LAPORAN TUGAS 1 DATA MINING

PREPROCESSING DATA

Kelompok 2:

1. Chlaudiah Julinar 1301150434
2. Ridea V. P. Siwabesy 1301154458
3. Septiana Putri 1301154463
4. Nadya Aditama 1301154556
5. Windias Friliani 1301154570

INFORMASI DATA

Sumber Data

Data set yang digunakan yaitu **irish.csv** . Dataset diambil dari salah satu situs penyedia data set yaitu OpenML dengan link menuju situs tersebut adalah www.openml.org , dimana link dataset irish yaitu www.openml.org/d/451 .

Deskripsi Data

Berikut deskripsi dataset yang ada di site OpenML dengan jumlah *record* sebanyak 500 *record*

dan jumlah atribut sebanyak 6 atribut.

Keterangan Atribut

Atribut irish dataset adalah sebanyak 6 atribut, dengan tiap atribut memiliki tipe atribut yang berbeda-beda yaitu sebagai berikut.

In [1]:
```
import pandas as pd
raw_atribut = {'Nama Atribut': ['Sex','DVRT','Educational Level','Leaving Certificate','Prestige Score', 'Type School'],
           'Tipe Atribut': ['Diskrit','Diskrit','Ordinal','Diskrit','Kontinyu','Ordinal']}
atribut= pd.DataFrame(raw_atribut,
                      columns=['Nama Atribut','Tipe Atribut'])
atribut
```

Out[1]:

	Nama Atribut	Tipe Atribut
0	Sex	Diskrit
1	DVRT	Diskrit
2	Educational Level	Ordinal
3	Leaving Certificate	Diskrit
4	Prestige Score	Kontinyu
5	Type School	Ordinal

Masalah Data

Masalah yang terjadi pada irish dataset adalah terdapatnya *missing values* sebanyak 32 *missing values* seperti yang tertera pada situs dengan rincian yaitu terdapat 26 *missing values* pada

atribut **Prestige Score** dan 6 *missing values* pada atribut **Educational Level**

Solusi Penyelesaian Masalah

Dapat dilihat bahwa *missing values* terjadi pada 2 atribut yang memiliki tipe atribut berbeda yaitu **Prestige Score** bertipe **Kontinyu** dan **Educational Level** bertipe **Ordinal**. Sehingga, kelompok kami mengusulkan tiga skenario yang dapat digunakan sebagai solusi penyelesaian masalah pada irish dataset yaitu:

In [2]:
```
raw_skenario = {'Nama Skenario': ['Skenario 1', 'Skenario 2', 'Skenario
 3'],
           'Prestige Score': ['Mean','Modus','Mean'],
           'Educational Level': ['Modus','Modus','Ignore']}
skenario= pd.DataFrame(raw_skenario,
                       columns=['Nama Skenario','Prestige Score','Educa
tional Level'])
skenario
```

Out[2]:

	Nama Skenario	Prestige Score	Educational Level
0	Skenario 1	Mean	Modus
1	Skenario 2	Modus	Modus
2	Skenario 3	Mean	Ignore

Diharapkan, dengan menggunakan 3 skenario yang berbeda untuk penanganan *missing values* pada tiap atribut maka kami dapat menentukan skenario terbaik yang dapat digunakan dengan tahap *preprocessing* yang sama untuk tiap skenario-nya.

IMPLEMENTASI

Read Data From CSV

Data dibaca dari file **irish.csv** menggunakan library python yaitu **pandas** biasa ditulis dengan **pd**.

In [3]:
```
file = pd.read_csv('Irish.csv')
file
```

Out[3]:

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	
0	male	113	Junior_cycle_incomplete-secondary_school	not_taken	28.0	seco
1	male	101	Primary_terminal_leaver	not_taken	28.0	prim
2	male	110	Senior_cycle_terminal_leaver-secondary_school	taken	69.0	seco
3	male	121	Junior_cycle_terminal_leaver-secondary_school	not_taken	57.0	seco
4	male	82	Junior_cycle_terminal_leaver-vocational_school	not_taken	18.0	voca
5	male	85	Junior_cycle_terminal_leaver-vocational_school	not_taken	28.0	voca
6	male	84	Primary_terminal_leaver	not_taken	NaN	prim
7	male	98	Junior_cycle_incomplete-vocational_school	not_taken	43.0	voca
8	male	92	Junior_cycle_terminal_leaver-vocational_school	not_taken	33.0	voca
9	male	90	Primary_terminal_leaver	not_taken	18.0	prim

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	
10	male	88	Junior_cycle_terminal_leaver-vocational_school	not_taken	28.0	voca
11	male	84	Junior_cycle_terminal_leaver-vocational_school	not_taken	18.0	voca
12	male	114	3rd_level_complete	taken	18.0	seco
13	male	70	Junior_cycle_terminal_leaver-vocational_school	not_taken	57.0	voca
14	male	109	Senior_cycle_terminal_leaver-secondary_school	taken	40.0	seco
15	male	83	Junior_cycle_terminal_leaver-vocational_school	not_taken	43.0	voca
16	male	108	Junior_cycle_terminal_leaver-vocational_school	not_taken	69.0	voca
17	male	95	Junior_cycle_incomplete-secondary_school	not_taken	43.0	seco
18	male	90	Junior_cycle_terminal_leaver-vocational_school	not_taken	28.0	voca
19	male	107	Senior_cycle_terminal_leaver-secondary_school	taken	28.0	seco
20	male	86	Primary_terminal_leaver	not_taken	18.0	prim
21	male	70	Junior_cycle_incomplete-vocational_school	not_taken	43.0	voca
22	male	114	Senior_cycle_terminal_leaver-secondary_school	taken	NaN	seco
23	male	117	Senior_cycle_terminal_leaver-secondary_school	taken	42.0	seco

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	
24	male	112	Senior_cycle_terminal_leaver-secondary_school	taken	43.0	seco
25	male	88	3rd_level_complete	taken	NaN	seco
26	male	106	3rd_level_complete	taken	42.0	seco
27	male	125	3rd_level_complete	taken	28.0	seco
28	male	94	Senior_cycle_terminal_leaver-secondary_school	taken	NaN	seco
29	male	103	Junior_cycle_terminal_leaver-secondary_school	not_taken	NaN	seco
...	...	...	...	...	...	...
470	male	129	Junior_cycle_terminal_leaver-secondary_school	not_taken	18.0	seco
471	male	122	Senior_cycle_terminal_leaver-secondary_school	taken	62.0	seco
472	male	121	3rd_level_incomplete	taken	37.0	seco
473	male	129	3rd_level_incomplete	taken	NaN	seco
474	male	122	Senior_cycle_terminal_leaver-secondary_school	taken	40.0	seco
475	male	126	Junior_cycle_terminal_leaver-vocational_school	not_taken	51.0	voca
476	male	122	3rd_level_complete	taken	35.0	seco
477	male	123	Senior_cycle_terminal_leaver-secondary_school	taken	65.0	seco
478	male	119	3rd_level_complete	taken	71.0	seco

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	
479	female	120	3rd_level_complete	taken	40.0	sec
480	female	127	Junior_cycle_incomplete-vocational_school	not_taken	35.0	voc
481	female	127	Senior_cycle_terminal_leaver-secondary_school	taken	62.0	sec
482	female	120	Senior_cycle_terminal_leaver-secondary_school	taken	61.0	sec
483	female	127	Senior_cycle_terminal_leaver-secondary_school	taken	58.0	sec
484	female	123	Senior_cycle_terminal_leaver-secondary_school	taken	37.0	sec
485	female	120	3rd_level_complete	taken	37.0	sec
486	female	123	Senior_cycle_terminal_leaver-secondary_school	taken	37.0	sec
487	female	122	Junior_cycle_terminal_leaver-vocational_school	not_taken	18.0	voc
488	female	119	3rd_level_complete	taken	37.0	sec
489	male	130	Senior_cycle_incomplete-vocational_school	not_taken	NaN	voc
490	male	134	3rd_level_incomplete	taken	62.0	sec
491	male	136	3rd_level_complete	taken	61.0	sec
492	male	135	3rd_level_complete	taken	61.0	sec
493	male	140	3rd_level_complete	taken	71.0	sec
494	male	131	Senior_cycle_terminal_leaver-secondary_school	taken	30.0	sec

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	
495	male	137	3rd_level_complete	taken	62.0	seco
496	male	136	3rd_level_complete	taken	18.0	seco
497	male	132	3rd_level_complete	taken	37.0	seco
498	female	135	3rd_level_complete	taken	62.0	seco
499	female	134	3rd_level_complete	taken	NaN	seco

500 rows × 6 columns

Setelah dataset telah dibaca kedalam program, selanjutnya adalah mengecek tipe tiap atribut yang terdefinisi oleh library **pd**

In [4]:
```
file.dtypes
```

Out[4]:
```
Sex                    object
DVRT                    int64
Educational_level      object
Leaving_Certificate    object
Prestige_score        float64
Type_school            object
dtype: object
```

DATA CLEANING

Skenario 1

Data cleaning pada *missing values* atribut dengan tipe kontinyu menggunakan pendekatan *mean*

- Melakukan pengecekan *missing values* atribut dengan tipe kontinyu

In [5]:
```
file1_float = file.select_dtypes(include=['float64']).copy()
file1_float[file1_float.isnull().any(axis=1)]
```

Out[5]:

	Prestige_score
6	NaN
22	NaN
25	NaN
28	NaN
29	NaN
35	NaN
66	NaN
89	NaN
111	NaN
112	NaN
137	NaN
147	NaN
198	NaN
217	NaN
219	NaN
291	NaN
300	NaN
301	NaN

	Prestige_score
310	NaN
341	NaN
347	NaN
348	NaN
349	NaN
473	NaN
489	NaN
499	NaN

- Mengisi *missing value* yang ditemukan dengan cara mengestimasi menggunakan *mean*

In [6]:
```
import math
for i, item in enumerate(file1_float['Prestige_score']):
    if math.isnan(item):
        sum_record = file1_float['Prestige_score'].sum()
        mean = sum_record/file1_float['Prestige_score'].count()
        missing_vall = mean*(file1_float['Prestige_score'].count()+1)-(
sum_record)
        file1_float.at[i, 'Prestige_score'] = missing_vall
file1_float
```

Out[6]:

