UNIVERSITÀ DEGLI STUDI
DI MILANO

# Supervised and Unsupervised Learning
# Fraud prediction & Clustering survivors

May 2024

Vladislav Kovalev. 44300A

UNIVERSITÀ
DEGLI STUDI
DI MILANO

# Goals and statement of the problem

This report focuses on the prediction of fraud using supervised learning techniques. It delves into the analysis of various variables to determine their influence on the target variable, which indicates fraudulent transactions.

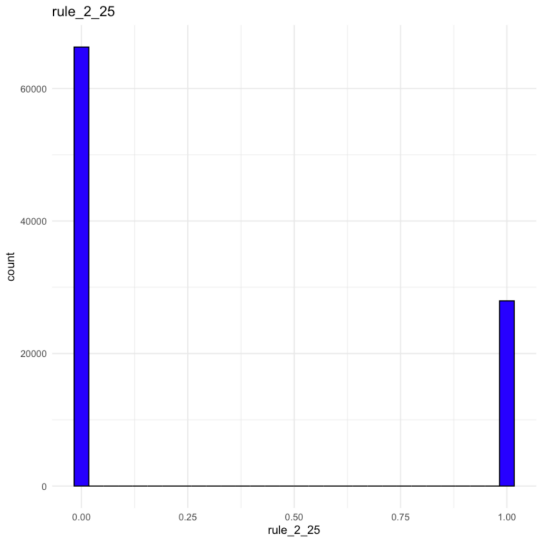
The goal of this analysis is to identify the factors influencing fraudulent transactions. The dataset includes variables such as ID, date, target (indicating fraud), income, age, binary rule variables, and others. There are gaps and imbalances in the classes of the target variable, requiring preprocessing steps like filling gaps in non-binary characteristics.

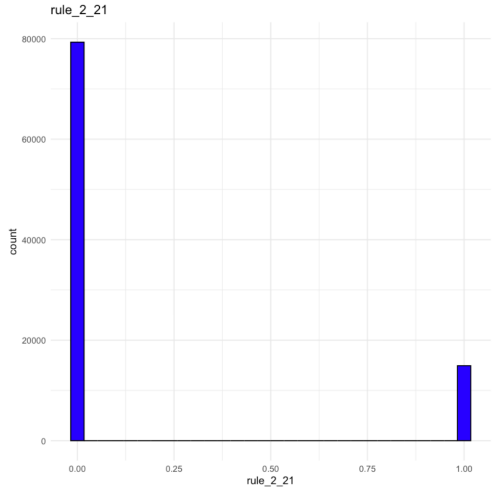
# Data

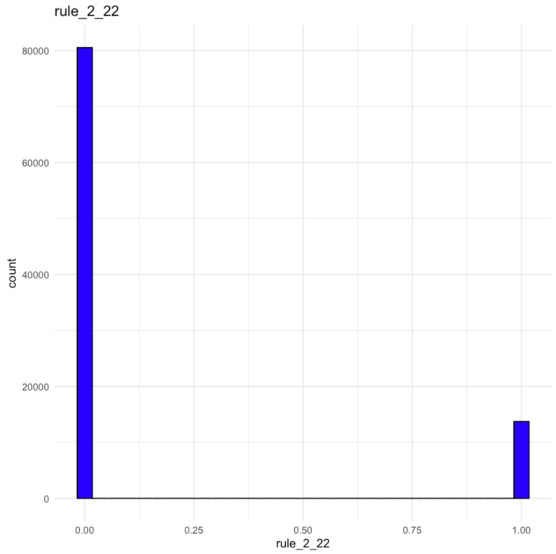
- id - entity identifier date – date
- target - target variable
- req_amt - income
- age - age
- rule_2_21 - rule 2_806 binary variables
- rule_combi_1, rule_combi_2 binary variables in the form of combinations flags of the previous paragraph
- rules_count - number of triggered flags
- score – score by banking system

```
[1] "Percentage of missing values"
              id             date           target          req_amt 
           0.000            0.000            0.000            0.000 
             age        rule_2_21        rule_2_22        rule_2_25 
           0.000            5.777            5.777            5.777 
       rule_2_27        rule_2_31        rule_2_32        rule_2_33 
           5.777            5.777            5.777            5.777 
       rule_2_34       rule_2_801       rule_2_802       rule_2_806 
           5.777            5.777            5.777            5.777 
    rule_combi_1     rule_combi_2      rules_count    score_bank_16 
           0.000            0.000            5.777           13.472 
   score_bank_50      score_nn_16      score_nn_50 score_bank_nn_16 
          13.472           13.472           13.472           13.472 
score_bank_nn_50      count_nulls 
          13.472            0.000 
```

