Supervised & Unsupervised Encoding Schemes of Binary Variables for Prediction Performance Enhancement

Institute of Industrial Engineering, NTU

Yun-Hao Yang, Jakey Blue

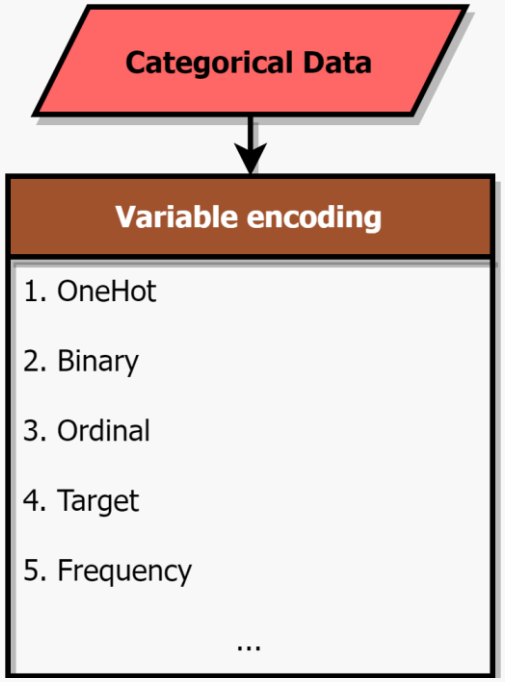
Overview

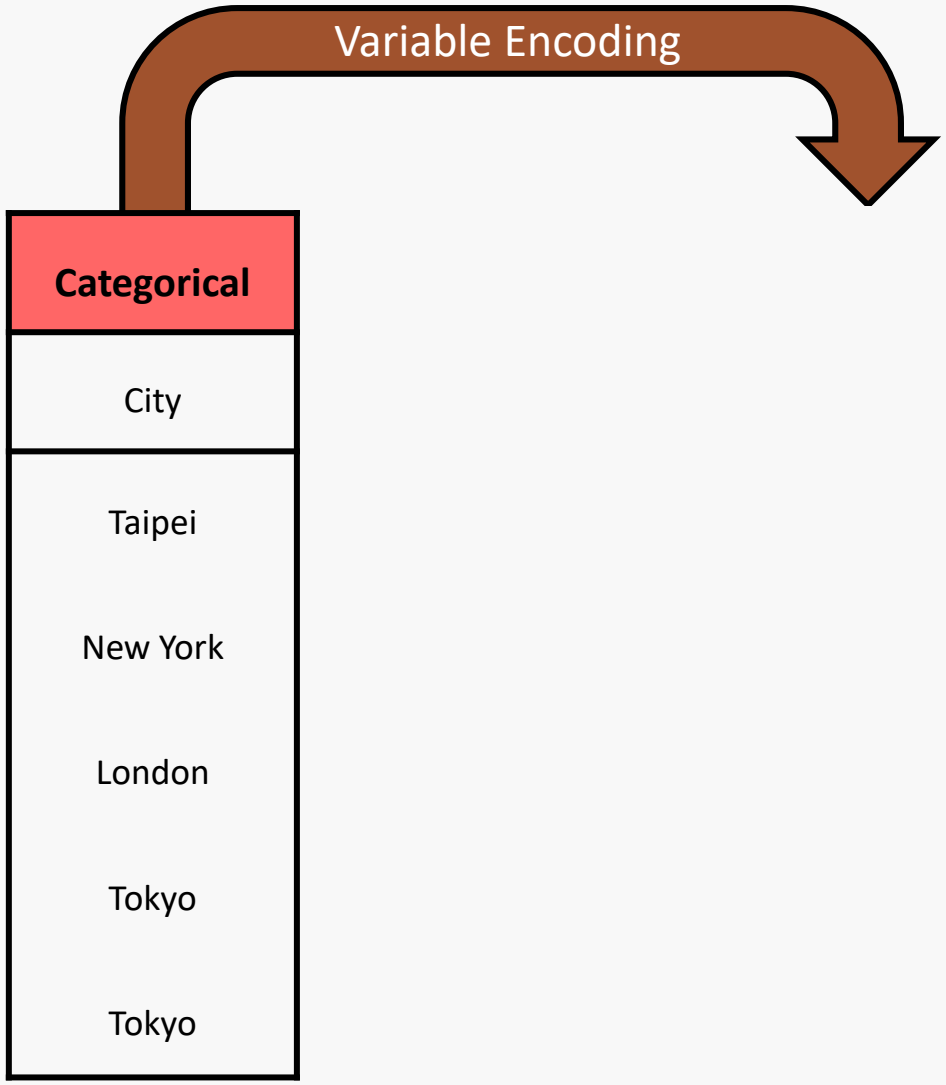
- Motivation
- Methodology
- Case Study
- Conclusion & Future Work

Motivation

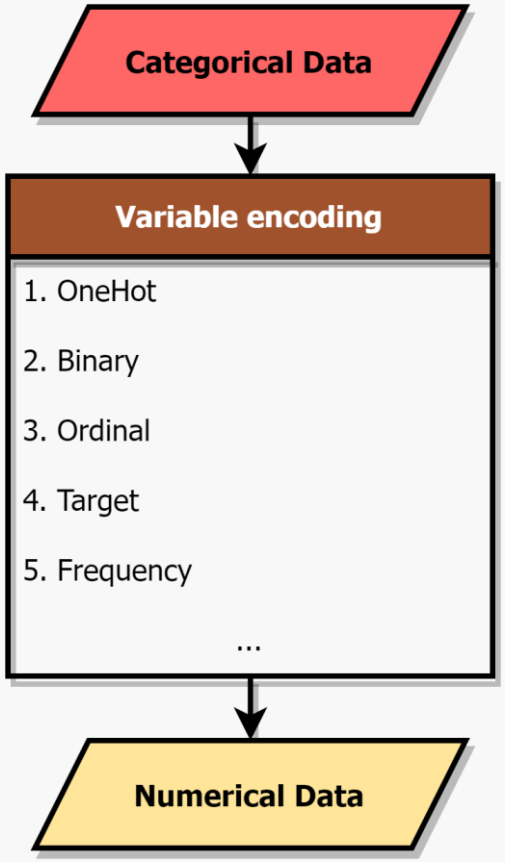
Categorical
City
Taipei
New York
London
Tokyo
Tokyo

Motivation


Categorical Data
Variable encoding
1. OneHot
2. Binary
3. Ordinal
4. Target
5. Frequency
...


Variable Encoding
Categorical
City
Taipei
New York
London
Tokyo
Tokyo

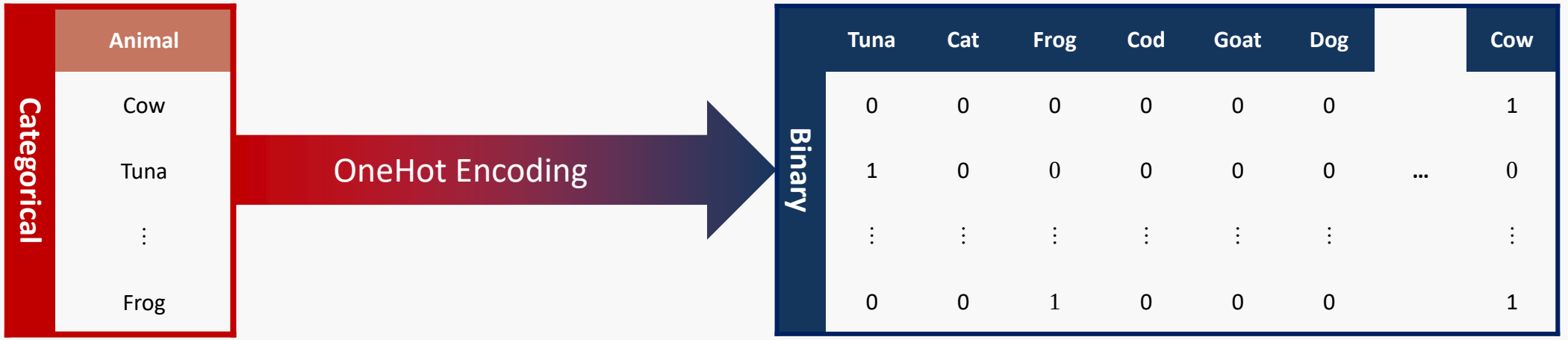
Motivation



Variable Encoding

Categorical	Numerical						
City	Ordinal	Binary		OneHot				Frequency
Taipei	0	0	0	0	0	0	1	0.2
New York	1	0	1	0	0	1	0	0.2
London	2	1	0	0	1	0	0	0.2
Tokyo	3	1	1	1	0	0	0	0.2
Tokyo	3	1	1	1	0	0	0	0.4

Motivation

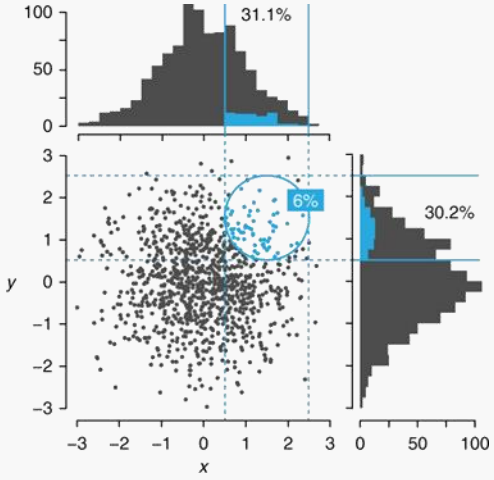

Categorical
Animal
Cow
Tuna
⋮
Frog
OneHot Encoding
Binary
Tuna
Cat
Frog
Cod
Goat
Dog
...
Cow
0 0 0 0 0 0 1
1 0 0 0 0 0 0
⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮
0 0 1 0 0 0 1

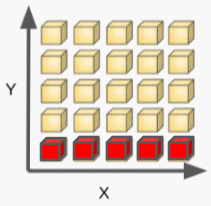
Motivation

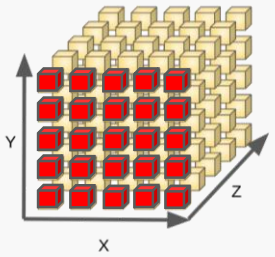
Categorical: Animal
Cow
Tuna
⋮
Frog

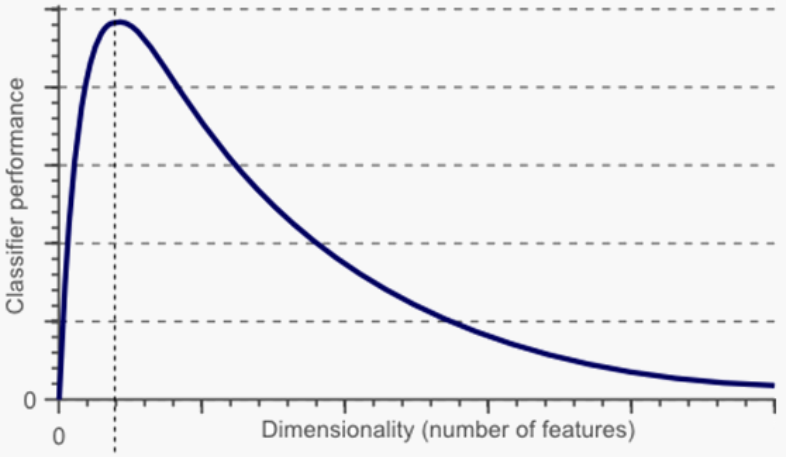
OneHot Encoding

Binary: Tuna	Cat	Frog	Cod	Goat	Dog		Cow
0	0	0	0	0	0		1
1	0	0	0	0	0	…	0
⋮	⋮	⋮	⋮	⋮	⋮		⋮
0	0	1	0	0	0		1









Motivation

Categorical

Animal	Size	Color
Cow	Big	White
Tuna	Med	Black
⋮	⋮	⋮
Frog	Tiny	Green

Other Encoding

Numerical

Animal	Size	Color
10	4	7
2	3	6
⋮	⋮	⋮
7	1	3

One Hot

Binary

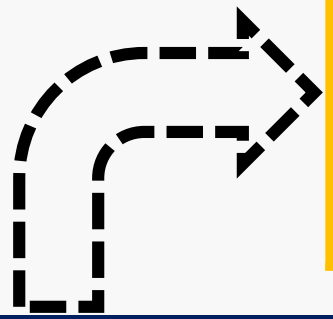
Animal								Size				Color					
Tuna	Cat	Frog	Cod	Goat	Dog	Toad	Cow	Large	Mid	Small	Tiny	White	Black	Red	Blue	Green	Gray
0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0
1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0

Motivation

Categorical

Animal	Size	Color
Cow	Big	White
Tuna	Med	Black
⋮	⋮	⋮
Frog	Tiny	Green

Other Encoding



Numerical

Animal	Size	Color
10	4	7
2	3	6
⋮	⋮	⋮
7	1	3

Binary

Tuna	Cat	Frog	Cod	Goat	Dog	Toad	Cow	Large	Mid	Small	Tiny	White	Black	Red	Blue	Green	Gray
0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0
1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0

Motivation

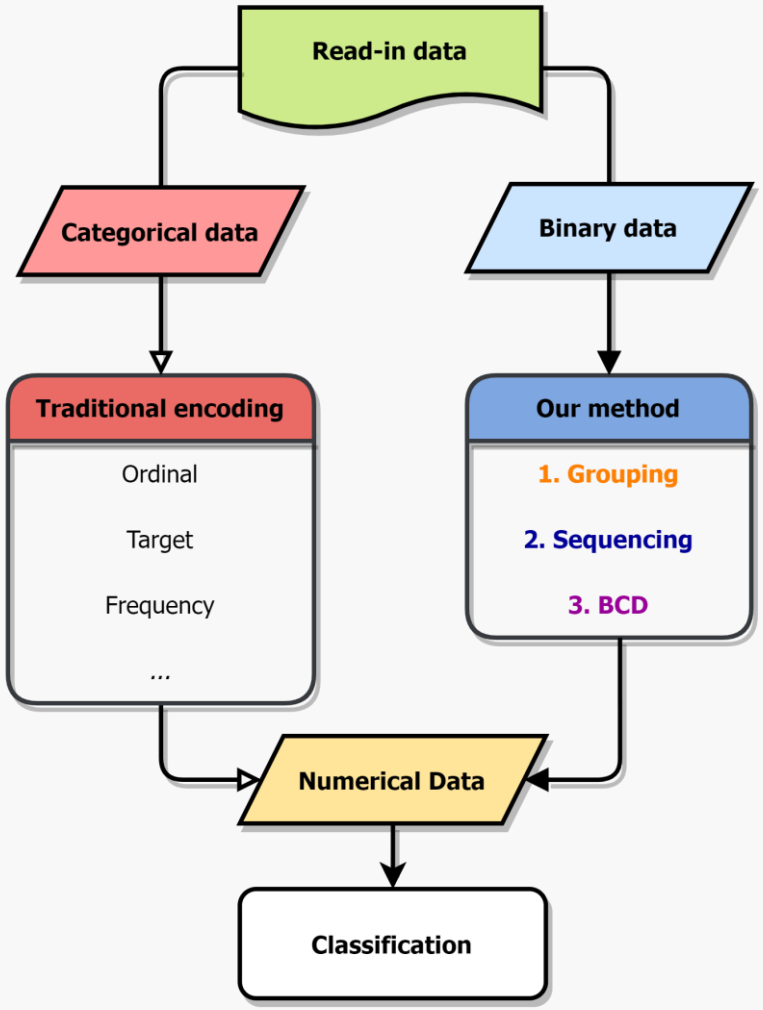
Numerical

Animal	Size	Color
10	4	7
2	3	6
⋮	⋮	⋮
7	1	3

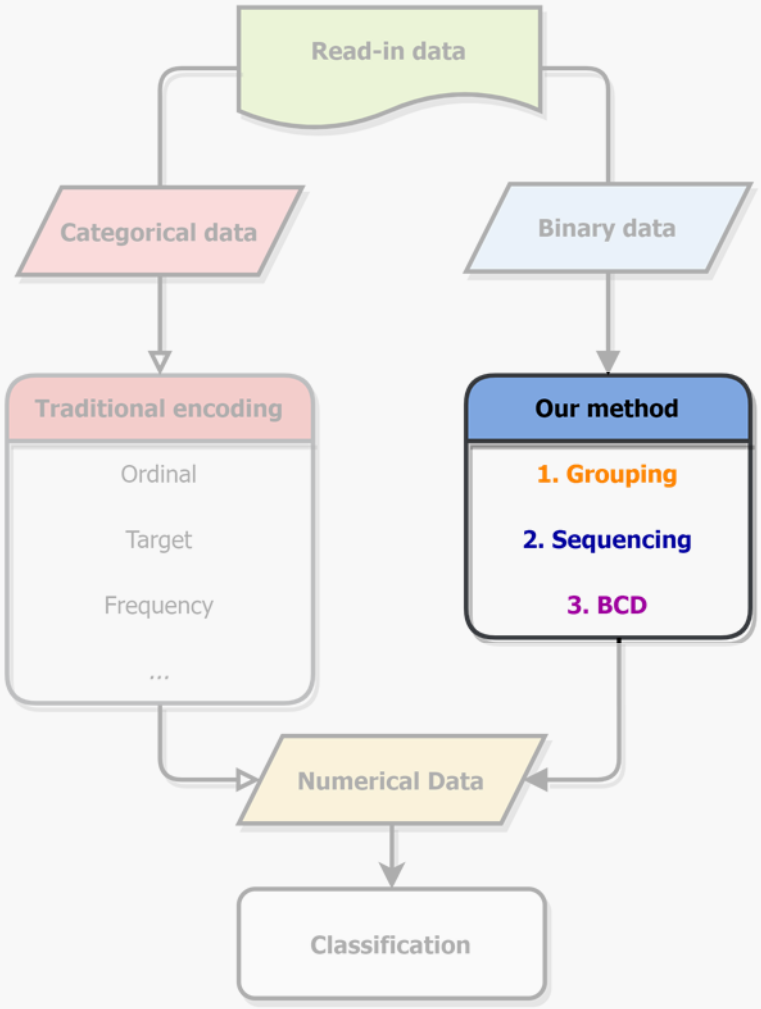
Binary

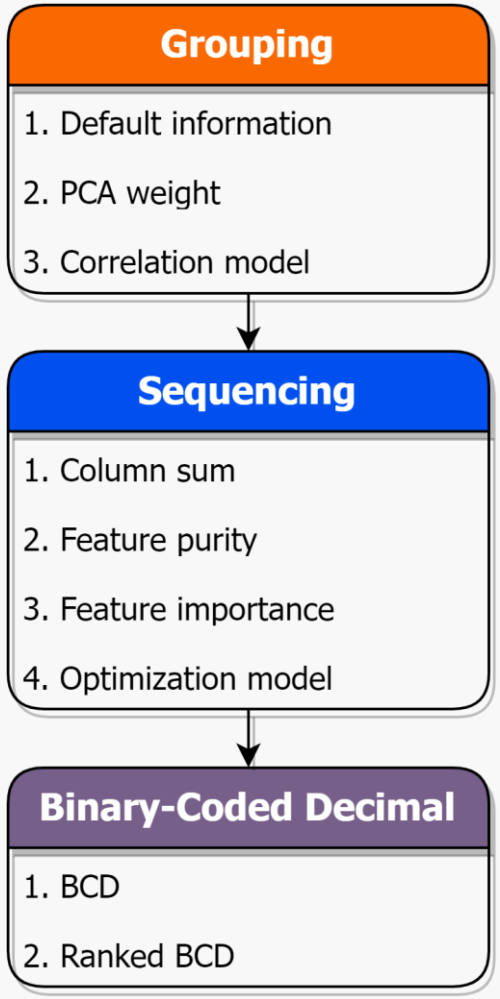
Tuna	Cat	Frog	Cod	Goat	Dog	Toad	Cow	Large	Mid	Small	Tiny	White	Black	Red	Blue	Green	Gray
0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0
1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0

Methodology



Methodology





Group
Sequence
BCD

Methodology

In this study, we propose a method of encoding binary data into numerical data. Through grouping, sequencing, and transforming the binary row data with BCD encode.

