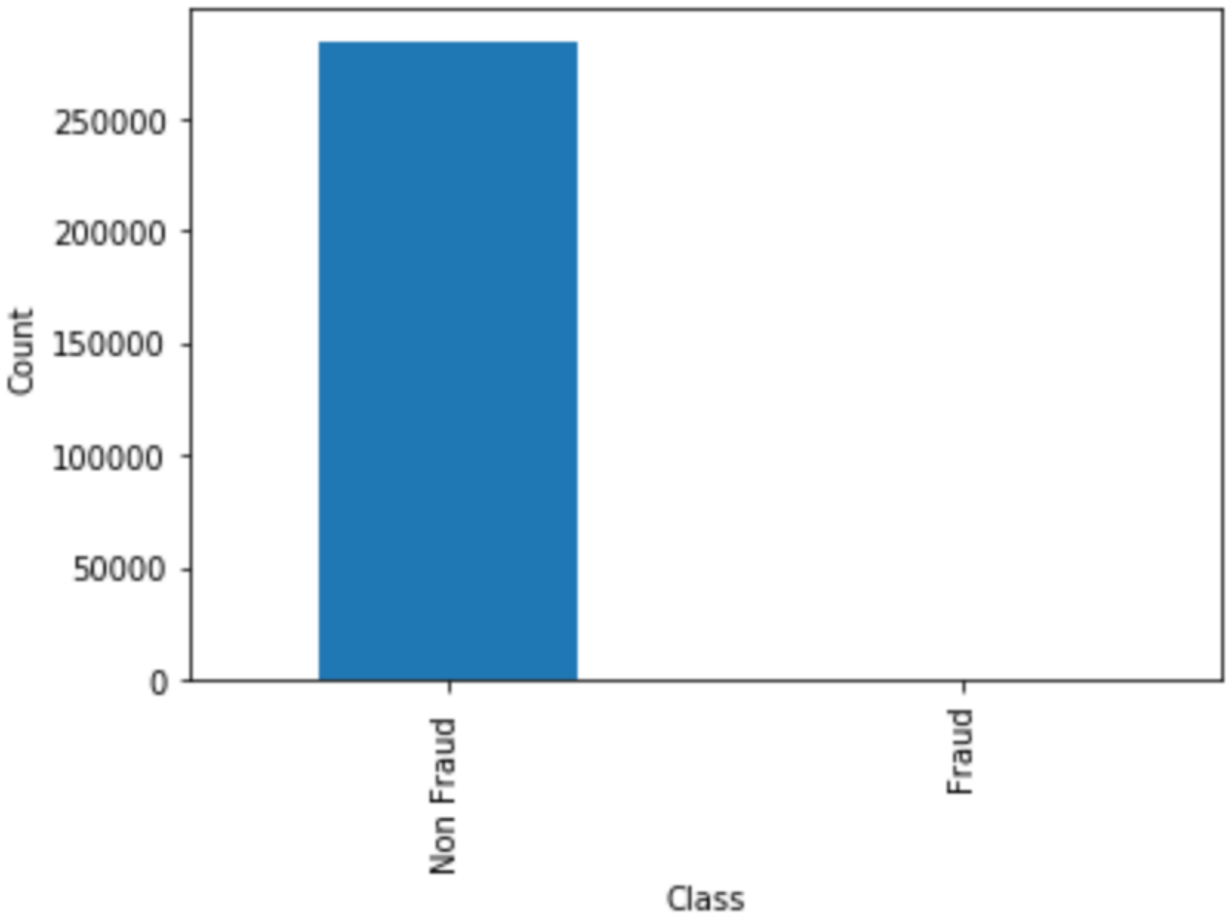
Credit card issuers: Reduce cost

EXPLORATORY DATA ANALYSIS

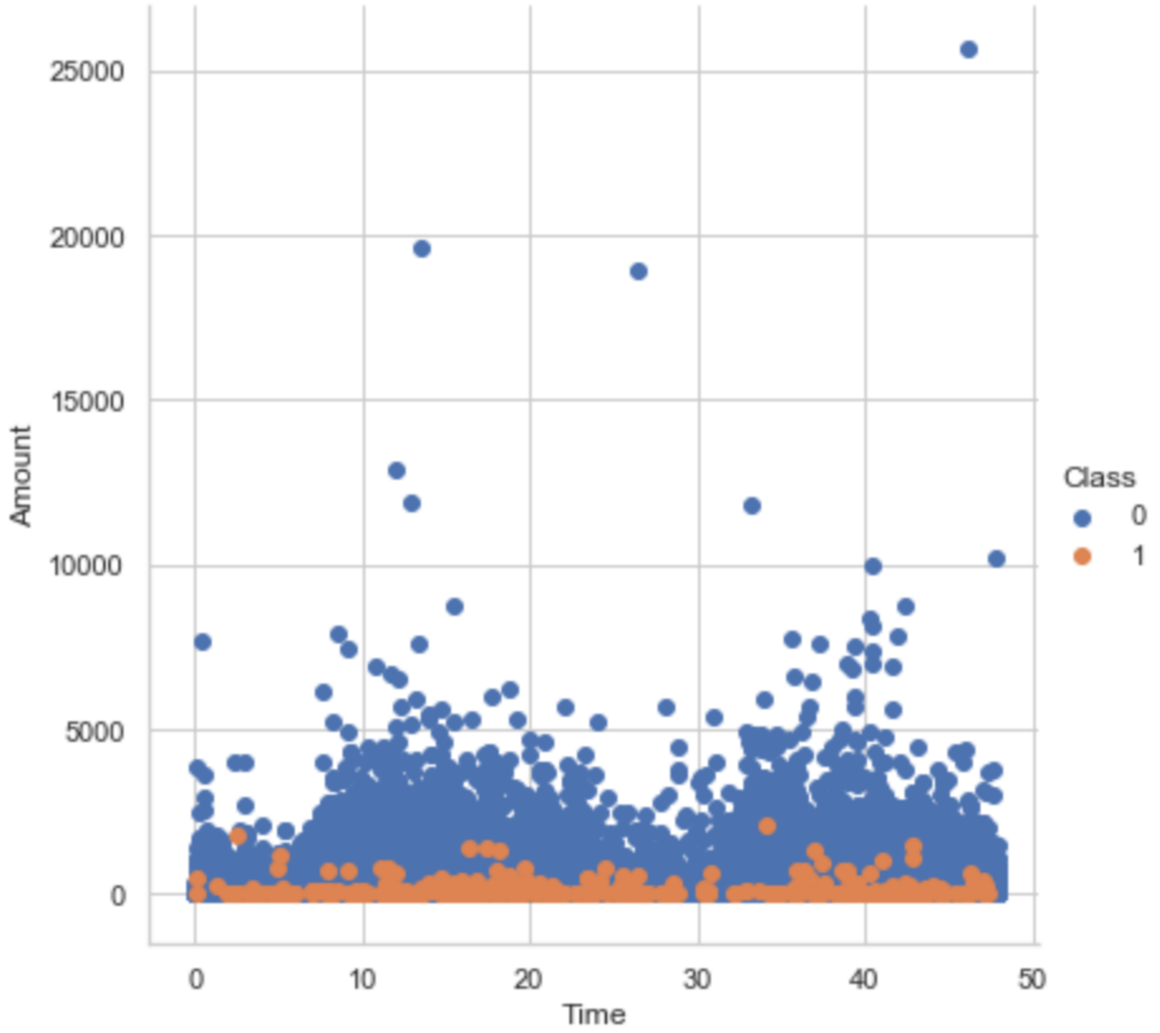
CLASS DISTRIBUTION

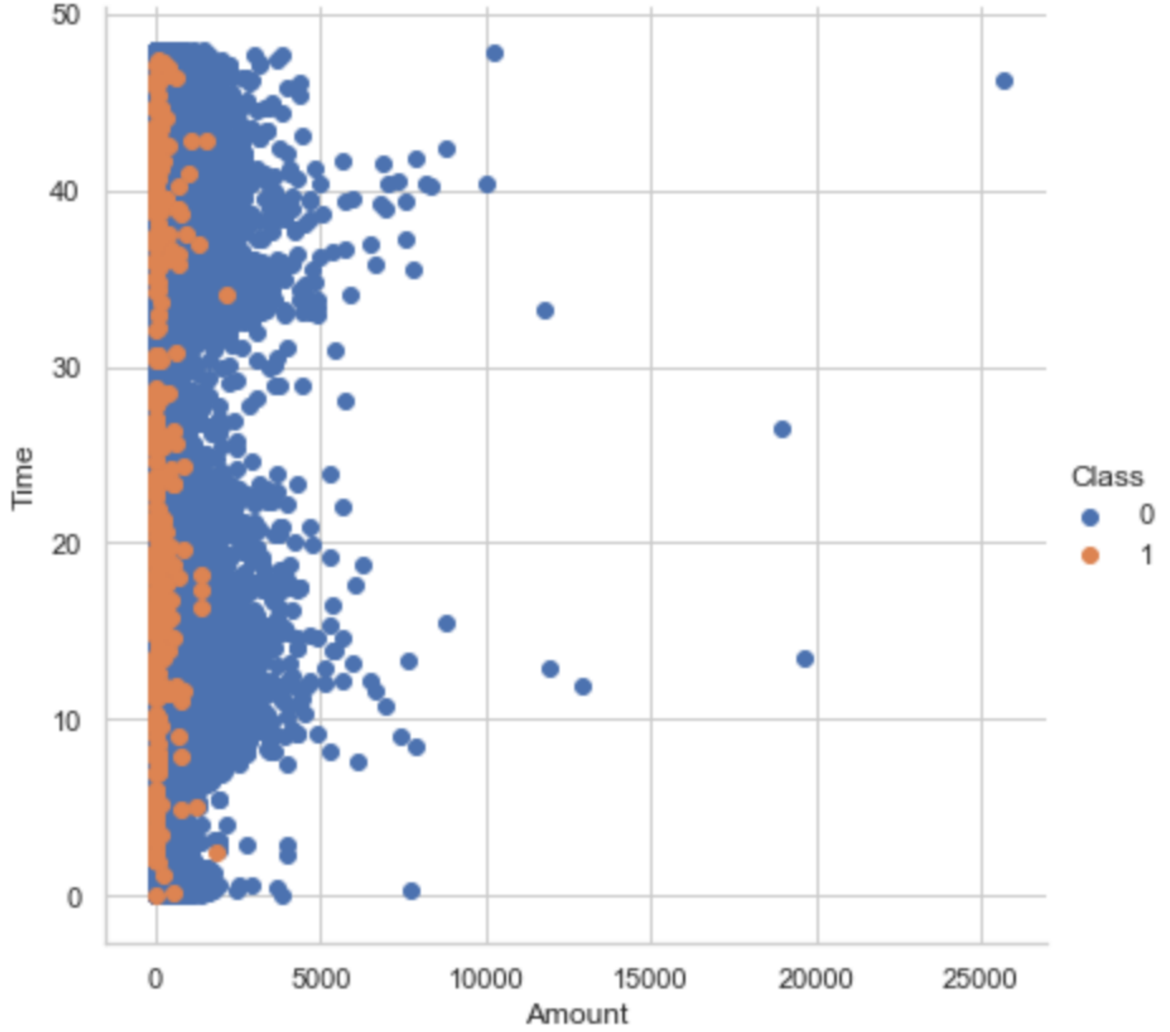
99.83%

0.17%



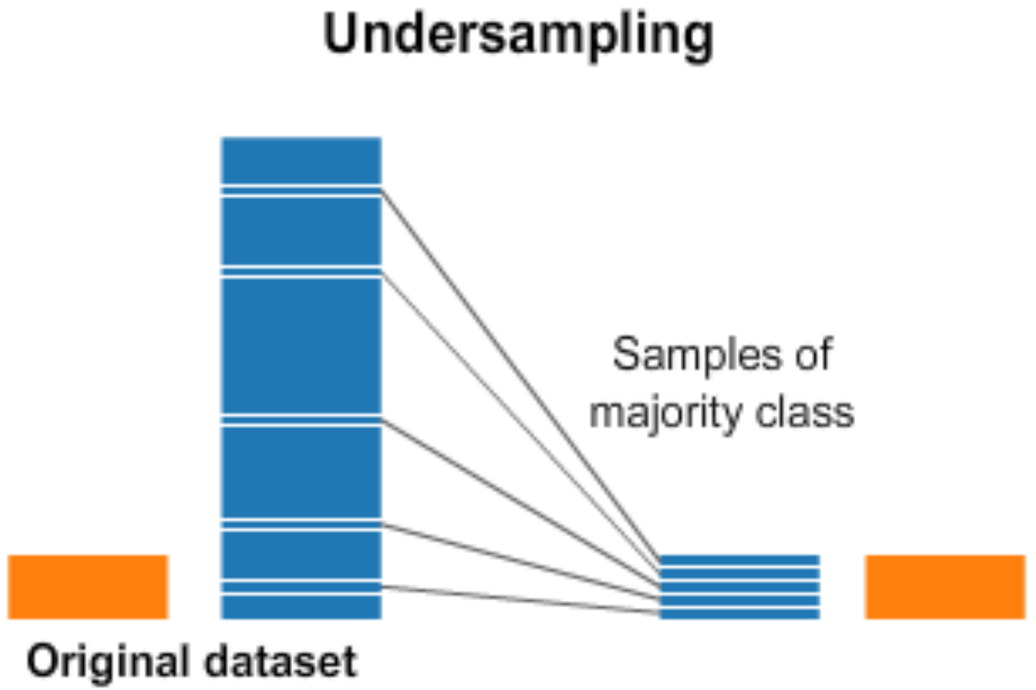
RELATIONSHIP BETWEEN AMOUNT & TIME

