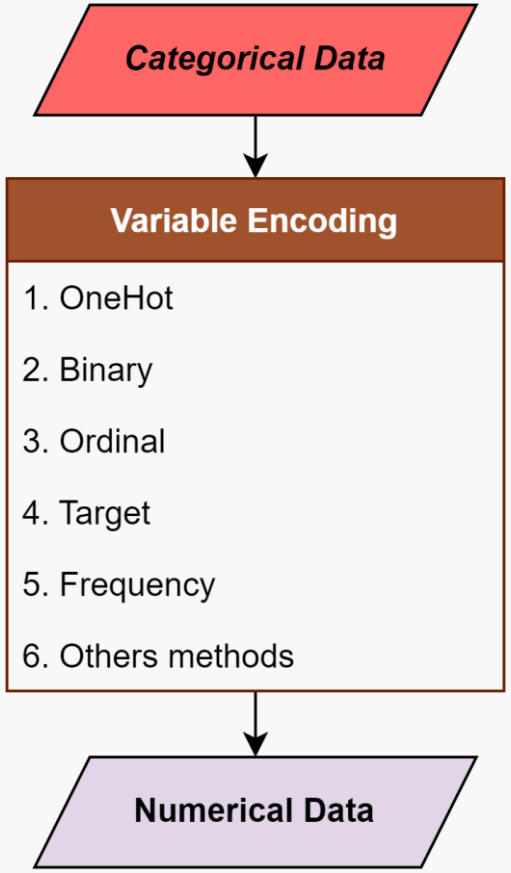
Supervised & Unsupervised Encoding Schemes of Binary Variables for Prediction Performance Enhancement

Institute of Industrial Engineering, NTU

Yun-Hao Yang, Jakey Blue

Motivation



Categorical	Numerical							
City	Ordinal	Binary		One Hot				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	0.4
Tokyo	3	1	1	1	0	0	0	0.4

Motivation

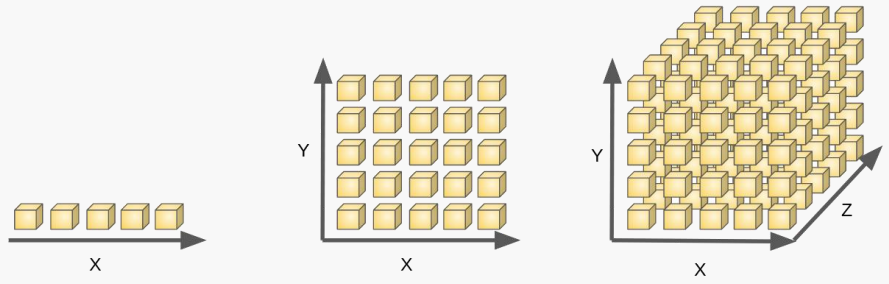
Categorical

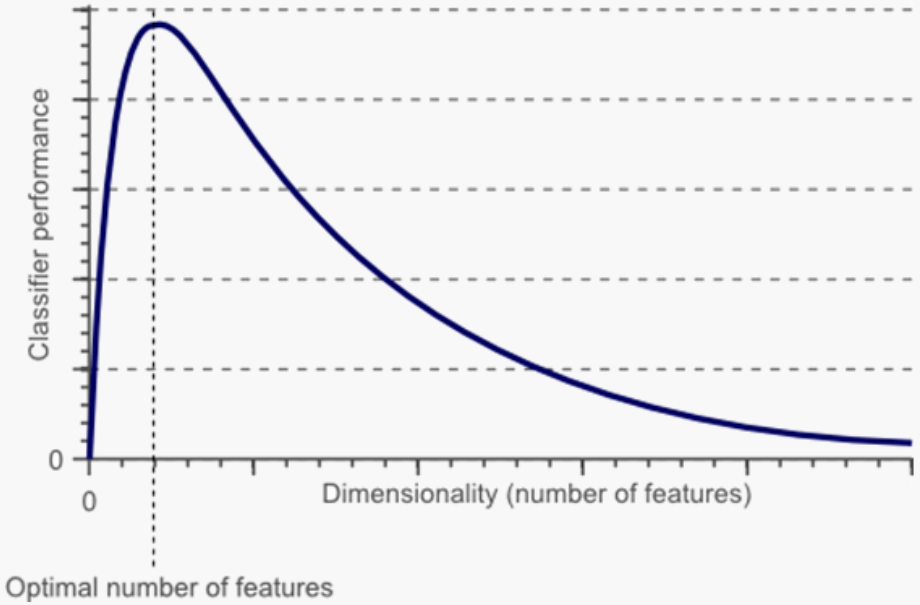
Animal
Cow
Tuna
⋮
Frog

One Hot

Binary

Animal							
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

One Hot

Numerical

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

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

Numerical

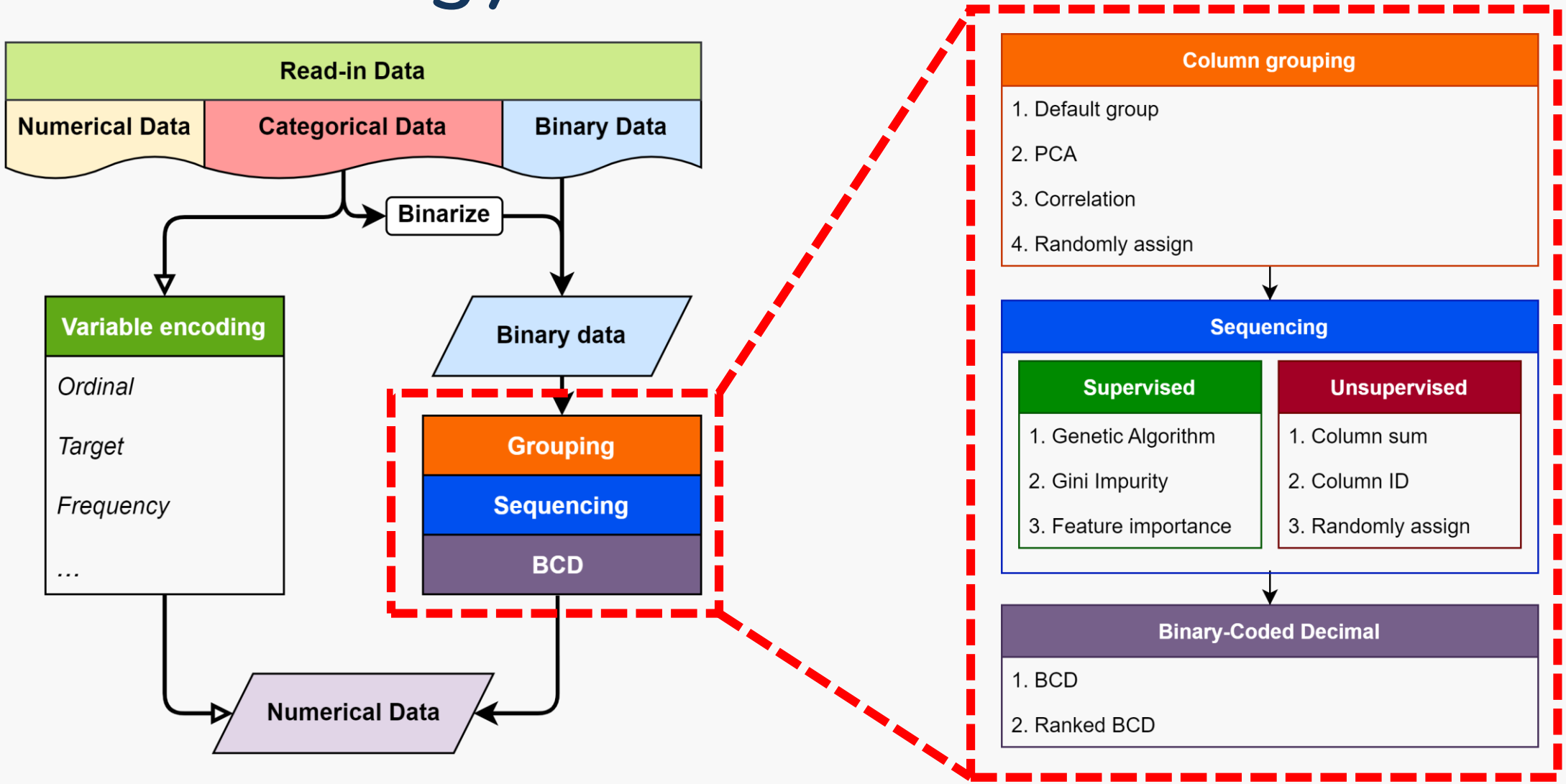
Group 1	Group 2	Group 3
10	4	7
2	3	6
⋮	⋮	⋮
7	1	3

When the data is not follow the One Hot rule, can't use other method to transform the binary data. If only can we find a way in representing the numerous binary features data in a much simple way…

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



Group
Sequence
BCD

Methodology

In the research, we propose a method to encode binary data in to numerical data. Via 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 the research, we propose a method to encode binary data in to numerical data. Via 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 the research, we propose a method to encode binary data in to numerical data. Via 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 the research, we propose a method to encode binary data in to numerical data. Via 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 the research, we propose a method to encode binary data in to numerical data. Via 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

Group
Sequence
BCD

Methodology - Grouping

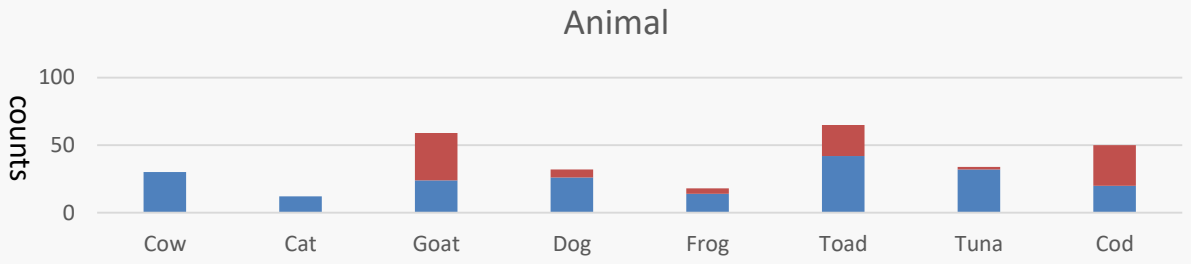
Animal							
Cow	Cat	Goat	Dog	Frog	Toad	Tuna	Cod
		V					