1. Grouping similar, correlated features
2. Sequencing features in each feature group
3. BCD encode on each feature group

Tiny	Cat	Large	Black	White	Cow
1	1	0	1	0	0
0	0	1	0	1	1
⋮	⋮	⋮	⋮	⋮	⋮
0	0	1	0	1	1

Methodology

Group
Sequence
BCD

In this study, we propose a method of encoding binary data into numerical data. Through grouping, sequencing, and transforming the binary row data with BCD encode.

1. Grouping similar, correlated features
2. Sequencing features in each feature group
3. BCD encode on each feature group

Tiny	Cat	Large	Black	White	Cow
1	1	0	1	0	0
0	0	1	0	1	1
⋮	⋮	⋮	⋮	⋮	⋮
0	0	1	0	1	1

Group
Sequence
BCD

Methodology

In this study, we propose a method of encoding binary data into numerical data. Through grouping, sequencing, and transforming the binary row data with BCD encode.

1. Grouping similar, correlated features
2. Sequencing features in each feature group
3. BCD encode on each feature group

Cat	Cow	Black	White	Tiny	Large
1	0	1	0	1	0
0	1	0	1	0	1
⋮	⋮	⋮	⋮	⋮	⋮
0	1	0	1	0	1

Group
Sequence
BCD

Methodology

In this study, we propose a method of encoding binary data into numerical data. Through grouping, sequencing, and transforming the binary row data with BCD encode.

1. Grouping similar, correlated features
2. Sequencing features in each feature group
3. BCD encode on each feature group

Cow	Cat	White	Black	Tiny	Large
0	1	0	1	1	0
1	0	1	0	0	1
⋮	⋮	⋮	⋮	⋮	⋮
1	0	1	0	0	1

Group
Sequence
BCD

Methodology

In this study, we propose a method of encoding binary data into numerical data. Through grouping, sequencing, and transforming the binary row data with BCD encode.

1. Grouping similar, correlated features
2. Sequencing features in each feature group
3. BCD encode on each feature group

Animal	Color	Size
1	1	2
2	2	1
⋮	⋮	⋮
2	2	1

Methodology

-problem definition

The main goal is to find the optimal G_j and S_j , such that the encoded numerical data would perform better in the classification task.

Symbol	Definition
X_i	The i^{th} feature of Binary data X, $0 \leq i \leq n$
$g(X)$	Clustering methods for X
G_j	The j^{th} clustered feature group
$s(G_j)$	Sequencing methods for G_j
S_j	Sequenced features of G_j
$BCD(S_j)$	BCD code of S_j
Y_j	The j^{th} feature of encoded numerical data

Methodology

-An animal health check

Animal							
Cow	Cat	Goat	Dog	Frog	Toad	Tuna	Cod
		V					

Size			
Large	Mid	Small	Tiny
	V		

Color					
White	Black	Gray	Red	Blue	Green
	V				

Health Status	
Good	Bad
	V

Methodology

-An animal health check

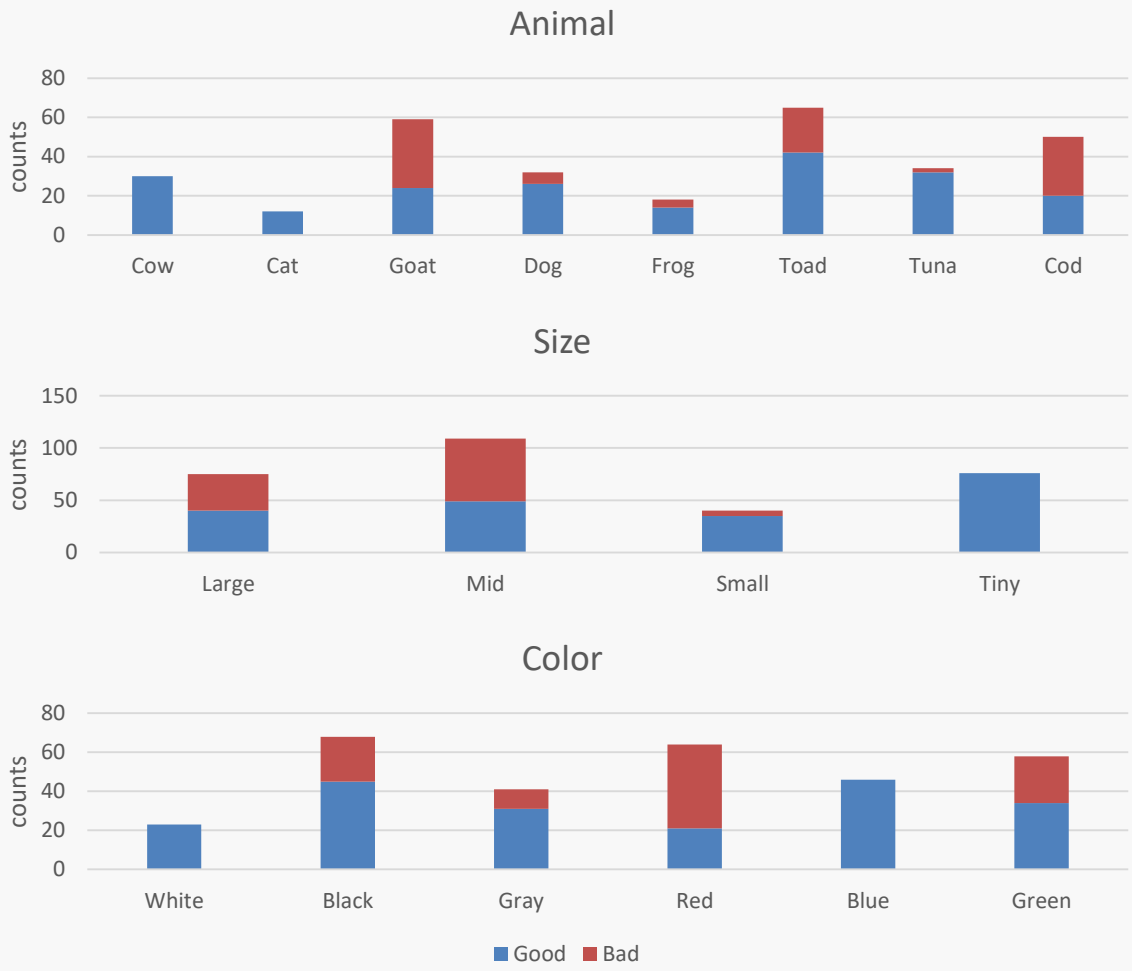
Animal							
Cow	Cat	Goat	Dog	Frog	Toad	Tuna	Cod
		V					

Size			
Large	Mid	Small	Tiny
	V		

Color					
White	Black	Gray	Red	Blue	Green
	V				

Health Status	
Good	Bad
	V

Binary Data (Total:300 samples, 200 good, 100 bad)



Methodology

-An animal health check

Animal							
Cow	Cat	Goat	Dog	Frog	Toad	Tuna	Cod
		V					

Size			
Large	Mid	Small	Tiny
	V		

Color					
White	Black	Gray	Red	Blue	Green
	V				

Health Status	
Good	Bad
	V

Binary Data (Total:300 samples, 200 good, 100 bad)

Animal	Cow	Cat	Goat	Dog	Frog	…	Blue	Green
#1	0	1	0	0	0	…	0	0
#2	1	0	0	0	0	…	0	0
#3	0	0	0	0	0	…	0	0
#4	0	0	0	0	0	…	0	0
#5	0	0	1	0	0	…	0	0
#6	0	0	0	0	1	…	0	0
#7	0	0	0	0	0	…	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
#300	0	0	0	1	0	…	0	0

Methodology

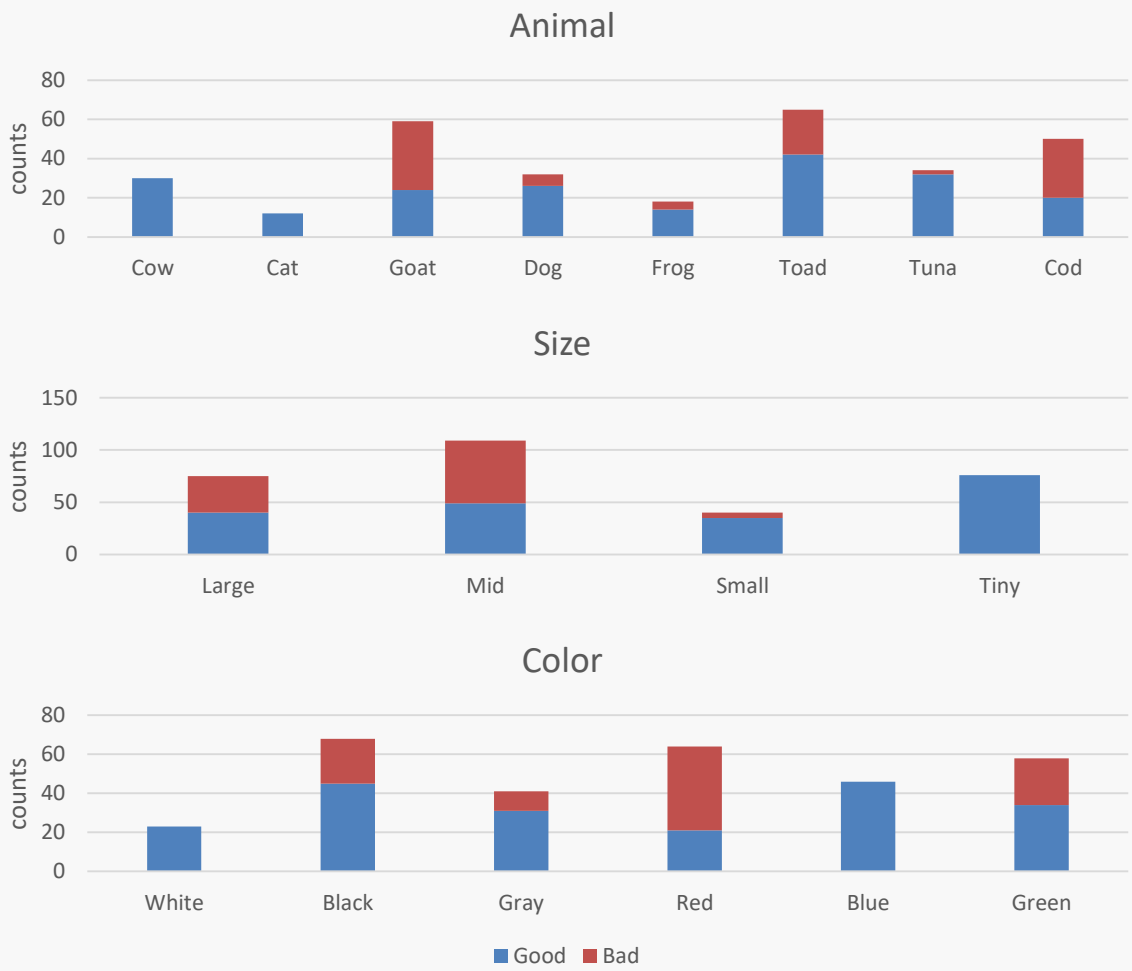
-Grouping

Group
Sequence
BCD

First, we try to find and cluster the related binary features.

1. Default information
2. PCA weight
3. Correlation model
   - K-means
   - Hierarchical clustering
   - Block modeling

Binary Data (Total:300 sample, 200 good, 100 bad)



Group
Sequence
BCD

Methodology

-PCA Grouping

No.	PC 1			PC 2			PC 3		
	Feature	Abs. weight	select	Feature	Abs. weight	select	Feature	Abs. weight	select
1.	Cow	0.562	O	Cat	0.642	X	Black	0.047	X
2.	Cat	0.486	O	Cow	0.496	X	Goat	0.035	X
3.	Large	0.348	O	Mid	0.402	X	Cat	0.032	X
4.	Mid	0.311	O	Black	0.351	X	Cow	0.030	X
5.	White	0.307	O	Goat	0.302	O	Mid	0.028	X
6.	Black	0.202	O	Dog	0.229	O	Gray	0.023	X
7.	Small	0.187	X	Large	0.199	X	White	0.021	X
8.	Goat	0.153	X	Frog	0.183	O	Dog	0.020	X
9.	Cod	0.101	X	Small	0.152	O	Large	0.018	X
10.	Tuna	0.091	X	Gray	0.132	O	Frog	0.016	X
11.	Dog	0.074	X	White	0.105	X	Small	0.014	X
12.	Gray	0.060	X	Red	0.008	O	Toad	0.014	O
13.	Frog	0.056	X	Frog	0.008	X	Tuna	0.012	O
14.	Toad	0.032	X	Red	0.007	X	Cod	0.007	O
15.	Red	0.029	X	Toad	0.005	X	Red	0.005	X
16.	Tiny	0.023	X	Green	0.001	X	Tiny	0.002	O
17.	Green	0.015	X	Tiny	0.001	X	Blue	0.001	O
18.	Blue	0.009	X	Blue	0.001	X	Green	0.001	O

Group
Sequence
BCD

Methodology

-PCA Grouping

No.	PC 1			PC 2			PC 3		
	Feature	Abs. weight	select	Feature	Abs. weight	select	Feature	Abs. weight	select
1.	Cow	0.562	O	Cat	0.642	X	Black	0.047	X
2.	Cat	0.486	O	Cow	0.496	X	Goat	0.035	X
3.	Large	0.348	O	Mid	0.402	X	Cat	0.032	X
4.	Mid	0.311	O	Black	0.351	X	Cow	0.030	X
5.	White	0.307	O	Goat	0.302	O	Mid	0.028	X
6.	Black	0.202	O	Dog	0.229	O	Gray	0.023	X
7.	Small	0.187	X	Large	0.199	X	White	0.021	X
8.	Goat	0.153	X	Frog	0.183	O	Dog	0.020	X
9.	Cod	0.101	X	Small	0.152	O	Large	0.018	X
10.	Tuna	0.091	X	Gray	0.132	O	Frog	0.016	X
11.	Dog	0.074	X	White	0.105	X	Small	0.014	X
12.	Gray	0.060	X	Red	0.008	O	Toad	0.014	O
13.	Frog	0.056	X	Frog	0.008	X	Tuna	0.012	O
14.	Toad	0.032	X	Red	0.007	X	Cod	0.007	O
15.	Red	0.029	X	Toad	0.005	X	Red	0.005	X
16.	Tiny	0.023	X	Green	0.001	X	Tiny	0.002	O
17.	Green	0.015	X	Tiny	0.001	X	Blue	0.001	O
18.	Blue	0.009	X	Blue	0.001	X	Green	0.001	O