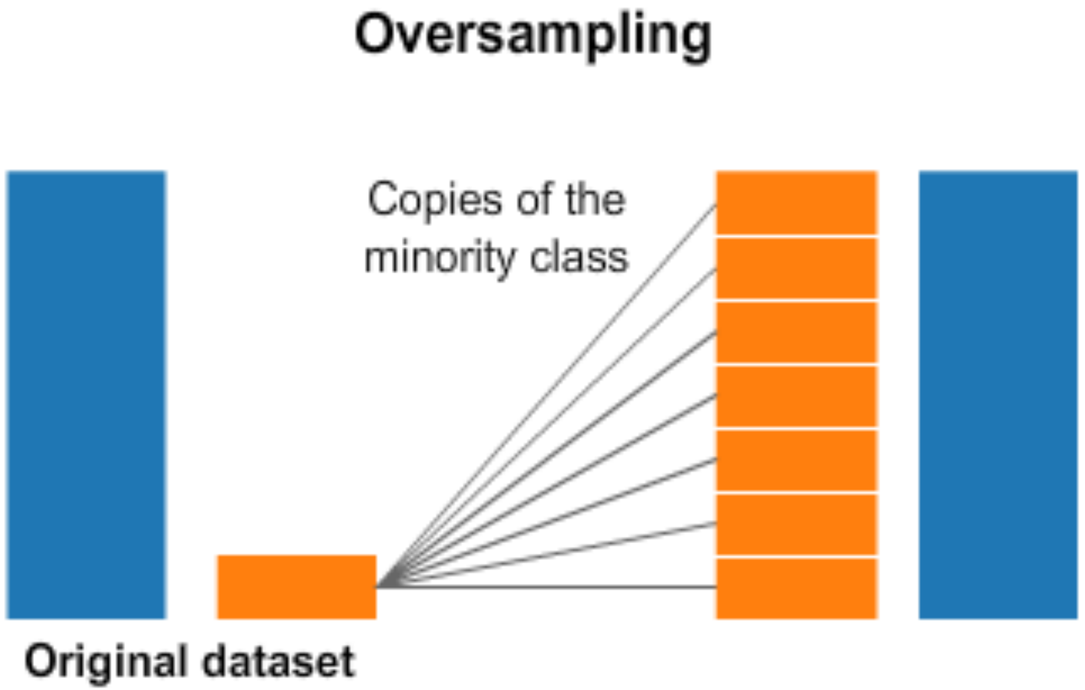




SUPERVISED LEARNING

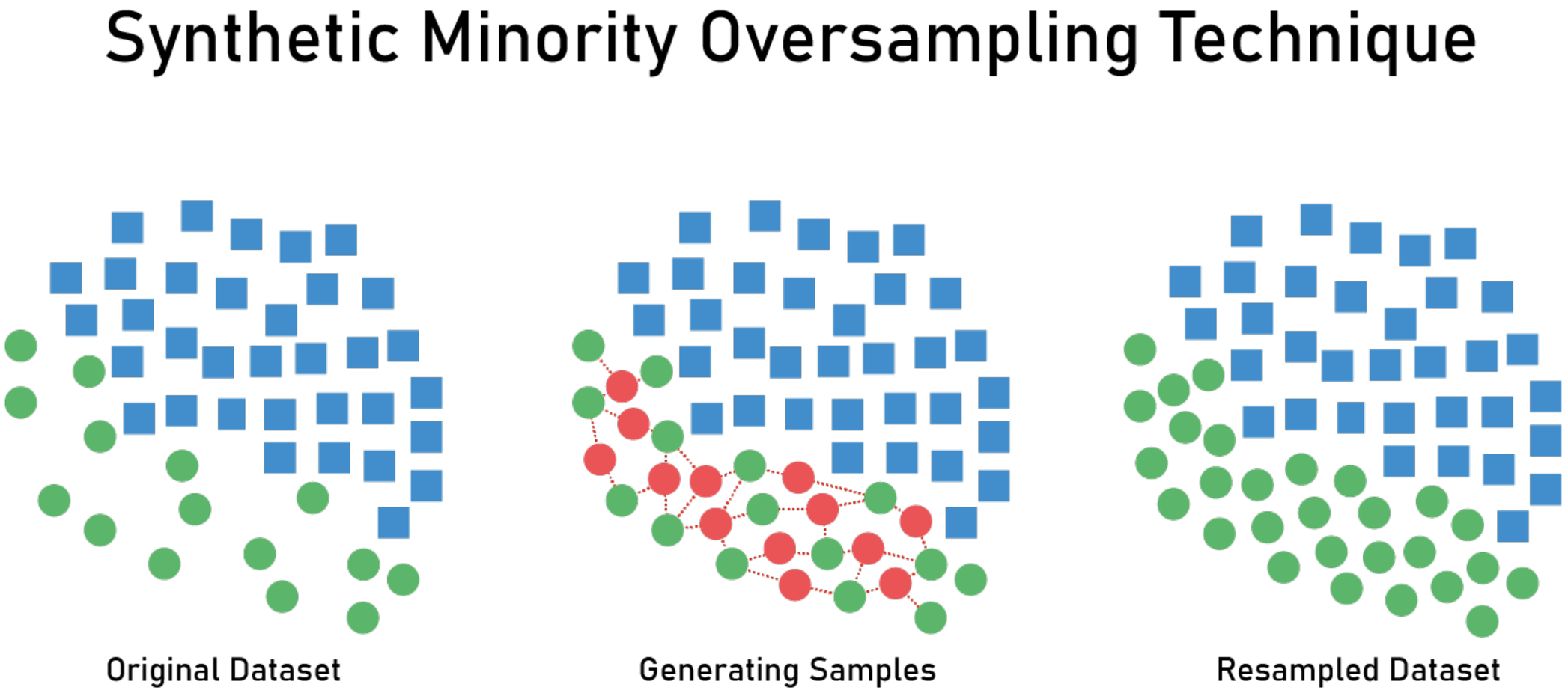
RESAMPLING METHODS





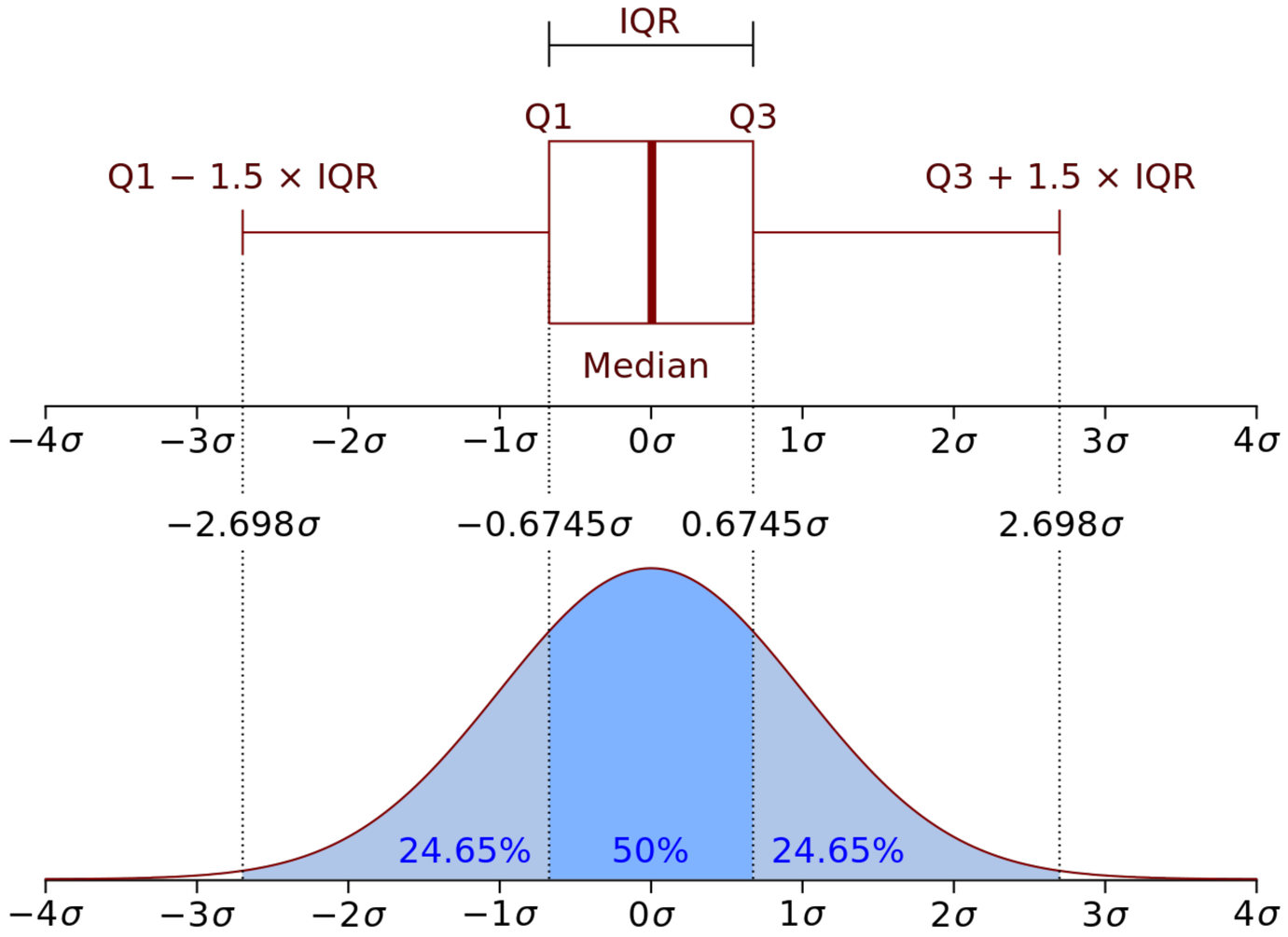
Source: https://www.kdnuggets.com/2020/01/5-most-useful-techniques-handle-imbalanced-datasets.html

SMOTE

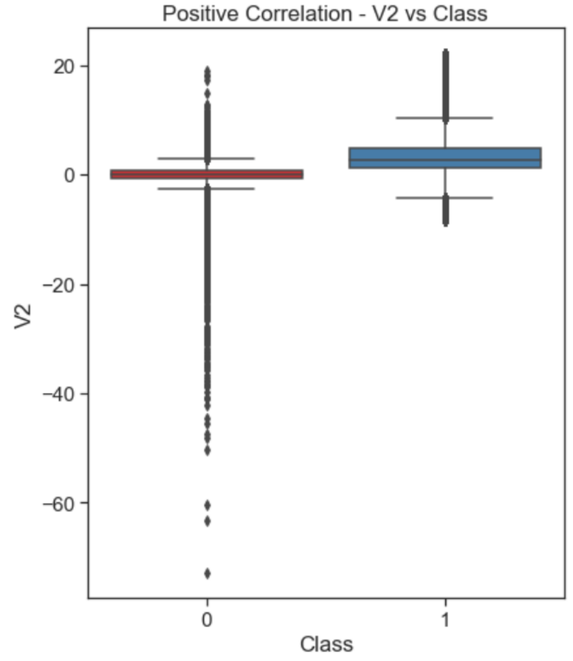