	Prestige_score
0	28.000000
1	28.000000
2	69.000000
3	57.000000

	Prestige_score
4	18.000000
5	28.000000
6	38.934599
7	43.000000
8	33.000000
9	18.000000
10	28.000000
11	18.000000
12	18.000000
13	57.000000
14	40.000000
15	43.000000
16	69.000000
17	43.000000
18	28.000000
19	28.000000
20	18.000000
21	43.000000
22	38.934599
23	42.000000
24	43.000000
25	38.934599

	Prestige_score
26	42.000000
27	28.000000
28	38.934599
29	38.934599
...	...
470	18.000000
471	62.000000
472	37.000000
473	38.934599
474	40.000000
475	51.000000
476	35.000000
477	65.000000
478	71.000000
479	40.000000
480	35.000000
481	62.000000
482	61.000000
483	58.000000
484	37.000000
485	37.000000
486	37.000000

	Prestige_score
487	18.000000
488	37.000000
489	38.934599
490	62.000000
491	61.000000
492	61.000000
493	71.000000
494	30.000000
495	62.000000
496	18.000000
497	37.000000
498	62.000000
499	38.934599

500 rows × 1 columns

Data cleaning pada *missing values* atribut dengan tipe *ordinal* menggunakan pendekatan *most frequency value*

- Melakukan pengecekan *missing values* atribut dengan tipe *ordinal*

In [7]:
```
file1_object = file.select_dtypes(include=['object']).copy()
file1_object[file1_object.isnull().any(axis=1)]
```

Out[7]:

	Sex	Educational_level	Leaving_Certificate	Type_school

	Sex	Educational_level	Leaving_Certificate	Type_school
63	male	NaN	not_taken	secondary
68	male	NaN	not_taken	secondary
144	male	NaN	not_taken	secondary
161	male	NaN	not_taken	secondary
261	male	NaN	not_taken	secondary
444	male	NaN	not_taken	secondary

- Mengisi *missing value* yang ditemukan dengan cara mengestimasi menggunakan *most frequency value*

In [8]:
```
print('Frekuensi tiap atribut: ')
print(file1_object['Educational_level'].value_counts())
file1_object = file1_object.fillna({'Educational_level':'Senior_cycle_terminal_leaver-secondary_school'})
file1_object
```

```
Frekuensi tiap atribut:
Senior_cycle_terminal_leaver-secondary_school     158
Junior_cycle_terminal_leaver-vocational_school     68
Junior_cycle_terminal_leaver-secondary_school      65
3rd_level_complete                                 57
Junior_cycle_incomplete-vocational_school          50
Primary_terminal_leaver                            37
Junior_cycle_incomplete-secondary_school           30
Senior_cycle_incomplete-vocational_school          13
Senior_cycle_incomplete-secondary_school            9
3rd_level_incomplete                                7
Name: Educational_level, dtype: int64
```

Out[8]:

	Sex	Educational_level	Leaving_Certificate	Type_school

	Sex	Educational_level	Leaving_Certificate	Type_school
0	male	Junior_cycle_incomplete-secondary_school	not_taken	secondary
1	male	Primary_terminal_leaver	not_taken	primary_terminal_leaver
2	male	Senior_cycle_terminal_leaver-secondary_school	taken	secondary
3	male	Junior_cycle_terminal_leaver-secondary_school	not_taken	secondary
4	male	Junior_cycle_terminal_leaver-vocational_school	not_taken	vocational
5	male	Junior_cycle_terminal_leaver-vocational_school	not_taken	vocational
6	male	Primary_terminal_leaver	not_taken	primary_terminal_leaver
7	male	Junior_cycle_incomplete-vocational_school	not_taken	vocational
8	male	Junior_cycle_terminal_leaver-vocational_school	not_taken	vocational
9	male	Primary_terminal_leaver	not_taken	primary_terminal_leaver
10	male	Junior_cycle_terminal_leaver-vocational_school	not_taken	vocational
11	male	Junior_cycle_terminal_leaver-vocational_school	not_taken	vocational
12	male	3rd_level_complete	taken	secondary
13	male	Junior_cycle_terminal_leaver-vocational_school	not_taken	vocational

	Sex	Educational_level	Leaving_Certificate	Type_school
14	male	Senior_cycle_terminal_leaver-secondary_school	taken	secondary
15	male	Junior_cycle_terminal_leaver-vocational_school	not_taken	vocational
16	male	Junior_cycle_terminal_leaver-vocational_school	not_taken	vocational
17	male	Junior_cycle_incomplete-secondary_school	not_taken	secondary
18	male	Junior_cycle_terminal_leaver-vocational_school	not_taken	vocational
19	male	Senior_cycle_terminal_leaver-secondary_school	taken	secondary
20	male	Primary_terminal_leaver	not_taken	primary_terminal_leaver
21	male	Junior_cycle_incomplete-vocational_school	not_taken	vocational
22	male	Senior_cycle_terminal_leaver-secondary_school	taken	secondary
23	male	Senior_cycle_terminal_leaver-secondary_school	taken	secondary
24	male	Senior_cycle_terminal_leaver-secondary_school	taken	secondary
25	male	3rd_level_complete	taken	secondary
26	male	3rd_level_complete	taken	secondary
27	male	3rd_level_complete	taken	secondary

	Sex	Educational_level	Leaving_Certificate	Type_school
28	male	Senior_cycle_terminal_leaver-secondary_school	taken	secondary
29	male	Junior_cycle_terminal_leaver-secondary_school	not_taken	secondary
...	...	...	...	...
470	male	Junior_cycle_terminal_leaver-secondary_school	not_taken	secondary
471	male	Senior_cycle_terminal_leaver-secondary_school	taken	secondary
472	male	3rd_level_incomplete	taken	secondary
473	male	3rd_level_incomplete	taken	secondary
474	male	Senior_cycle_terminal_leaver-secondary_school	taken	secondary
475	male	Junior_cycle_terminal_leaver-vocational_school	not_taken	vocational
476	male	3rd_level_complete	taken	secondary
477	male	Senior_cycle_terminal_leaver-secondary_school	taken	secondary
478	male	3rd_level_complete	taken	secondary
479	female	3rd_level_complete	taken	secondary
480	female	Junior_cycle_incomplete-vocational_school	not_taken	vocational
481	female	Senior_cycle_terminal_leaver-secondary_school	taken	secondary

	Sex	Educational_level	Leaving_Certificate	Type_school
482	female	Senior_cycle_terminal_leaver-secondary_school	taken	secondary
483	female	Senior_cycle_terminal_leaver-secondary_school	taken	secondary
484	female	Senior_cycle_terminal_leaver-secondary_school	taken	secondary
485	female	3rd_level_complete	taken	secondary
486	female	Senior_cycle_terminal_leaver-secondary_school	taken	secondary
487	female	Junior_cycle_terminal_leaver-vocational_school	not_taken	vocational
488	female	3rd_level_complete	taken	secondary
489	male	Senior_cycle_incomplete-vocational_school	not_taken	vocational
490	male	3rd_level_incomplete	taken	secondary
491	male	3rd_level_complete	taken	secondary
492	male	3rd_level_complete	taken	secondary
493	male	3rd_level_complete	taken	secondary
494	male	Senior_cycle_terminal_leaver-secondary_school	taken	secondary
495	male	3rd_level_complete	taken	secondary
496	male	3rd_level_complete	taken	secondary
497	male	3rd_level_complete	taken	secondary
498	female	3rd_level_complete	taken	secondary

	Sex	Educational_level	Leaving_Certificate	Type_school
499	female	3rd_level_complete	taken	secondary

500 rows × 4 columns

Data cleaning pada *missing values* atribut dengan tipe diskrit

- Melakukan pengecekan *missing values* atribut dengan tipe diskrit

In [9]:
```
file1_int = file.select_dtypes(include=['int64']).copy()
file1_int[file1_int.isnull().any(axis=1)]
```

Out[9]:

	DVRT

Dikarenakan tidak terdapat *missing value* pada atribut DVRT, maka tidak perlu dilakukan data cleaning

Skenario 2

Data Cleaning pada *Missing Values* Atribut dengan Tipe *Ordinal* menggunakan Pendetakan *Most Frequent Value*

- Melakukan pengecekan *missing values* pada atribut dengan tipe ordinal

In [10]:
```
file2_object = file.select_dtypes(include=['object']).copy()
file2_object[file2_object.isnull().any(axis=1)]
```

Out[10]:

	Sex	Educational_level	Leaving_Certificate	Type_school
63	male	NaN	not_taken	secondary

	Sex	Educational_level	Leaving_Certificate	Type_school
68	male	NaN	not_taken	secondary
144	male	NaN	not_taken	secondary
161	male	NaN	not_taken	secondary
261	male	NaN	not_taken	secondary
444	male	NaN	not_taken	secondary

- Mengisi *missing values* yang ditemukan dengan *value* yang memiliki frekuensi kemunculan terbanyak pada atribut tersebut

In [11]:
```
print('Frekuensi tiap atribut: ')
print(file2_object['Educational_level'].value_counts())
file2_object = file2_object.fillna({'Educational_level':'Senior_cycle_terminal_leaver-secondary_school'})
file2_object
```

```
Frekuensi tiap atribut:
Senior_cycle_terminal_leaver-secondary_school     158
Junior_cycle_terminal_leaver-vocational_school     68
Junior_cycle_terminal_leaver-secondary_school      65
3rd_level_complete                                 57
Junior_cycle_incomplete-vocational_school          50
Primary_terminal_leaver                            37
Junior_cycle_incomplete-secondary_school           30
Senior_cycle_incomplete-vocational_school          13
Senior_cycle_incomplete-secondary_school            9
3rd_level_incomplete                                7
Name: Educational_level, dtype: int64
```

Out[11]:

	Sex	Educational_level	Leaving_Certificate	Type_school
0	male	Junior_cycle_incomplete-secondary_school	not_taken	secondary