Group
Sequence
BCD

Methodology

-*Block modeling Grouping*

	Tuna	Cat	Frog	Cod	Goat	Dog	Toad	Cow	Large	Mid	Small	Tiny	White	Black	Red	Blue	Green	Gray
Tuna	1.0	0.2	0.2	0.8	0.2	0.2	0.8	0.2	0.1	0.1	0.1	0.3	0.0	0.0	0.0	0.5	0.5	0.0
Cat	0.2	1.0	0.2	0.2	0.2	0.2	0.2	0.8	0.3	0.3	0.1	0.1	0.2	0.2	0.0	0.0	0.0	0.0
Frog	0.2	0.2	1.0	0.2	0.8	0.8	0.2	0.2	0.1	0.1	0.3	0.1	0.0	0.0	0.3	0.0	0.0	0.3
Cod	0.8	0.2	0.2	1.0	0.2	0.2	0.8	0.8	0.1	0.1	0.1	0.3	0.0	0.0	0.0	0.3	0.3	0.0
Goat	0.2	0.2	0.8	0.2	1.0	0.8	0.2	0.2	0.1	0.1	0.3	0.1	0.0	0.0	0.3	0.0	0.0	0.3
Dog	0.2	0.2	0.8	0.2	0.8	1.0	0.2	0.2	0.1	0.1	0.3	0.1	0.0	0.0	0.3	0.0	0.0	0.3
Toad	0.8	0.2	0.2	0.8	0.2	0.2	1.0	0.2	0.1	0.1	0.1	0.3	0.0	0.0	0.0	0.3	0.0	0.0
Cow	0.2	0.8	0.2	0.8	0.2	0.2	0.2	1.0	0.3	0.3	0.1	0.1	0.3	0.3	0.0	0.0	0.0	0.0
Large	0.1	0.3	0.1	0.1	0.1	0.1	0.1	0.3	1.0	0.5	0.1	0.1	0.6	0.6	0.2	0.2	0.2	0.2
Mid	0.1	0.3	0.1	0.1	0.1	0.1	0.1	0.3	0.5	1.0	0.1	0.1	0.6	0.6	0.2	0.2	0.2	0.2
Small	0.1	0.1	0.3	0.1	0.3	0.3	0.1	0.1	0.1	0.1	1.0	0.1	0.2	0.2	0.6	0.2	0.2	0.6
Tiny	0.3	0.1	0.1	0.3	0.1	0.1	0.3	0.1	0.1	0.1	0.1	1.0	0.2	0.2	0.2	0.6	0.2	0.2
White	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.3	0.6	0.6	0.2	0.2	1.0	0.8	0.4	0.4	0.4	0.4
Black	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.3	0.6	0.6	0.2	0.2	0.8	1.0	0.4	0.4	0.4	0.8
Red	0.0	0.0	0.3	0.0	0.3	0.3	0.0	0.0	0.2	0.2	0.6	0.2	0.4	0.4	1.0	0.4	0.4	0.4
Blue	0.5	0.0	0.0	0.3	0.0	0.0	0.3	0.0	0.2	0.2	0.2	0.6	0.4	0.4	0.4	1.0	0.8	0.8
Green	0.5	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.2	0.2	0.2	0.2	0.4	0.4	0.4	0.8	1.0	0.4
Gray	0.0	0.0	0.3	0.0	0.3	0.3	0.0	0.0	0.2	0.2	0.6	0.2	0.4	0.8	0.4	0.8	0.4	1.0

Group
Sequence
BCD

Methodology

-*Block modeling Grouping*

	Cow	Cat	Large	Mid	White	Black	Goat	Dog	Frog	Small	Gray	Red	Toad	Tuna	Cod	Tiny	Blue	Green
Cow	1.0	0.8	0.3	0.3	0.5	0.5	0.2	0.2	0.2	0.1	0.0	0.0	0.2	0.2	0.2	0.1	0.0	0.0
Cat	0.8	1.0	0.3	0.3	0.5	0.5	0.2	0.2	0.2	0.1	0.0	0.0	0.2	0.2	0.2	0.1	0.0	0.0
Large	0.3	0.3	1.0	0.5	0.6	0.6	0.1	0.1	0.1	0.1	0.2	0.2	0.1	0.1	0.1	0.1	0.2	0.2
Mid	0.3	0.3	0.5	1.0	0.6	0.6	0.1	0.1	0.1	0.1	0.2	0.2	0.1	0.1	0.1	0.1	0.2	0.2
White	0.5	0.5	0.6	0.6	1.0	0.8	0.0	0.0	0.0	0.2	0.4	0.4	0.0	0.0	0.0	0.2	0.4	0.4
Black	0.5	0.5	0.6	0.6	0.8	1.0	0.0	0.0	0.0	0.2	0.4	0.4	0.0	0.0	0.0	0.2	0.4	0.4
Goat	0.2	0.2	0.1	0.1	0.0	0.0	1.0	0.8	0.8	0.3	0.3	0.3	0.2	0.2	0.2	0.1	0.0	0.0
Dog	0.2	0.2	0.1	0.1	0.0	0.0	0.8	1.0	0.8	0.3	0.5	0.5	0.2	0.2	0.2	0.1	0.0	0.0
Frog	0.2	0.2	0.1	0.1	0.0	0.0	0.8	0.8	1.0	0.3	0.5	0.5	0.2	0.2	0.2	0.1	0.0	0.0
Small	0.1	0.1	0.1	0.1	0.2	0.2	0.3	0.3	0.3	1.0	0.6	0.6	0.1	0.1	0.1	0.1	0.2	0.2
Gray	0.0	0.0	0.2	0.2	0.4	0.4	0.3	0.5	0.5	0.6	1.0	0.8	0.0	0.0	0.0	0.2	0.4	0.4
Red	0.0	0.0	0.2	0.2	0.4	0.4	0.3	0.5	0.5	0.6	0.8	1.0	0.0	0.0	0.0	0.2	0.4	0.4
Toad	0.2	0.2	0.1	0.1	0.0	0.0	0.2	0.2	0.2	0.1	0.0	0.0	1.0	0.8	0.8	0.3	0.5	0.5
Tuna	0.2	0.2	0.1	0.1	0.0	0.0	0.2	0.2	0.2	0.1	0.0	0.0	0.8	1.0	0.8	0.3	0.5	0.5
Cod	0.2	0.2	0.1	0.1	0.0	0.0	0.2	0.2	0.2	0.1	0.0	0.0	0.8	0.8	1.0	0.3	0.5	0.5
Tiny	0.1	0.1	0.1	0.1	0.2	0.2	0.1	0.1	0.1	0.1	0.2	0.2	0.3	0.3	0.3	1.0	0.6	0.6
Blue	0.0	0.0	0.2	0.2	0.4	0.4	0.0	0.0	0.0	0.2	0.4	0.4	0.5	0.5	0.5	0.6	1.0	0.8
Green	0.0	0.0	0.2	0.2	0.4	0.4	0.0	0.0	0.0	0.2	0.4	0.4	0.5	0.5	0.5	0.6	0.8	1.0

Methodology

-Sequencing

Group

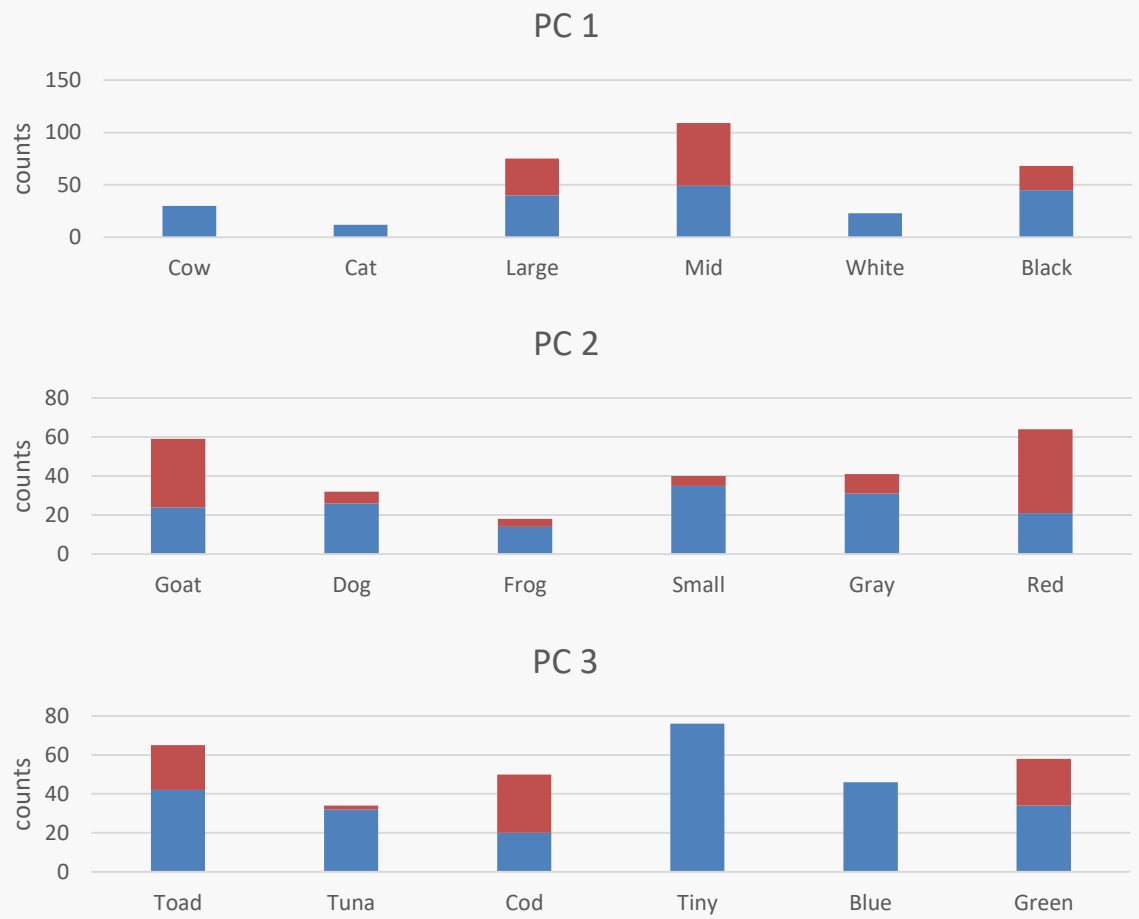
Sequence

BCD

Secondly, sequencing features in each group by the column properties, for outputting more diversified BCD values after encoding.

1. Column sum
2. Type purity
3. Feature importance

Grouped by PCA



Group
Sequence
BCD

Methodology

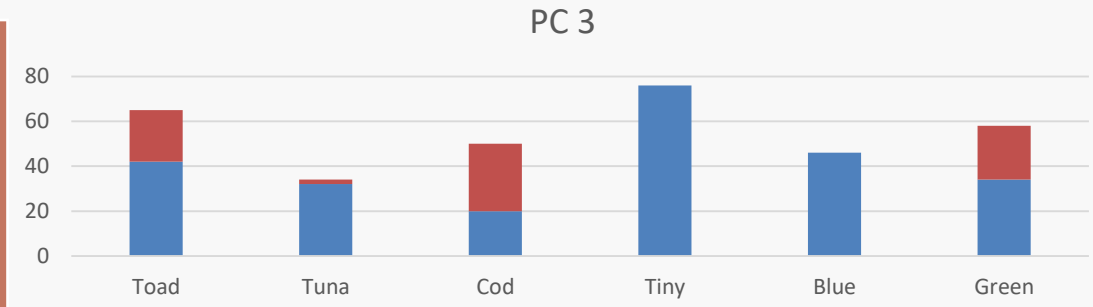
-Sequencing

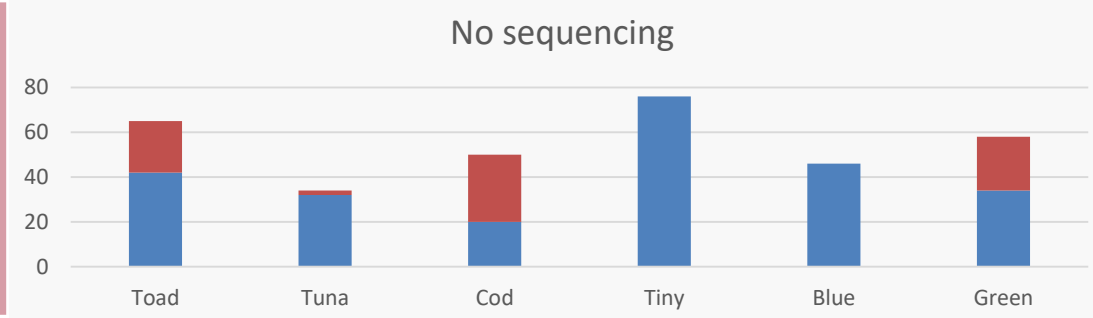
Group

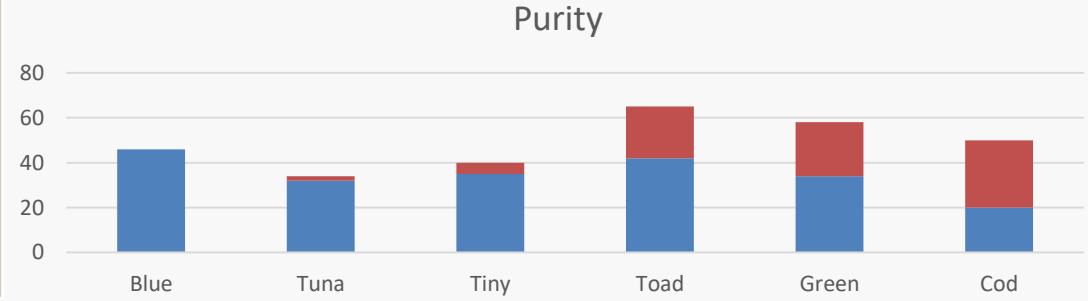
Sequence

No Seq.