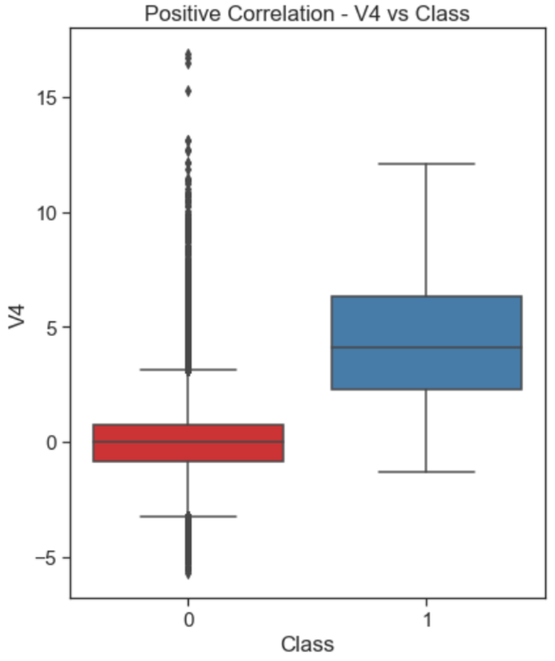
Source: https://medium.com/analytics-vidhya/bank-data-smote-b5cb01a5e0a2

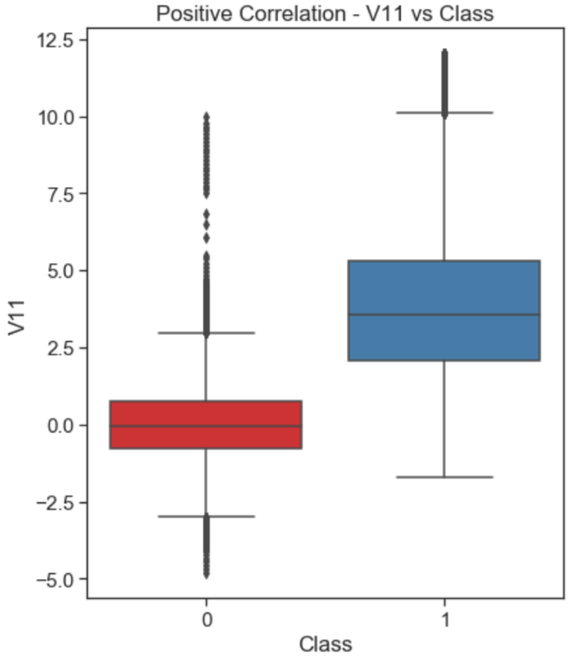
INTERQUARTILE RANGE (IQR)

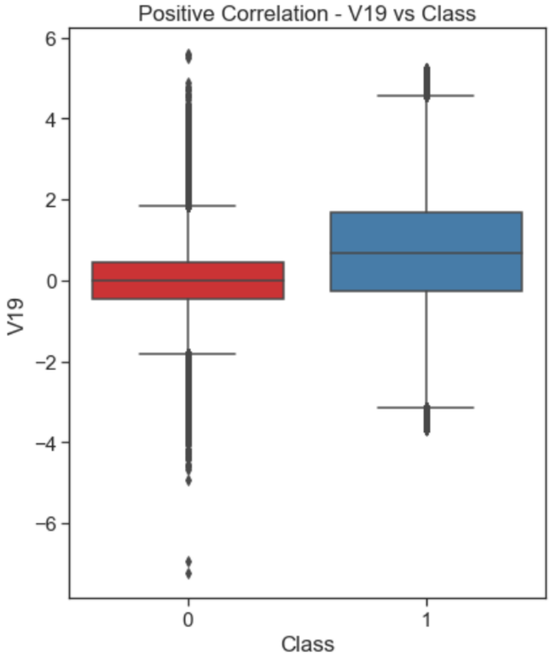


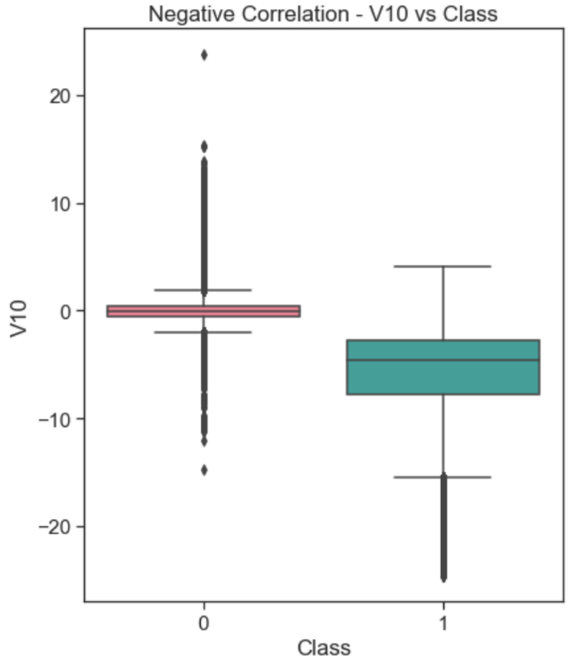
Source: https://en.wikipedia.org/wiki/Interquartile_range