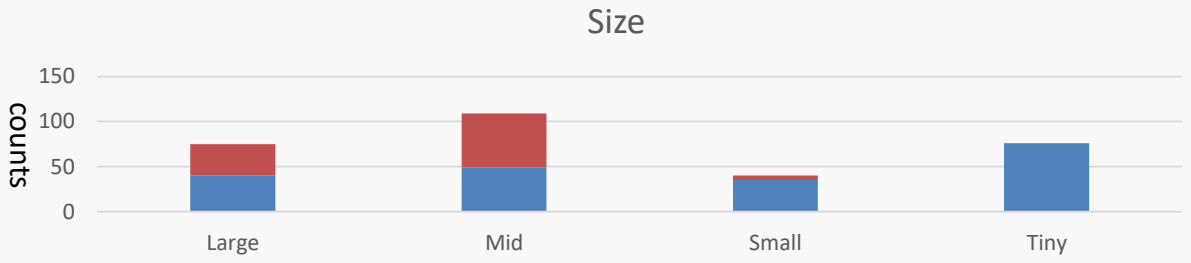
Size			
Large	Mid	Small	Tiny
	V		

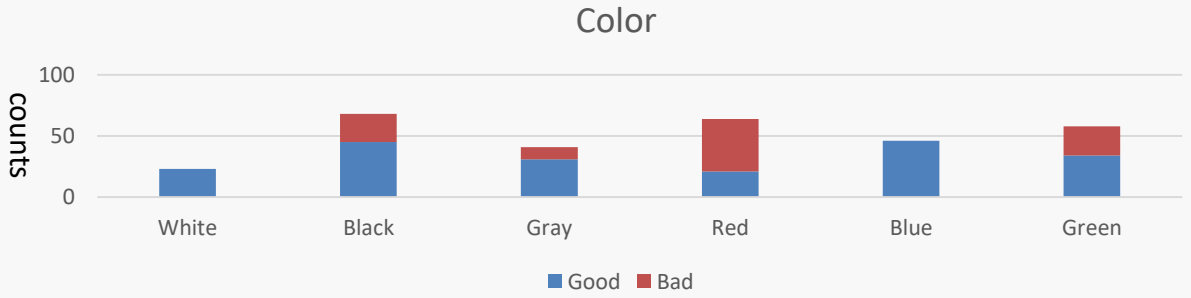
Color					
White	Black	Gray	Red	Blue	Green
	V				

Health Status	
Good	Bad
	V

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

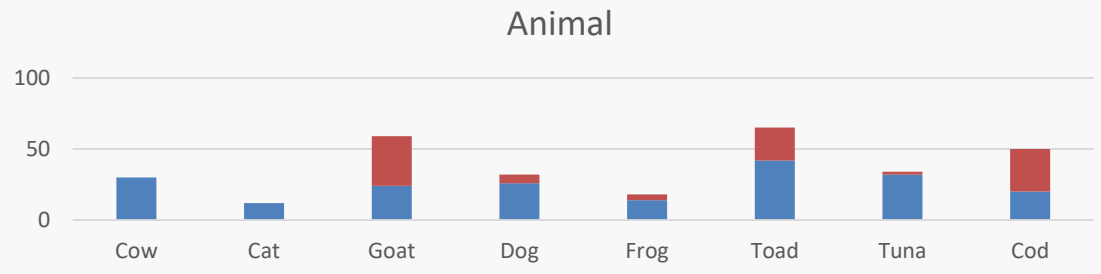


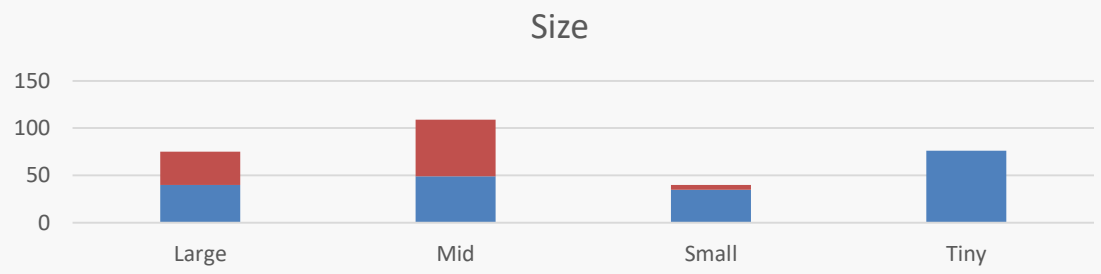


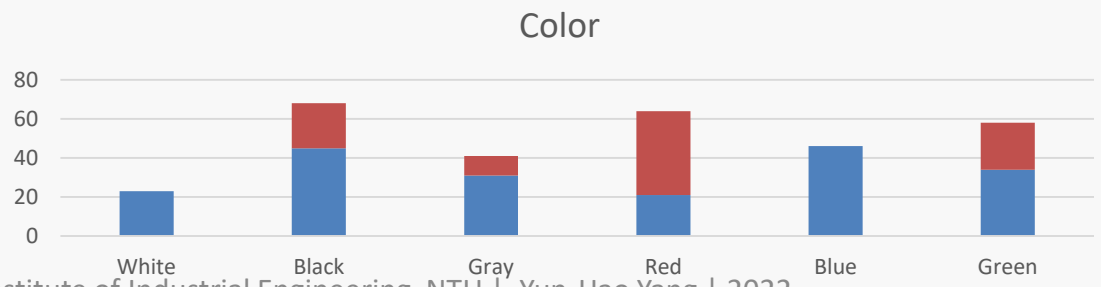


Methodology - Grouping

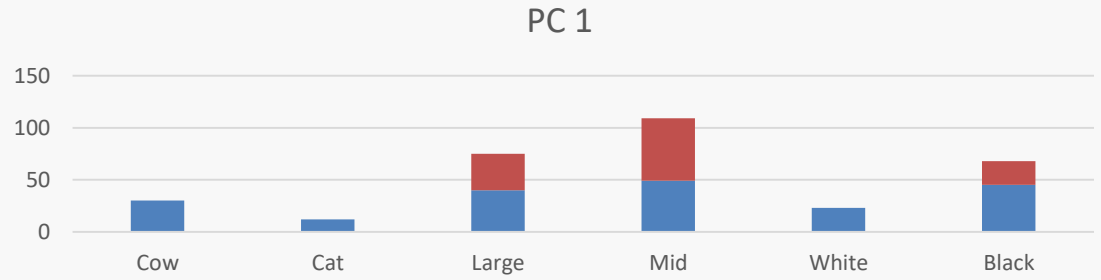
Binary Data

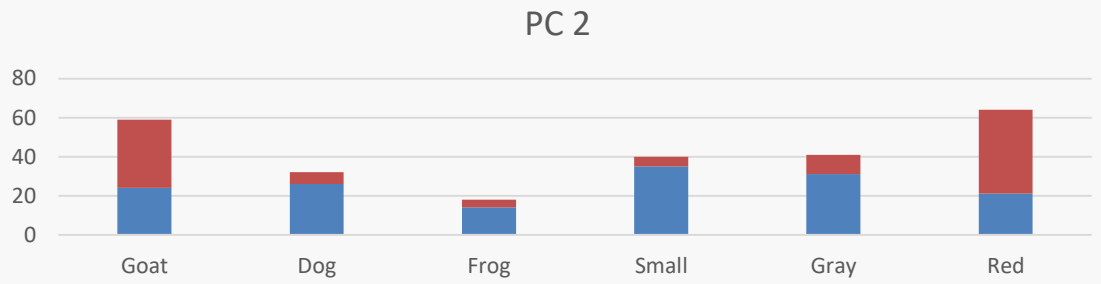


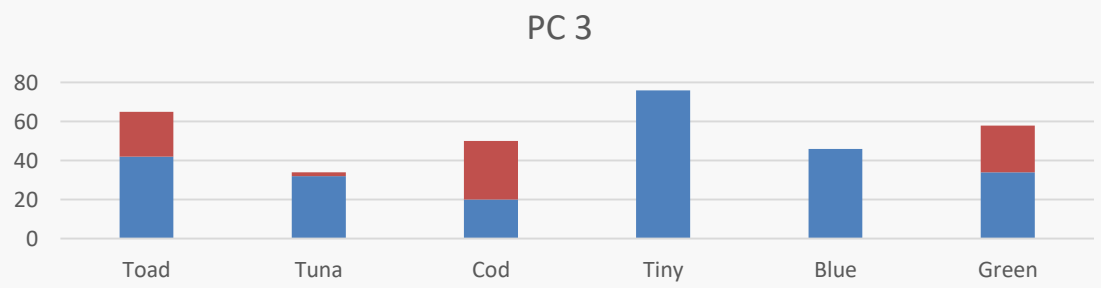




Grouped by PCA







Methodology - Sequencing

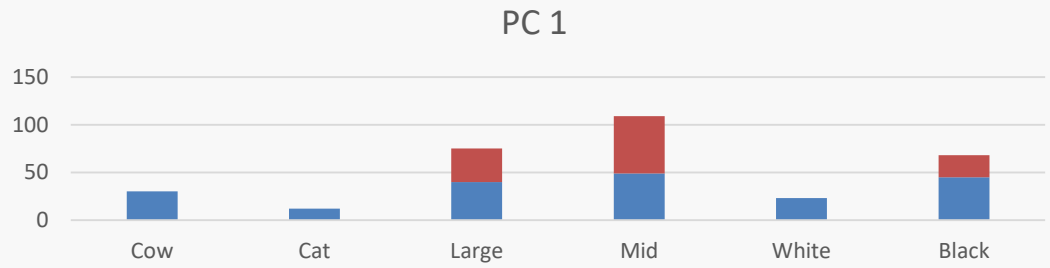
Secondly, sequencing features in each group by columns' attributes, for outputting better BCD values after encode.