# Distribution of features

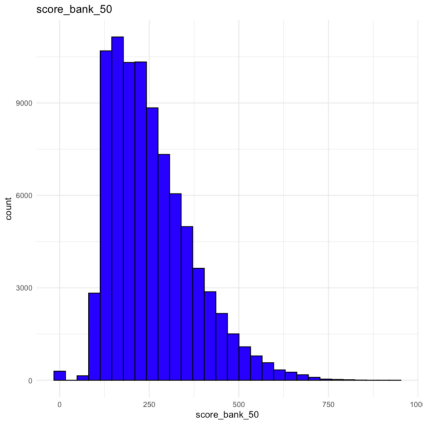


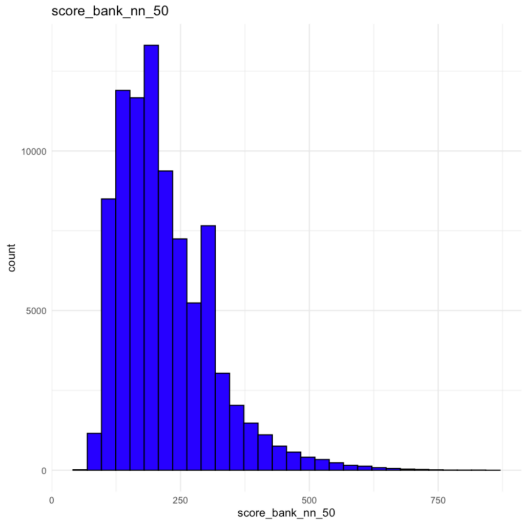



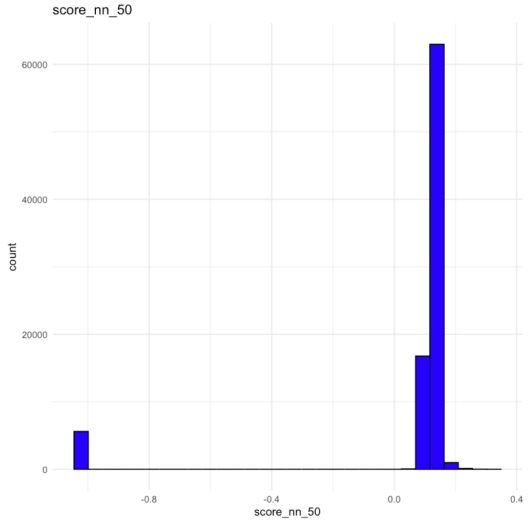




Rules features are binary.
The rule features should be calculated by frequency of fraud

# Distribution of features




In connection with this distribution of data, the strategy of filling voids was chosen Mean in the case of Score, in the case of rules features filling with mode

# Working with hypotheses

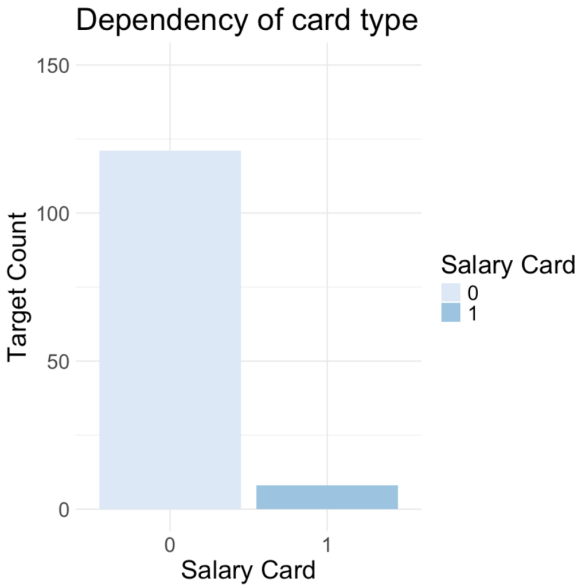
Based on the initial analysis of the data, it became clear that the data represents separate transactions. The main hypotheses regarding transactional fraud are:

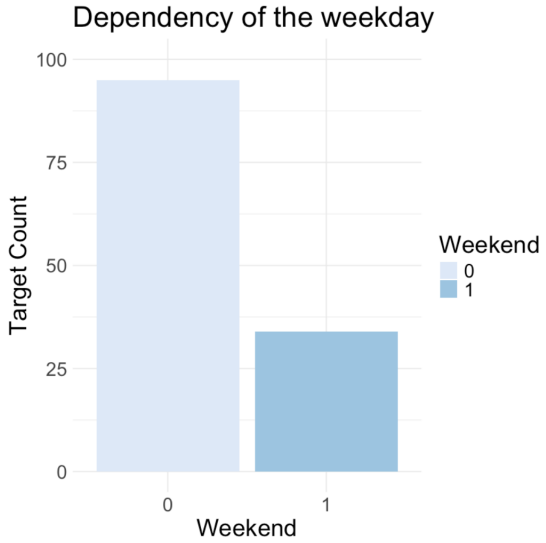
- The frequency of fraudulent transactions from the card.
- The discrepancy between the card data and the passport data of its holder.

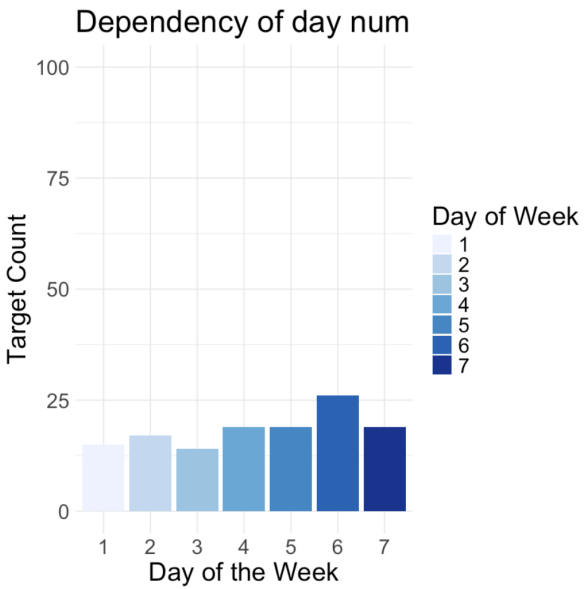
However, since the data is anonymized, decryption attempts are necessary. Good hypotheses, based on statistics on the data, include:

- Repeated transactions with the same “yield” (repeated transactions within a short period of time).
- Identifying salary accounts based on the date of age and amount, with the introduction of an additional indicator for weekends (as wages are typically calculated on weekdays).
- One of the rules may involve detecting the first transfer from someone else’s card.