Purity



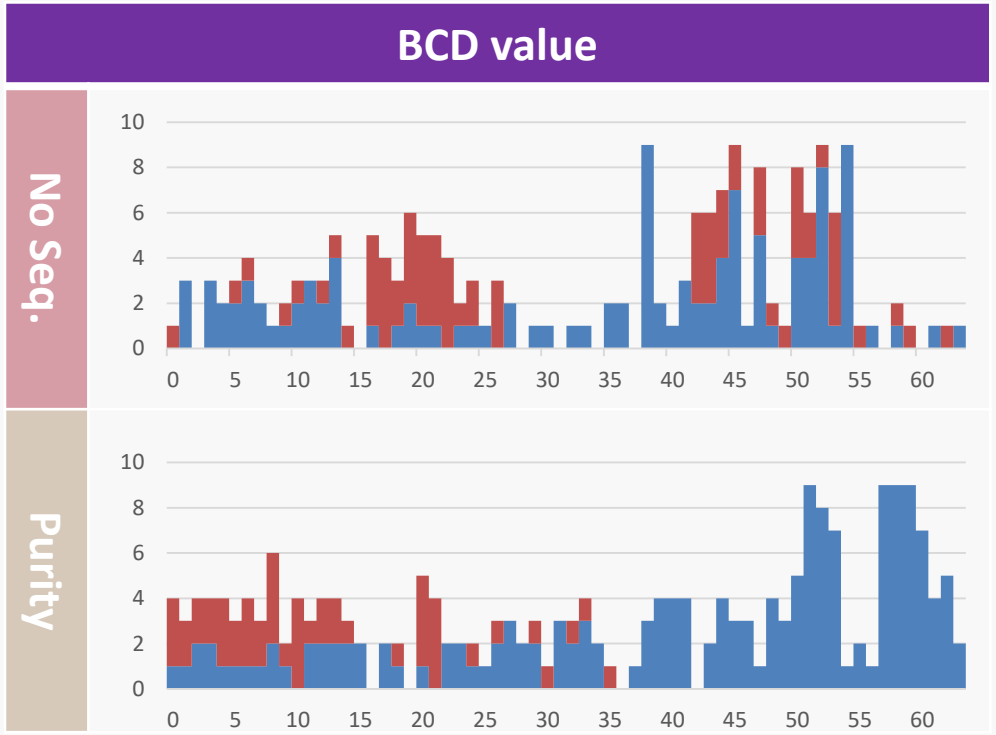




Methodology

-Sequencing

Group
Sequence
BCD



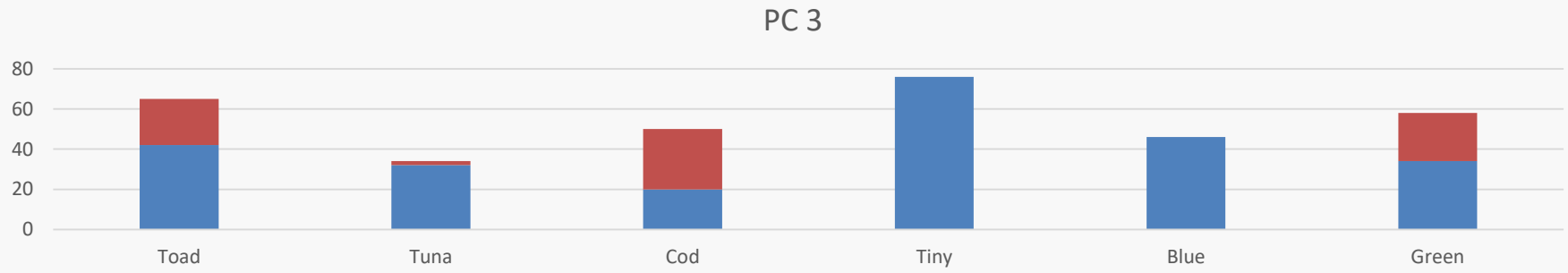
Group
Sequence
BCD

Methodology

-Sequencing

Group

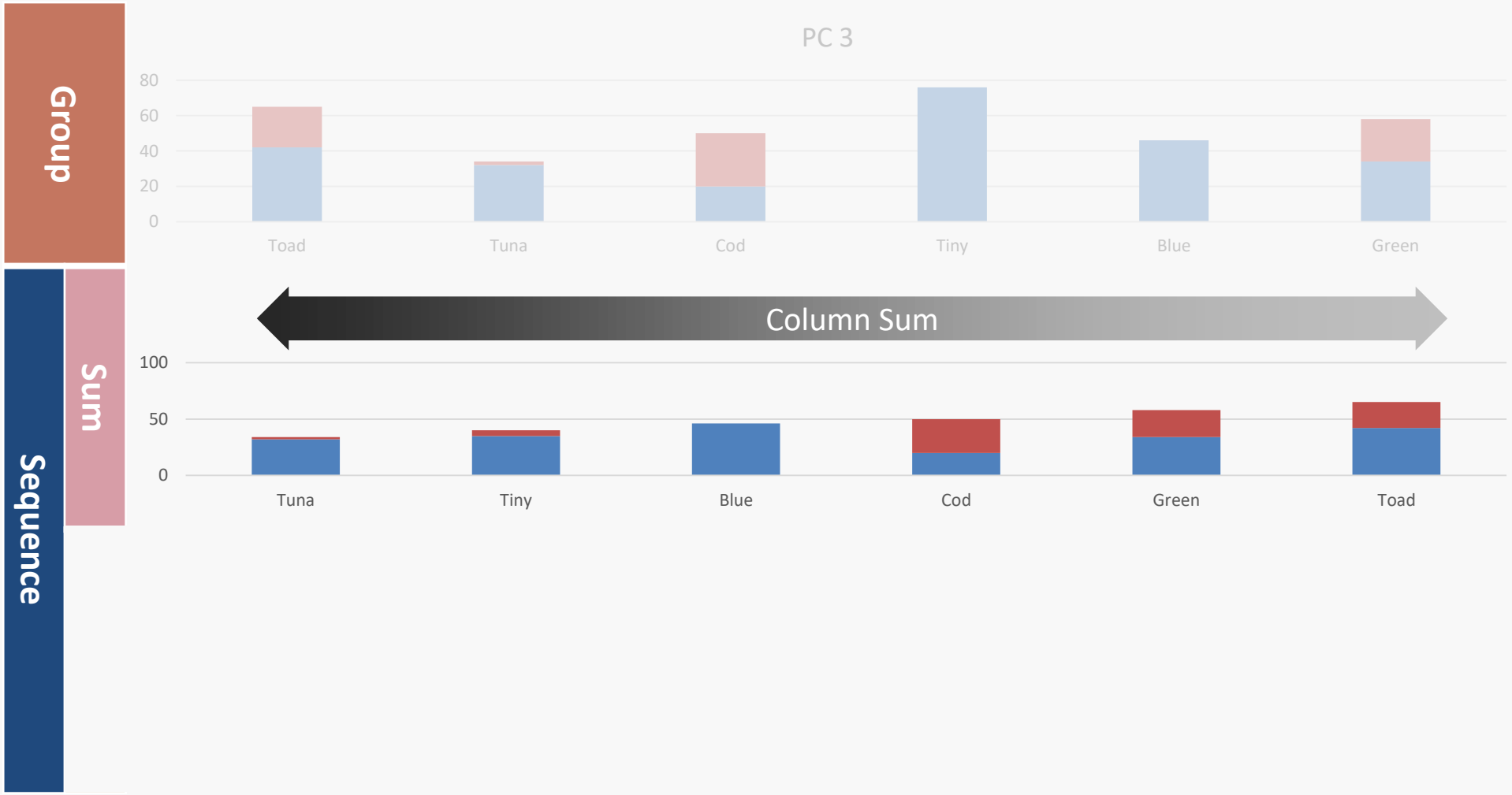
Sequence



Group
Sequence
BCD

Methodology

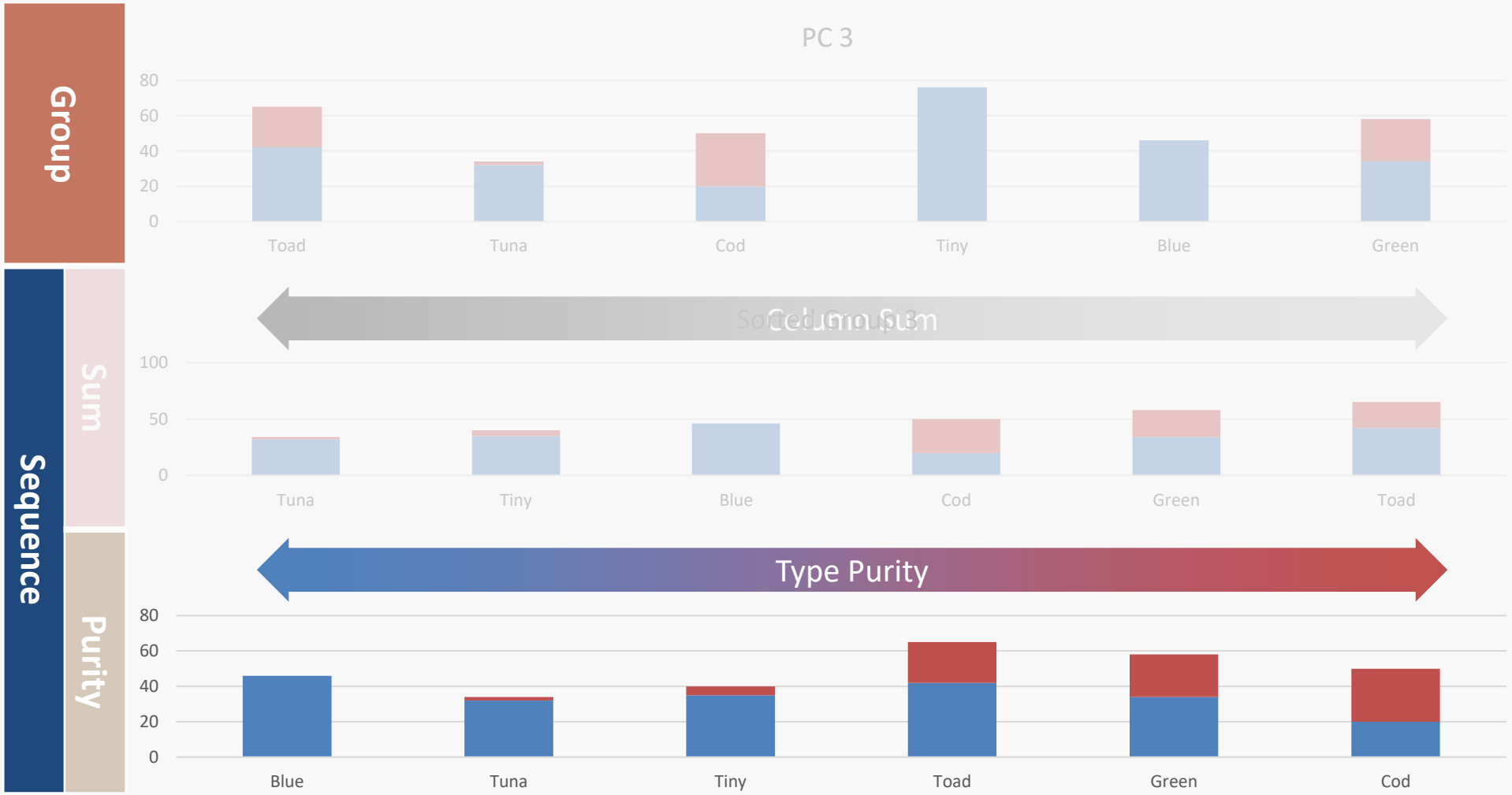
-Sequencing



Methodology

-Sequencing

Group
Sequence
BCD



Group
Sequence
BCD

Methodology

-BCD code

Finally, using Binary Coded Decimal to compute the numerical values representing the feature groups.

1. BCD
2. Ranked BCD

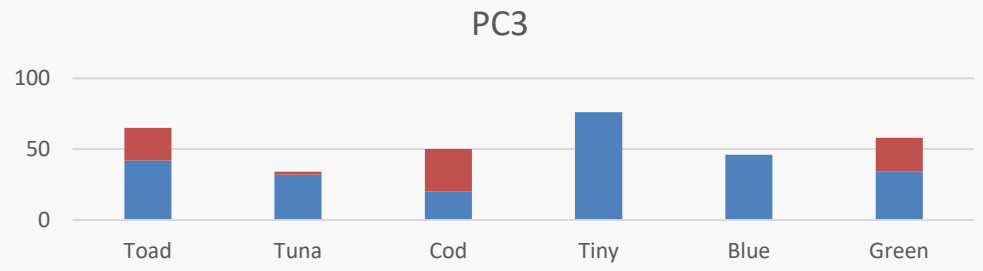
Decimal digit	BCD			
	8	4	2	1
0	0	0	0	0
1	0	0	0	1
2	0	0	1	0
3	0	0	1	1
4	0	1	0	0
5	0	1	0	1
6	0	1	1	0
7	0	1	1	1
8	1	0	0	0
9	1	0	0	1

Group
Sequence
BCD

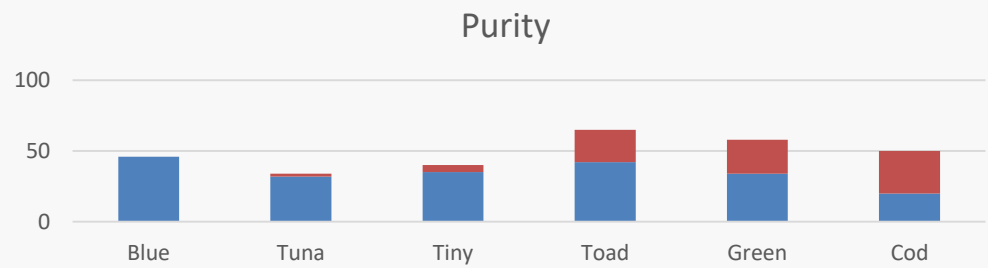
Methodology

-BCD code

Group



Sequence

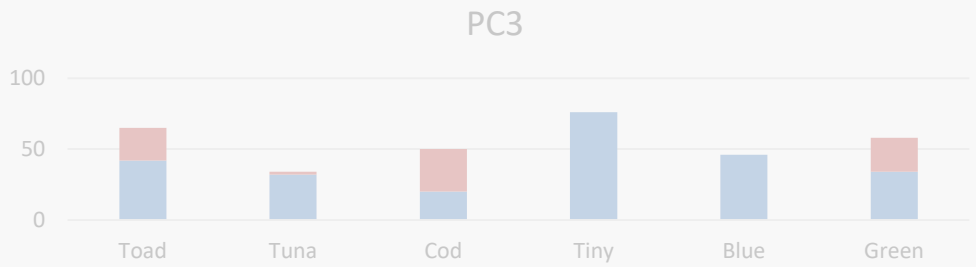


Group
Sequence
BCD

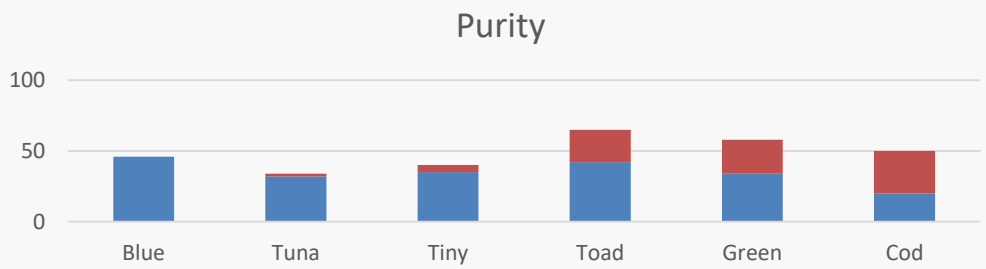
Methodology

-BCD code

Group



Sequence



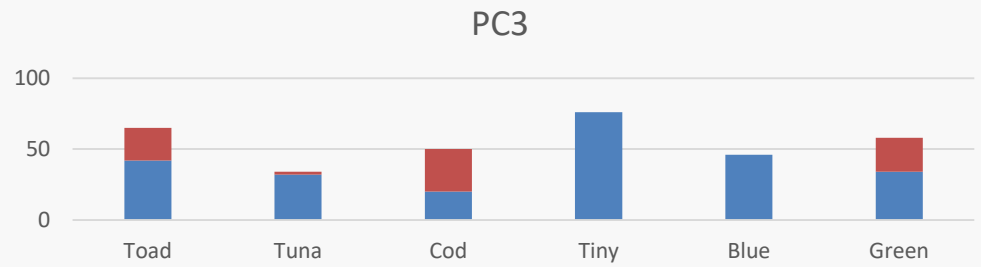
BCD	Sequenced features						Type
	Blue	Tuna	Tiny	Toad	Green	Cod	
33	1	0	0	0	0	1	Good
36	1	0	0	1	0	0	Good
26	0	1	1	0	1	0	Bad
14	0	0	1	1	1	0	Bad

Group
Sequence
BCD

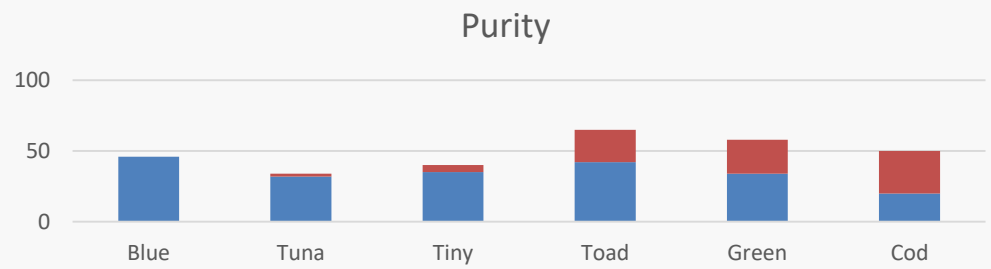
Methodology

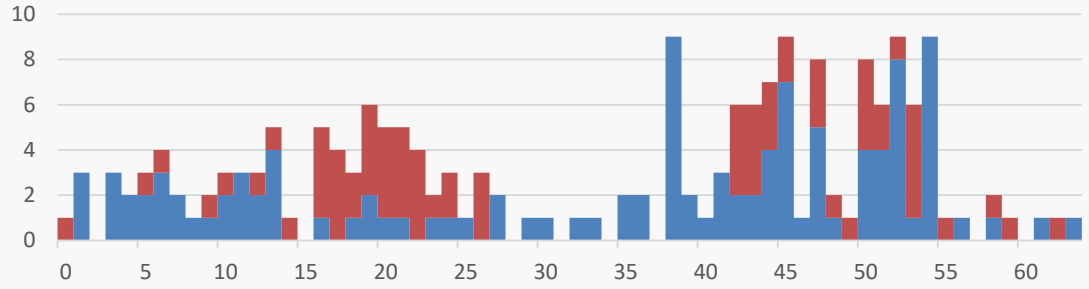
-BCD code

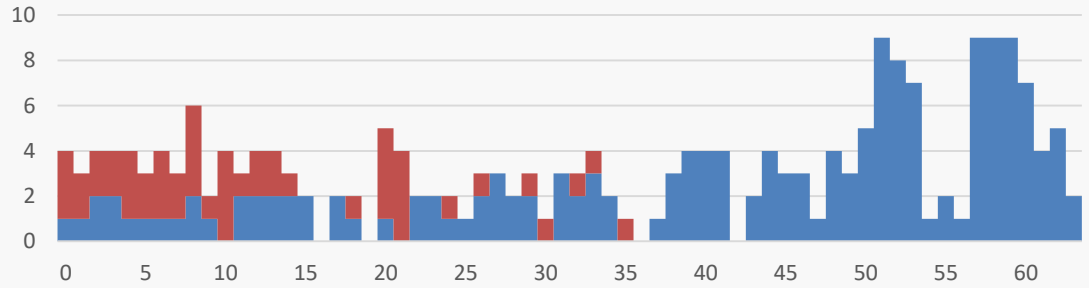
Group



Sequence





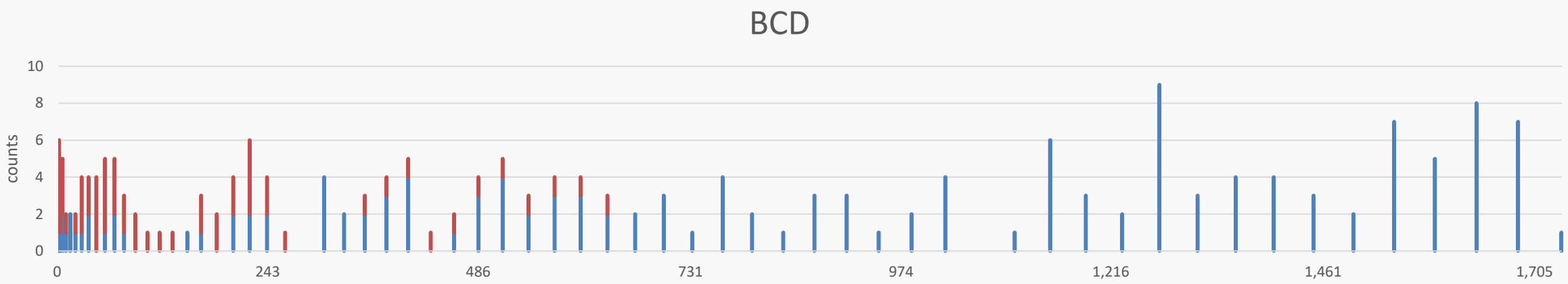


BCD	Sequenced features						Type
	Blue	Tuna	Tiny	Toad	Green	Cod	
33	1	0	0	0	0	1	Good
36	1	0	0	1	0	0	Good
26	0	1	1	0	1	0	Bad
14	0	0	1	1	1	0	Bad

Group
Sequence
BCD

Methodology

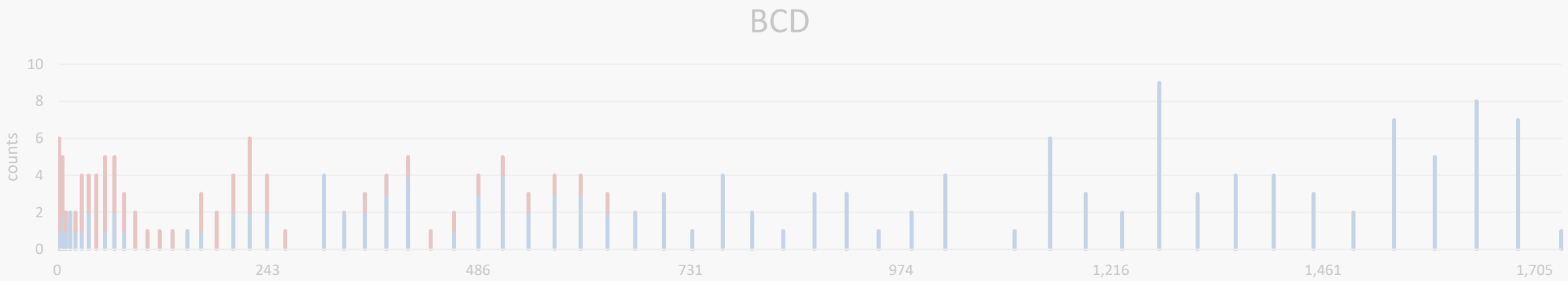
-BCD code

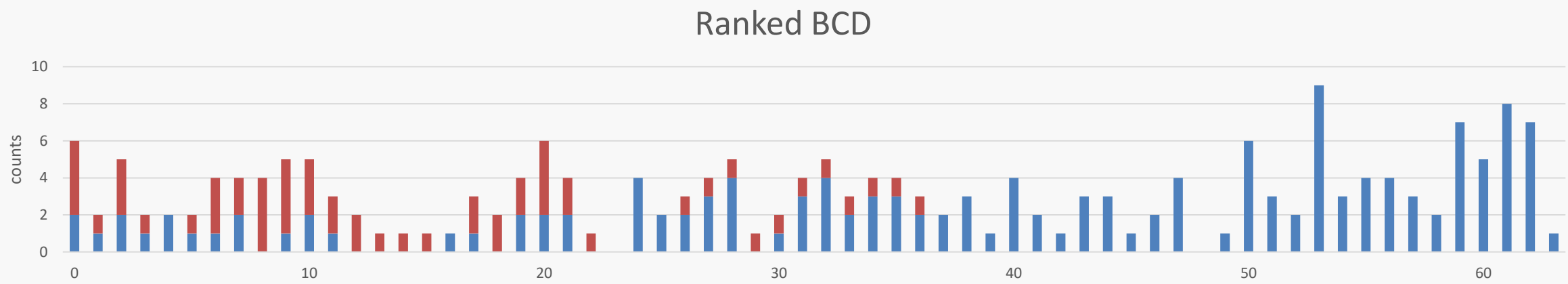


Group
Sequence
BCD

Methodology

-BCD code





Methodology

-sum up

Grouping

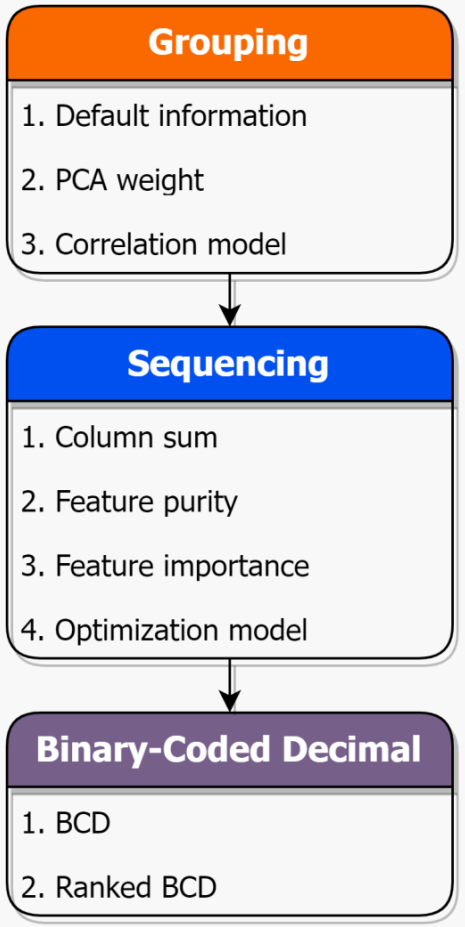
1. Clustering related or similar columns into groups.