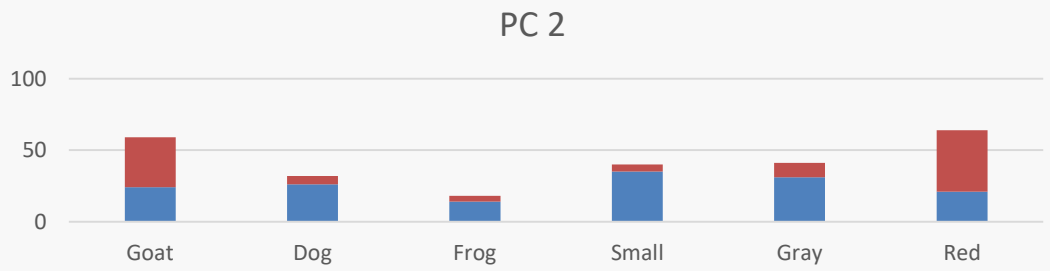
1. Column sum
2. Type impurity
3. Feature importance

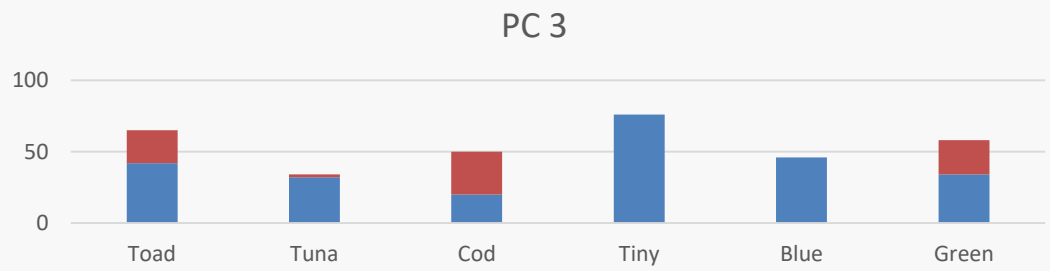
Grouped by PCA

Methodology - Sequencing

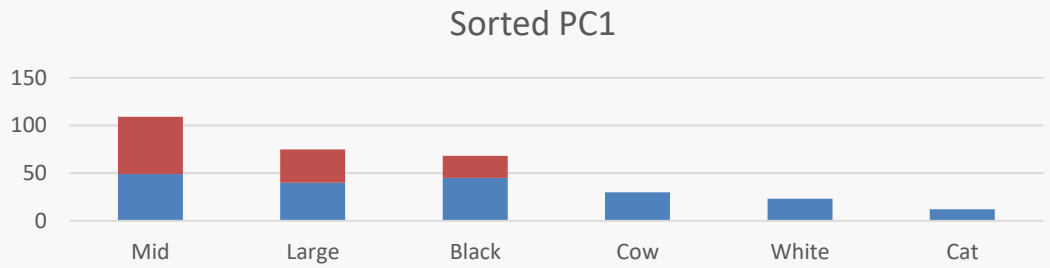
Grouped by PCA

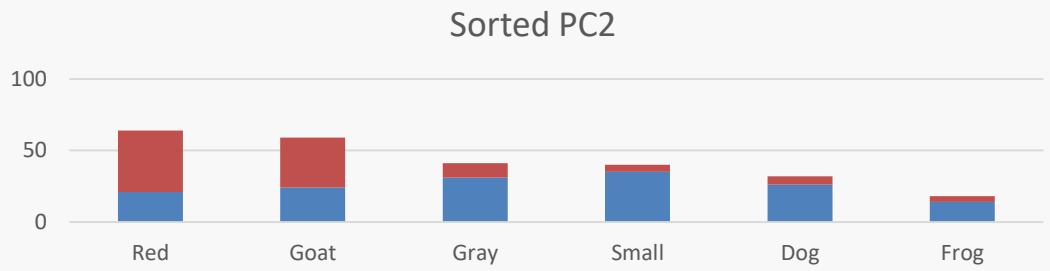


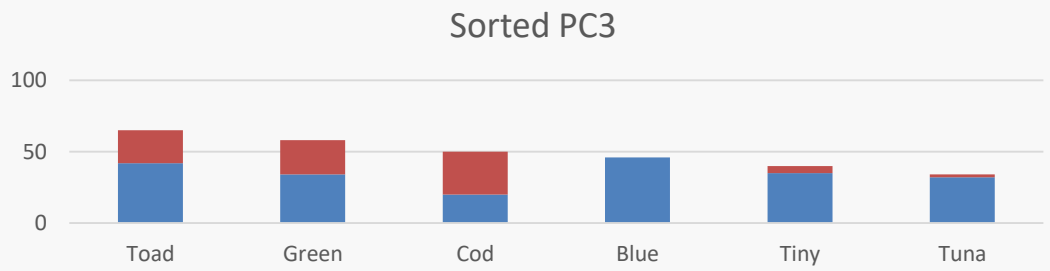




Sequenced by column sum

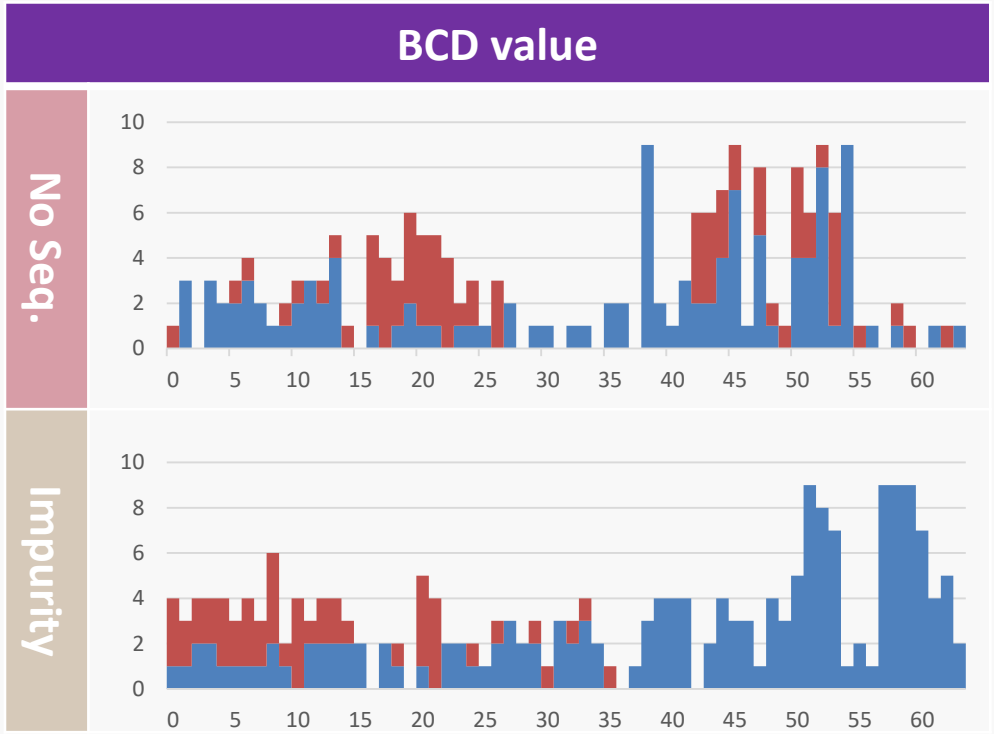






Group
Sequence
BCD

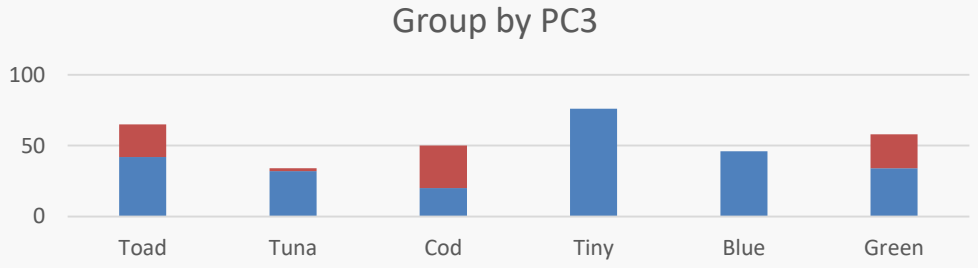
Methodology - Sequencing



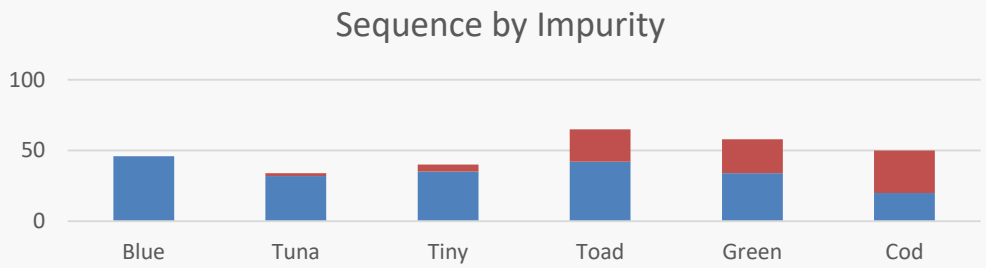
Group
Sequence
BCD

Methodology - BCD code

Group



Sequence



BCD	Sorted Column Group						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

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

Methodology - sum up

1. Column Grouping

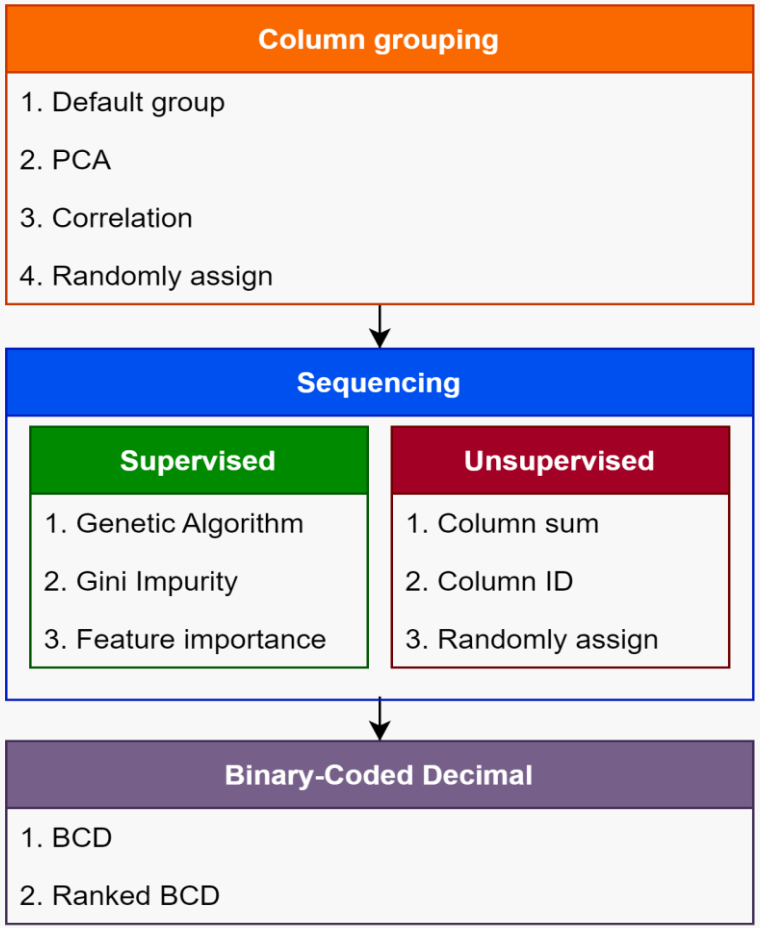
Clustering related or similar columns together into groups.