	Sex	Educational_level	Leaving_Certificate	Type_school
1	male	Primary_terminal_leaver	not_taken	primary_terminal_leaver
2	male	Senior_cycle_terminal_leaver-secondary_school	taken	secondary
3	male	Junior_cycle_terminal_leaver-secondary_school	not_taken	secondary
4	male	Junior_cycle_terminal_leaver-vocational_school	not_taken	vocational
5	male	Junior_cycle_terminal_leaver-vocational_school	not_taken	vocational
6	male	Primary_terminal_leaver	not_taken	primary_terminal_leaver
7	male	Junior_cycle_incomplete-vocational_school	not_taken	vocational
8	male	Junior_cycle_terminal_leaver-vocational_school	not_taken	vocational
9	male	Primary_terminal_leaver	not_taken	primary_terminal_leaver
10	male	Junior_cycle_terminal_leaver-vocational_school	not_taken	vocational
11	male	Junior_cycle_terminal_leaver-vocational_school	not_taken	vocational
12	male	3rd_level_complete	taken	secondary
13	male	Junior_cycle_terminal_leaver-vocational_school	not_taken	vocational
14	male	Senior_cycle_terminal_leaver-secondary_school	taken	secondary

	Sex	Educational_level	Leaving_Certificate	Type_school
15	male	Junior_cycle_terminal_leaver-vocational_school	not_taken	vocational
16	male	Junior_cycle_terminal_leaver-vocational_school	not_taken	vocational
17	male	Junior_cycle_incomplete-secondary_school	not_taken	secondary
18	male	Junior_cycle_terminal_leaver-vocational_school	not_taken	vocational
19	male	Senior_cycle_terminal_leaver-secondary_school	taken	secondary
20	male	Primary_terminal_leaver	not_taken	primary_terminal_leaver
21	male	Junior_cycle_incomplete-vocational_school	not_taken	vocational
22	male	Senior_cycle_terminal_leaver-secondary_school	taken	secondary
23	male	Senior_cycle_terminal_leaver-secondary_school	taken	secondary
24	male	Senior_cycle_terminal_leaver-secondary_school	taken	secondary
25	male	3rd_level_complete	taken	secondary
26	male	3rd_level_complete	taken	secondary
27	male	3rd_level_complete	taken	secondary
28	male	Senior_cycle_terminal_leaver-secondary_school	taken	secondary

	Sex	Educational_level	Leaving_Certificate	Type_school
29	male	Junior_cycle_terminal_leaver-secondary_school	not_taken	secondary
...	...	...	...	...
470	male	Junior_cycle_terminal_leaver-secondary_school	not_taken	secondary
471	male	Senior_cycle_terminal_leaver-secondary_school	taken	secondary
472	male	3rd_level_incomplete	taken	secondary
473	male	3rd_level_incomplete	taken	secondary
474	male	Senior_cycle_terminal_leaver-secondary_school	taken	secondary
475	male	Junior_cycle_terminal_leaver-vocational_school	not_taken	vocational
476	male	3rd_level_complete	taken	secondary
477	male	Senior_cycle_terminal_leaver-secondary_school	taken	secondary
478	male	3rd_level_complete	taken	secondary
479	female	3rd_level_complete	taken	secondary
480	female	Junior_cycle_incomplete-vocational_school	not_taken	vocational
481	female	Senior_cycle_terminal_leaver-secondary_school	taken	secondary
482	female	Senior_cycle_terminal_leaver-secondary_school	taken	secondary

	Sex	Educational_level	Leaving_Certificate	Type_school
483	female	Senior_cycle_terminal_leaver-secondary_school	taken	secondary
484	female	Senior_cycle_terminal_leaver-secondary_school	taken	secondary
485	female	3rd_level_complete	taken	secondary
486	female	Senior_cycle_terminal_leaver-secondary_school	taken	secondary
487	female	Junior_cycle_terminal_leaver-vocational_school	not_taken	vocational
488	female	3rd_level_complete	taken	secondary
489	male	Senior_cycle_incomplete-vocational_school	not_taken	vocational
490	male	3rd_level_incomplete	taken	secondary
491	male	3rd_level_complete	taken	secondary
492	male	3rd_level_complete	taken	secondary
493	male	3rd_level_complete	taken	secondary
494	male	Senior_cycle_terminal_leaver-secondary_school	taken	secondary
495	male	3rd_level_complete	taken	secondary
496	male	3rd_level_complete	taken	secondary
497	male	3rd_level_complete	taken	secondary
498	female	3rd_level_complete	taken	secondary
499	female	3rd_level_complete	taken	secondary

500 rows × 4 columns

Data Cleaning pada Missing Values Atribut dengan Tipe Diskrit menggunakan Pendetakan Most Frequent Value

- Melakukan pengecekan *missing values* pada atribut dengan tipe diskrit

In [12]:
```
file2_int = file.select_dtypes(include=['int64']).copy()
file2_int[file2_int.isnull().any(axis=1)]
```

Out[12]:

	DVRT

Atribut DVRT memiliki tipe atribut yaitu integer. Karena, pada atribut DVRT tidak terdapat *missing values* maka tidak perlu dilakukan *data cleaning* pada atribut DVRT.

Data Cleaning pada *Missing Values* Atribut dengan Tipe Kontinyu menggunakan Pendetakan *Most Frequent Value*

- Melakukan pengecekan *missing values* pada atribut dengan tipe kontinyu

In [13]:
```
file2_float = file.select_dtypes(include=['float64']).copy()
file2_float[file2_float.isnull().any(axis=1)]
```

Out[13]:

	Prestige_score
6	NaN
22	NaN
25	NaN
28	NaN
29	NaN

	Prestige_score
35	NaN
66	NaN
89	NaN
111	NaN
112	NaN
137	NaN
147	NaN
198	NaN
217	NaN
219	NaN
291	NaN
300	NaN
301	NaN
310	NaN
341	NaN
347	NaN
348	NaN
349	NaN
473	NaN
489	NaN
499	NaN

- Mengisi missing values yang ditemukan dengan value yang memiliki frekuensi kemunculan terbanyak pada atribut tersebut

In [14]:
```
print('Frekuensi kemunculan tiap atribut: ')
print(file2_float['Prestige_score'].value_counts())
file2_float = file2_float.fillna({'Prestige_score':18.0})
file2_float
```

```
Frekuensi kemunculan tiap atribut:
18.0    91
37.0    89
43.0    46
28.0    39
58.0    23
40.0    23
57.0    22
35.0    18
61.0    15
31.0    12
38.0    12
75.0     9
69.0     9
62.0     8
71.0     8
51.0     7
46.0     6
27.0     6
33.0     6
36.0     5
65.0     5
42.0     4
66.0     3
30.0     3
47.0     2
64.0     1
48.0     1
53.0     1
Name: Prestige_score, dtype: int64
```

Out[14]:

	Prestige_score
0	28.0
1	28.0
2	69.0
3	57.0
4	18.0
5	28.0
6	18.0
7	43.0
8	33.0
9	18.0
10	28.0
11	18.0
12	18.0
13	57.0
14	40.0
15	43.0
16	69.0
17	43.0
18	28.0
19	28.0
20	18.0
21	43.0

	Prestige_score
22	18.0
23	42.0
24	43.0
25	18.0
26	42.0
27	28.0
28	18.0
29	18.0
...	...
470	18.0
471	62.0
472	37.0
473	18.0
474	40.0
475	51.0
476	35.0
477	65.0
478	71.0
479	40.0
480	35.0
481	62.0
482	61.0

	Prestige_score
483	58.0
484	37.0
485	37.0
486	37.0
487	18.0
488	37.0
489	18.0
490	62.0
491	61.0
492	61.0
493	71.0
494	30.0
495	62.0
496	18.0
497	37.0
498	62.0
499	18.0

500 rows × 1 columns

Skenario 3

Data cleaning pada *Missing Value* atribut bertipe kontinyu menggunakan *Mean/Rata-Rata*

- Melakukan pengecekan *Missing value* pada atribut bertipe kontinyu

In [15]:
```
file_copy = file.copy()
file_copy
```

Out[15]:

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	
0	male	113	Junior_cycle_incomplete-secondary_school	not_taken	28.0	sec
1	male	101	Primary_terminal_leaver	not_taken	28.0	prim
2	male	110	Senior_cycle_terminal_leaver-secondary_school	taken	69.0	sec
3	male	121	Junior_cycle_terminal_leaver-secondary_school	not_taken	57.0	sec
4	male	82	Junior_cycle_terminal_leaver-vocational_school	not_taken	18.0	voc
5	male	85	Junior_cycle_terminal_leaver-vocational_school	not_taken	28.0	voc
6	male	84	Primary_terminal_leaver	not_taken	NaN	prim
7	male	98	Junior_cycle_incomplete-vocational_school	not_taken	43.0	voc
8	male	92	Junior_cycle_terminal_leaver-vocational_school	not_taken	33.0	voc
9	male	90	Primary_terminal_leaver	not_taken	18.0	prim
10	male	88	Junior_cycle_terminal_leaver-vocational_school	not_taken	28.0	voc
11	male	84	Junior_cycle_terminal_leaver-vocational_school	not_taken	18.0	voc

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	
12	male	114	3rd_level_complete	taken	18.0	seco
13	male	70	Junior_cycle_terminal_leaver-vocational_school	not_taken	57.0	voca
14	male	109	Senior_cycle_terminal_leaver-secondary_school	taken	40.0	seco
15	male	83	Junior_cycle_terminal_leaver-vocational_school	not_taken	43.0	voca
16	male	108	Junior_cycle_terminal_leaver-vocational_school	not_taken	69.0	voca
17	male	95	Junior_cycle_incomplete-secondary_school	not_taken	43.0	seco
18	male	90	Junior_cycle_terminal_leaver-vocational_school	not_taken	28.0	voca
19	male	107	Senior_cycle_terminal_leaver-secondary_school	taken	28.0	seco
20	male	86	Primary_terminal_leaver	not_taken	18.0	prim
21	male	70	Junior_cycle_incomplete-vocational_school	not_taken	43.0	voca
22	male	114	Senior_cycle_terminal_leaver-secondary_school	taken	NaN	seco
23	male	117	Senior_cycle_terminal_leaver-secondary_school	taken	42.0	seco
24	male	112	Senior_cycle_terminal_leaver-secondary_school	taken	43.0	seco
25	male	88	3rd_level_complete	taken	NaN	seco

