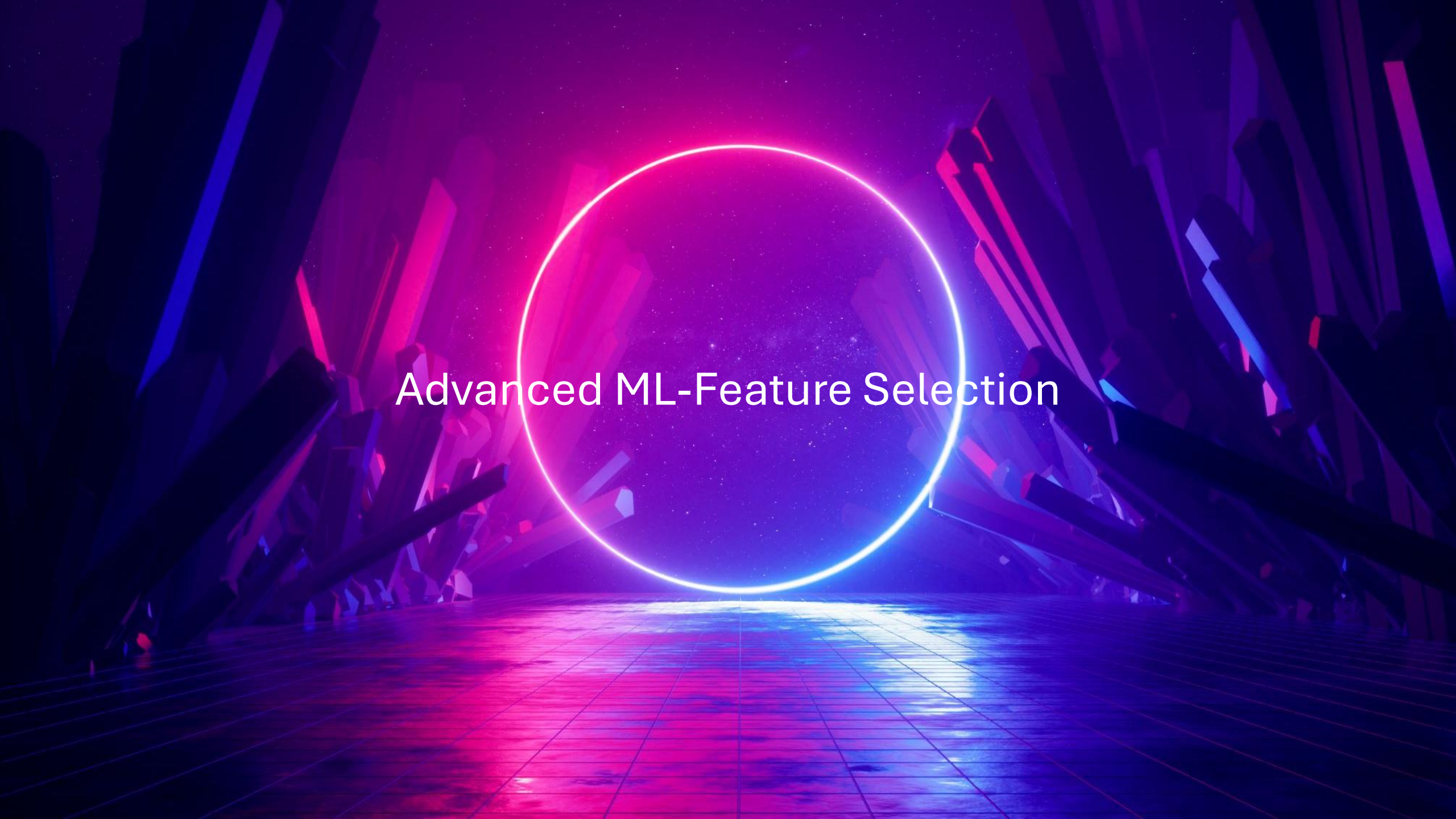
Advanced ML-Feature Selection

Why feature Selection?