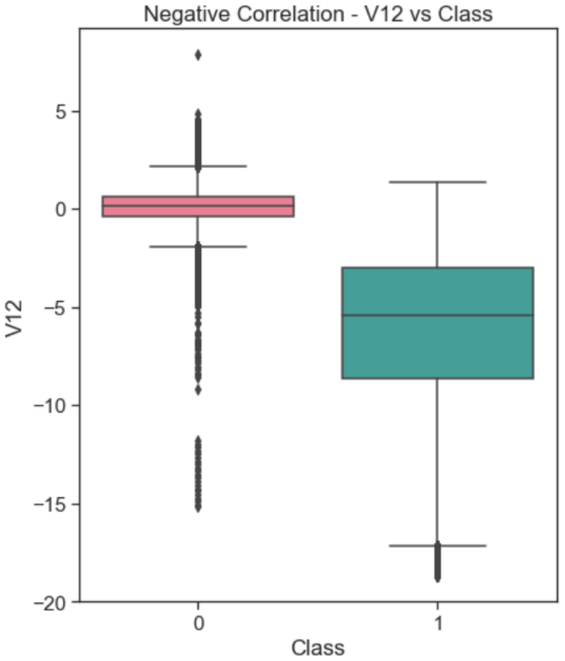

Positive Correlation - V2 vs Class
V2
Class

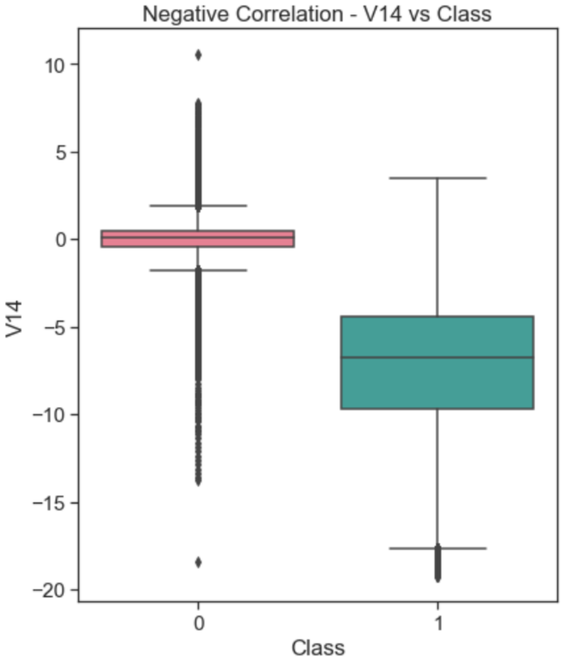

Positive Correlation - V4 vs Class
V4
Class

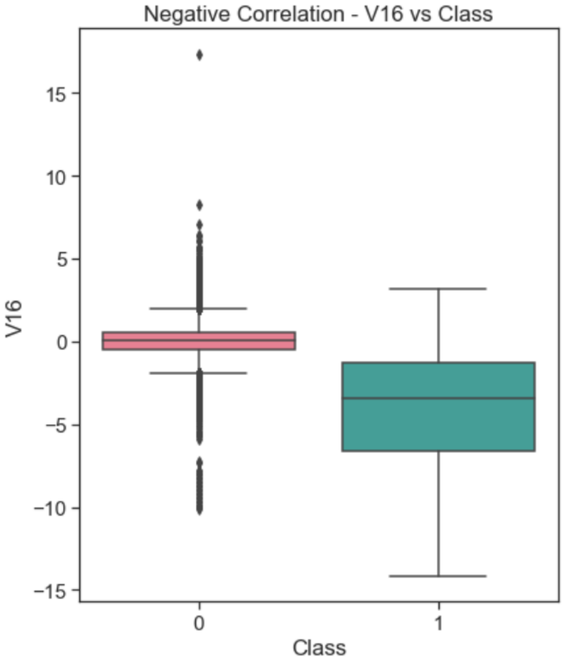

Positive Correlation - V11 vs Class
V11
Class


Positive Correlation - V19 vs Class
V19
Class


Negative Correlation - V10 vs Class
V10
Class


Negative Correlation - V12 vs Class
V12
Class

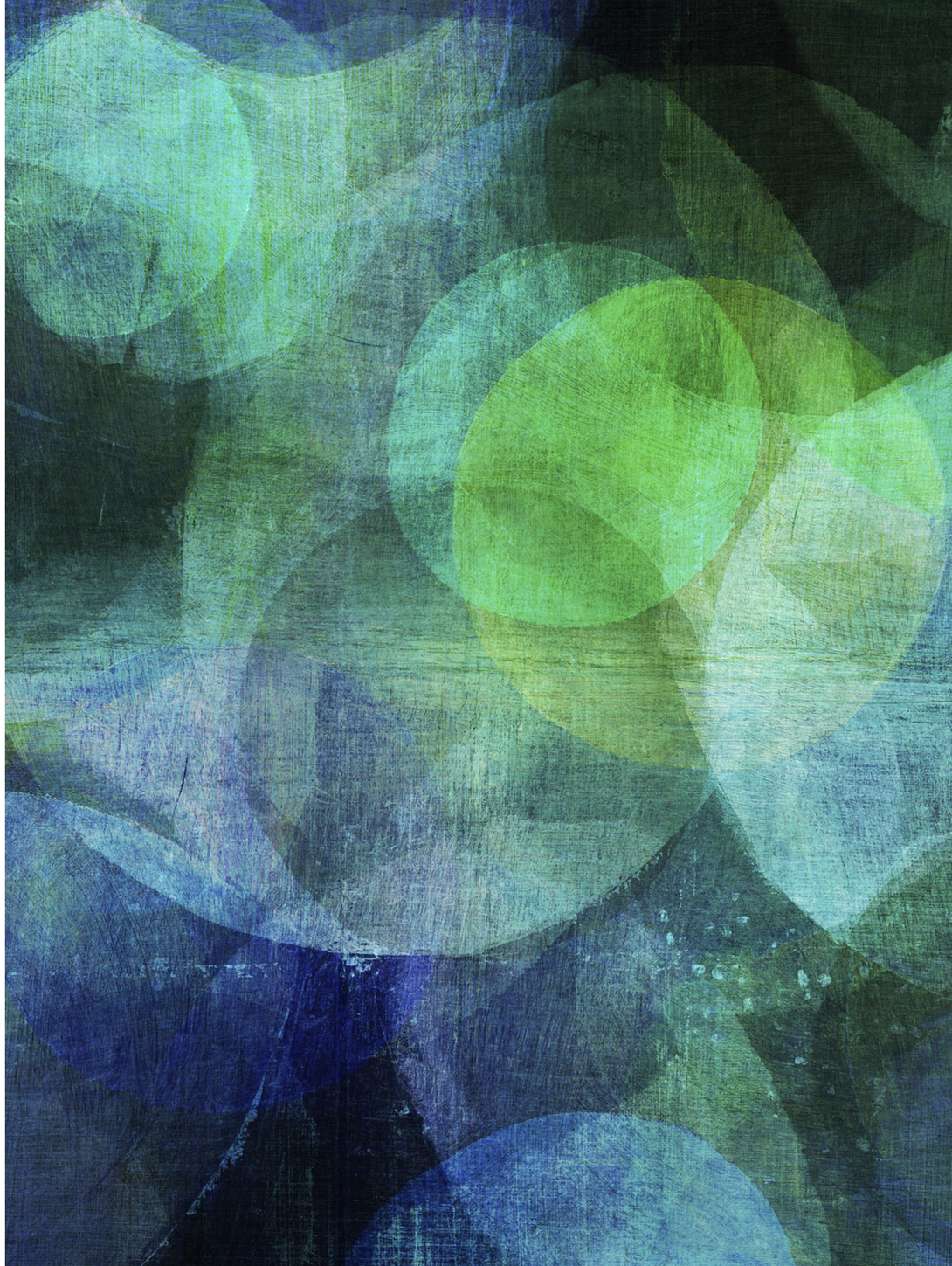

Negative Correlation - V14 vs Class
V14
Class


Negative Correlation - V16 vs Class
V16
Class