2. Sequencing

Rearrange columns in group, for a more distinctive /competitive BCD value.

3. Binary-Coded Decimal

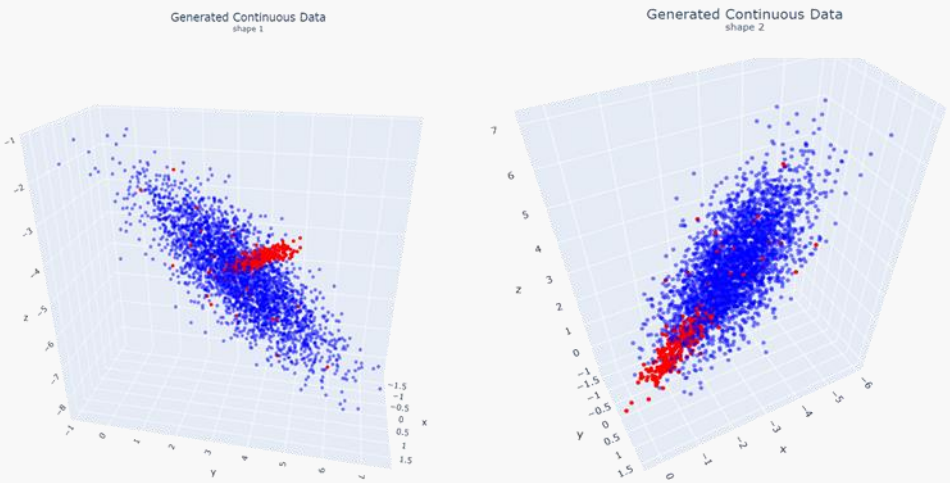
Transform column groups into a numerical value with BCD.



Case study

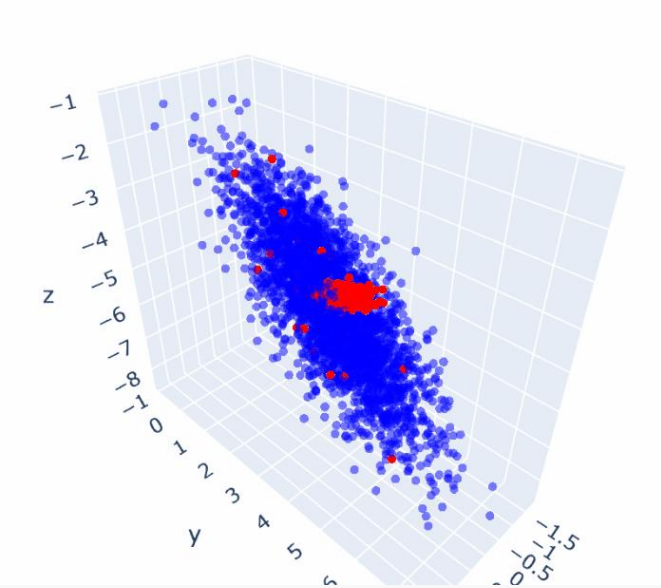
In Case study, we compare classification results with the commonly used variable encoding method under different datasets.

1. Simulated datasets
2. Kaggle dataset

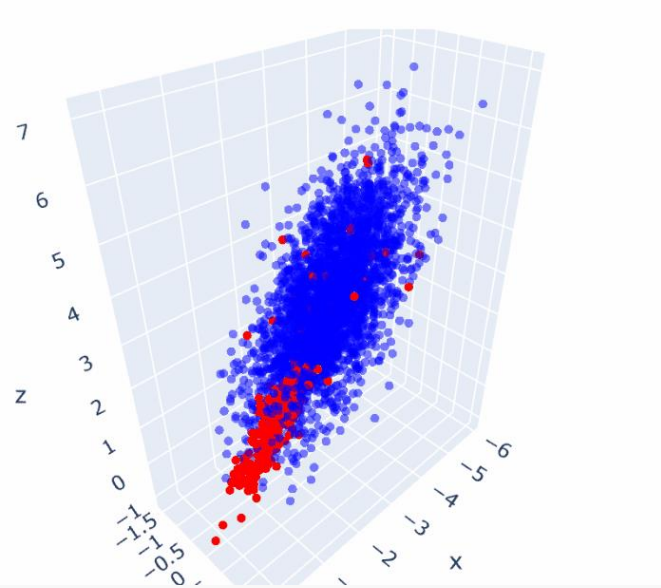


Simulated data (3300 samples, 3000 Good, 300 Bad)