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	
26	male	106	3rd_level_complete	taken	42.0	seco
27	male	125	3rd_level_complete	taken	28.0	seco
28	male	94	Senior_cycle_terminal_leaver-secondary_school	taken	NaN	seco
29	male	103	Junior_cycle_terminal_leaver-secondary_school	not_taken	NaN	seco
...	...	...	...	...	...	...
470	male	129	Junior_cycle_terminal_leaver-secondary_school	not_taken	18.0	seco
471	male	122	Senior_cycle_terminal_leaver-secondary_school	taken	62.0	seco
472	male	121	3rd_level_incomplete	taken	37.0	seco
473	male	129	3rd_level_incomplete	taken	NaN	seco
474	male	122	Senior_cycle_terminal_leaver-secondary_school	taken	40.0	seco
475	male	126	Junior_cycle_terminal_leaver-vocational_school	not_taken	51.0	voca
476	male	122	3rd_level_complete	taken	35.0	seco
477	male	123	Senior_cycle_terminal_leaver-secondary_school	taken	65.0	seco
478	male	119	3rd_level_complete	taken	71.0	seco
479	female	120	3rd_level_complete	taken	40.0	seco
480	female	127	Junior_cycle_incomplete-vocational_school	not_taken	35.0	voca

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	
481	female	127	Senior_cycle_terminal_leaver-secondary_school	taken	62.0	seco
482	female	120	Senior_cycle_terminal_leaver-secondary_school	taken	61.0	seco
483	female	127	Senior_cycle_terminal_leaver-secondary_school	taken	58.0	seco
484	female	123	Senior_cycle_terminal_leaver-secondary_school	taken	37.0	seco
485	female	120	3rd_level_complete	taken	37.0	seco
486	female	123	Senior_cycle_terminal_leaver-secondary_school	taken	37.0	seco
487	female	122	Junior_cycle_terminal_leaver-vocational_school	not_taken	18.0	voca
488	female	119	3rd_level_complete	taken	37.0	seco
489	male	130	Senior_cycle_incomplete-vocational_school	not_taken	NaN	voca
490	male	134	3rd_level_incomplete	taken	62.0	seco
491	male	136	3rd_level_complete	taken	61.0	seco
492	male	135	3rd_level_complete	taken	61.0	seco
493	male	140	3rd_level_complete	taken	71.0	seco
494	male	131	Senior_cycle_terminal_leaver-secondary_school	taken	30.0	seco
495	male	137	3rd_level_complete	taken	62.0	seco
496	male	136	3rd_level_complete	taken	18.0	seco

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	
497	male	132	3rd_level_complete	taken	37.0	seco
498	female	135	3rd_level_complete	taken	62.0	seco
499	female	134	3rd_level_complete	taken	NaN	seco

500 rows × 6 columns

In [16]:
```
file3_float = file_copy.select_dtypes(include=['float64']).copy()
file3_float[file3_float.isnull().any(axis=1)]
```

Out[16]:

	Prestige_score
6	NaN
22	NaN
25	NaN
28	NaN
29	NaN
35	NaN
66	NaN
89	NaN
111	NaN
112	NaN
137	NaN
147	NaN
198	NaN

	Prestige_score
217	NaN
219	NaN
291	NaN
300	NaN
301	NaN
310	NaN
341	NaN
347	NaN
348	NaN
349	NaN
473	NaN
489	NaN
499	NaN

- Mengisi *missing values* yang ditemukan dengan menghitung *Mean/Rata-Rata*

In [17]:
```
for i,item in enumerate(file_copy['Prestige_score']):
    if math.isnan(item):
        jumlah_semua = file_copy['Prestige_score'].sum()
        rata2 = jumlah_semua/file_copy['Prestige_score'].count()
        missing_vall = rata2*(file_copy['Prestige_score'].count()+1)-(j
umlah_semua)
        file_copy.at[i, 'Prestige_score'] = missing_vall
file_copy
```

Out[17]:

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	
0	male	113	Junior_cycle_incomplete-secondary_school	not_taken	28.000000	seco
1	male	101	Primary_terminal_leaver	not_taken	28.000000	prim
2	male	110	Senior_cycle_terminal_leaver-secondary_school	taken	69.000000	seco
3	male	121	Junior_cycle_terminal_leaver-secondary_school	not_taken	57.000000	seco
4	male	82	Junior_cycle_terminal_leaver-vocational_school	not_taken	18.000000	voca
5	male	85	Junior_cycle_terminal_leaver-vocational_school	not_taken	28.000000	voca
6	male	84	Primary_terminal_leaver	not_taken	38.934599	prim
7	male	98	Junior_cycle_incomplete-vocational_school	not_taken	43.000000	voca
8	male	92	Junior_cycle_terminal_leaver-vocational_school	not_taken	33.000000	voca
9	male	90	Primary_terminal_leaver	not_taken	18.000000	prim
10	male	88	Junior_cycle_terminal_leaver-vocational_school	not_taken	28.000000	voca
11	male	84	Junior_cycle_terminal_leaver-vocational_school	not_taken	18.000000	voca
12	male	114	3rd_level_complete	taken	18.000000	seco
13	male	70	Junior_cycle_terminal_leaver-vocational_school	not_taken	57.000000	voca

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	
14	male	109	Senior_cycle_terminal_leaver-secondary_school	taken	40.000000	seco
15	male	83	Junior_cycle_terminal_leaver-vocational_school	not_taken	43.000000	voca
16	male	108	Junior_cycle_terminal_leaver-vocational_school	not_taken	69.000000	voca
17	male	95	Junior_cycle_incomplete-secondary_school	not_taken	43.000000	seco
18	male	90	Junior_cycle_terminal_leaver-vocational_school	not_taken	28.000000	voca
19	male	107	Senior_cycle_terminal_leaver-secondary_school	taken	28.000000	seco
20	male	86	Primary_terminal_leaver	not_taken	18.000000	prim
21	male	70	Junior_cycle_incomplete-vocational_school	not_taken	43.000000	voca
22	male	114	Senior_cycle_terminal_leaver-secondary_school	taken	38.934599	seco
23	male	117	Senior_cycle_terminal_leaver-secondary_school	taken	42.000000	seco
24	male	112	Senior_cycle_terminal_leaver-secondary_school	taken	43.000000	seco
25	male	88	3rd_level_complete	taken	38.934599	seco
26	male	106	3rd_level_complete	taken	42.000000	seco
27	male	125	3rd_level_complete	taken	28.000000	seco

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	
28	male	94	Senior_cycle_terminal_leaver-secondary_school	taken	38.934599	seco
29	male	103	Junior_cycle_terminal_leaver-secondary_school	not_taken	38.934599	seco
...	...	...	...	...	...	...
470	male	129	Junior_cycle_terminal_leaver-secondary_school	not_taken	18.000000	seco
471	male	122	Senior_cycle_terminal_leaver-secondary_school	taken	62.000000	seco
472	male	121	3rd_level_incomplete	taken	37.000000	seco
473	male	129	3rd_level_incomplete	taken	38.934599	seco
474	male	122	Senior_cycle_terminal_leaver-secondary_school	taken	40.000000	seco
475	male	126	Junior_cycle_terminal_leaver-vocational_school	not_taken	51.000000	voca
476	male	122	3rd_level_complete	taken	35.000000	seco
477	male	123	Senior_cycle_terminal_leaver-secondary_school	taken	65.000000	seco
478	male	119	3rd_level_complete	taken	71.000000	seco
479	female	120	3rd_level_complete	taken	40.000000	seco
480	female	127	Junior_cycle_incomplete-vocational_school	not_taken	35.000000	voca
481	female	127	Senior_cycle_terminal_leaver-secondary_school	taken	62.000000	seco

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	
482	female	120	Senior_cycle_terminal_leaver-secondary_school	taken	61.000000	seco
483	female	127	Senior_cycle_terminal_leaver-secondary_school	taken	58.000000	seco
484	female	123	Senior_cycle_terminal_leaver-secondary_school	taken	37.000000	seco
485	female	120	3rd_level_complete	taken	37.000000	seco
486	female	123	Senior_cycle_terminal_leaver-secondary_school	taken	37.000000	seco
487	female	122	Junior_cycle_terminal_leaver-vocational_school	not_taken	18.000000	voca
488	female	119	3rd_level_complete	taken	37.000000	seco
489	male	130	Senior_cycle_incomplete-vocational_school	not_taken	38.934599	voca
490	male	134	3rd_level_incomplete	taken	62.000000	seco
491	male	136	3rd_level_complete	taken	61.000000	seco
492	male	135	3rd_level_complete	taken	61.000000	seco
493	male	140	3rd_level_complete	taken	71.000000	seco
494	male	131	Senior_cycle_terminal_leaver-secondary_school	taken	30.000000	seco
495	male	137	3rd_level_complete	taken	62.000000	seco
496	male	136	3rd_level_complete	taken	18.000000	seco
497	male	132	3rd_level_complete	taken	37.000000	seco
498	female	135	3rd_level_complete	taken	62.000000	seco

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	
499	female	134	3rd_level_complete	taken	38.934599	sec

500 rows × 6 columns

Data cleaning pada masalah *Missing Value* atribut bertipe *Ordinal* dengan *Mengabaikan Missing Value Pada Atribut*

- Melakukan pengecekan *Missing Value* pada atribut bertipe *Ordinal*

In [18]:
```
file3_object = file_copy.select_dtypes(include=['object']).copy()
file3_object[file3_object.isnull().any(axis=1)]
```

Out[18]:

	Sex	Educational_level	Leaving_Certificate	Type_school
63	male	NaN	not_taken	secondary
68	male	NaN	not_taken	secondary
144	male	NaN	not_taken	secondary
161	male	NaN	not_taken	secondary
261	male	NaN	not_taken	secondary
444	male	NaN	not_taken	secondary

- Menghapus record yang terdapat missing value pada atributnya

In [19]:
```
file_copy.dropna(inplace=True)
file_copy
```

Out[19]:

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	
0	male	113	Junior_cycle_incomplete-secondary_school	not_taken	28.000000	seco
1	male	101	Primary_terminal_leaver	not_taken	28.000000	prim
2	male	110	Senior_cycle_terminal_leaver-secondary_school	taken	69.000000	seco
3	male	121	Junior_cycle_terminal_leaver-secondary_school	not_taken	57.000000	seco
4	male	82	Junior_cycle_terminal_leaver-vocational_school	not_taken	18.000000	voca
5	male	85	Junior_cycle_terminal_leaver-vocational_school	not_taken	28.000000	voca
6	male	84	Primary_terminal_leaver	not_taken	38.934599	prim
7	male	98	Junior_cycle_incomplete-vocational_school	not_taken	43.000000	voca
8	male	92	Junior_cycle_terminal_leaver-vocational_school	not_taken	33.000000	voca
9	male	90	Primary_terminal_leaver	not_taken	18.000000	prim
10	male	88	Junior_cycle_terminal_leaver-vocational_school	not_taken	28.000000	voca
11	male	84	Junior_cycle_terminal_leaver-vocational_school	not_taken	18.000000	voca
12	male	114	3rd_level_complete	taken	18.000000	seco
13	male	70	Junior_cycle_terminal_leaver-vocational_school	not_taken	57.000000	voca