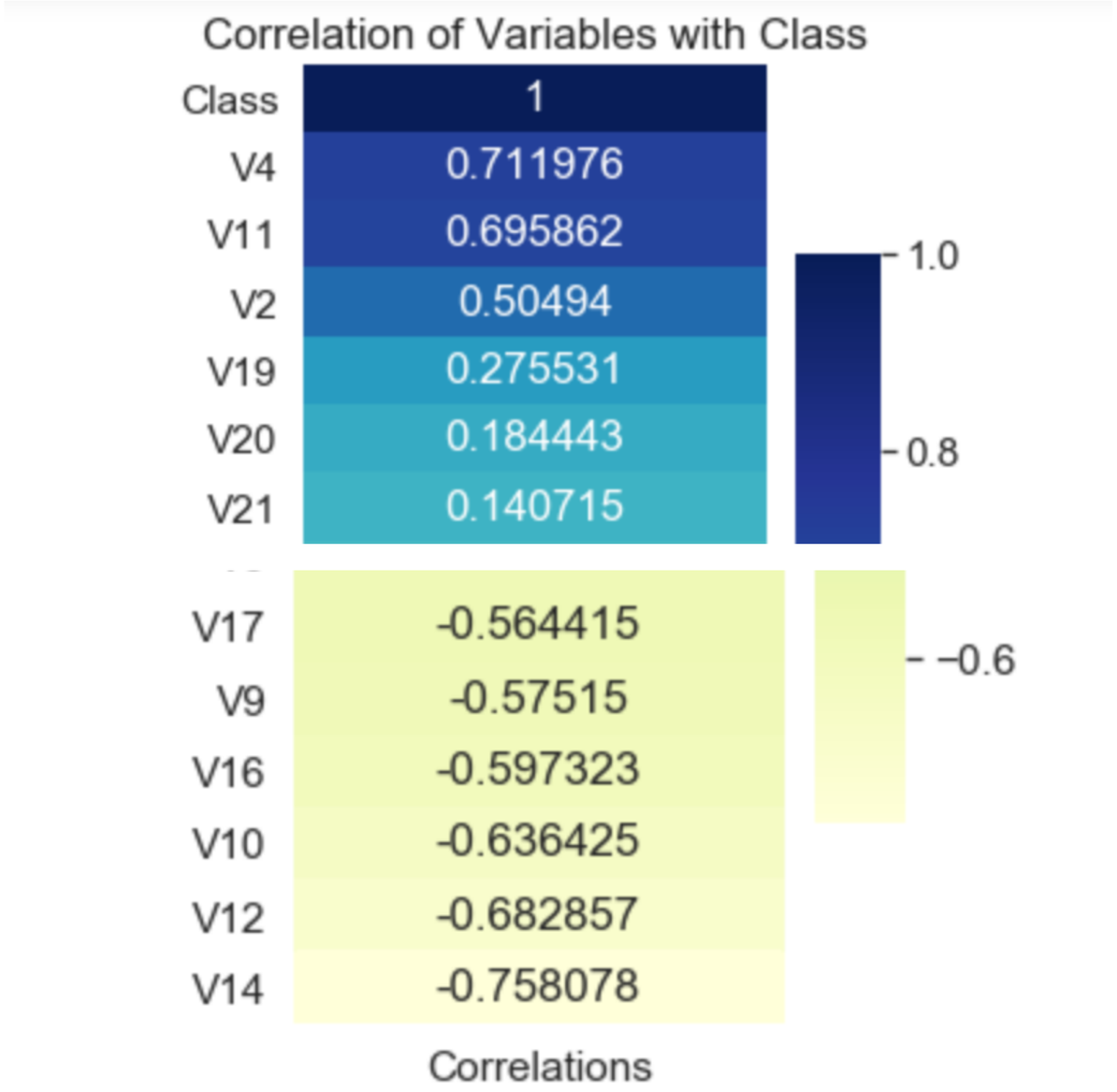
REMOVAL OF EXTREME OUTLIERS

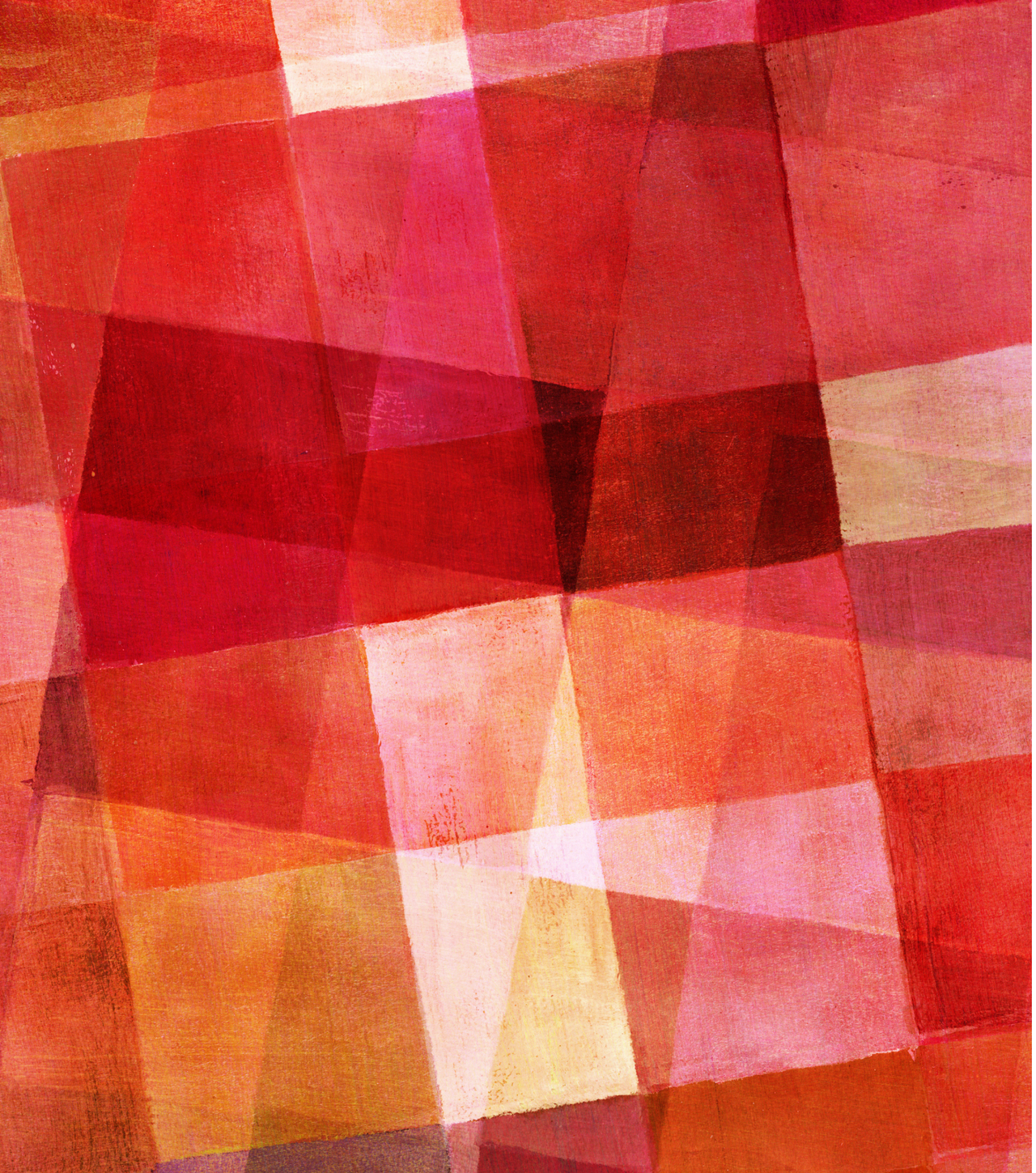
FROM TOP 4 FEATURES WHICH ARE POSITIVELY & NEGATIVELY CORRELATED WITH CLASS

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



HIGHEST POSITIVE AND NEGATIVE CORRELATIONS



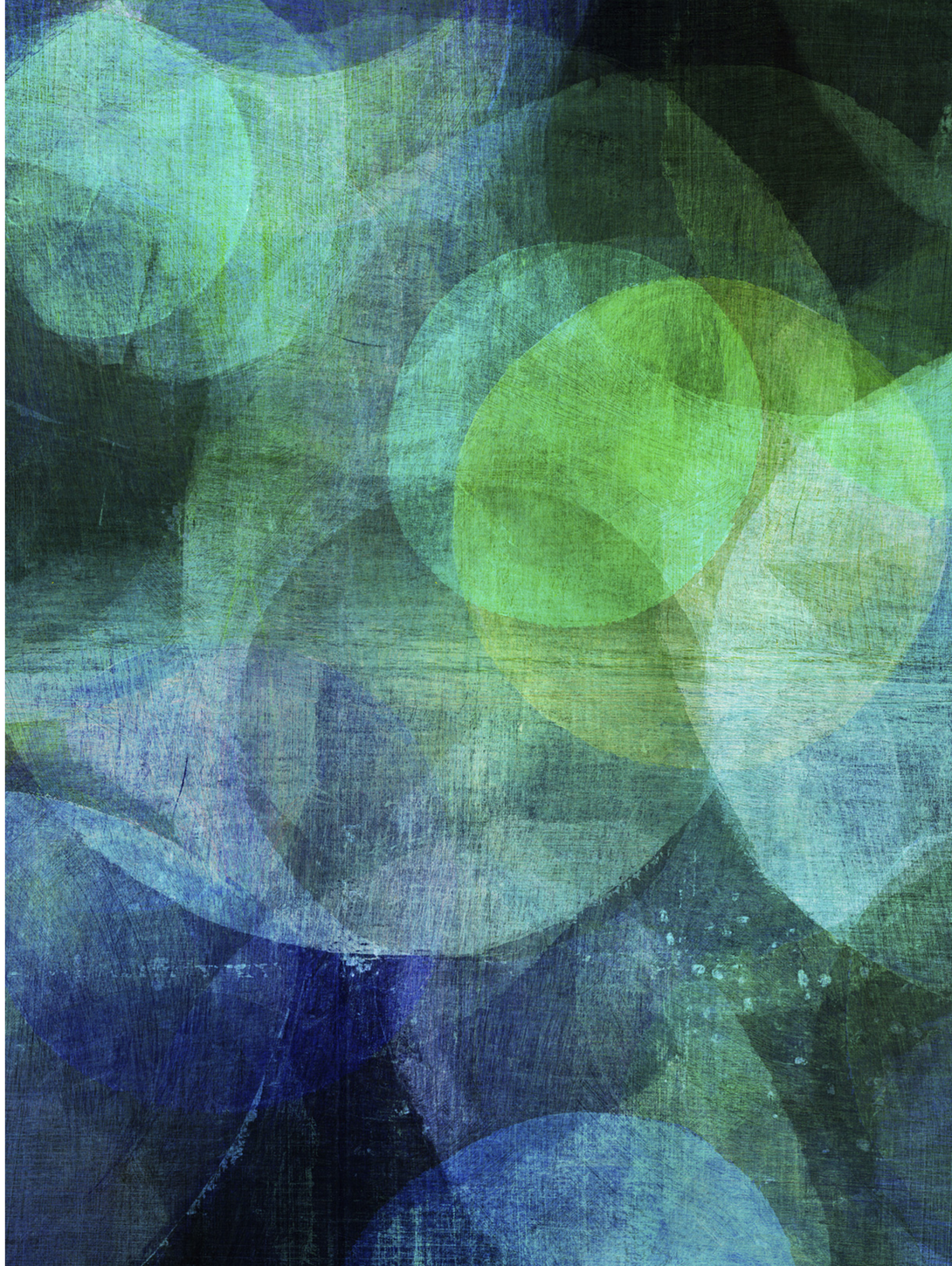


5 CLASSIFIERS

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

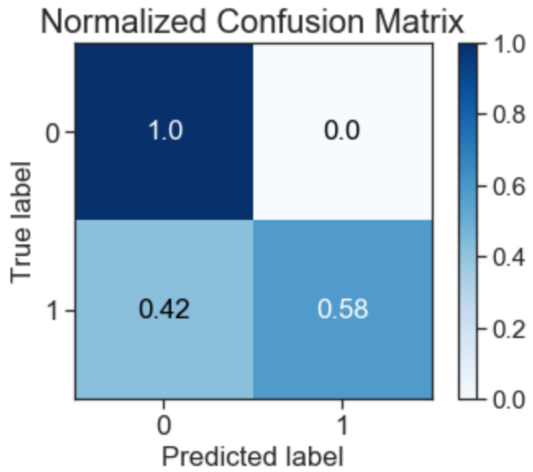
1. Classification report
2. Confusion matrix
3. Precision-Recall Curve
4. ROC Curve and AUC