# Distribution of target by hypotisis values




There is a noticeable dependence of fraud on the day of the week

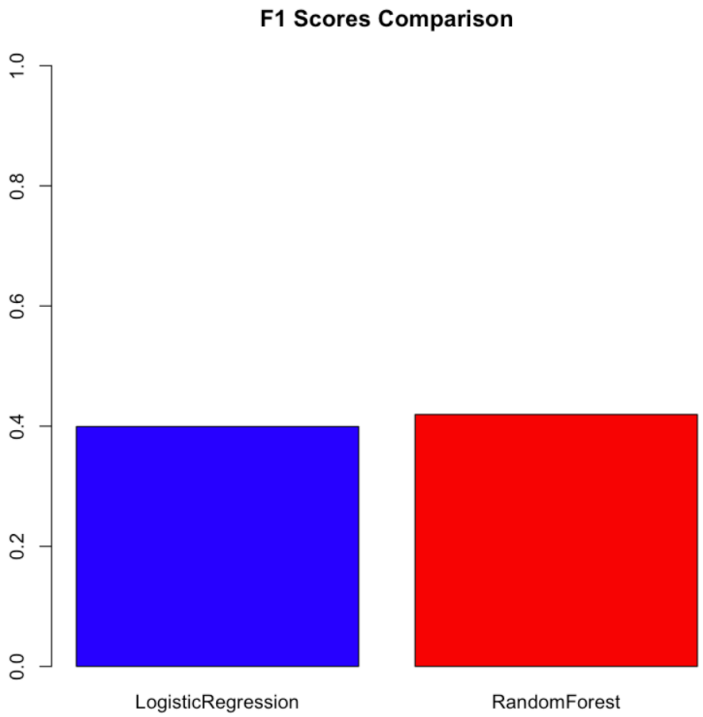
# Logistic Regression Model Fitting

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -12.525889555 | 66.47876302 | -0.18841941 | 8.505479e-01 |
| score_nn_16 | 9.839042114 | 3.14039583 | 3.13305794 | 1.729953e-03 |
| min_fraud_amt | 5.906455202 | 127.49183015 | 0.04632811 | 9.630487e-01 |
| score_nn_50 | -3.629514817 | 3.82343707 | -0.94928065 | 3.424779e-01 |
| score_bank_nn_50 | 1.334315672 | 0.28223257 | 4.72771682 | 2.270586e-06 |
| score_bank_nn_16 | -1.003122069 | 0.33910280 | -2.95816507 | 3.094764e-03 |
| year | -0.951570893 | 0.14146414 | -6.72658715 | 1.736886e-11 |
| score_bank_16 | 0.905169232 | 0.51838546 | 1.74613160 | 8.078809e-02 |
| req_amt | 0.813045037 | 0.09366984 | 8.67990244 | 3.961085e-18 |
| rule_2_33 | 0.760446121 | 0.34253205 | 2.22007293 | 2.641382e-02 |
| score_bank_50 | -0.740444980 | 0.45093595 | -1.64201806 | 1.005863e-01 |
| rule_2_21 | -0.695463969 | 0.38312695 | -1.81523114 | 6.948840e-02 |
| age_group | 0.268534365 | 0.20456625 | 1.31270120 | 1.892837e-01 |
| rule_2_31 | 0.215574041 | 0.12206164 | 1.76610798 | 7.737773e-02 |
| rule_2_32 | -0.210970387 | 0.16220483 | -1.30064185 | 1.933811e-01 |
| rule_2_25 | 0.189972914 | 0.15381781 | 1.23505148 | 2.168113e-01 |
| weekend | -0.182683701 | 0.10594115 | -1.72438845 | 8.463774e-02 |
| age | -0.159646390 | 0.19389429 | -0.82336818 | 4.102987e-01 |
| rule_combi_1 | 0.130706350 | 0.13509655 | 0.96750326 | 3.332925e-01 |
| rule_2_802 | -0.117864448 | 0.14913227 | -0.79033495 | 4.293322e-01 |
| number_of_day | 0.097100798 | 0.10651173 | 0.91164415 | 3.619561e-01 |
| rule_combi_2 | 0.062441140 | 0.10554044 | 0.59163236 | 5.540968e-01 |
| rule_2_22 | -0.060003403 | 0.09817667 | -0.61117779 | 5.410819e-01 |
| day | 0.052828755 | 0.09632188 | 0.54846056 | 5.833757e-01 |
| rule_2_806 | 0.044955350 | 0.16164727 | 0.27810769 | 7.809297e-01 |
| salary_card | 0.040953489 | 0.04476529 | 0.91484910 | 3.602709e-01 |
| month | -0.032631991 | 0.11058989 | -0.29507211 | 7.679388e-01 |
| rule_2_801 | 0.014972038 | 0.17021742 | 0.08795832 | 9.299098e-01 |
| rule_2_27 | -0.013774517 | 0.14428796 | -0.09546546 | 9.239451e-01 |
| rule_2_34 | 0.007341811 | 0.16637330 | 0.04412854 | 9.648019e-01 |

Features importance by regularisations

# Random Forest vs Logistic regression

Due to the fact that the samples are not balanced with respect to the target variable, the F1 measure was chosen to compare models




LogisticRegression 0.399374577840042
RandomForest 0.4193785222042

# Clustering and dimensionality reduction

We will look at the quality of division into clusters using the Adjusted Rand Index metric. This metric evaluates the quality of the existing markup. In addition, we will try to evaluate the quality of clustering using the interia_ and Silhouette metrics, which do not require markings. Detailed descriptions of the metrics are given below.

## Statement of the Problem

The objective is to cluster Titanic passengers based on features such as ticket class, age, and fare to discern any patterns related to survival. The dataset contains information about passengers including their survival status (survived or not), ticket class, age, sex, and fare.