Sequencing

2. Sort the columns in the group with different column properties.

Binary-Coded Decimal

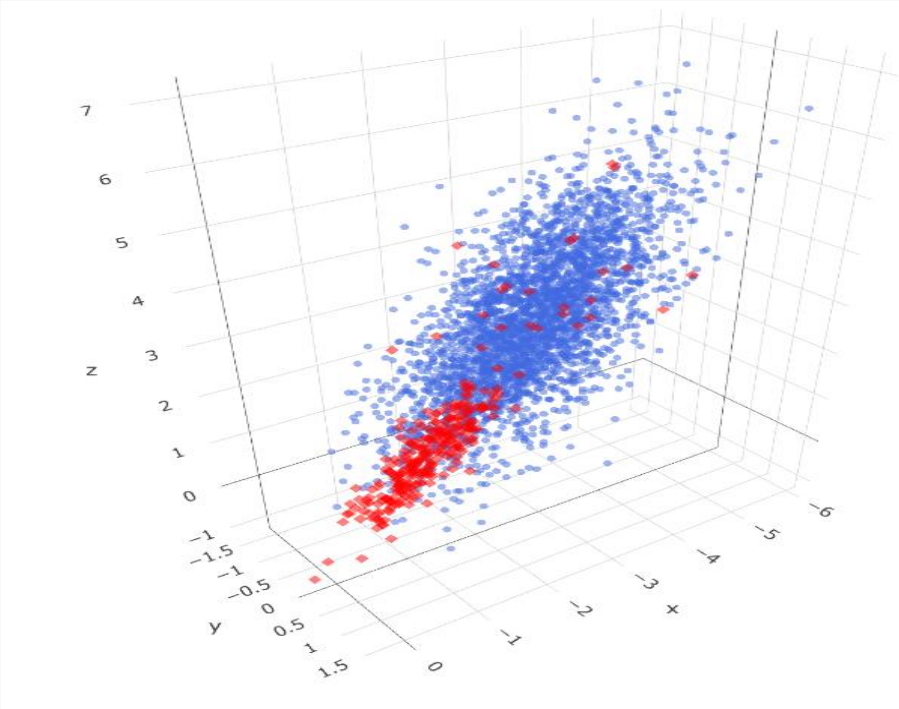
3. Transform column groups into a numerical value with BCD.

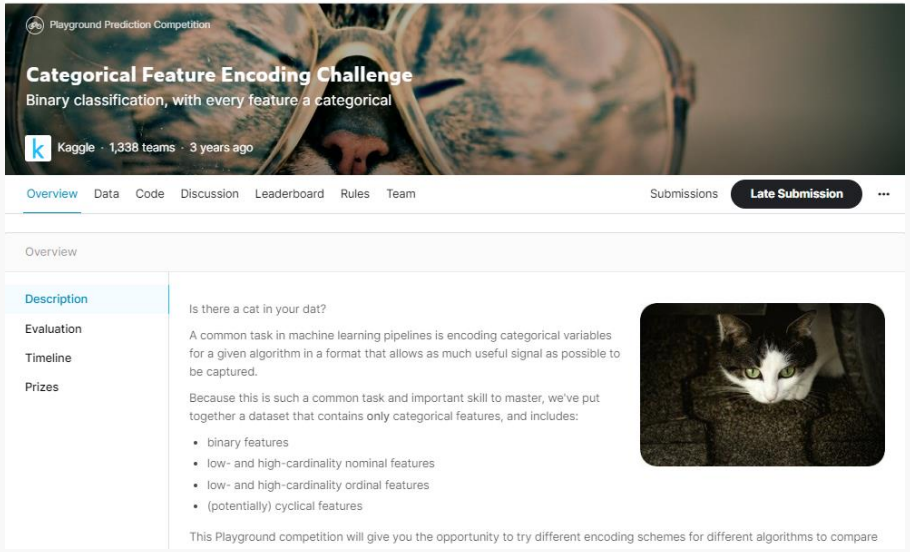


Case study

We compare classification results with the commonly used variable encoding method under different datasets.

1. Simulated continuous datasets
2. Kaggle-Feature encoding challenge dataset

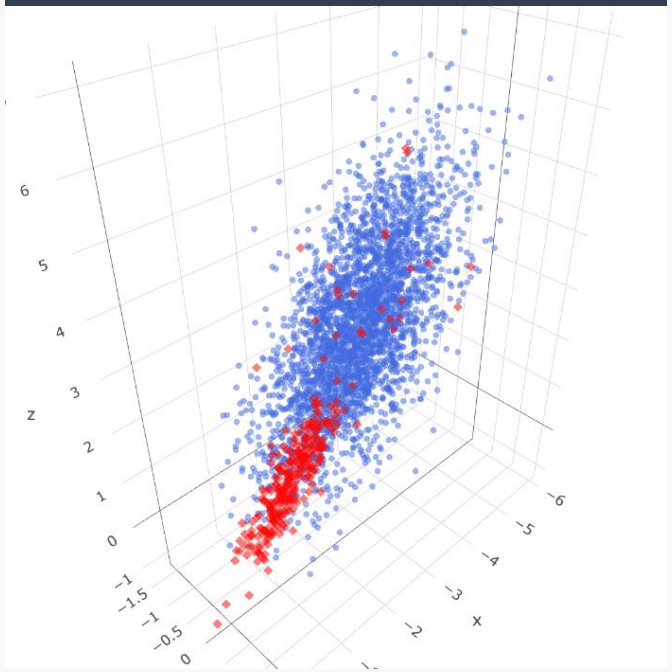




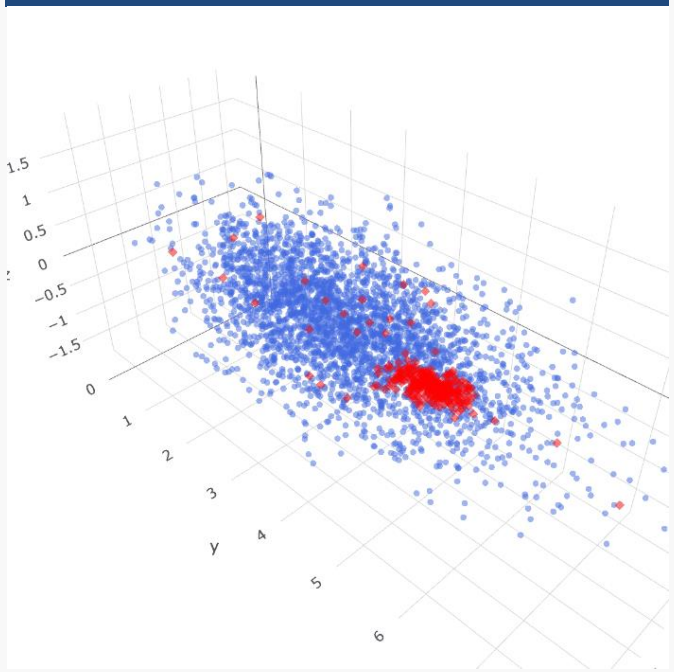
Simulated data

-(3300 samples, 3000 Good, 300 Bad)