PERFORMANCE EVALUATION

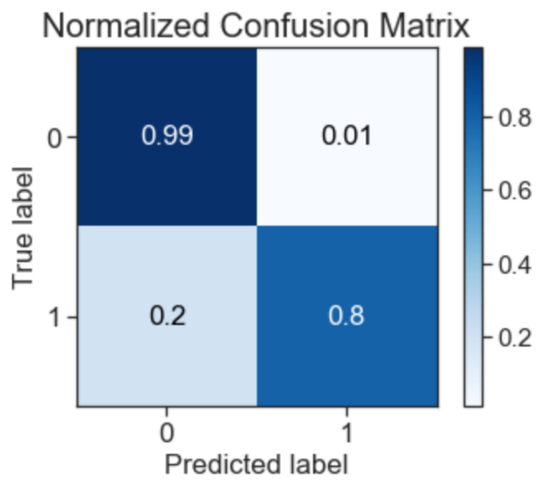


CONFUSION MATRIX

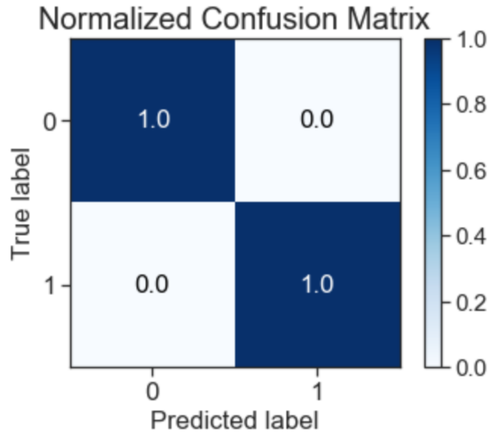
Confusion Matrix for ['Baseline Logistic Regression']



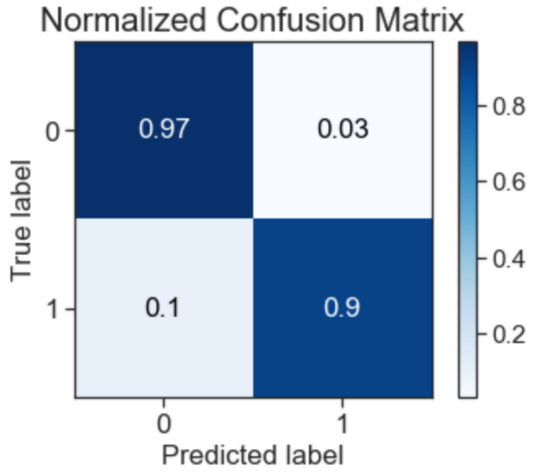
Confusion Matrix for ['Naive Bayes']



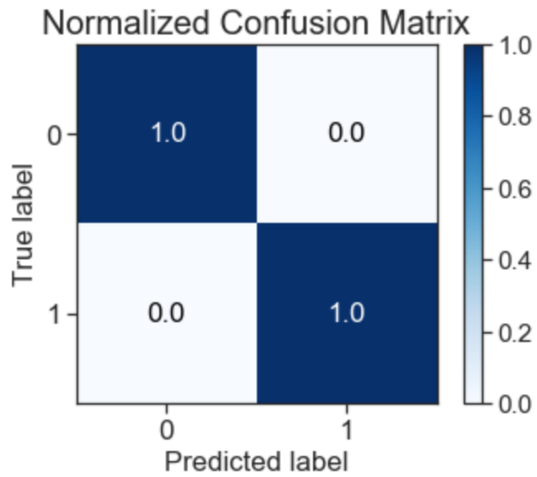
Confusion Matrix for RandomForestClassifier()



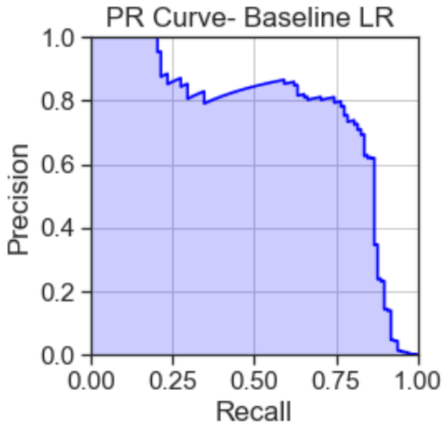
Confusion Matrix for ['Optimized Logistic Regression']

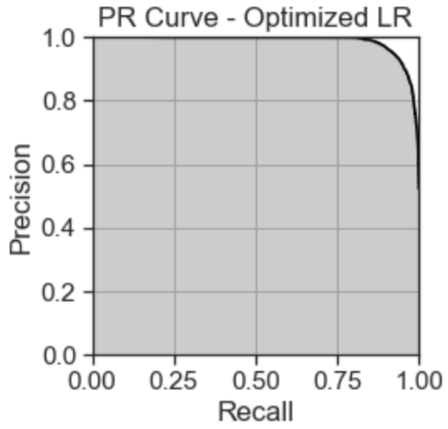


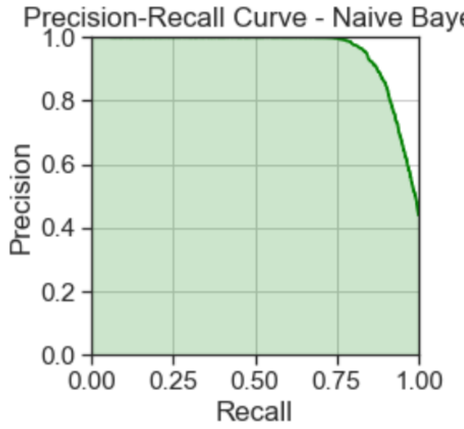
Confusion Matrix for ['Decision Tree']