Feature Selection Method Vs Dimensionality Reduction

Feature Selection Method

It selects the important feature among the input features variable.

Based on the user input it selects the number of feature

No change in the data value

Dimensionality Reduction

The input variable values are converted into cluster feature.

Based on the user input it selects the number of clusters

Change in the data value.

Feature Selection Method

It selects the important feature among the input features variable.

Based on the user input it selects the number of feature

No change in the data value

All the feature Variable

	I_1	I_2	I_3	I_4
	House	Sale price (100$)	Size (sqft)	Age (years)
	Avalon	2050	2650	13
	Cross Winds	2080	2600	*
	The White House	2150	2554	6
	The Rectory	2150	2921	3
	Larchwood	1999	2580	4
	Orchard House	1900	2580	4
	Shangri-La	1800	2774	2
	The Stables	1560	1920	1
	Cobweb Cottage	1450	2150	*
	Nairn House	1449	1710	1

After the feature selection method

Size (sqft)	Age (years)
2650	13
2600	*
2554	6
2921	3
2580	4
2580	4
2774	2
1920	1
2150	*
1710	1

All the feature Variable

I_1 I_2 I_3 I_4

Dimensionality Reduction

The input variable values are converted into cluster feature.

Based on the user input it selects the number of clusters

hange in the data value.

House	Sale price (100$)	Size (sqft)	Age (years)
Avalon	2050	2650	13
Cross Winds	2080	2600	*
The White House	2150	2554	6
The Rectory	2150	2921	3
Larchwood	1999	2580	4
Orchard House	1900	2580	4
Shangri-La	1800	2774	2
The Stables	1560	1920	1
Cobweb Cottage	1450	2150	*
Nairn House	1449	1710	1

Wpc1	Wpc2	Wpc3	Wpc4
7	26	6	60
1	29	15	52
11	56	8	20
11	31	8	47
7	52	6	33
11	55	9	22
3	71	17	6
1	31	22	44
2	54	18	22
21	47	4	26
1	40	23	34

After the PCA method

Feature Selection Method

- SelectK Algorithm → Chi Square, RMS
- Recursive Feature Elimination → Each Variable with output
- Feature Importance → Changing Each Variable Value ,changes in the output
- Forward Selection → Based on Correlation
- Backward Selection → Based on Correlation

Dimensionality Reduction

- Principle Component Analysis
- Linear Discriminant Analysis
- Kernel PCA

Dataset

	sl_no	gender	ssc_p	ssc_b	hsc_p	hsc_b	hsc_s	degree_p	degree_t	workex	etest_p	specialisation	mba_p	status	salary
0	1	M	67.00	Others	91.00	Others	Commerce	58.00	Sci&Tech	No	55.0	Mkt&HR	58.80	Placed	270000.0
1	2	M	79.33	Central	78.33	Others	Science	77.48	Sci&Tech	Yes	86.5	Mkt&Fin	66.28	Placed	200000.0
2	3	M	65.00	Central	68.00	Central	Arts	64.00	Comm&Mgmt	No	75.0	Mkt&Fin	57.80	Placed	250000.0
3	4	M	56.00	Central	52.00	Central	Science	52.00	Sci&Tech	No	66.0	Mkt&HR	59.43	Not Placed	NaN
4	5	M	85.80	Central	73.60	Central	Commerce	73.30	Comm&Mgmt	No	96.8	Mkt&Fin	55.50	Placed	425000.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
210	211	M	80.60	Others	82.00	Others	Commerce	77.60	Comm&Mgmt	No	91.0	Mkt&Fin	74.49	Placed	400000.0
211	212	M	58.00	Others	60.00	Others	Science	72.00	Sci&Tech	No	74.0	Mkt&Fin	53.62	Placed	275000.0
212	213	M	67.00	Others	67.00	Others	Commerce	73.00	Comm&Mgmt	Yes	59.0	Mkt&Fin	69.72	Placed	295000.0
213	214	F	74.00	Others	66.00	Others	Commerce	58.00	Comm&Mgmt	No	70.0	Mkt&HR	60.23	Placed	204000.0
214	215	M	62.00	Central	58.00	Others	Science	53.00	Comm&Mgmt	No	89.0	Mkt&HR	60.22	Not Placed	NaN