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	
14	male	109	Senior_cycle_terminal_leaver-secondary_school	taken	40.000000	seco
15	male	83	Junior_cycle_terminal_leaver-vocational_school	not_taken	43.000000	voca
16	male	108	Junior_cycle_terminal_leaver-vocational_school	not_taken	69.000000	voca
17	male	95	Junior_cycle_incomplete-secondary_school	not_taken	43.000000	seco
18	male	90	Junior_cycle_terminal_leaver-vocational_school	not_taken	28.000000	voca
19	male	107	Senior_cycle_terminal_leaver-secondary_school	taken	28.000000	seco
20	male	86	Primary_terminal_leaver	not_taken	18.000000	prim
21	male	70	Junior_cycle_incomplete-vocational_school	not_taken	43.000000	voca
22	male	114	Senior_cycle_terminal_leaver-secondary_school	taken	38.934599	seco
23	male	117	Senior_cycle_terminal_leaver-secondary_school	taken	42.000000	seco
24	male	112	Senior_cycle_terminal_leaver-secondary_school	taken	43.000000	seco
25	male	88	3rd_level_complete	taken	38.934599	seco
26	male	106	3rd_level_complete	taken	42.000000	seco
27	male	125	3rd_level_complete	taken	28.000000	seco

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	
28	male	94	Senior_cycle_terminal_leaver-secondary_school	taken	38.934599	seco
29	male	103	Junior_cycle_terminal_leaver-secondary_school	not_taken	38.934599	seco
...	...	...	...	...	...	...
470	male	129	Junior_cycle_terminal_leaver-secondary_school	not_taken	18.000000	seco
471	male	122	Senior_cycle_terminal_leaver-secondary_school	taken	62.000000	seco
472	male	121	3rd_level_incomplete	taken	37.000000	seco
473	male	129	3rd_level_incomplete	taken	38.934599	seco
474	male	122	Senior_cycle_terminal_leaver-secondary_school	taken	40.000000	seco
475	male	126	Junior_cycle_terminal_leaver-vocational_school	not_taken	51.000000	voca
476	male	122	3rd_level_complete	taken	35.000000	seco
477	male	123	Senior_cycle_terminal_leaver-secondary_school	taken	65.000000	seco
478	male	119	3rd_level_complete	taken	71.000000	seco
479	female	120	3rd_level_complete	taken	40.000000	seco
480	female	127	Junior_cycle_incomplete-vocational_school	not_taken	35.000000	voca
481	female	127	Senior_cycle_terminal_leaver-secondary_school	taken	62.000000	seco

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	
482	female	120	Senior_cycle_terminal_leaver-secondary_school	taken	61.000000	seco
483	female	127	Senior_cycle_terminal_leaver-secondary_school	taken	58.000000	seco
484	female	123	Senior_cycle_terminal_leaver-secondary_school	taken	37.000000	seco
485	female	120	3rd_level_complete	taken	37.000000	seco
486	female	123	Senior_cycle_terminal_leaver-secondary_school	taken	37.000000	seco
487	female	122	Junior_cycle_terminal_leaver-vocational_school	not_taken	18.000000	voca
488	female	119	3rd_level_complete	taken	37.000000	seco
489	male	130	Senior_cycle_incomplete-vocational_school	not_taken	38.934599	voca
490	male	134	3rd_level_incomplete	taken	62.000000	seco
491	male	136	3rd_level_complete	taken	61.000000	seco
492	male	135	3rd_level_complete	taken	61.000000	seco
493	male	140	3rd_level_complete	taken	71.000000	seco
494	male	131	Senior_cycle_terminal_leaver-secondary_school	taken	30.000000	seco
495	male	137	3rd_level_complete	taken	62.000000	seco
496	male	136	3rd_level_complete	taken	18.000000	seco
497	male	132	3rd_level_complete	taken	37.000000	seco
498	female	135	3rd_level_complete	taken	62.000000	seco

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	
499	female	134	3rd_level_complete	taken	38.934599	sec

494 rows × 6 columns

PREPROCESSING DATA

Langkah-Langkah Preprocessing

1. Binarization : pengubahan atribut sex dan leaving certificate menjadi bentuk biner dengan ketentukan pada atribut sex, male: 1 dan female:0
2. Encoding Categorical Feature : Mengubah atribut Educational_level dan type school.

   Educational_level :

   - Primary_terminal_leaver = 1
   - Junior_cycle_incomplete-vocational_school = 2
   - Junior_cycle_incomplete-secondary_school = 3
   - Junior_cycle_terminal_leaver-vocational_school = 4
   - Junior_cycle_terminal_leaver-secondary_school = 5
   - Senior_cycle_incomplete-vocational_school = 6
   - Senior_cycle_incomplete-secondary_school = 7
   - Senior_cycle_terminal_leaver-secondary_school = 8
   - 3rd_level_incomplete = 9
   - 3rd_level_complete = 10

   Type_school :

   - Secondary =1
   - Vocational =2
   - Primary_terminal_leaver = 3

3. Feature Creation :
   Nama Atribut : Indeks

   Rumus :

$$indeks = (\frac{DVRT+Prestige}{maxValueOf(DVRT)+maxValueOf(Prestige)}) * 100$$

4.Equal Interval : Mengelompokkan indeks menjadi kaategoi, terdiri dari :

- A = X >= 80
- B = 60 =< X <= 80
- C = 40 =< X <= 60
- D = 20 =< X <= 40
- E = X < 20

Skenario 1

- *Binarization* pada Atribut *Sex* dan *Leaving Certificate*

In [20]:
```
biner1 = {'Sex': {'male':1, 'female':0},
          'Leaving_Certificate': {'taken':1, 'not_taken':0}}
file1_object.replace(biner1,inplace=True)
file1_object
```

Out[20]:

	Sex	Educational_level	Leaving_Certificate	Type_school
0	1	Junior_cycle_incomplete-secondary_school	0	secondary
1	1	Primary_terminal_leaver	0	primary_terminal_leaver
2	1	Senior_cycle_terminal_leaver-secondary_school	1	secondary

	Sex	Educational_level	Leaving_Certificate	Type_school
3	1	Junior_cycle_terminal_leaver-secondary_school	0	secondary
4	1	Junior_cycle_terminal_leaver-vocational_school	0	vocational
5	1	Junior_cycle_terminal_leaver-vocational_school	0	vocational
6	1	Primary_terminal_leaver	0	primary_terminal_leaver
7	1	Junior_cycle_incomplete-vocational_school	0	vocational
8	1	Junior_cycle_terminal_leaver-vocational_school	0	vocational
9	1	Primary_terminal_leaver	0	primary_terminal_leaver
10	1	Junior_cycle_terminal_leaver-vocational_school	0	vocational
11	1	Junior_cycle_terminal_leaver-vocational_school	0	vocational
12	1	3rd_level_complete	1	secondary
13	1	Junior_cycle_terminal_leaver-vocational_school	0	vocational
14	1	Senior_cycle_terminal_leaver-secondary_school	1	secondary
15	1	Junior_cycle_terminal_leaver-vocational_school	0	vocational
16	1	Junior_cycle_terminal_leaver-vocational_school	0	vocational

	Sex	Educational_level	Leaving_Certificate	Type_school
17	1	Junior_cycle_incomplete-secondary_school	0	secondary
18	1	Junior_cycle_terminal_leaver-vocational_school	0	vocational
19	1	Senior_cycle_terminal_leaver-secondary_school	1	secondary
20	1	Primary_terminal_leaver	0	primary_terminal_leaver
21	1	Junior_cycle_incomplete-vocational_school	0	vocational
22	1	Senior_cycle_terminal_leaver-secondary_school	1	secondary
23	1	Senior_cycle_terminal_leaver-secondary_school	1	secondary
24	1	Senior_cycle_terminal_leaver-secondary_school	1	secondary
25	1	3rd_level_complete	1	secondary
26	1	3rd_level_complete	1	secondary
27	1	3rd_level_complete	1	secondary
28	1	Senior_cycle_terminal_leaver-secondary_school	1	secondary
29	1	Junior_cycle_terminal_leaver-secondary_school	0	secondary
...	...	...	...	...
470	1	Junior_cycle_terminal_leaver-secondary_school	0	secondary

	Sex	Educational_level	Leaving_Certificate	Type_school
471	1	Senior_cycle_terminal_leaver-secondary_school	1	secondary
472	1	3rd_level_incomplete	1	secondary
473	1	3rd_level_incomplete	1	secondary
474	1	Senior_cycle_terminal_leaver-secondary_school	1	secondary
475	1	Junior_cycle_terminal_leaver-vocational_school	0	vocational
476	1	3rd_level_complete	1	secondary
477	1	Senior_cycle_terminal_leaver-secondary_school	1	secondary
478	1	3rd_level_complete	1	secondary
479	0	3rd_level_complete	1	secondary
480	0	Junior_cycle_incomplete-vocational_school	0	vocational
481	0	Senior_cycle_terminal_leaver-secondary_school	1	secondary
482	0	Senior_cycle_terminal_leaver-secondary_school	1	secondary
483	0	Senior_cycle_terminal_leaver-secondary_school	1	secondary
484	0	Senior_cycle_terminal_leaver-secondary_school	1	secondary
485	0	3rd_level_complete	1	secondary

	Sex	Educational_level	Leaving_Certificate	Type_school
486	0	Senior_cycle_terminal_leaver-secondary_school	1	secondary
487	0	Junior_cycle_terminal_leaver-vocational_school	0	vocational
488	0	3rd_level_complete	1	secondary
489	1	Senior_cycle_incomplete-vocational_school	0	vocational
490	1	3rd_level_incomplete	1	secondary
491	1	3rd_level_complete	1	secondary
492	1	3rd_level_complete	1	secondary
493	1	3rd_level_complete	1	secondary
494	1	Senior_cycle_terminal_leaver-secondary_school	1	secondary
495	1	3rd_level_complete	1	secondary
496	1	3rd_level_complete	1	secondary
497	1	3rd_level_complete	1	secondary
498	0	3rd_level_complete	1	secondary
499	0	3rd_level_complete	1	secondary