# Data

```
  PassengerId       Survived          Pclass          Name
Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891
1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character
Median :446.0   Median :0.0000   Median :3.000   Mode  :character

  Mean   :446.0   Mean   :0.3838   Mean   :2.309
  3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
  Max.   :891.0   Max.   :1.0000   Max.   :3.000

     Sex                 Age           SibSp           Parch
  Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000
  Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
  Mode  :character   Median :28.00   Median :0.000   Median :0.0000
                     Mean   :29.70   Mean   :0.523   Mean   :0.3816
                     3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
                     Max.   :80.00   Max.   :8.000   Max.   :6.0000
                     NA's   :177
    Ticket               Fare           Cabin             Embarked
  Length:891         Min.   :  0.00   Length:891         Length:891
  Class :character   1st Qu.:  7.91   Class :character   Class :character
  Mode  :character   Median : 14.45   Mode  :character   Mode  :character
                     Mean   : 32.20
                     3rd Qu.: 31.00
                     Max.   :512.33
```

# Adjusted Rand Index (ARI)

It is assumed that the true object labels are known. This measure does not depend on the label values themselves, but only on the division of the sample into clusters. Let N be the number of objects in the sample. Let us denote by a the number of pairs of objects that have the same labels and are in the same cluster, and by b the number of pairs of objects that have different labels and are in different clusters. Then the Rand Index is
RI= 2(a+b). n(n−1)

That is, this is the proportion of objects for which these partitions (the original and those obtained as a result of clustering ) “agreed”. The Rand Index (RI) expresses the similarity of two different clusterings of the same sample. In order for this index to give values close to zero for random clusterings for any N and number of clusters, it is necessary to normalize it.

This measure is symmetric and does not depend on the values and permutations of labels. Thus, this index is a measure of the distance between different sample splits. ARI takes values in the range [−1,1]. Negative values correspond to “independent” cluster splits, values close to zero indicate random splits, and positive values indicate that the two splits are similar (the same when ARI = 1).

[1] 0.04287034 The metric is close to 0 – our clustering is close to random

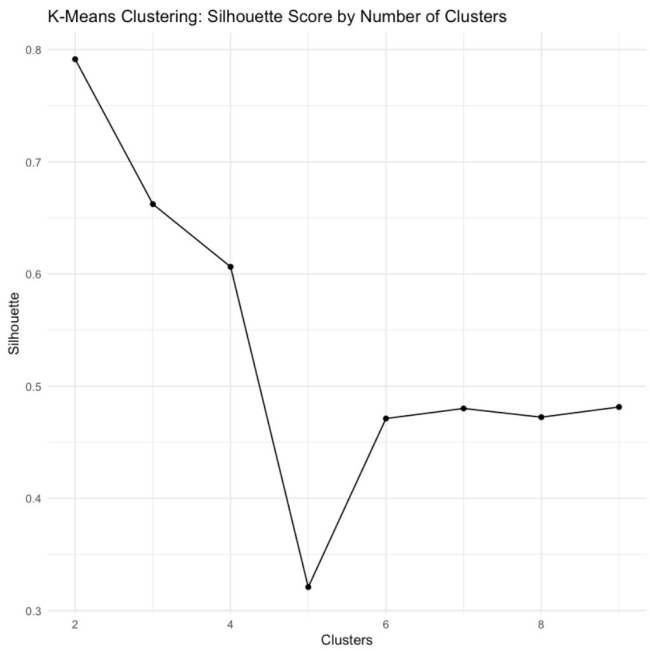
# Silhouette coefficient

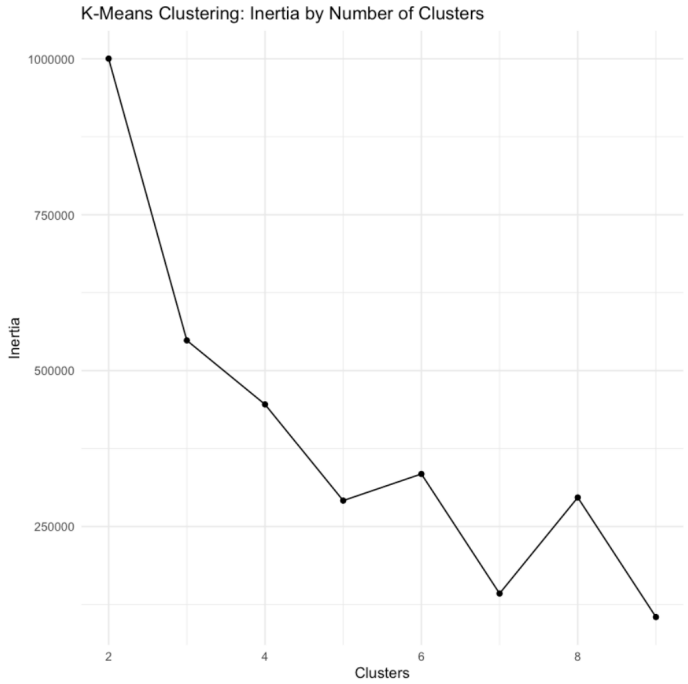
Using the silhouette, you can select the optimal number of clusters k (if it is unknown in advance) - the number of clusters that maximizes the value of the silhouette is selected. Unlike previous metrics, the silhouette depends on the shape of the clusters, and reaches higher values on more convex clusters obtained using algorithms based on density distribution reconstruction.