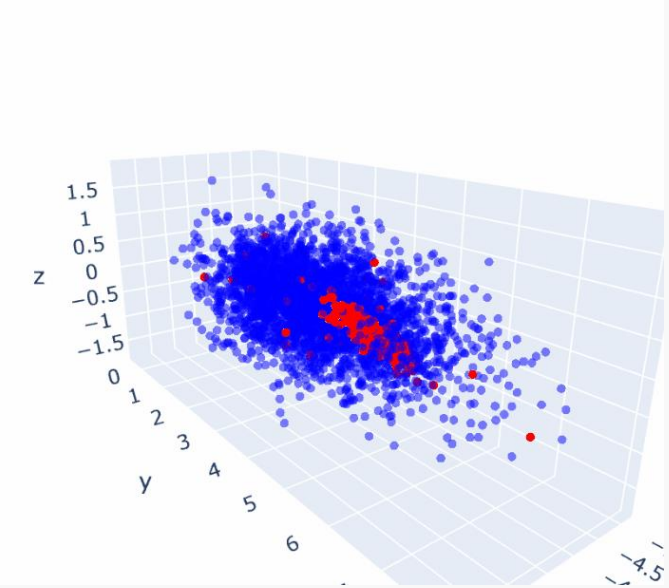
Dataset 1



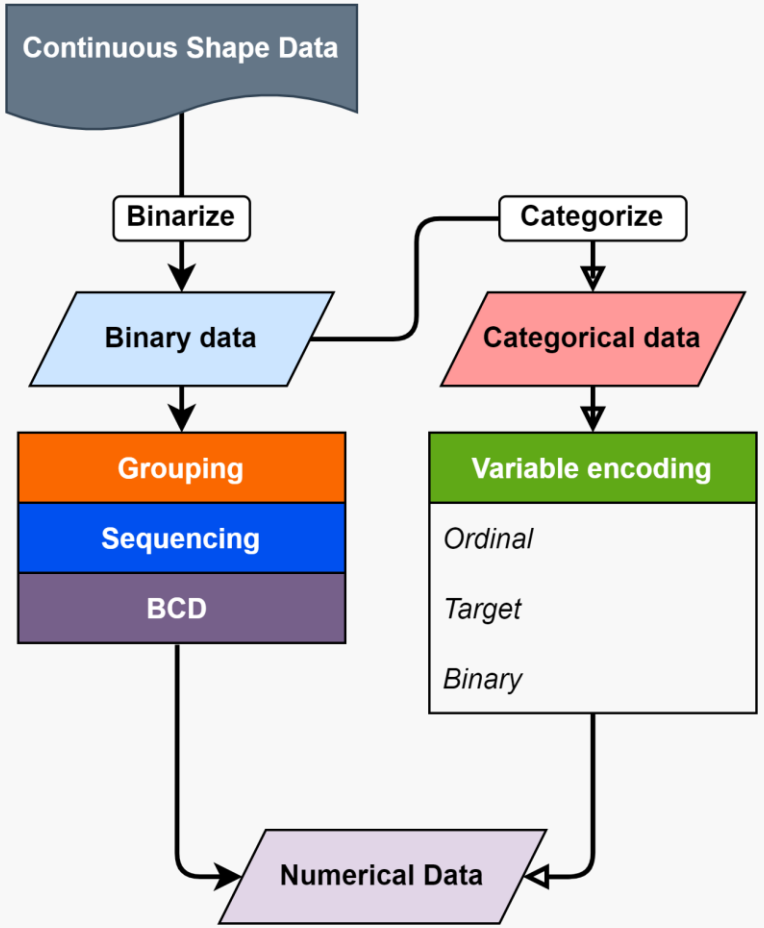
Dataset 2

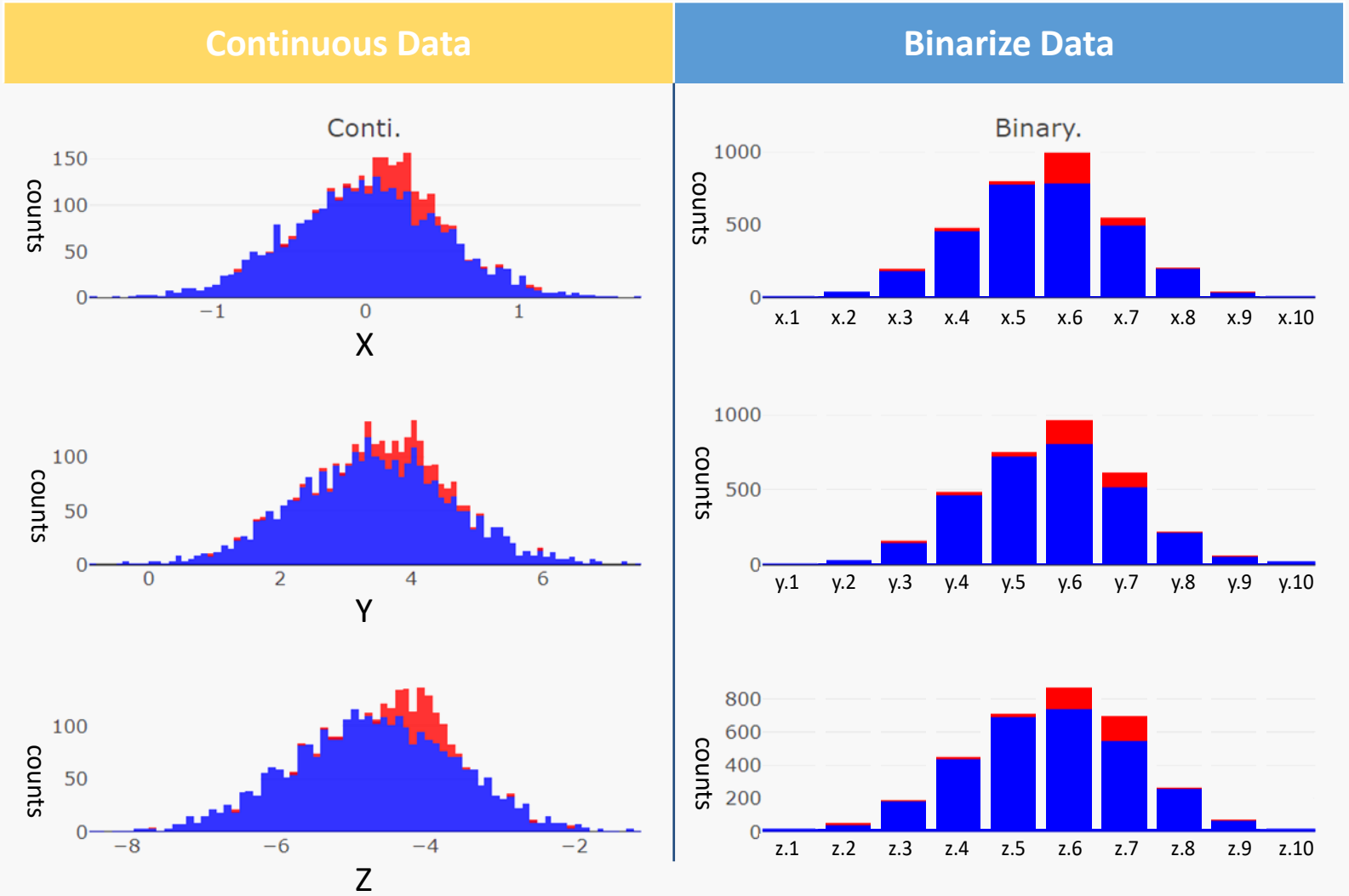


Dataset 3



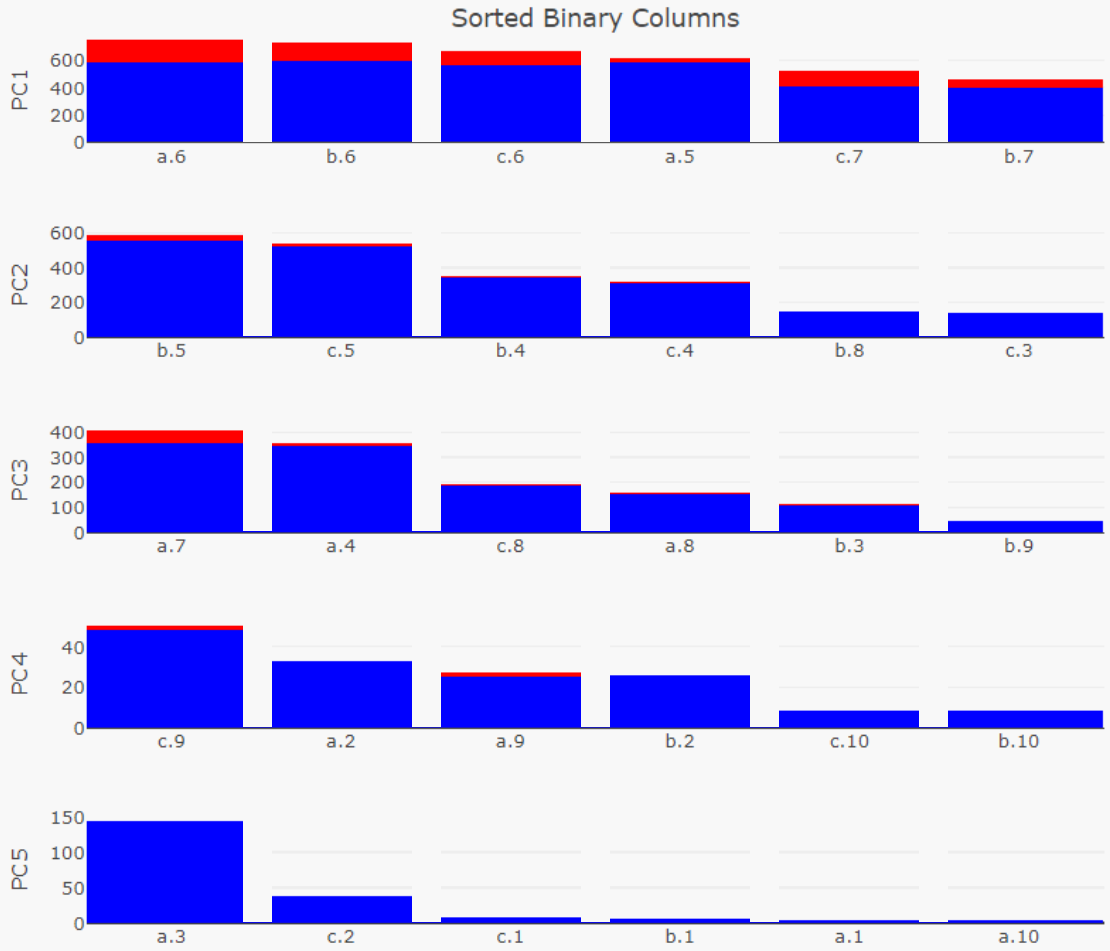
Simulated data - binarize