500 rows × 4 columns

- *Encoding Categorical Features* pada atribut *Educational Level* dan *Type School*

In [21]:
```
kategori1 = {'Educational_level': {'Primary_terminal_leaver':1, 'Junior
_cycle_incomplete-vocational_school':2,
                                   'Junior_cycle_incomplete-secondary_sch
```

```
ool':3, 'Junior_cycle_terminal_leaver-vocational_school':4,
                                    'Junior_cycle_terminal_leaver-secondar
y_school':5, 'Senior_cycle_incomplete-vocational_school':6,
                                    'Senior_cycle_incomplete-secondary_sch
ool':7, 'Senior_cycle_terminal_leaver-secondary_school':8,
                                    '3rd_level_incomplete':9, '3rd_level_c
omplete':10},
              'Type_school': {'secondary':1, 'vocational':2, 'primary_term
inal_leaver':3}}
file1_object.replace(kategori1,inplace=True)
file1_object
```

Out[21]:

	Sex	Educational_level	Leaving_Certificate	Type_school
0	1	3	0	1
1	1	1	0	3
2	1	8	1	1
3	1	5	0	1
4	1	4	0	2
5	1	4	0	2
6	1	1	0	3
7	1	2	0	2
8	1	4	0	2
9	1	1	0	3
10	1	4	0	2
11	1	4	0	2
12	1	10	1	1
13	1	4	0	2
14	1	8	1	1

	Sex	Educational_level	Leaving_Certificate	Type_school
15	1	4	0	2
16	1	4	0	2
17	1	3	0	1
18	1	4	0	2
19	1	8	1	1
20	1	1	0	3
21	1	2	0	2
22	1	8	1	1
23	1	8	1	1
24	1	8	1	1
25	1	10	1	1
26	1	10	1	1
27	1	10	1	1
28	1	8	1	1
29	1	5	0	1
...	...	...	...	...
470	1	5	0	1
471	1	8	1	1
472	1	9	1	1
473	1	9	1	1
474	1	8	1	1
475	1	4	0	2

	Sex	Educational_level	Leaving_Certificate	Type_school
476	1	10	1	1
477	1	8	1	1
478	1	10	1	1
479	0	10	1	1
480	0	2	0	2
481	0	8	1	1
482	0	8	1	1
483	0	8	1	1
484	0	8	1	1
485	0	10	1	1
486	0	8	1	1
487	0	4	0	2
488	0	10	1	1
489	1	6	0	2
490	1	9	1	1
491	1	10	1	1
492	1	10	1	1
493	1	10	1	1
494	1	8	1	1
495	1	10	1	1
496	1	10	1	1
497	1	10	1	1

	Sex	Educational_level	Leaving_Certificate	Type_school
498	0	10	1	1
499	0	10	1	1

500 rows × 4 columns

- *Feature Creation* sebagai atribut baru menggunakan *equal interval* untuk menggambarkan tingkat kemampuan

Atribut baru ini dinamakan dengan atribut **indeks** dimana *values* pada atribut indeks akan diperoleh dari rumus, sebagai berikut.

$$indeks = (\frac{DVRT + Prestige}{maxValueOf(DVRT) + maxValueOf(Prestige)}) * 100$$

In [22]:
```
for i in file1_float:
    for j in file1_int:
        counter1 = file1_float[i] + file1_int[j]
        count1_indeks = (counter1/215)*100
    print(count1_indeks)
```

```
0      65.581395
1      60.000000
2      83.255814
3      82.790698
4      46.511628
5      52.558140
6      57.178883
7      65.581395
8      58.139535
9      50.232558
10     53.953488
11     47.441860
12     61.395349
13     59.069767
14     69.302326
```

```
15     58.604651
16     82.325581
17     64.186047
18     54.883721
19     62.790698
20     48.372093
21     52.558140
22     71.132372
23     73.953488
24     72.093023
25     59.039348
26     68.837209
27     71.162791
28     61.830046
29     66.016093
         ...
470    68.372093
471    85.581395
472    73.488372
473    78.109116
474    75.348837
475    82.325581
476    73.023256
477    87.441860
478    88.372093
479    74.418605
480    75.348837
481    87.906977
482    84.186047
483    86.046512
484    74.418605
485    73.023256
486    74.418605
487    65.116279
488    72.558140
489    78.574232
490    91.162791
491    91.627907
492    91.162791
```

```
493    98.139535
494    74.883721
495    92.558140
496    71.627907
497    78.604651
498    91.627907
499    80.434697
Length: 500, dtype: float64
```

- Menambahkan hasil count indeks sebagai suatu atribut dan menyimpannya dalam bentuk csv

In [23]:
```
raw_data1 = {
    'Sex': file1_object['Sex'].values,
    'DVRT': file1_int['DVRT'].values,
    'Educational_level': file1_object['Educational_level'].values,
    'Leaving_Certificate': file1_object['Leaving_Certificate'].values,
    'Prestige_score': file1_float['Prestige_score'].values,
    'Type_school': file1_object['Type_school'].values,
    'Count_Indeks': count1_indeks
}
df1 = pd.DataFrame(raw_data1,
                   columns = ['Sex', 'DVRT', 'Educational_level', 'Leavin
g_Certificate', 'Prestige_score', 'Type_school', 'Count_Indeks'])
df1
```

Out[23]:

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	Type_school	Co
0	1	113	3	0	28.000000	1	65.
1	1	101	1	0	28.000000	3	60.
2	1	110	8	1	69.000000	1	83.
3	1	121	5	0	57.000000	1	82.
4	1	82	4	0	18.000000	2	46.
5	1	85	4	0	28.000000	2	52.

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	Type_school	Co
6	1	84	1	0	38.934599	3	57.
7	1	98	2	0	43.000000	2	65.
8	1	92	4	0	33.000000	2	58.
9	1	90	1	0	18.000000	3	50.
10	1	88	4	0	28.000000	2	53.
11	1	84	4	0	18.000000	2	47.
12	1	114	10	1	18.000000	1	61.
13	1	70	4	0	57.000000	2	59.
14	1	109	8	1	40.000000	1	69.
15	1	83	4	0	43.000000	2	58.
16	1	108	4	0	69.000000	2	82.
17	1	95	3	0	43.000000	1	64.
18	1	90	4	0	28.000000	2	54.
19	1	107	8	1	28.000000	1	62.
20	1	86	1	0	18.000000	3	48.
21	1	70	2	0	43.000000	2	52.
22	1	114	8	1	38.934599	1	71.
23	1	117	8	1	42.000000	1	73.
24	1	112	8	1	43.000000	1	72.
25	1	88	10	1	38.934599	1	59.
26	1	106	10	1	42.000000	1	68.
27	1	125	10	1	28.000000	1	71.

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	Type_school	Co
28	1	94	8	1	38.934599	1	61.
29	1	103	5	0	38.934599	1	66.
...	...	...	...	...	...	...	...
470	1	129	5	0	18.000000	1	68.
471	1	122	8	1	62.000000	1	85.
472	1	121	9	1	37.000000	1	73.
473	1	129	9	1	38.934599	1	78.
474	1	122	8	1	40.000000	1	75.
475	1	126	4	0	51.000000	2	82.
476	1	122	10	1	35.000000	1	73.
477	1	123	8	1	65.000000	1	87.
478	1	119	10	1	71.000000	1	88.
479	0	120	10	1	40.000000	1	74.
480	0	127	2	0	35.000000	2	75.
481	0	127	8	1	62.000000	1	87.
482	0	120	8	1	61.000000	1	84.
483	0	127	8	1	58.000000	1	86.
484	0	123	8	1	37.000000	1	74.
485	0	120	10	1	37.000000	1	73.
486	0	123	8	1	37.000000	1	74.
487	0	122	4	0	18.000000	2	65.
488	0	119	10	1	37.000000	1	72.

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	Type_school	Co
489	1	130	6	0	38.934599	2	78.
490	1	134	9	1	62.000000	1	91.
491	1	136	10	1	61.000000	1	91.
492	1	135	10	1	61.000000	1	91.
493	1	140	10	1	71.000000	1	98.
494	1	131	8	1	30.000000	1	74.
495	1	137	10	1	62.000000	1	92.
496	1	136	10	1	18.000000	1	71.
497	1	132	10	1	37.000000	1	78.
498	0	135	10	1	62.000000	1	91.
499	0	134	10	1	38.934599	1	80.

500 rows × 7 columns

- Melakukan proses *binning* (pembagian interval) untuk menentukan indeks berdasarkan count indeks

Pembagian interval sebagai berikut :

A = 81-100
B = 61-80
C = 41-60
D = 21-40
E = 0-20

In [24]:
```
bins1 = [0, 20, 40, 60, 80, 100]
group1_names = ['E', 'D', 'C', 'B', 'A']
```

In [25]:
```
df1['Indeks'] = pd.cut(df1['Count_Indeks'], bins1, labels=group1_names)
df1.to_csv('Irish Preprocessing Skenario 1.csv')
df1
```

Out[25]:

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	Type_school	Co
0	1	113	3	0	28.000000	1	65.
1	1	101	1	0	28.000000	3	60.
2	1	110	8	1	69.000000	1	83.
3	1	121	5	0	57.000000	1	82.
4	1	82	4	0	18.000000	2	46.
5	1	85	4	0	28.000000	2	52.
6	1	84	1	0	38.934599	3	57.
7	1	98	2	0	43.000000	2	65.
8	1	92	4	0	33.000000	2	58.
9	1	90	1	0	18.000000	3	50.
10	1	88	4	0	28.000000	2	53.
11	1	84	4	0	18.000000	2	47.
12	1	114	10	1	18.000000	1	61.
13	1	70	4	0	57.000000	2	59.
14	1	109	8	1	40.000000	1	69.
15	1	83	4	0	43.000000	2	58.
16	1	108	4	0	69.000000	2	82.
17	1	95	3	0	43.000000	1	64.
18	1	90	4	0	28.000000	2	54.