[1] 0.7914459

By themselves, these metrics do not say anything, because strongly depend on the dimension of space and the scale of features. Maybe there are more than 2 clusters in our data?

# Inertia




The smaller the interia, the better our clustering. It is clear that the more clusters, the smaller the sum of squared distances to their centers. So it would be useful to compare our plot with the one obtained for data that is distributed evenly or, on the contrary, has a clear cluster structure

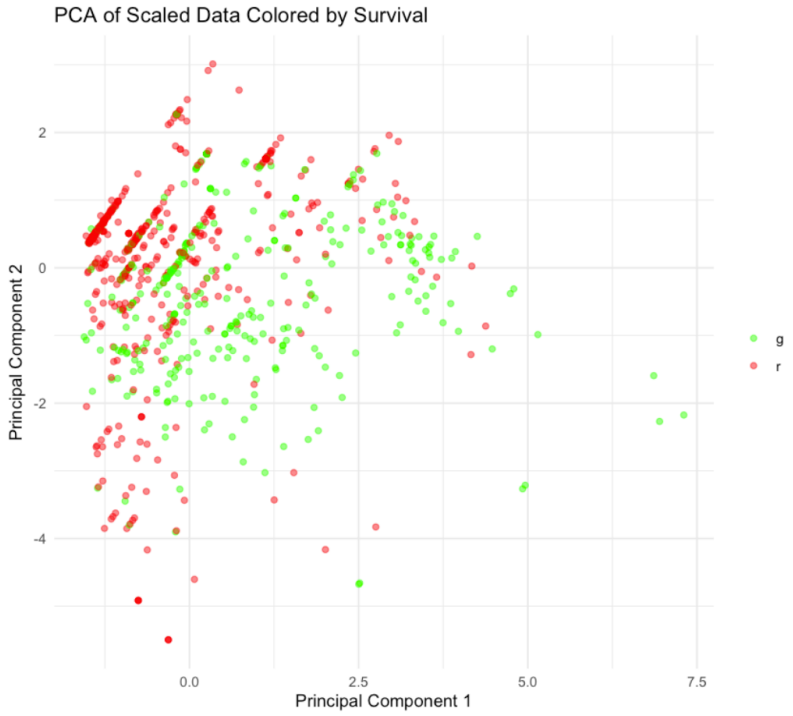
For silhouette the situation is not so clear-cut.

# PCA

What we can see in this graph:

- Green and red dots are distributed in different areas of the graph. Probably, supervised algorithms could cope with the classification.

- If we remove the color, no clustering algorithm will separate the red and green points into 2 clusters.



# ARI after scaling and PCA

[1] 0.1280429

The metrics ARI have not improved significantly. kMeans fraction of the 2-class case tries to draw a line between clusters of points. Those. will divide the space into 2 parts with a line. Our clusters have a more cunning structure. For such a case of such data, the DBSCAN algorithm is well suited.

In addition, visually there are 4 clusters, not 2. So we can’t do without partial data marking (to determine whether our cluster is “green” or “red”)

# Conclusion

Clustering analysis on the Titanic dataset using KMeans reveals potential for more than two clusters, supported by various evaluation metrics.

Scaling and normalization improve clustering performance, albeit marginally.

PCA aids in visualizing high-dimensional data but does not significantly enhance clustering quality in this context.

Further exploration with alternative clustering algorithms such as DBSCAN may be beneficial, especially considering the dataset's complex cluster structure. Additionally, incorporating partial data labeling could enhance clustering accuracy.