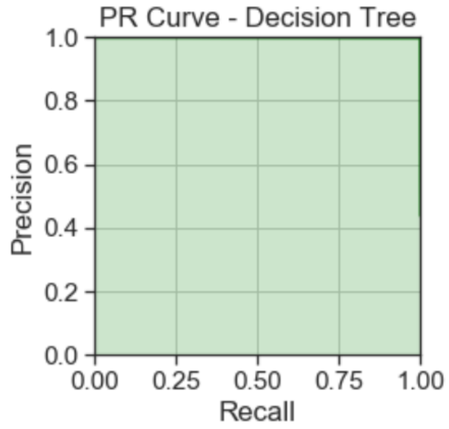


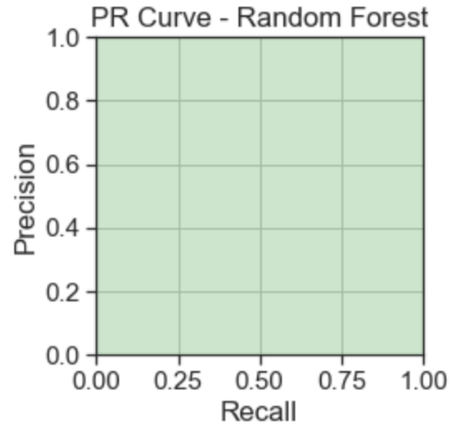
PRECISION-RECALL CURVE, ROC CURVE

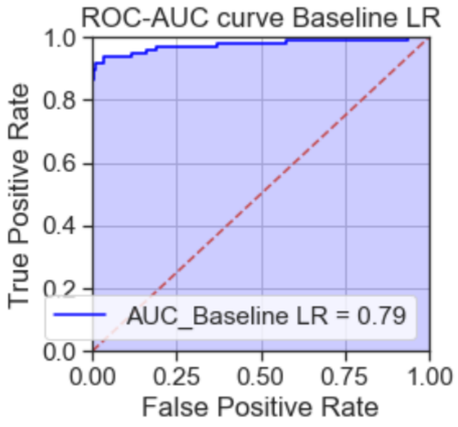


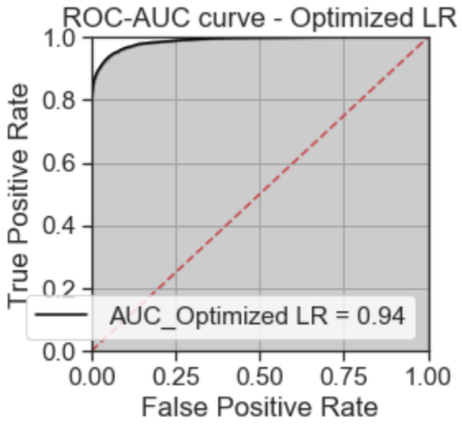


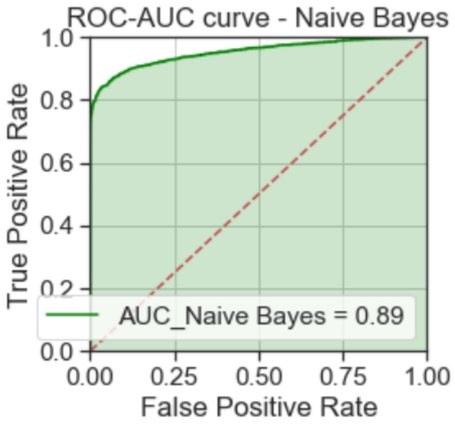


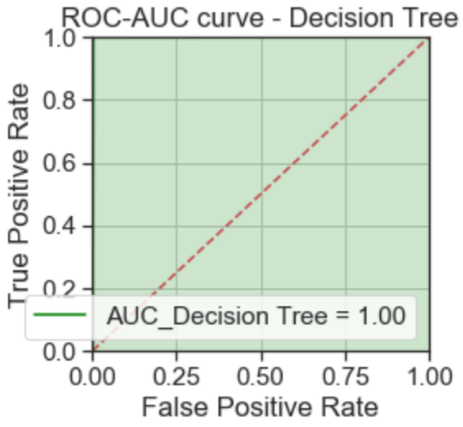


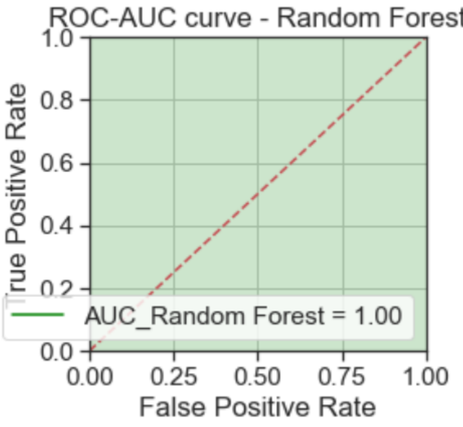








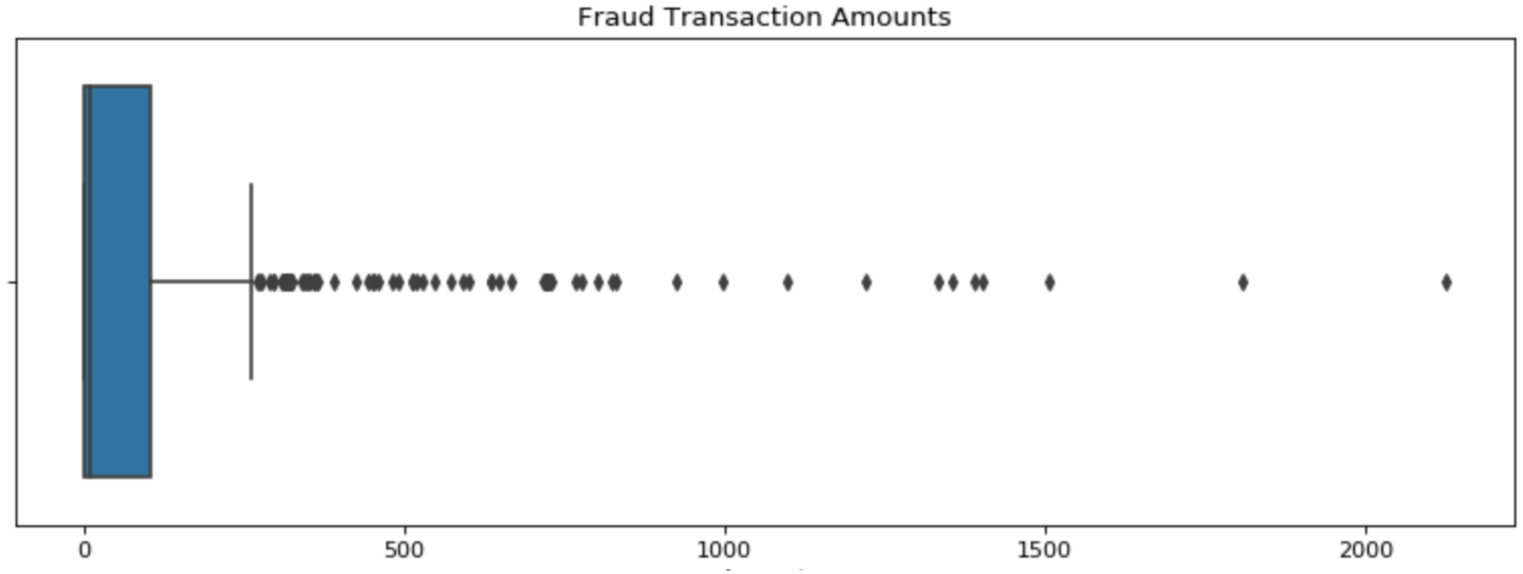


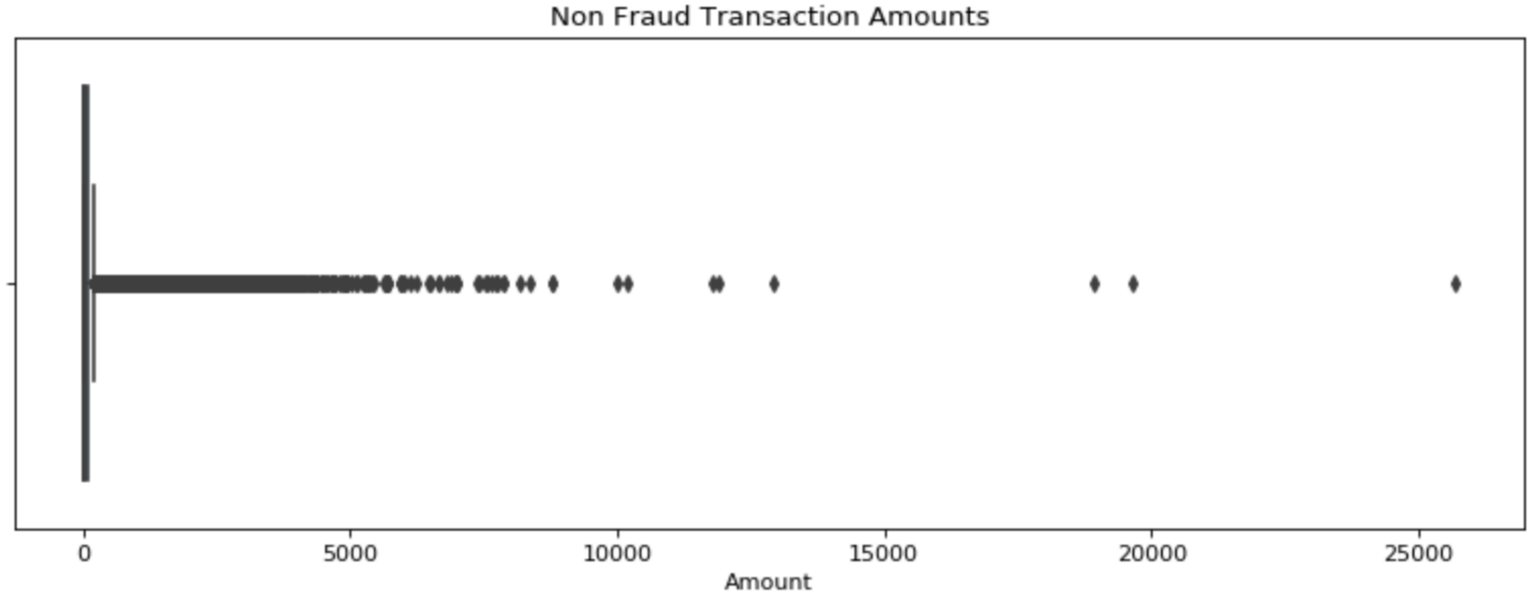


Classifier	SMOTE - Accuracy rate(%)	RUS - Accuracy rate(%)
Baseline Logistic Regression	Train 99.92, Test 99.91	Train 99.92, Test 99.91
Optimized Logistic Regression	Train 98.75, Test 98.78	Train 98.17, Test 97.65
Naive Bayes	Train 90.28, Test 90.37	Train 91.38, Test 91.11
Decision Tree	Train 100, Test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