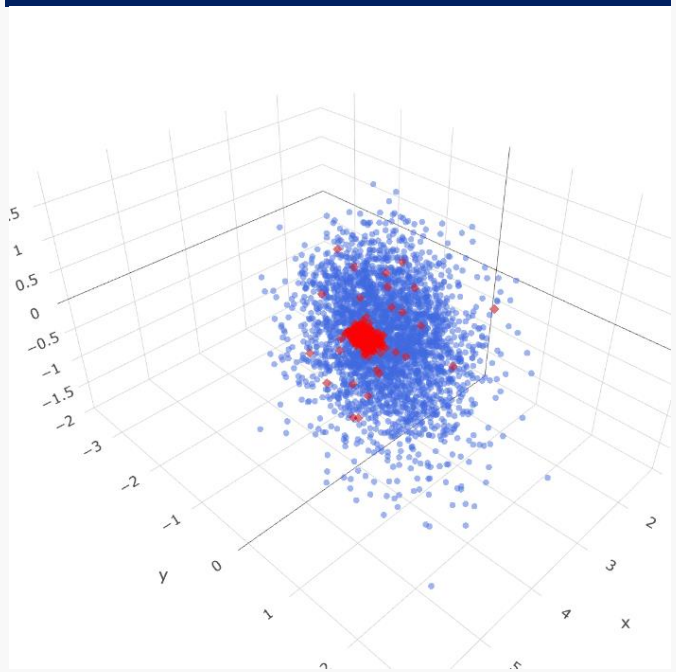
Dataset 1



Dataset 2

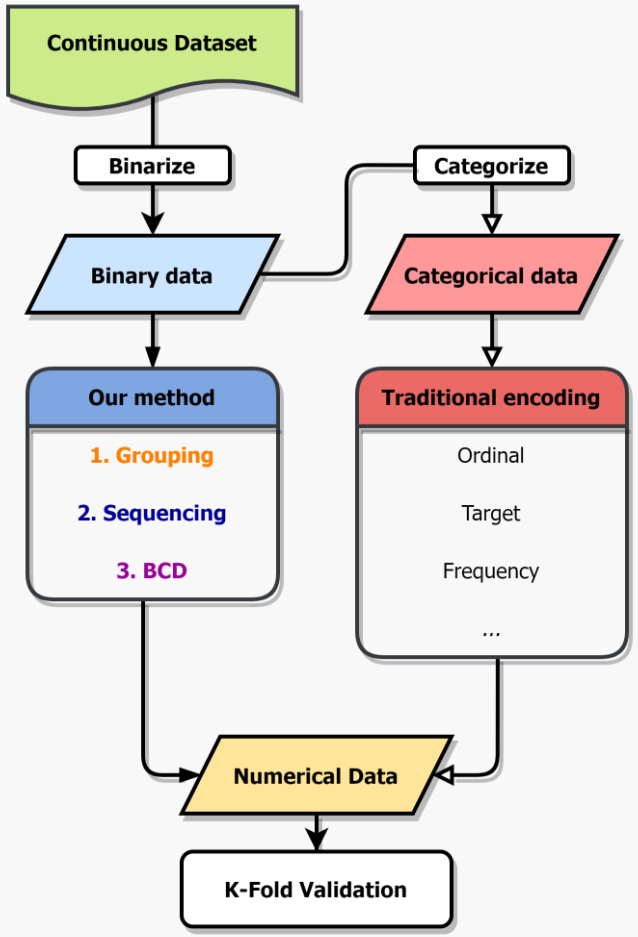


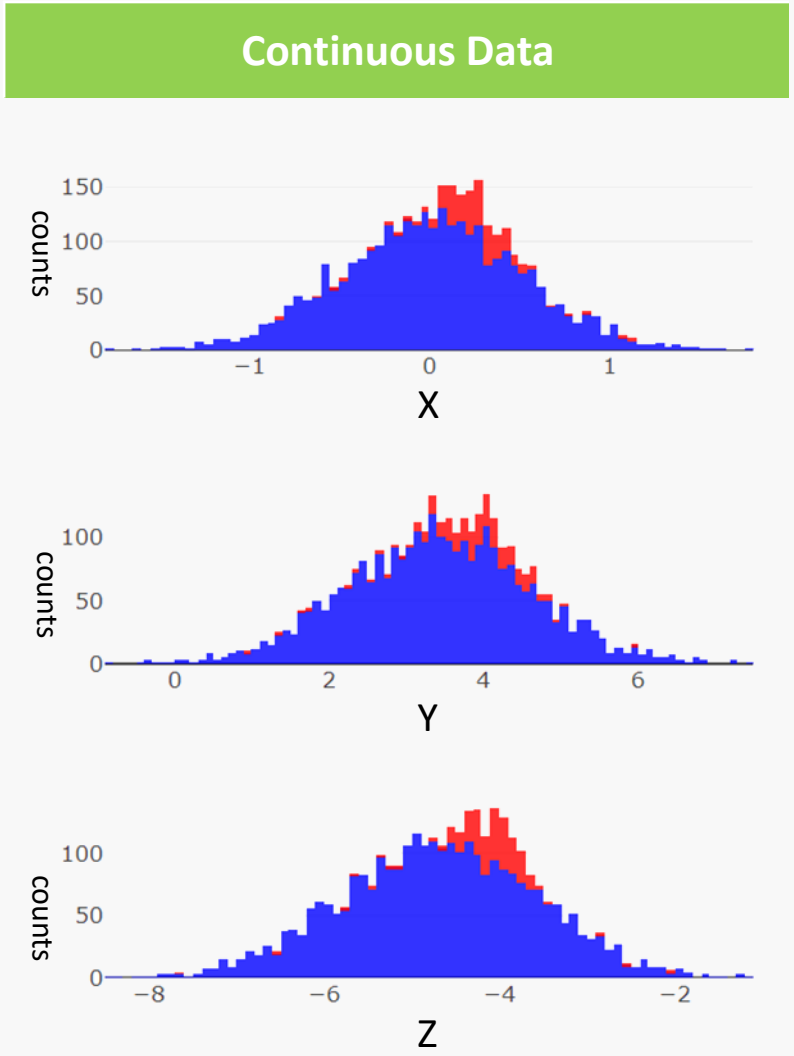
Dataset 3



Simulated data

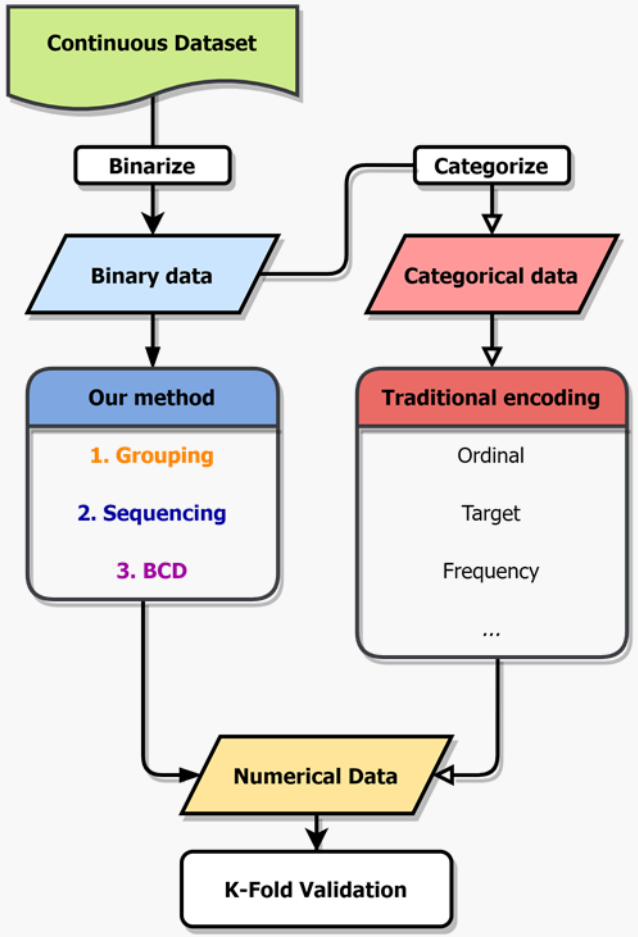
- binarize

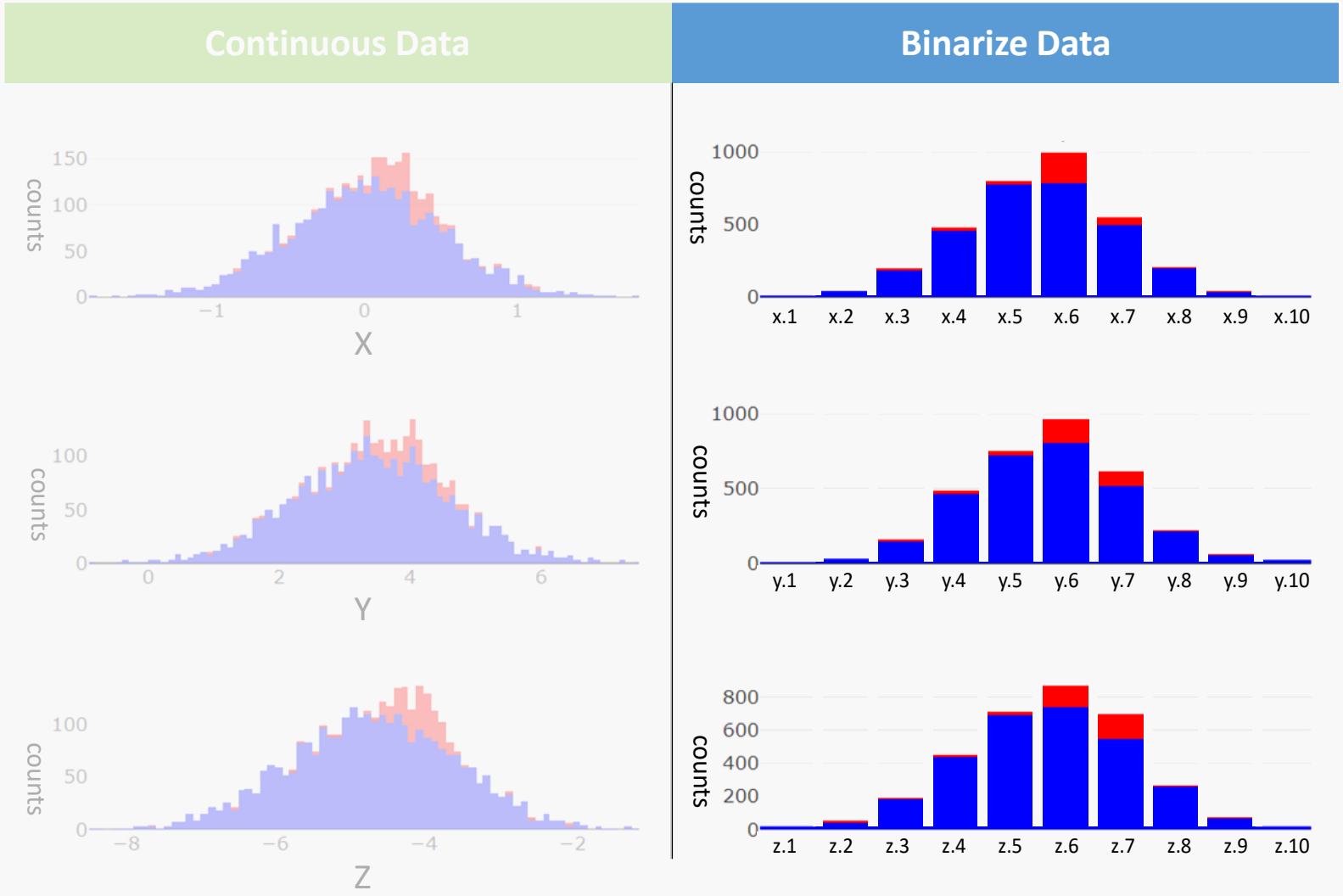




Simulated data

- *binarize*





Simulated data

- Categorize

Binarize Data

Instances	x.1	x.2	x.3	x.4	x.5	x.6	…	z.9	z.10
1	0	1	0	0	0	0	…	0	0
2	1	0	0	0	0	0	…	0	0
3	0	0	0	0	0	0	…	0	0
4	0	0	0	0	1	0	…	0	0
5	0	0	1	0	0	0	…	0	0
6	0	0	0	0	1	0	…	0	0
7	0	0	0	0	0	0	…	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3300	0	0	0	1	0	0	…	0	0

Simulated data

- Categorize

Categorical Data

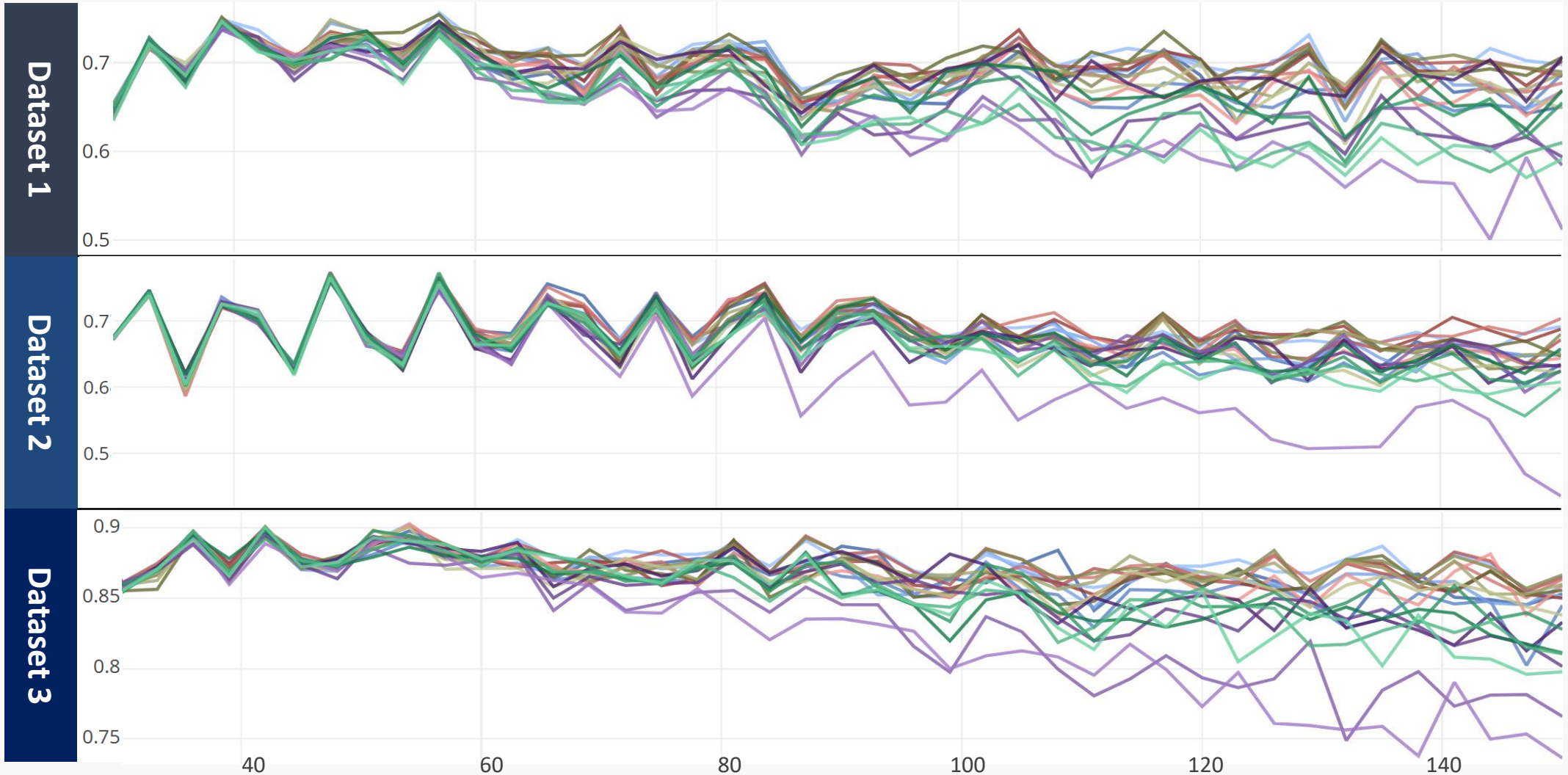
Instances	X	Y	Z
1	x.2	y.9	z.4
2	x.1	y.8	z.2
3	x.9	y.1	z.6
4	x.5	y.7	z.8
5	x.3	y.1	z.8
6	x.5	y.6	z.3
7	x.7	y.5	z.4
⋮	⋮	⋮	⋮
3300	x.4	y.8	z.1

Binarize Data

Instances	x.1	x.2	x.3	x.4	x.5	x.6	…	z.9	z.10
1	0	1	0	0	0	0	…	0	0
2	1	0	0	0	0	0	…	0	0
3	0	0	0	0	0	0	…	0	0
4	0	0	0	0	1	0	…	0	0
5	0	0	1	0	0	0	…	0	0
6	0	0	0	0	1	0	…	0	0
7	0	0	0	0	0	0	…	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3300	0	0	0	1	0	0	…	0	0

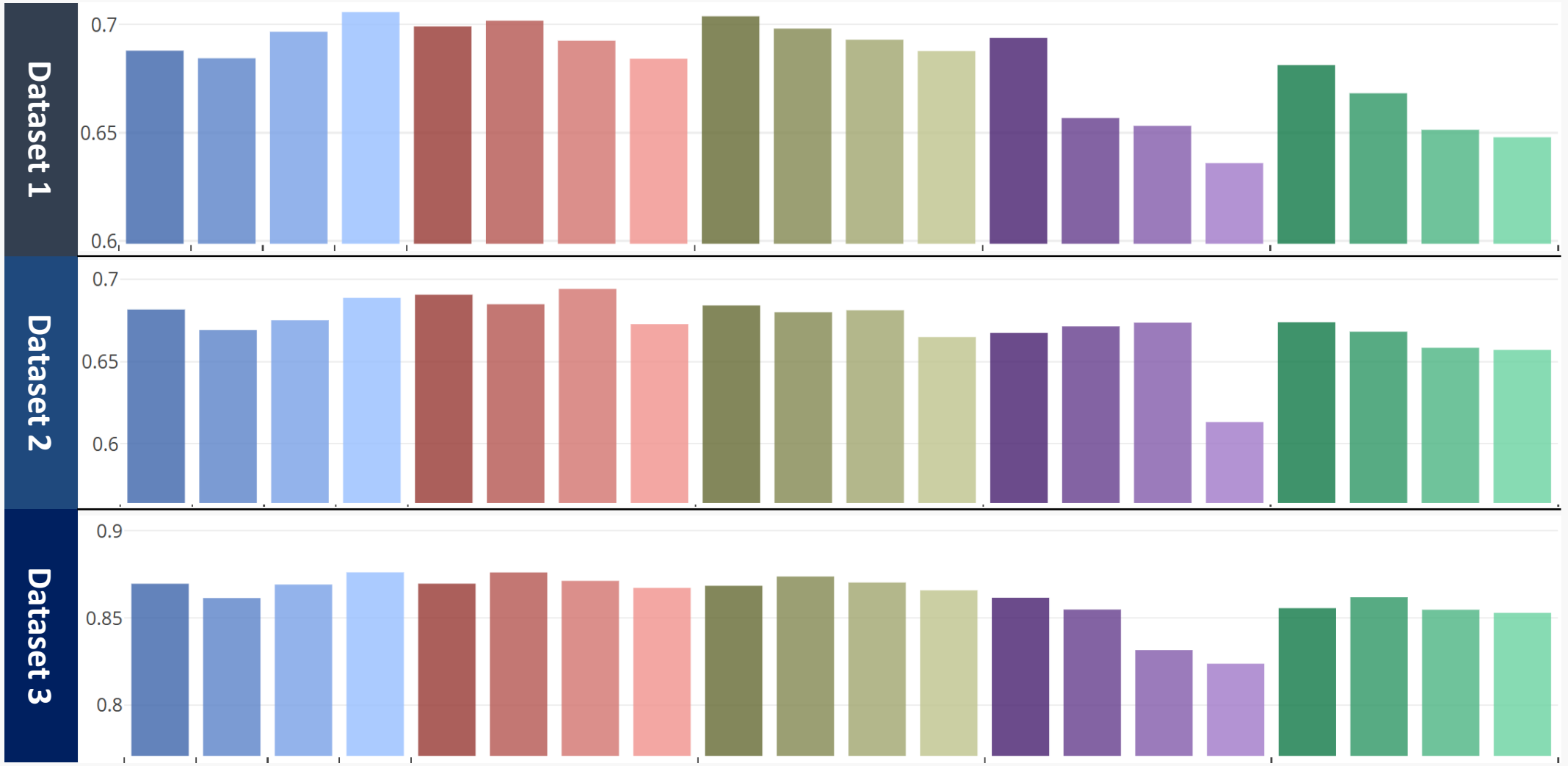
Simulated data

-Classification



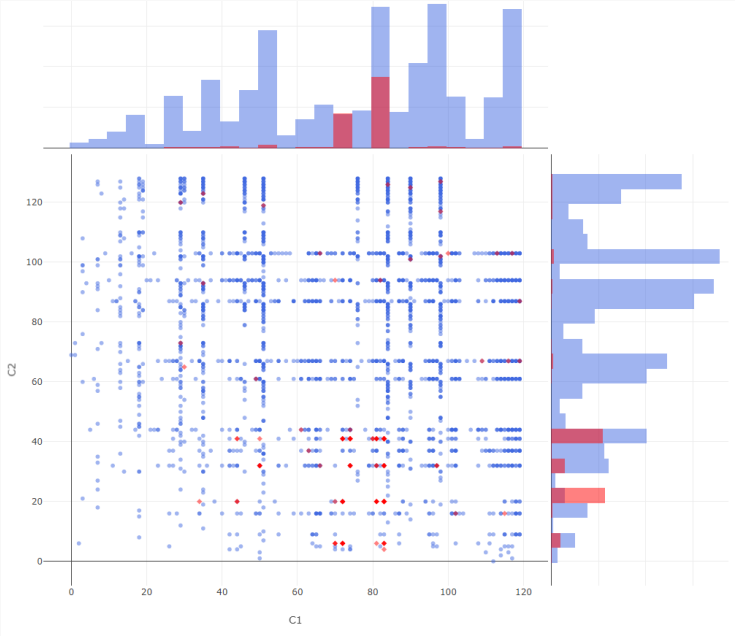
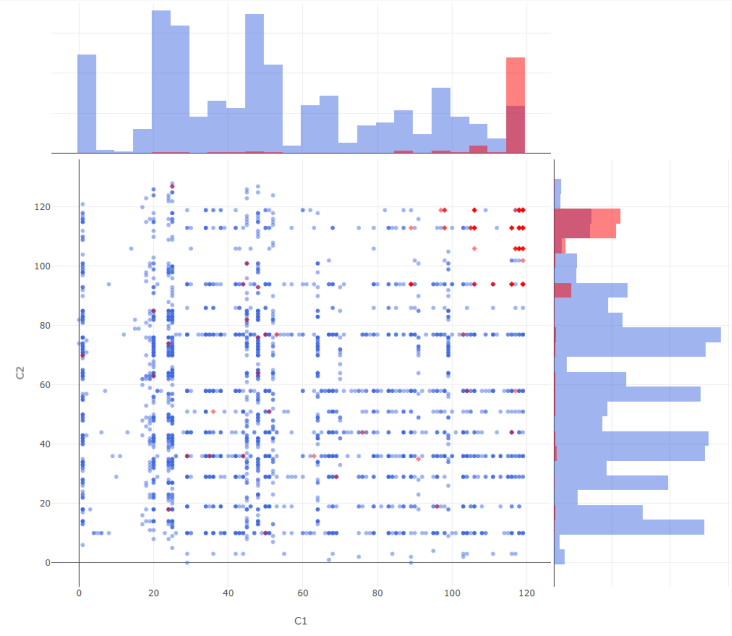
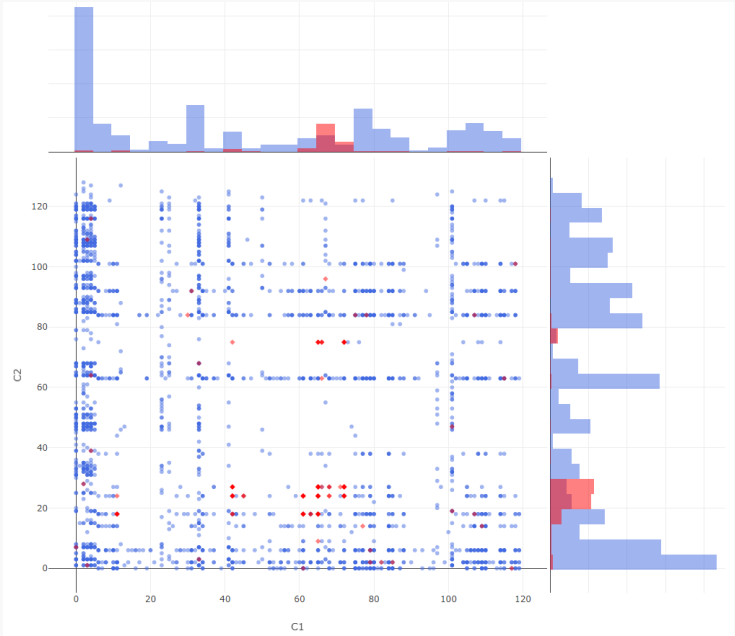
Simulated data