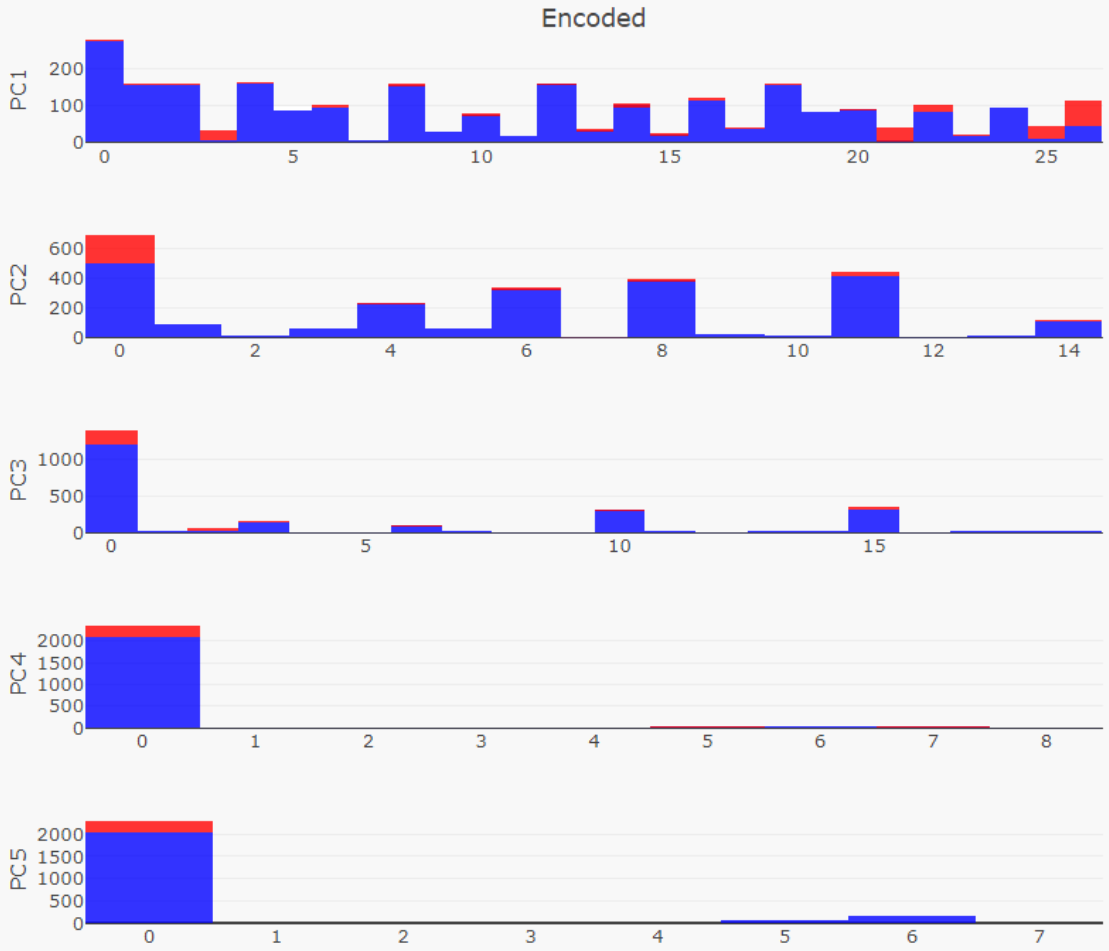




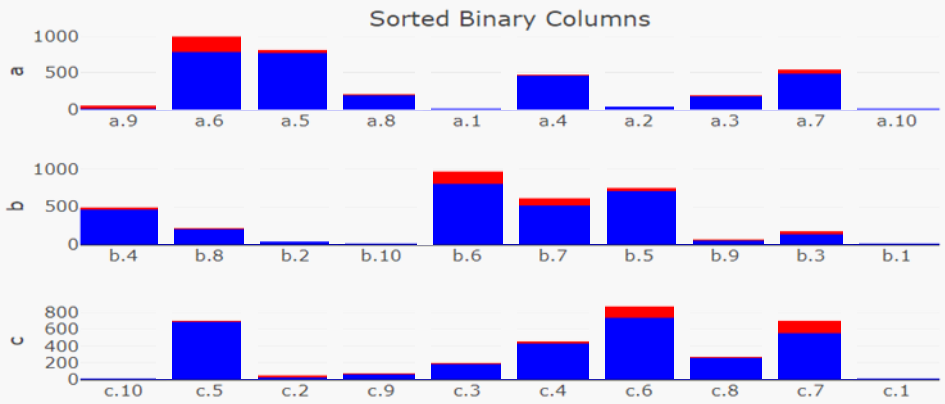
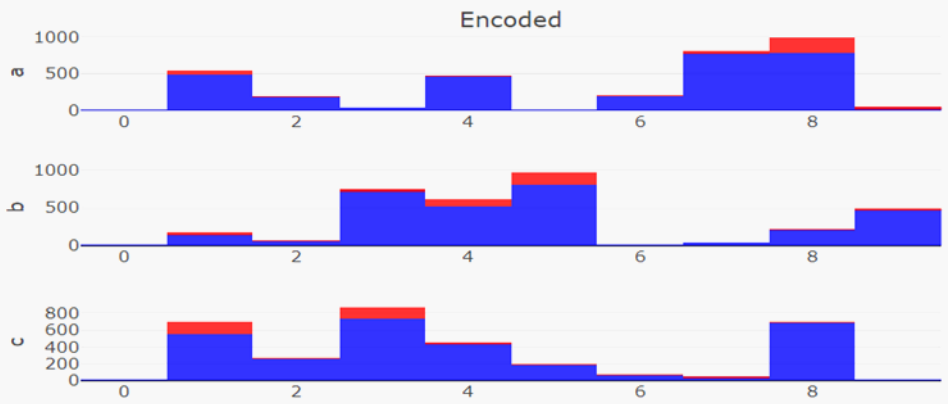
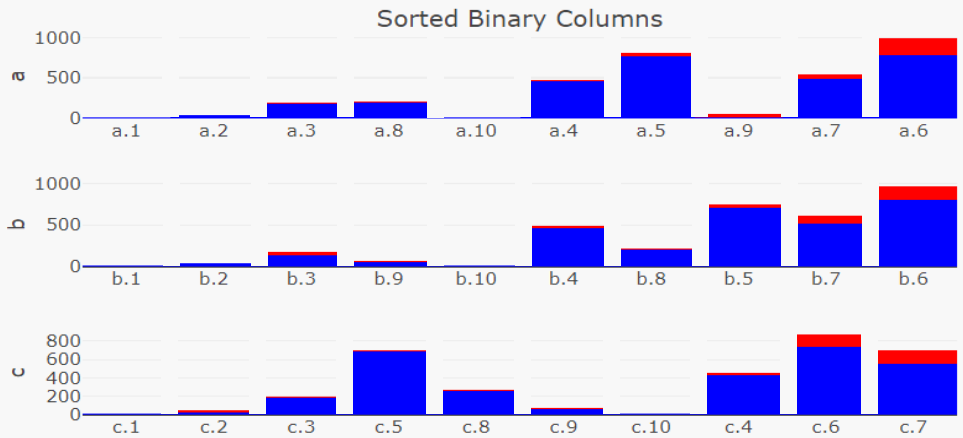
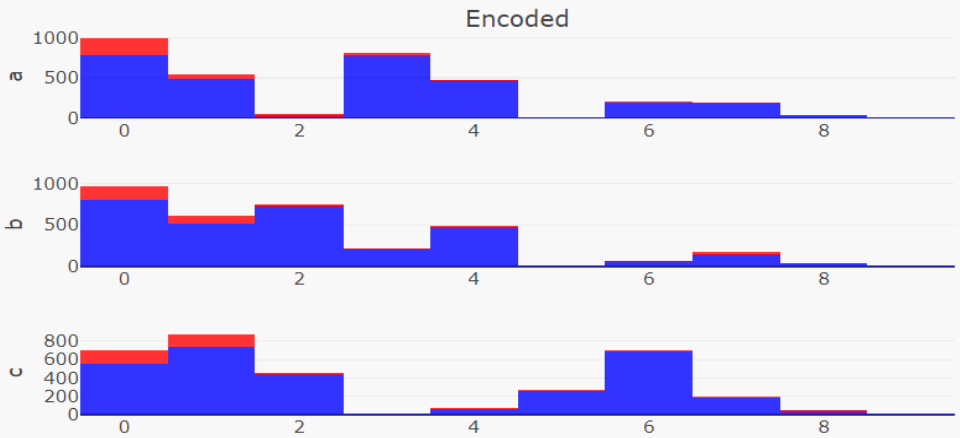
Simulated data - seq. by sum