215 rows × 15 columns

ssc_p
67.00
79.33
65.00
56.00
85.80
...
80.60
58.00
67.00
74.00
62.00

salary
270000.0
200000.0
250000.0
NaN
425000.0
...
400000.0
275000.0
295000.0
204000.0
NaN

Chi Square= 0.8

SelectK Algorithm → Chi Square, RMS

ssc_p	hsc_p
67.00	91.00
79.33	78.33
65.00	68.00
56.00	52.00
85.80	73.60
...	...
80.60	82.00
58.00	60.00
67.00	67.00
74.00	66.00
62.00	58.00

salary
270000.0
200000.0
250000.0
NaN
425000.0
...
400000.0
275000.0
295000.0
204000.0
NaN

Chi Square= 0.85

SelectK Algorithm → Chi Square, RMS

ssc_p	hsc_p	degree_p	salary
67.00	91.00	58.00	270000.0
79.33	78.33	77.48	200000.0
65.00	68.00	64.00	250000.0
56.00	52.00	52.00	NaN
85.80	73.60	73.30	425000.0
...	...	...	...
80.60	82.00	77.60	400000.0
58.00	60.00	72.00	275000.0
67.00	67.00	73.00	295000.0
74.00	66.00	58.00	204000.0
62.00	58.00	53.00	NaN

K = N.F column

Chi Square=0.75

Less K
↑ Chi Square

Select K Best

K=3

Classification

result - DataFrame

Index	Logistic	SVMl	SVMnl	KNN	Navie	Decision	Random
ChiSquare	0.94	0.94	0.95	0.89	0.83	0.96	0.95

Regression

result - DataFrame

Index	Linear	SVMl	SVMnl	Decision	Random
ChiSquare	0.632365	0.597382	0.934341	0.565972	0.895399

#learnaiwithramisha

ssc_p
67.00
79.33
65.00
56.00
85.80
...
80.60
58.00
67.00
74.00
62.00

salary
270000.0
200000.0
250000.0
NaN
425000.0
...
400000.0
275000.0
295000.0
204000.0
NaN

machine learning algorithm 1=0.78
machine learning algorithm 2=0.80

ssc_p	hsc_p	salary
67.00	91.00	270000.0
79.33	78.33	200000.0
65.00	68.00	250000.0
56.00	52.00	NaN
85.80	73.60	425000.0
...	...	...
80.60	82.00	400000.0
58.00	60.00	275000.0
67.00	67.00	295000.0
74.00	66.00	204000.0
62.00	58.00	NaN

machine learning algorithm 1=0.78
machine learning algorithm 2=0.80

Classification

Index	Logistic	SVMl	SVMnl	KNN	Navie	Decision	Random
Logistic	0.94	0.94	0.94	0.94	0.94	0.94	0.94
SVC	0.87	0.87	0.87	0.87	0.87	0.87	0.87
Random	0.94	0.94	0.94	0.94	0.9	0.91	0.92
DecisionTree	0.97	0.97	0.97	0.96	0.84	0.96	0.97

Recursive Feature Elimination

Each Variable with output

- Regression

result - DataFrame

Index	Linear	SVMI	Decision	Random
Linear	0.441961	0.262153	0.441961	0.441816
SVC	0.441961	0.262153	0.441961	0.441816
Random	0.664893	0.609652	0.965961	0.916304
DecisionTree	0.676174	0.670691	0.933504	0.887256