-Classification



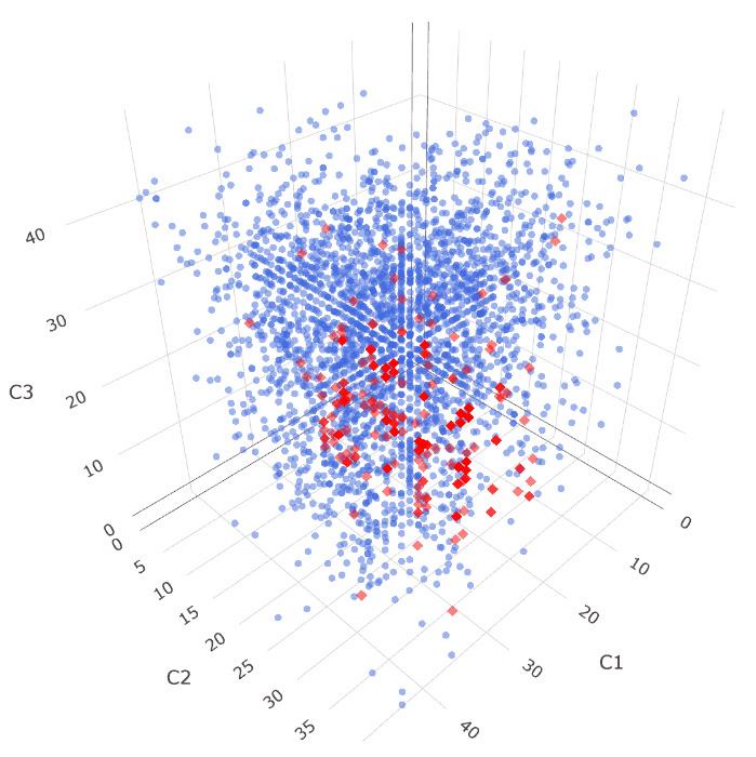
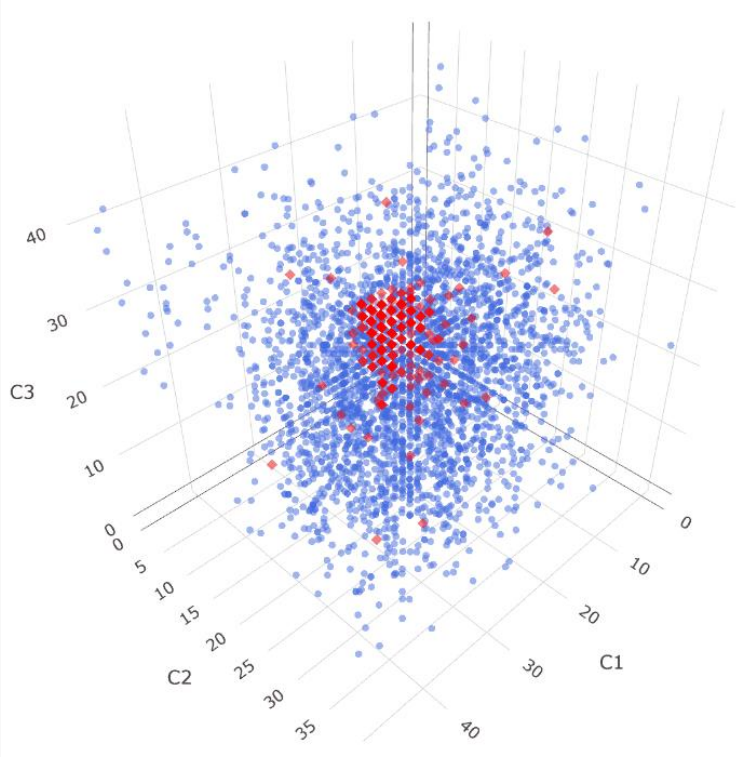
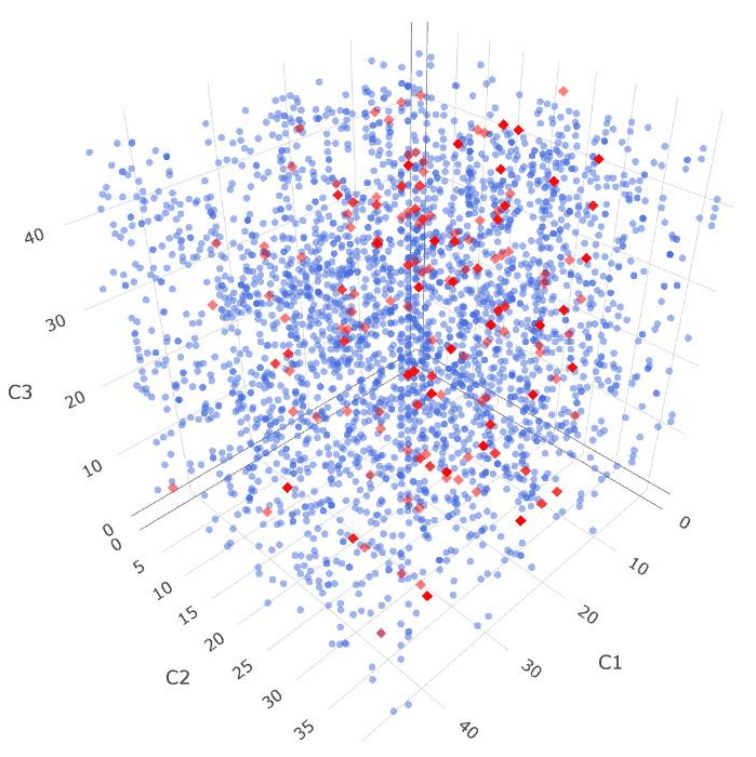
Simulated data

-Dimension reduction

Sum	Impurity	RND
		

Simulated data

-Dimension reduction

Sum	Impurity	RND
		

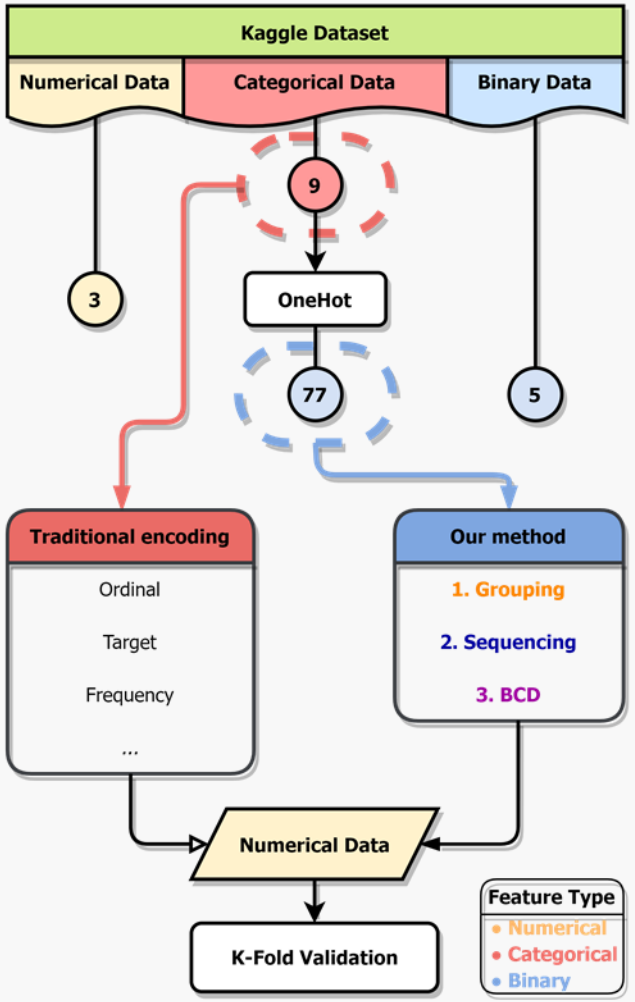
Kaggle dataset

Numerical feature uniqueness of Kaggle CFEC dataset

bin0	bin1	bin2	bin3	bin4	ord0	day	month	Target
0	0	0	0	0	1	1	1	0
1	1	1	1	1	2	2	2	1
					3	3	3	
						⋮	⋮	
						7	12	

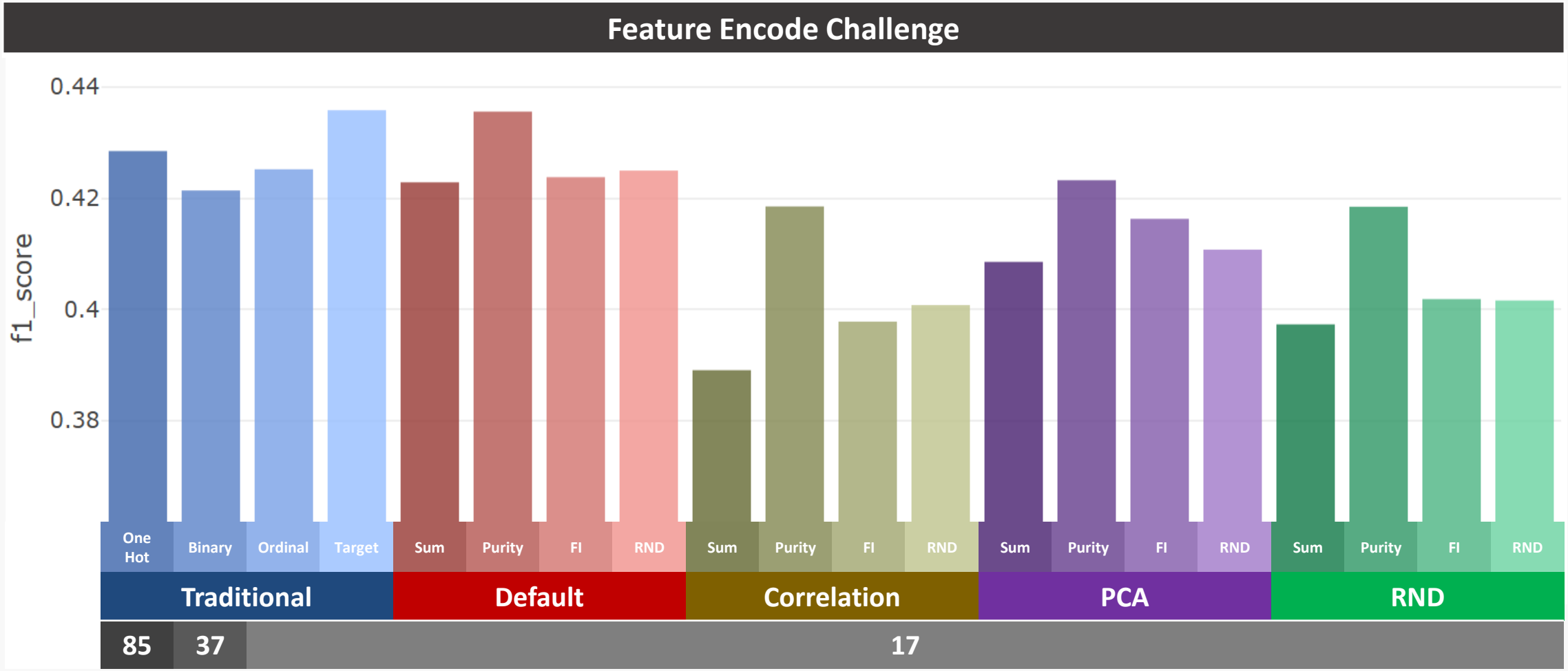
Categorical feature uniqueness of Kaggle CFEC dataset

nom0	nom1	nom2	nom3	nom4	ord1	ord2	ord3	ord4
Green	Triangle	Snake	Finland	Bassoon	Grandmaster	Cold	a	A
Blue	Trapezoid	Hamster	Russia	Piano	Expert	Hot	b	B
Red	Polygon	Lion	Canada	Theremin	Novice	Lava Hot	c	C
	Square	Cat	Costa Rica	Oboe	Contributor	Boiling Hot	d	D
	Star	Dog	China		Master	Freezing	⋮	⋮
	Circle	Axolotl	India			Warm	o	Z