Sequence by column sum	BCD



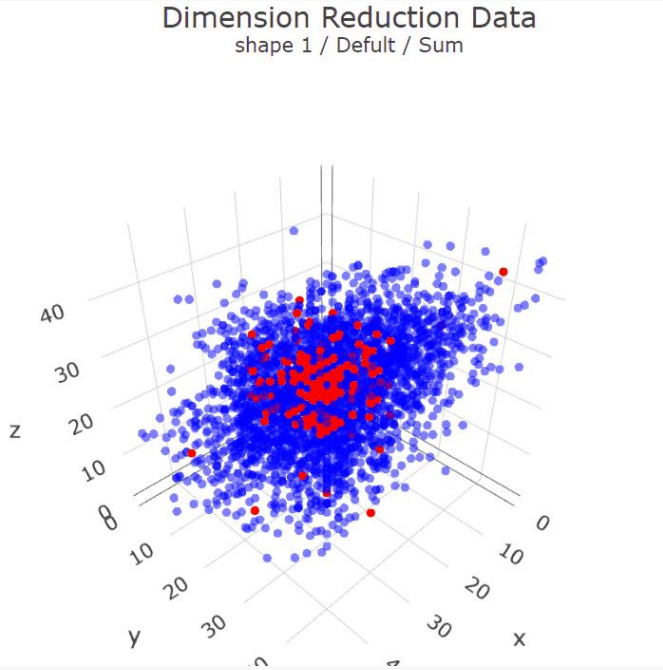


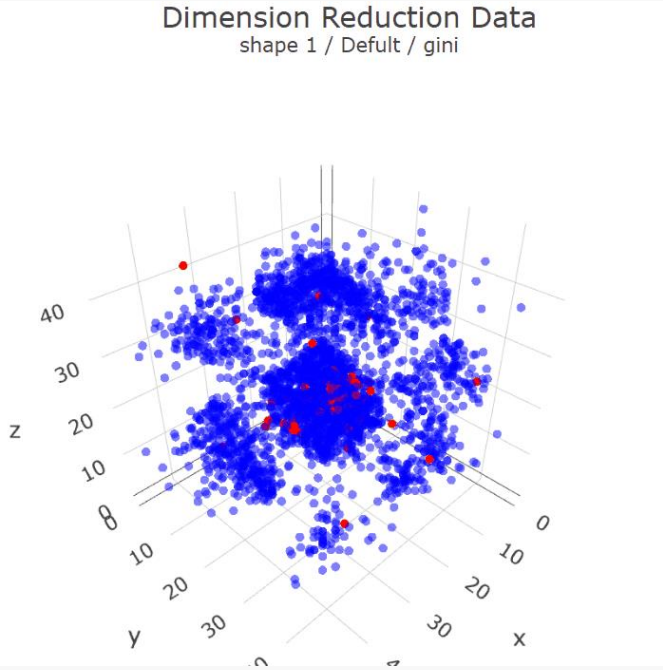
Simulated data - seq. comparison

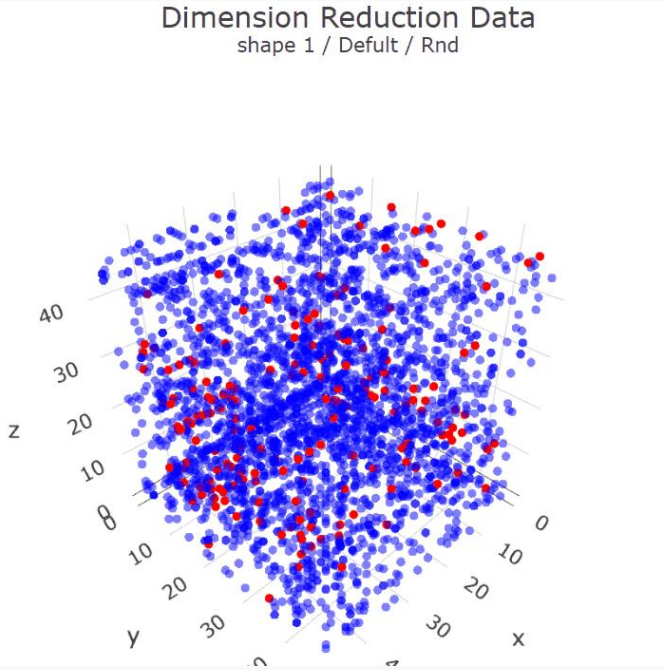
	Sequenced	BCD
Randomly		
By Impurity		

Simulated data - Dimension reduction

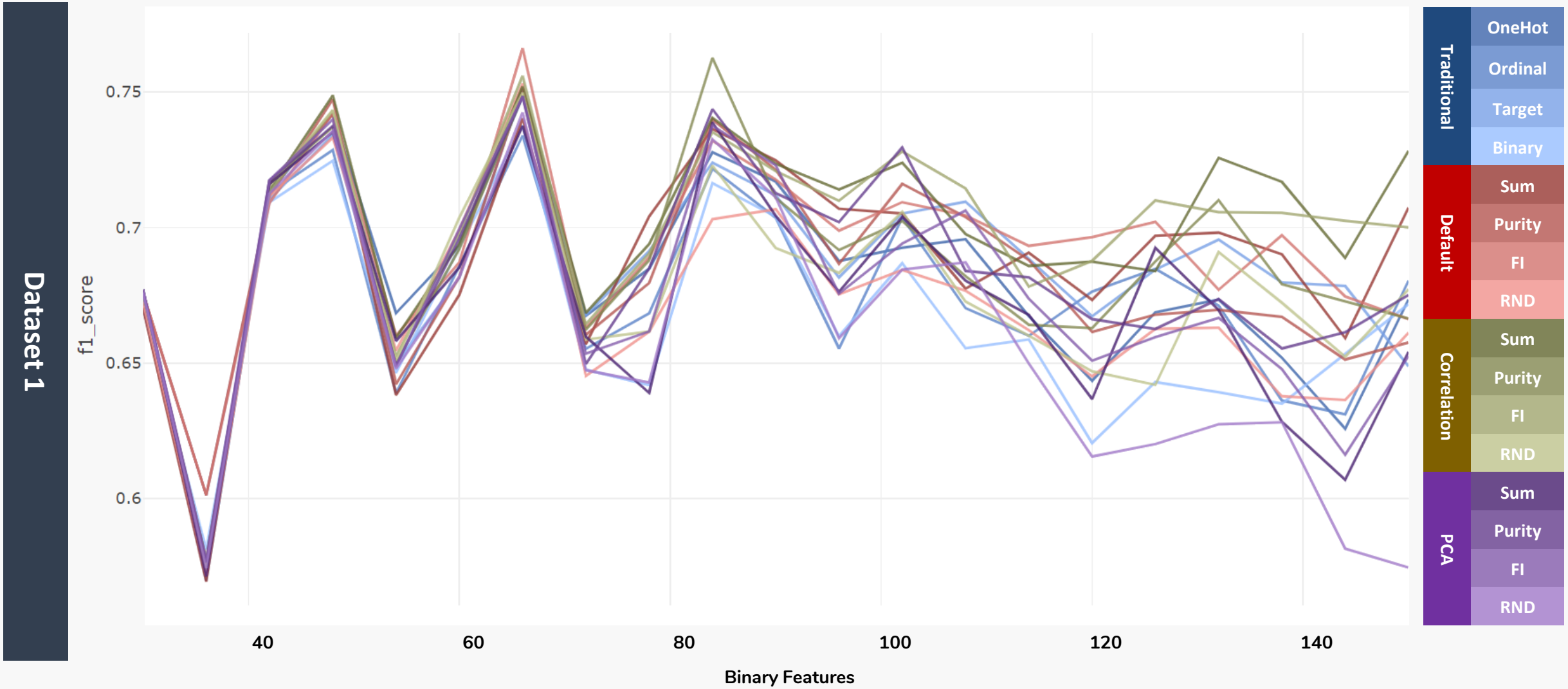
Sum	Impurity	RND



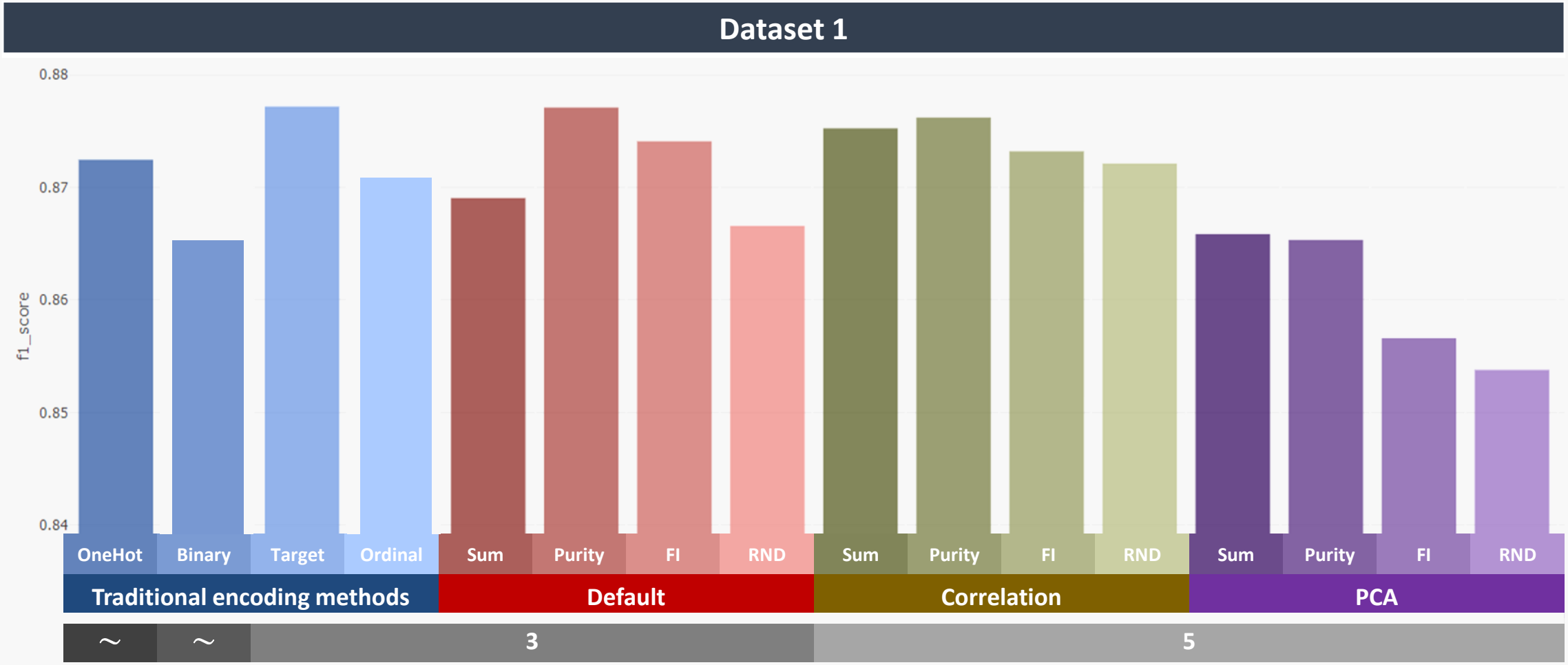




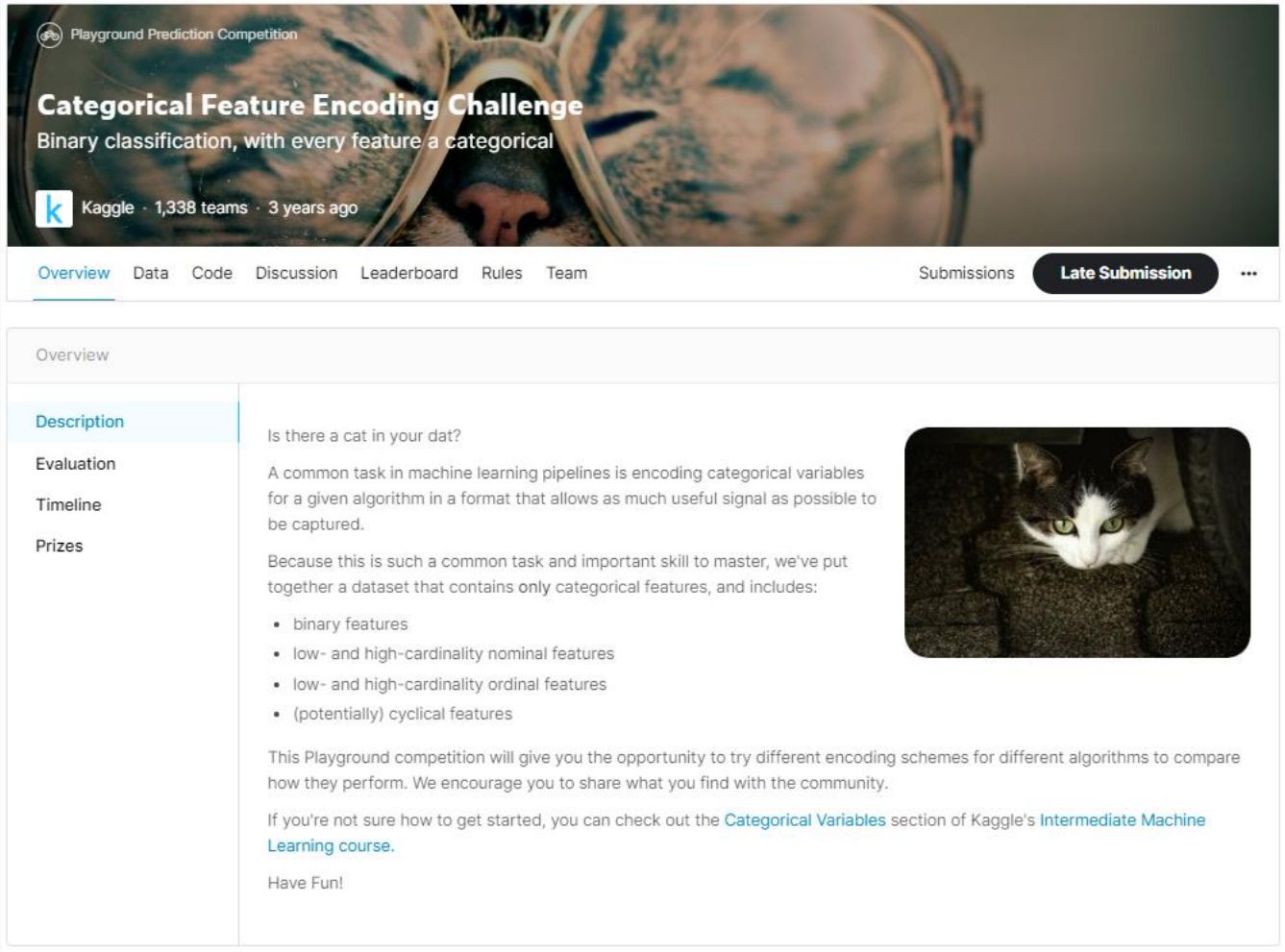
Continuous data - Classification

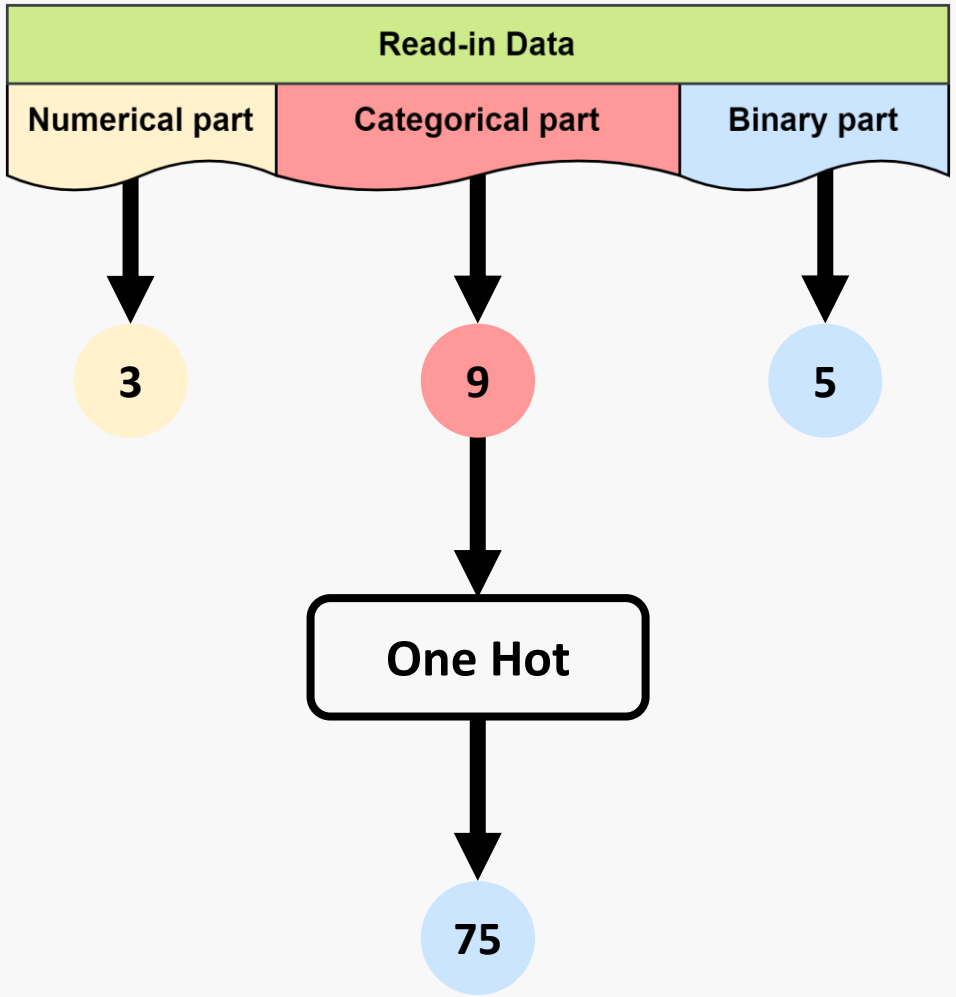


Simulated data - Classification

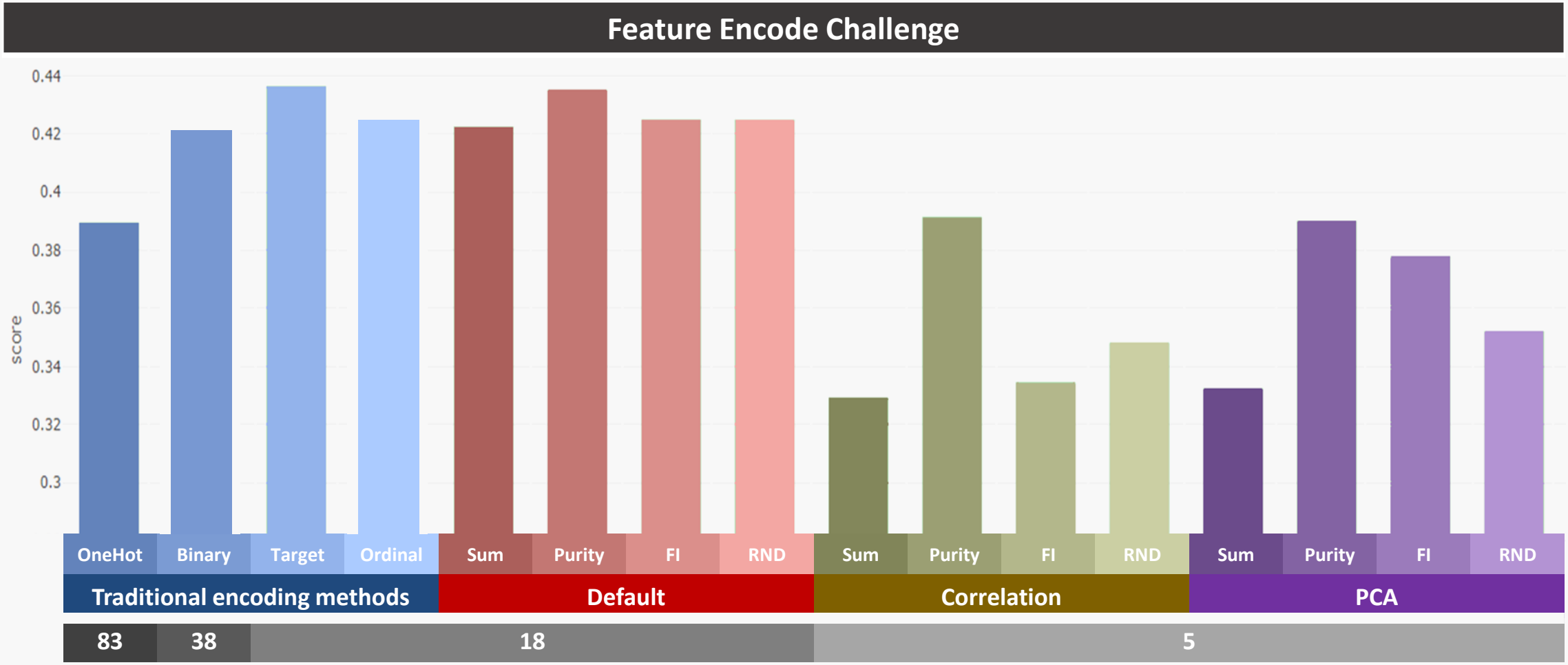


Kaggle dataset

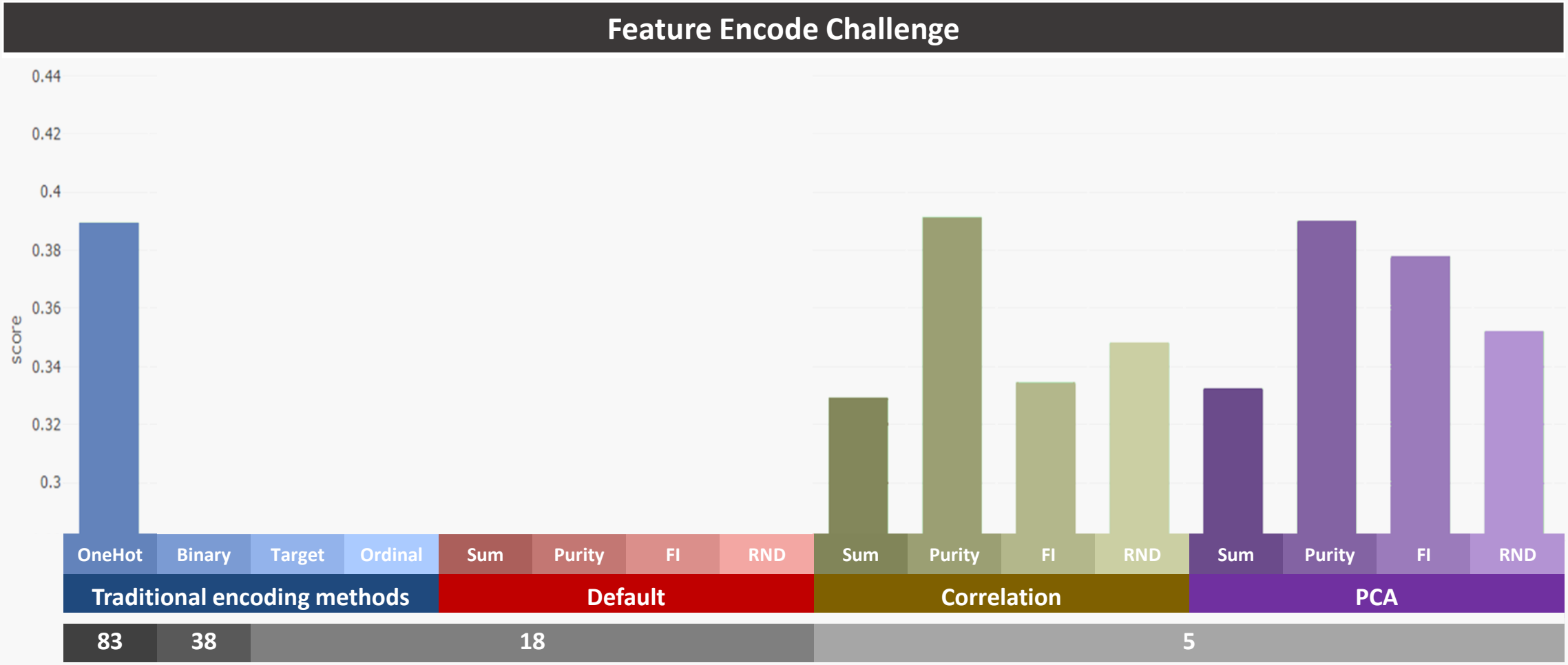




Kaggle dataset - Classification



Kaggle dataset - Classification



Conclusion

In this research, we develop a method to encode binary feature data into numerical data, which can compress the binary features information to reduce dimension while remaining a certain level in ML model classification performance. The main advantage of this method compare to traditional encoding methods are:

1. Encoding Binary features without group knowledge to numerical data.
2. Adjust the number of features of encoded numerical data.
3. Remain ML model performance with certain sequencing method.