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	Type_school	Co
19	1	107	8	1	28.000000	1	62.
20	1	86	1	0	18.000000	3	48.
21	1	70	2	0	43.000000	2	52.
22	1	114	8	1	38.934599	1	71.
23	1	117	8	1	42.000000	1	73.
24	1	112	8	1	43.000000	1	72.
25	1	88	10	1	38.934599	1	59.
26	1	106	10	1	42.000000	1	68.
27	1	125	10	1	28.000000	1	71.
28	1	94	8	1	38.934599	1	61.
29	1	103	5	0	38.934599	1	66.
...	...	...	...	...	...	...	...
470	1	129	5	0	18.000000	1	68.
471	1	122	8	1	62.000000	1	85.
472	1	121	9	1	37.000000	1	73.
473	1	129	9	1	38.934599	1	78.
474	1	122	8	1	40.000000	1	75.
475	1	126	4	0	51.000000	2	82.
476	1	122	10	1	35.000000	1	73.
477	1	123	8	1	65.000000	1	87.
478	1	119	10	1	71.000000	1	88.
479	0	120	10	1	40.000000	1	74.

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	Type_school	Co
480	0	127	2	0	35.000000	2	75.
481	0	127	8	1	62.000000	1	87.
482	0	120	8	1	61.000000	1	84.
483	0	127	8	1	58.000000	1	86.
484	0	123	8	1	37.000000	1	74.
485	0	120	10	1	37.000000	1	73.
486	0	123	8	1	37.000000	1	74.
487	0	122	4	0	18.000000	2	65.
488	0	119	10	1	37.000000	1	72.
489	1	130	6	0	38.934599	2	78.
490	1	134	9	1	62.000000	1	91.
491	1	136	10	1	61.000000	1	91.
492	1	135	10	1	61.000000	1	91.
493	1	140	10	1	71.000000	1	98.
494	1	131	8	1	30.000000	1	74.
495	1	137	10	1	62.000000	1	92.
496	1	136	10	1	18.000000	1	71.
497	1	132	10	1	37.000000	1	78.
498	0	135	10	1	62.000000	1	91.
499	0	134	10	1	38.934599	1	80.

500 rows × 8 columns

Skenario 2

- *Binarization* pada Atribut *Sex* dan *Leaving Certificate*

In [26]:
```
biner2 = {'Sex': {'male':1, 'female':0},
          'Leaving_Certificate': {'taken':1, 'not_taken':0}}
file2_object.replace(biner2,inplace=True)
file2_object
```

Out[26]:

	Sex	Educational_level	Leaving_Certificate	Type_school
0	1	Junior_cycle_incomplete-secondary_school	0	secondary
1	1	Primary_terminal_leaver	0	primary_terminal_leaver
2	1	Senior_cycle_terminal_leaver-secondary_school	1	secondary
3	1	Junior_cycle_terminal_leaver-secondary_school	0	secondary
4	1	Junior_cycle_terminal_leaver-vocational_school	0	vocational
5	1	Junior_cycle_terminal_leaver-vocational_school	0	vocational
6	1	Primary_terminal_leaver	0	primary_terminal_leaver
7	1	Junior_cycle_incomplete-vocational_school	0	vocational
8	1	Junior_cycle_terminal_leaver-vocational_school	0	vocational
9	1	Primary_terminal_leaver	0	primary_terminal_leaver

	Sex	Educational_level	Leaving_Certificate	Type_school
10	1	Junior_cycle_terminal_leaver-vocational_school	0	vocational
11	1	Junior_cycle_terminal_leaver-vocational_school	0	vocational
12	1	3rd_level_complete	1	secondary
13	1	Junior_cycle_terminal_leaver-vocational_school	0	vocational
14	1	Senior_cycle_terminal_leaver-secondary_school	1	secondary
15	1	Junior_cycle_terminal_leaver-vocational_school	0	vocational
16	1	Junior_cycle_terminal_leaver-vocational_school	0	vocational
17	1	Junior_cycle_incomplete-secondary_school	0	secondary
18	1	Junior_cycle_terminal_leaver-vocational_school	0	vocational
19	1	Senior_cycle_terminal_leaver-secondary_school	1	secondary
20	1	Primary_terminal_leaver	0	primary_terminal_leaver
21	1	Junior_cycle_incomplete-vocational_school	0	vocational
22	1	Senior_cycle_terminal_leaver-secondary_school	1	secondary
23	1	Senior_cycle_terminal_leaver-secondary_school	1	secondary

	Sex	Educational_level	Leaving_Certificate	Type_school
24	1	Senior_cycle_terminal_leaver-secondary_school	1	secondary
25	1	3rd_level_complete	1	secondary
26	1	3rd_level_complete	1	secondary
27	1	3rd_level_complete	1	secondary
28	1	Senior_cycle_terminal_leaver-secondary_school	1	secondary
29	1	Junior_cycle_terminal_leaver-secondary_school	0	secondary
...	...	...	...	...
470	1	Junior_cycle_terminal_leaver-secondary_school	0	secondary
471	1	Senior_cycle_terminal_leaver-secondary_school	1	secondary
472	1	3rd_level_incomplete	1	secondary
473	1	3rd_level_incomplete	1	secondary
474	1	Senior_cycle_terminal_leaver-secondary_school	1	secondary
475	1	Junior_cycle_terminal_leaver-vocational_school	0	vocational
476	1	3rd_level_complete	1	secondary
477	1	Senior_cycle_terminal_leaver-secondary_school	1	secondary
478	1	3rd_level_complete	1	secondary

	Sex	Educational_level	Leaving_Certificate	Type_school
479	0	3rd_level_complete	1	secondary
480	0	Junior_cycle_incomplete-vocational_school	0	vocational
481	0	Senior_cycle_terminal_leaver-secondary_school	1	secondary
482	0	Senior_cycle_terminal_leaver-secondary_school	1	secondary
483	0	Senior_cycle_terminal_leaver-secondary_school	1	secondary
484	0	Senior_cycle_terminal_leaver-secondary_school	1	secondary
485	0	3rd_level_complete	1	secondary
486	0	Senior_cycle_terminal_leaver-secondary_school	1	secondary
487	0	Junior_cycle_terminal_leaver-vocational_school	0	vocational
488	0	3rd_level_complete	1	secondary
489	1	Senior_cycle_incomplete-vocational_school	0	vocational
490	1	3rd_level_incomplete	1	secondary
491	1	3rd_level_complete	1	secondary
492	1	3rd_level_complete	1	secondary
493	1	3rd_level_complete	1	secondary
494	1	Senior_cycle_terminal_leaver-secondary_school	1	secondary

	Sex	Educational_level	Leaving_Certificate	Type_school
495	1	3rd_level_complete	1	secondary
496	1	3rd_level_complete	1	secondary
497	1	3rd_level_complete	1	secondary
498	0	3rd_level_complete	1	secondary
499	0	3rd_level_complete	1	secondary

500 rows × 4 columns

- *Encoding Categorical Features* pada Atribut *Educational Level* dan *Type School*

In [27]:
```
kategori2 = {'Educational_level': {'3rd_level_complete':10, '3rd_level_
incomplete':9, 'Senior_cycle_terminal_leaver-secondary_school':8,
                                   'Senior_cycle_incomplete-secondary_sch
ool':7, 'Senior_cycle_incomplete-vocational_school':6,
                                    'Junior_cycle_terminal_leaver-seconda
ry_school':5, 'Junior_cycle_terminal_leaver-vocational_school':4,
                                   'Junior_cycle_incomplete-secondary_sch
ool':3,'Junior_cycle_incomplete-vocational_school':2,
                                   'Primary_terminal_leaver':1},
                       'Type_school':{'secondary':1,'vocational':2,'primar
y_terminal_leaver':3}}
file2_object.replace(kategori2,inplace=True)
file2_object
```

Out[27]:

	Sex	Educational_level	Leaving_Certificate	Type_school
0	1	3	0	1
1	1	1	0	3
2	1	8	1	1
3	1	5	0	1

	Sex	Educational_level	Leaving_Certificate	Type_school
4	1	4	0	2
5	1	4	0	2
6	1	1	0	3
7	1	2	0	2
8	1	4	0	2
9	1	1	0	3
10	1	4	0	2
11	1	4	0	2
12	1	10	1	1
13	1	4	0	2
14	1	8	1	1
15	1	4	0	2
16	1	4	0	2
17	1	3	0	1
18	1	4	0	2
19	1	8	1	1
20	1	1	0	3
21	1	2	0	2
22	1	8	1	1
23	1	8	1	1
24	1	8	1	1
25	1	10	1	1

	Sex	Educational_level	Leaving_Certificate	Type_school
26	1	10	1	1
27	1	10	1	1
28	1	8	1	1
29	1	5	0	1
...	...	...	...	...
470	1	5	0	1
471	1	8	1	1
472	1	9	1	1
473	1	9	1	1
474	1	8	1	1
475	1	4	0	2
476	1	10	1	1
477	1	8	1	1
478	1	10	1	1
479	0	10	1	1
480	0	2	0	2
481	0	8	1	1
482	0	8	1	1
483	0	8	1	1
484	0	8	1	1
485	0	10	1	1
486	0	8	1	1

	Sex	Educational_level	Leaving_Certificate	Type_school
487	0	4	0	2
488	0	10	1	1
489	1	6	0	2
490	1	9	1	1
491	1	10	1	1
492	1	10	1	1
493	1	10	1	1
494	1	8	1	1
495	1	10	1	1
496	1	10	1	1
497	1	10	1	1
498	0	10	1	1
499	0	10	1	1

500 rows × 4 columns

- *Feature Creation* sebagai Atribut Baru yang Menggambarkan Tingkat Kemampuan

Atribut baru ini dinamakan dengan atribut **indeks** dimana *values* pada atribut indeks akan diperoleh dari rumus, sebagai berikut.

$$indeks = (\frac{DVRT + Prestige}{maxValueOf(DVRT) + maxValueOf(Prestige)}) * 100$$