Kaggle dataset

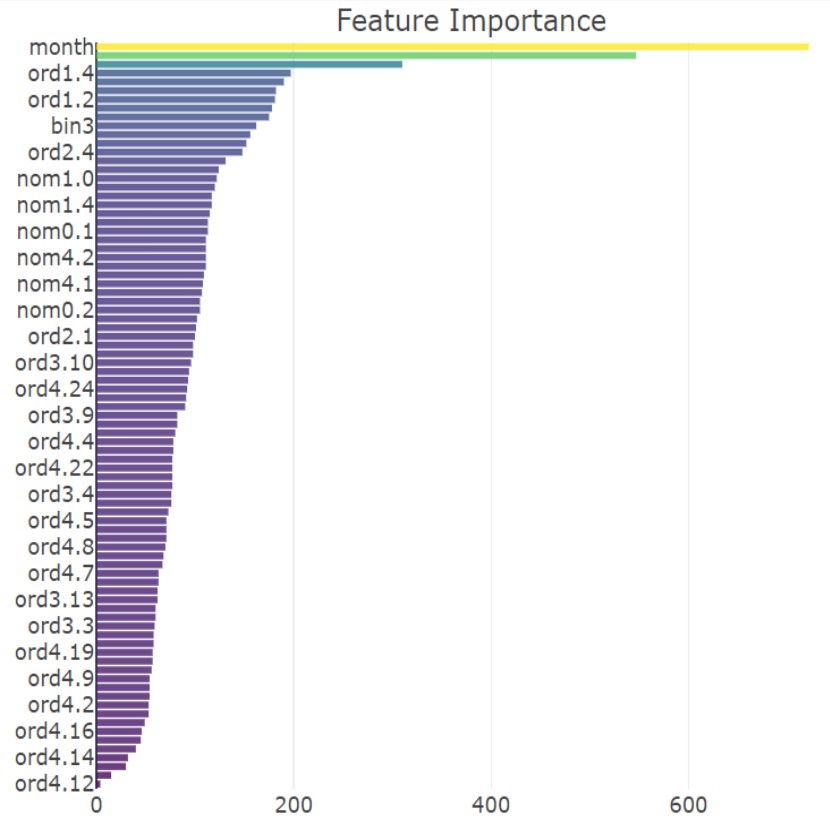
-Classification



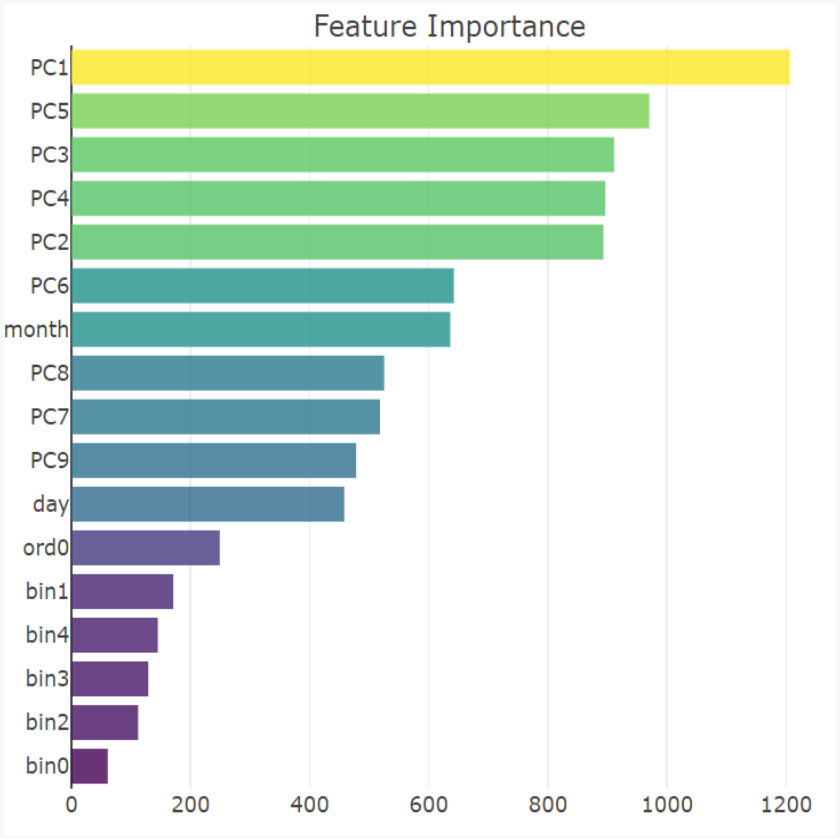
Kaggle dataset

-Feature importance changes

Before Encoding



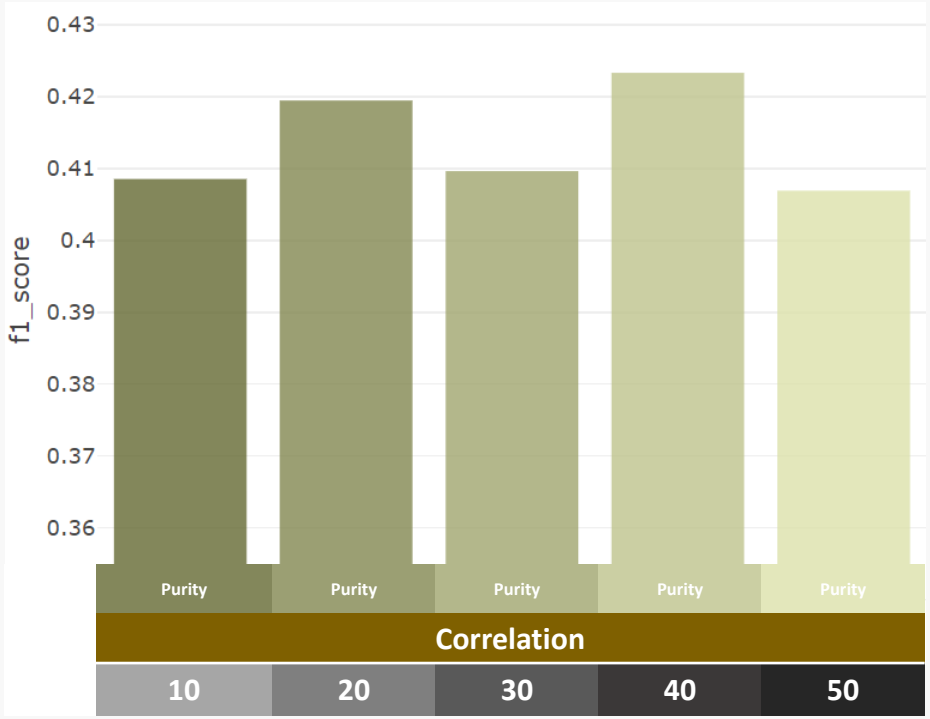
PCA-Purity



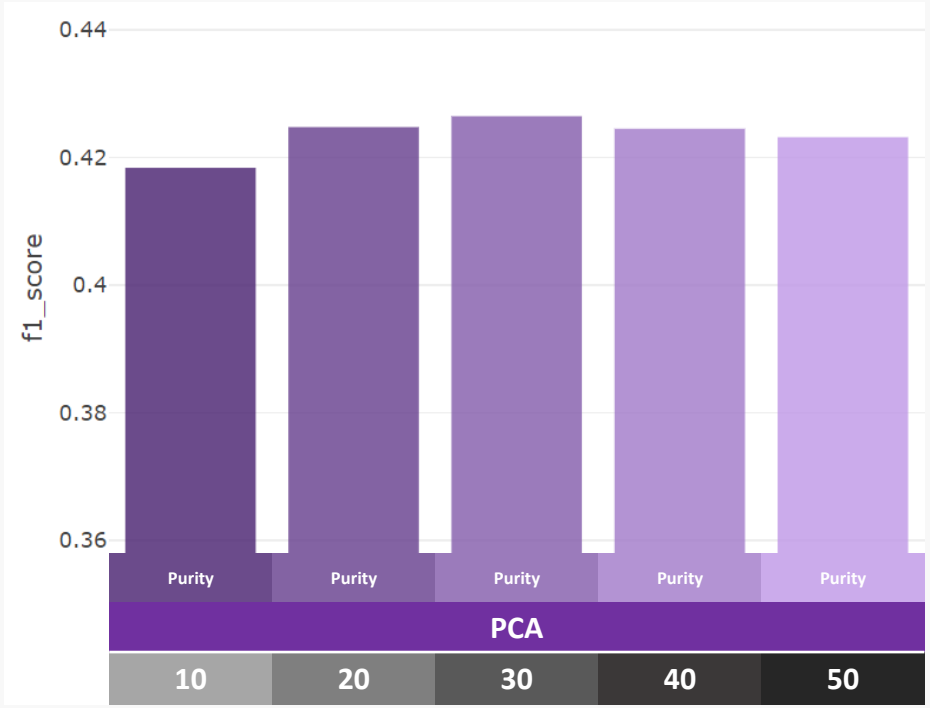
Kaggle dataset

-Different grouping seizes

Correlation

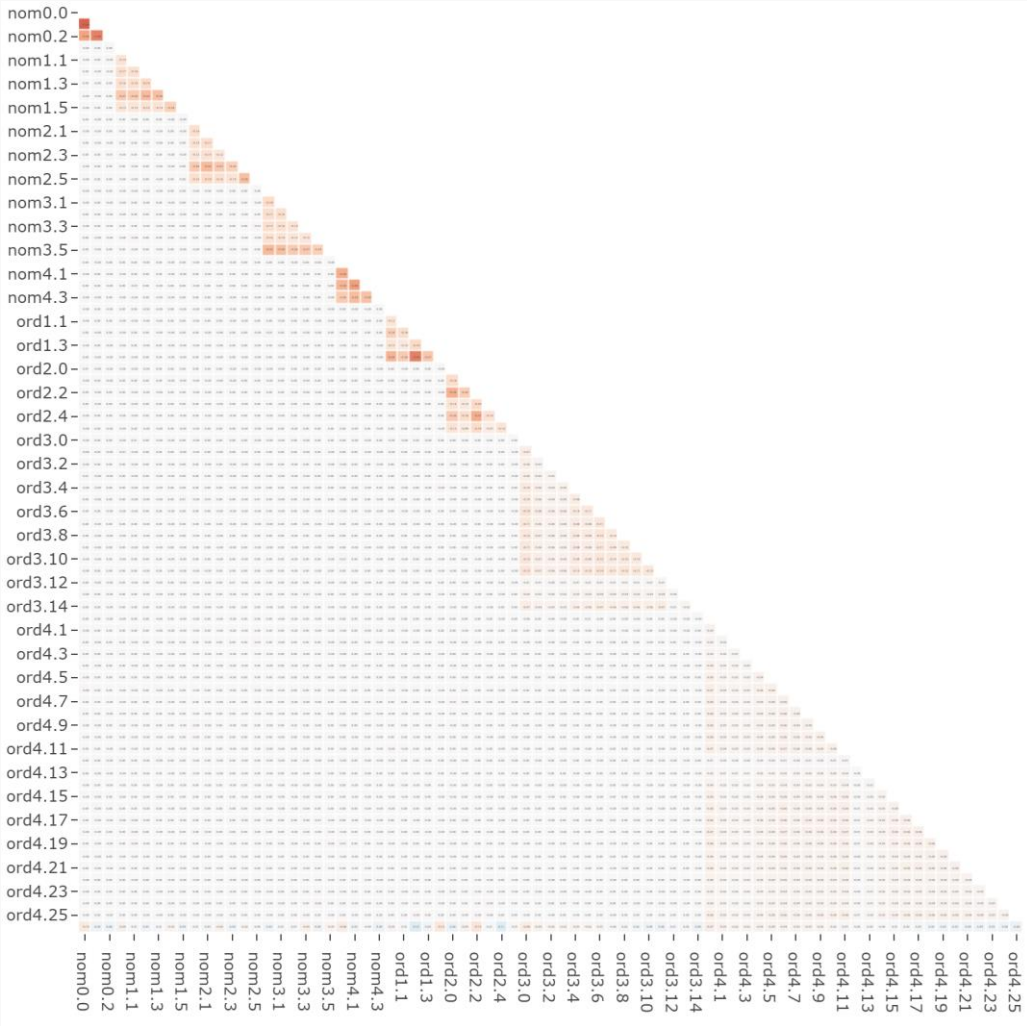


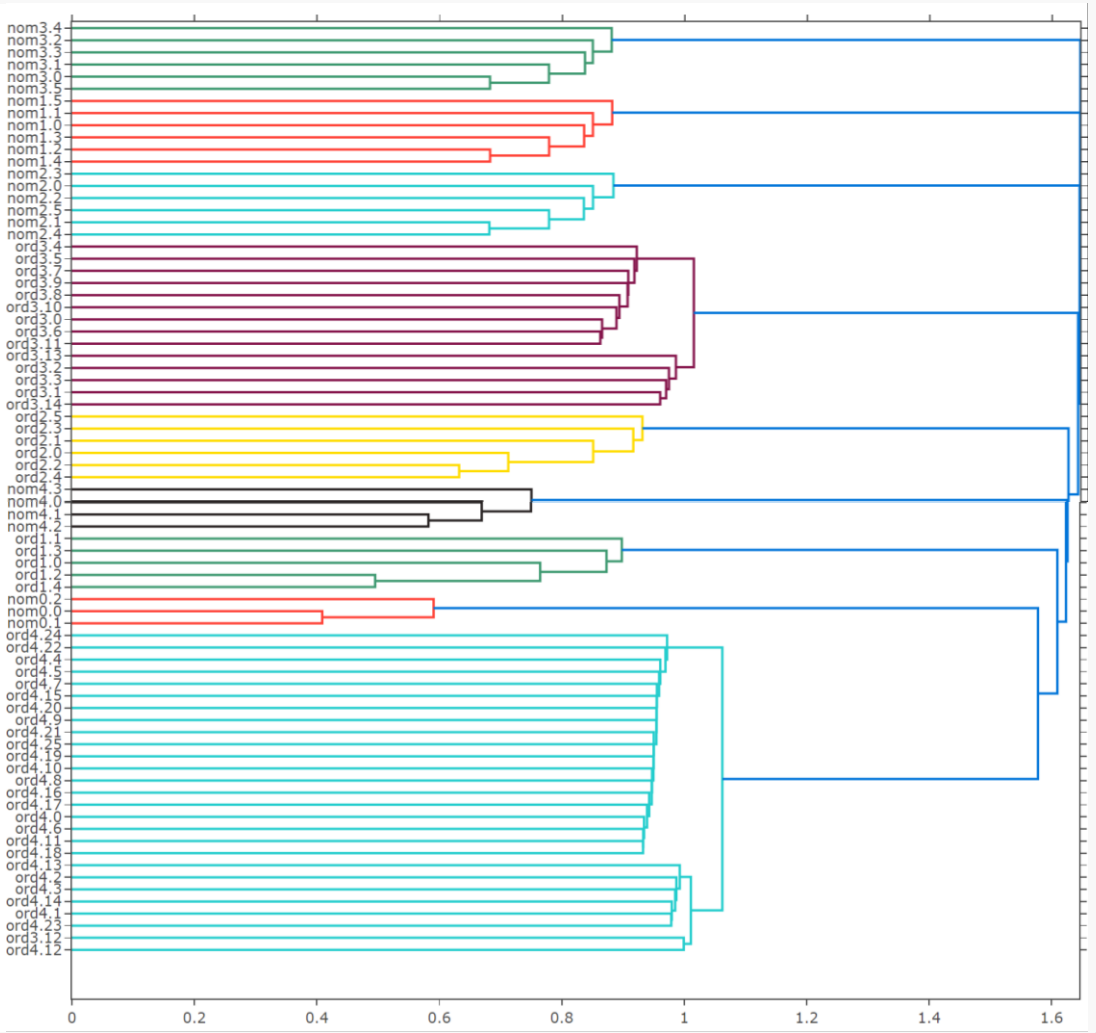
PCA



Kaggle dataset

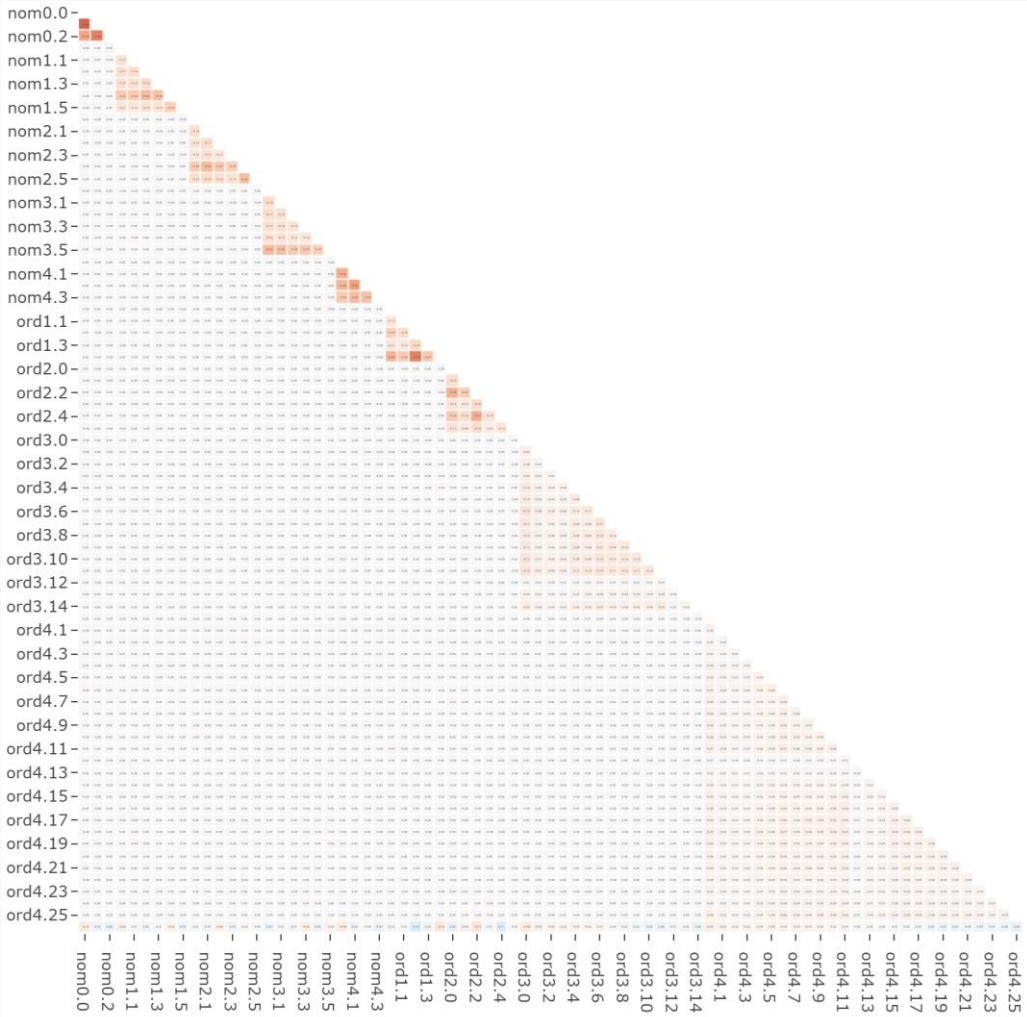
-Group by hierarchical clustering

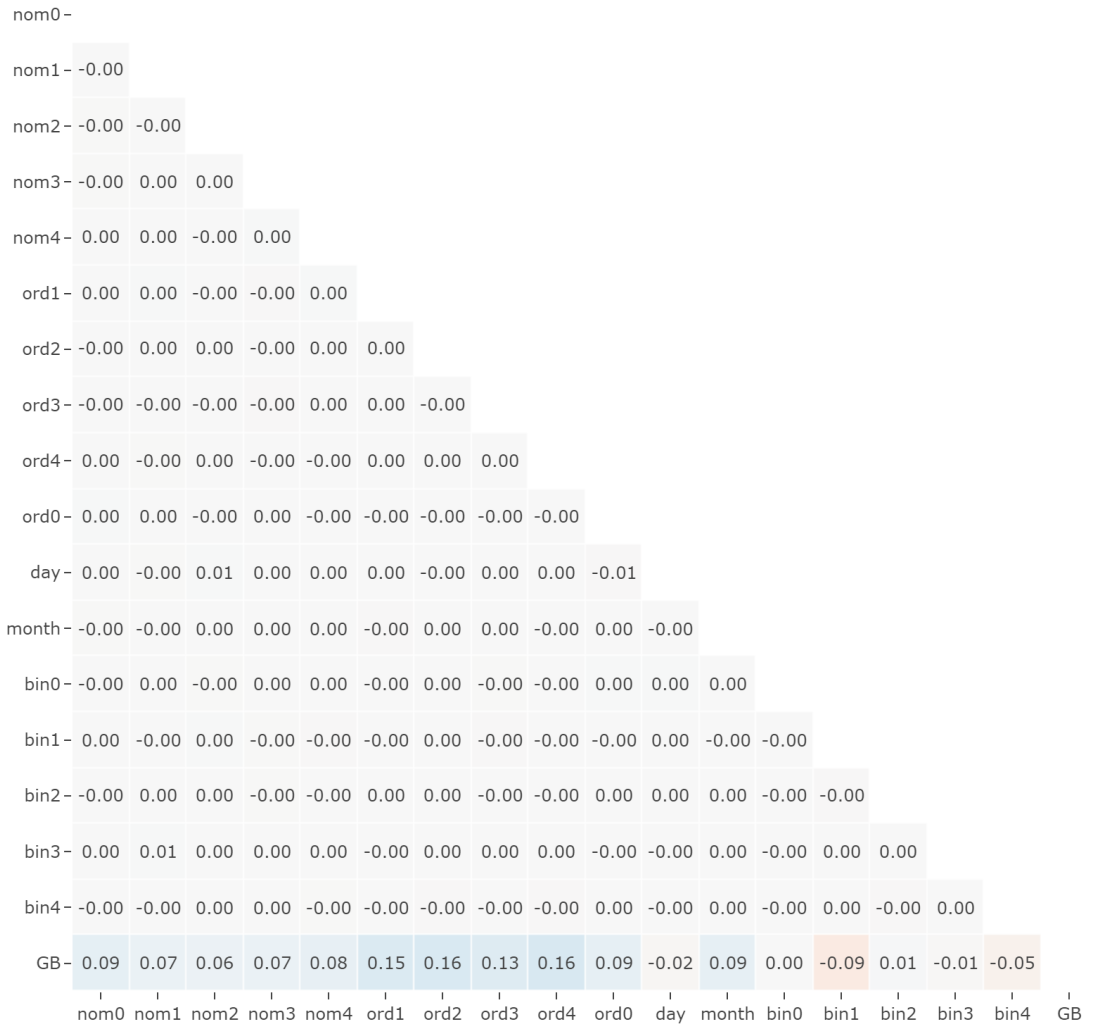




Kaggle dataset

-Group by hierarchical clustering

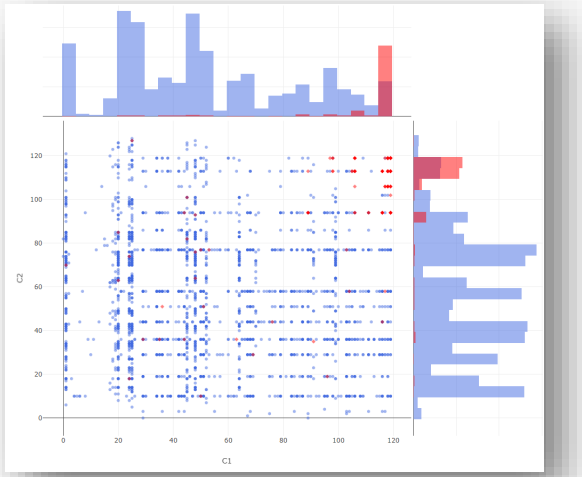


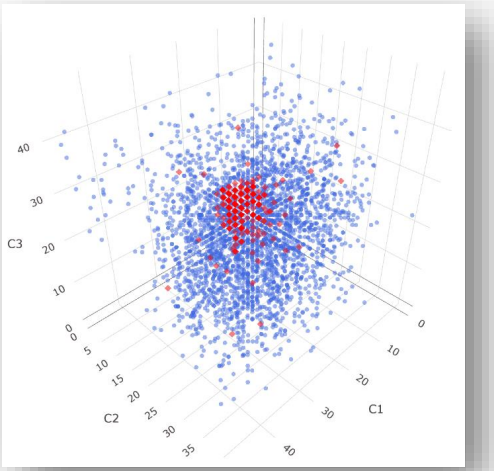


Conclusion

1. A new encoding scheme for binary data.
2. Compress binary features information into integers.
3. Preserve classification/regression performance with proposed grouping & sequencing techniques.
4. A dimension reduction method of high dimensional binary data.

Instances	x.1	x.2	x.3	x.4	x.5	x.6	…	z.9	z.10
1	0	1	0	0	0	0	…	0	0
2	1	0	0	0	0	0	…	0	0
3	0	0	0	0	0	0	…	0	0
4	0	0	0	0	1	0	…	0	0
5	0	0	1	0	0	0	…	0	0
6	0	0	0	0	1	0	…	0	0
7	0	0	0	0	0	0	…	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3300	0	0	0	1	0	0	…	0	0





Future work

1. Grouping techniques can be further enhanced.
2. Finding the optimal dimension of the data.
3. Sequencing problem may be solved with optimization algorithms.
4. Rigorous mathematical/statistical derivations are needed.