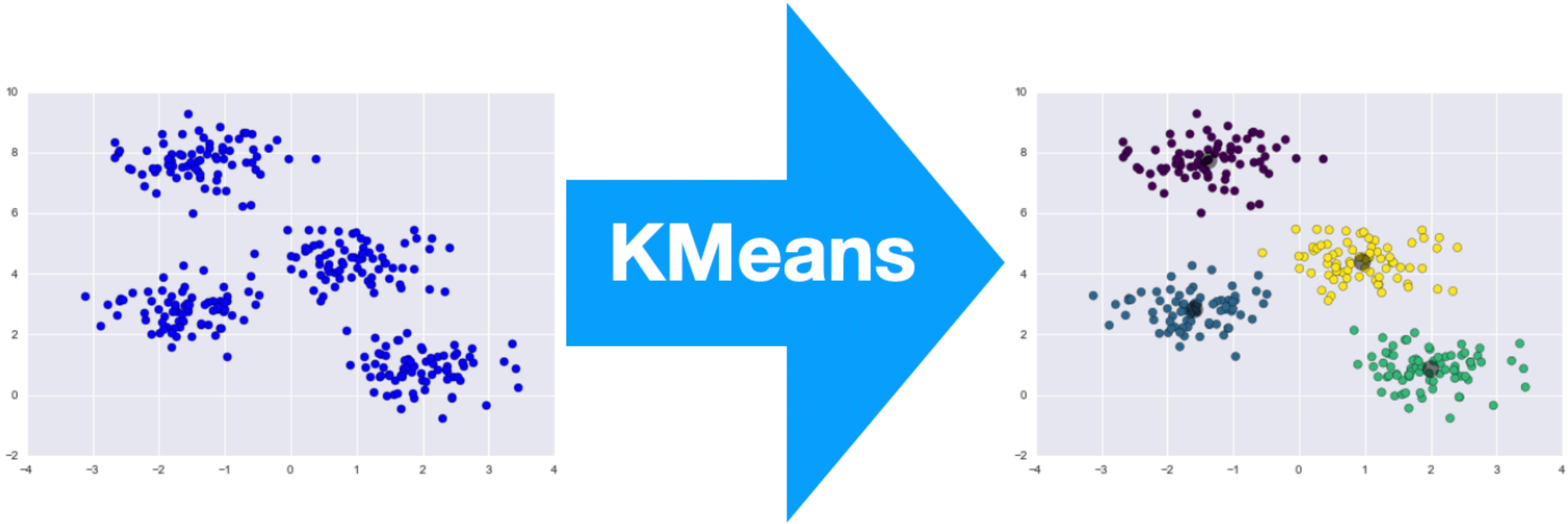
UNSUPERVISED LEARNING

ABNORMAL BEHAVIOURS





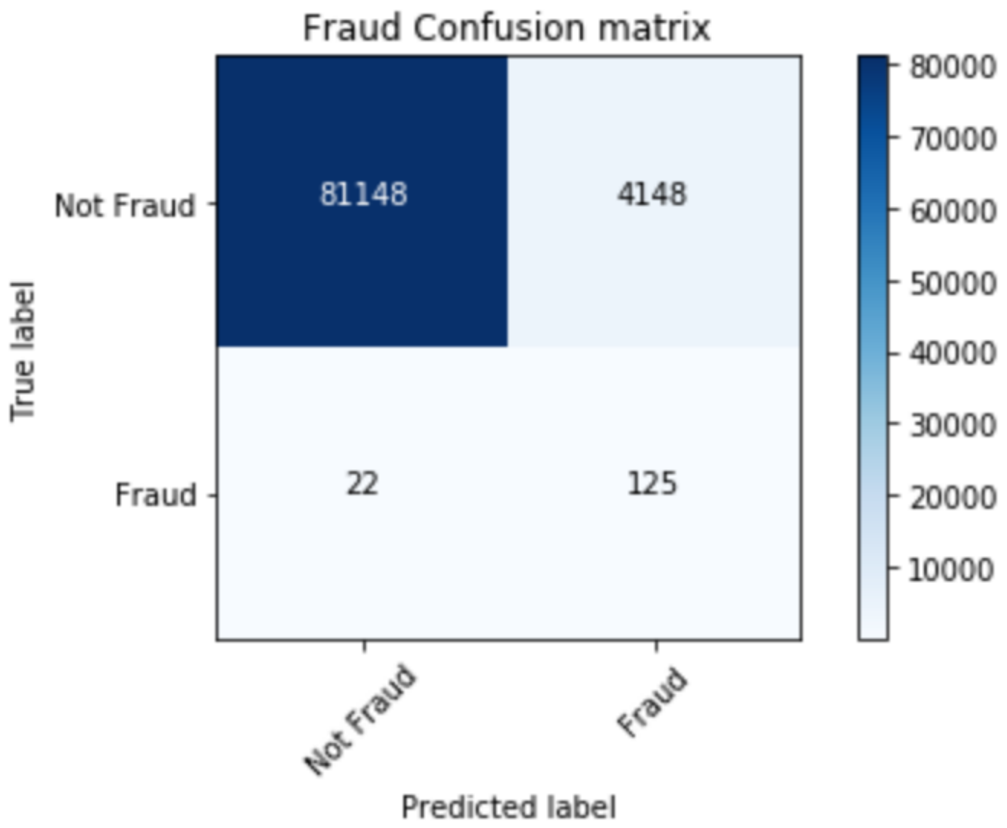
K-MEANS CLUSTERING

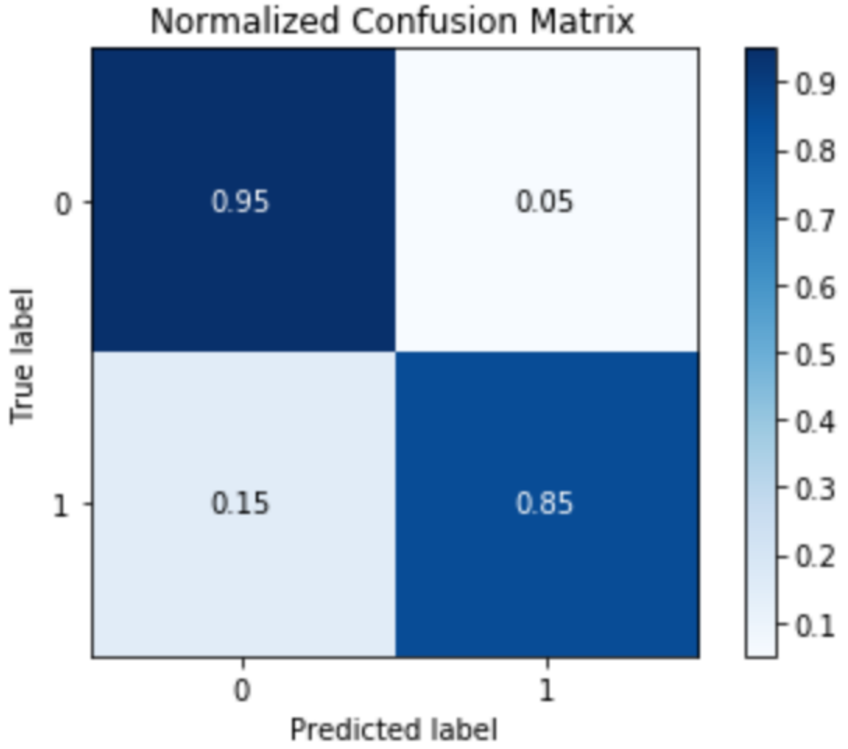


Source: https://aihub.cloud.google.com/u/0/p/products%2F0e0d2ed0-5563-4639-b348-53a83ac4ff4e

K-MEANS CLUSTERING RESULT

```
Confusion matrix, without normalization
```





“

0.9

ROC score

K-MEANS V.S. DBSCAN

DBSCAN

k-means

“

100%

DBSCAN

CONCLUSION

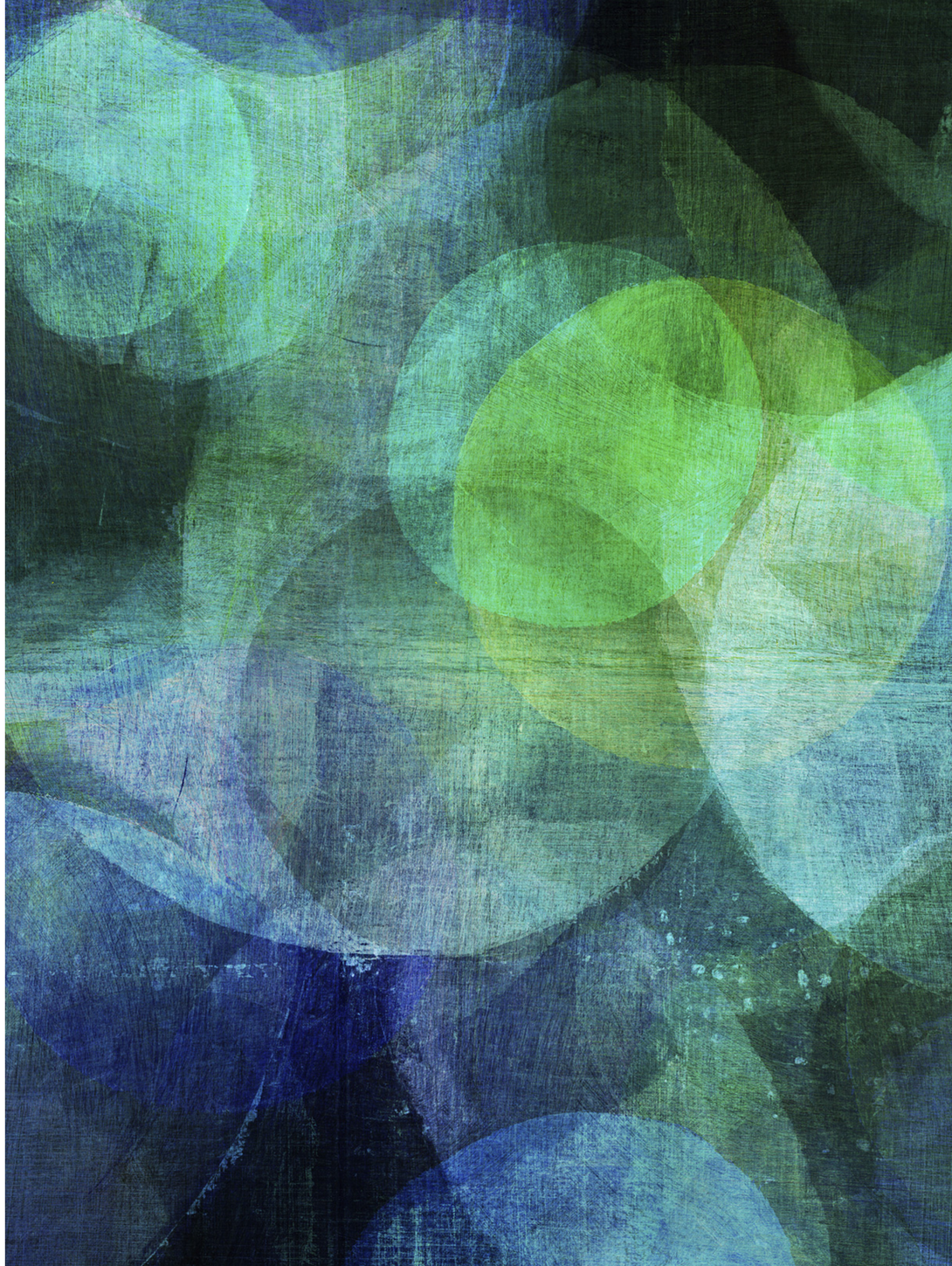
SUPERVISED V.S. UNSUPERVISED

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

1. GRID SEARCH CV

2. PCV-TRANSFORMED DATA

RECOMMENDATIONS




Revolut
VISA