In [28]:
```
count2 = 0
count_indeks2 = 0
for i in file2_float:
    for j in file2_int:
```

```
        count2 = file2_float[i]+file2_int[j]
        count2_indeks = (count2/215)*100
    print(count2_indeks)
```

```
0      65.581395
1      60.000000
2      83.255814
3      82.790698
4      46.511628
5      52.558140
6      47.441860
7      65.581395
8      58.139535
9      50.232558
10     53.953488
11     47.441860
12     61.395349
13     59.069767
14     69.302326
15     58.604651
16     82.325581
17     64.186047
18     54.883721
19     62.790698
20     48.372093
21     52.558140
22     61.395349
23     73.953488
24     72.093023
25     49.302326
26     68.837209
27     71.162791
28     52.093023
29     56.279070
         ...
470    68.372093
471    85.581395
472    73.488372
473    68.372093
474    75.348837
```

```
475    82.325581
476    73.023256
477    87.441860
478    88.372093
479    74.418605
480    75.348837
481    87.906977
482    84.186047
483    86.046512
484    74.418605
485    73.023256
486    74.418605
487    65.116279
488    72.558140
489    68.837209
490    91.162791
491    91.627907
492    91.162791
493    98.139535
494    74.883721
495    92.558140
496    71.627907
497    78.604651
498    91.627907
499    70.697674
Length: 500, dtype: float64
```

In [29]:

```
raw_data2={'Sex': file2_object['Sex'].values,
           'DVRT': file2_int['DVRT'].values,
           'Educational_level': file2_object['Educational_level'].values
,
           'Leaving_Certificate' : file2_object['Leaving_Certificate'].v
alues ,
           'Prestige_score' : file2_float['Prestige_score'].values,
             'Type_school': file2_object['Type_school'].values,
           'Count Indeks': count2_indeks}
df2 = pd.DataFrame(raw_data2,
                   columns=['Sex','DVRT','Educational_level','Leaving_C
```

```
ertificate','Prestige_score','Type_school','Count Indeks'])
df2
```

Out[29]:

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	Type_school	
0	1	113	3	0	28.0	1	65.
1	1	101	1	0	28.0	3	60.
2	1	110	8	1	69.0	1	83.
3	1	121	5	0	57.0	1	82.
4	1	82	4	0	18.0	2	46.
5	1	85	4	0	28.0	2	52.
6	1	84	1	0	18.0	3	47.
7	1	98	2	0	43.0	2	65.
8	1	92	4	0	33.0	2	58.
9	1	90	1	0	18.0	3	50.
10	1	88	4	0	28.0	2	53.
11	1	84	4	0	18.0	2	47.
12	1	114	10	1	18.0	1	61.
13	1	70	4	0	57.0	2	59.
14	1	109	8	1	40.0	1	69.
15	1	83	4	0	43.0	2	58.
16	1	108	4	0	69.0	2	82.
17	1	95	3	0	43.0	1	64.
18	1	90	4	0	28.0	2	54.

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	Type_school	
19	1	107	8	1	28.0	1	62.
20	1	86	1	0	18.0	3	48.
21	1	70	2	0	43.0	2	52.
22	1	114	8	1	18.0	1	61.
23	1	117	8	1	42.0	1	73.
24	1	112	8	1	43.0	1	72.
25	1	88	10	1	18.0	1	49.
26	1	106	10	1	42.0	1	68.
27	1	125	10	1	28.0	1	71.
28	1	94	8	1	18.0	1	52.
29	1	103	5	0	18.0	1	56.
...	...	...	...	...	...	...	...
470	1	129	5	0	18.0	1	68.
471	1	122	8	1	62.0	1	85.
472	1	121	9	1	37.0	1	73.
473	1	129	9	1	18.0	1	68.
474	1	122	8	1	40.0	1	75.
475	1	126	4	0	51.0	2	82.
476	1	122	10	1	35.0	1	73.
477	1	123	8	1	65.0	1	87.
478	1	119	10	1	71.0	1	88.

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	Type_school	
479	0	120	10	1	40.0	1	74.
480	0	127	2	0	35.0	2	75.
481	0	127	8	1	62.0	1	87.
482	0	120	8	1	61.0	1	84.
483	0	127	8	1	58.0	1	86.
484	0	123	8	1	37.0	1	74.
485	0	120	10	1	37.0	1	73.
486	0	123	8	1	37.0	1	74.
487	0	122	4	0	18.0	2	65.
488	0	119	10	1	37.0	1	72.
489	1	130	6	0	18.0	2	68.
490	1	134	9	1	62.0	1	91.
491	1	136	10	1	61.0	1	91.
492	1	135	10	1	61.0	1	91.
493	1	140	10	1	71.0	1	98.
494	1	131	8	1	30.0	1	74.
495	1	137	10	1	62.0	1	92.
496	1	136	10	1	18.0	1	71.
497	1	132	10	1	37.0	1	78.
498	0	135	10	1	62.0	1	91.
499	0	134	10	1	18.0	1	70.

500 rows × 7 columns

- Melakukan proses *binning* (pembagian interval) untuk menentukan indeks berdasarkan count indeks

Pembagian interval sebagai berikut :

A = 81-100
B = 61-80
C = 41-60
D = 21-40
E = 0-20

In [30]:
```
bins2 = [0,20,40,60,80,100]
```

In [31]:
```
group2_names = ['E','D','C','B','A']
```

In [32]:
```
df2['Indeks']  = pd.cut(df2['Count Indeks'], bins2, labels=group2_names
)
df2.to_csv('Irish Preprocessing Skenario 2.csv')
df2
```

Out[32]:

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	Type_school	
0	1	113	3	0	28.0	1	65.
1	1	101	1	0	28.0	3	60.
2	1	110	8	1	69.0	1	83.
3	1	121	5	0	57.0	1	82.
4	1	82	4	0	18.0	2	46.
5	1	85	4	0	28.0	2	52.

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	Type_school	
6	1	84	1	0	18.0	3	47.
7	1	98	2	0	43.0	2	65.
8	1	92	4	0	33.0	2	58.
9	1	90	1	0	18.0	3	50.
10	1	88	4	0	28.0	2	53.
11	1	84	4	0	18.0	2	47.
12	1	114	10	1	18.0	1	61.
13	1	70	4	0	57.0	2	59.
14	1	109	8	1	40.0	1	69.
15	1	83	4	0	43.0	2	58.
16	1	108	4	0	69.0	2	82.
17	1	95	3	0	43.0	1	64.
18	1	90	4	0	28.0	2	54.
19	1	107	8	1	28.0	1	62.
20	1	86	1	0	18.0	3	48.
21	1	70	2	0	43.0	2	52.
22	1	114	8	1	18.0	1	61.
23	1	117	8	1	42.0	1	73.
24	1	112	8	1	43.0	1	72.
25	1	88	10	1	18.0	1	49.
26	1	106	10	1	42.0	1	68.

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	Type_school	
27	1	125	10	1	28.0	1	71.
28	1	94	8	1	18.0	1	52.
29	1	103	5	0	18.0	1	56.
...	...	...	...	...	...	...	...
470	1	129	5	0	18.0	1	68.
471	1	122	8	1	62.0	1	85.
472	1	121	9	1	37.0	1	73.
473	1	129	9	1	18.0	1	68.
474	1	122	8	1	40.0	1	75.
475	1	126	4	0	51.0	2	82.
476	1	122	10	1	35.0	1	73.
477	1	123	8	1	65.0	1	87.
478	1	119	10	1	71.0	1	88.
479	0	120	10	1	40.0	1	74.
480	0	127	2	0	35.0	2	75.
481	0	127	8	1	62.0	1	87.
482	0	120	8	1	61.0	1	84.
483	0	127	8	1	58.0	1	86.
484	0	123	8	1	37.0	1	74.
485	0	120	10	1	37.0	1	73.
486	0	123	8	1	37.0	1	74.

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	Type_school	
487	0	122	4	0	18.0	2	65.
488	0	119	10	1	37.0	1	72.
489	1	130	6	0	18.0	2	68.
490	1	134	9	1	62.0	1	91.
491	1	136	10	1	61.0	1	91.
492	1	135	10	1	61.0	1	91.
493	1	140	10	1	71.0	1	98.
494	1	131	8	1	30.0	1	74.
495	1	137	10	1	62.0	1	92.
496	1	136	10	1	18.0	1	71.
497	1	132	10	1	37.0	1	78.
498	0	135	10	1	62.0	1	91.
499	0	134	10	1	18.0	1	70.

500 rows × 8 columns

Skenario 3

- Mengambil nilai pada atribut *DVRT* dan *Prestige Score* untuk keperluan preprocessing

In [33]:
```
file_ps = file_copy.select_dtypes(include=['float64']).copy()
file_dvrt = file_copy.select_dtypes(include=['int64']).copy()
```

- *Binarization* pada Atribut *Sex* dan *Leaving Certificate*

In [34]:
```
biner3 = {'Sex': {'male':1, 'female':0},
          'Leaving_Certificate': {'taken':1, 'not_taken':0}}
file_copy.replace(biner3,inplace=True)
file_copy
```

Out[34]:

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	
0	1	113	Junior_cycle_incomplete-secondary_school	0	28.000000	seconc
1	1	101	Primary_terminal_leaver	0	28.000000	primary
2	1	110	Senior_cycle_terminal_leaver-secondary_school	1	69.000000	seconc
3	1	121	Junior_cycle_terminal_leaver-secondary_school	0	57.000000	seconc
4	1	82	Junior_cycle_terminal_leaver-vocational_school	0	18.000000	vocatic
5	1	85	Junior_cycle_terminal_leaver-vocational_school	0	28.000000	vocatic
6	1	84	Primary_terminal_leaver	0	38.934599	primary
7	1	98	Junior_cycle_incomplete-vocational_school	0	43.000000	vocatic
8	1	92	Junior_cycle_terminal_leaver-vocational_school	0	33.000000	vocatic
9	1	90	Primary_terminal_leaver	0	18.000000	primary
10	1	88	Junior_cycle_terminal_leaver-vocational_school	0	28.000000	vocatic

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	
11	1	84	Junior_cycle_terminal_leaver-vocational_school	0	18.000000	vocatic
12	1	114	3rd_level_complete	1	18.000000	seconc
13	1	70	Junior_cycle_terminal_leaver-vocational_school	0	57.000000	vocatic
14	1	109	Senior_cycle_terminal_leaver-secondary_school	1	40.000000	seconc
15	1	83	Junior_cycle_terminal_leaver-vocational_school	0	43.000000	vocatic
16	1	108	Junior_cycle_terminal_leaver-vocational_school	0	69.000000	vocatic
17	1	95	Junior_cycle_incomplete-secondary_school	0	43.000000	seconc
18	1	90	Junior_cycle_terminal_leaver-vocational_school	0	28.000000	vocatic
19	1	107	Senior_cycle_terminal_leaver-secondary_school	1	28.000000	seconc
20	1	86	Primary_terminal_leaver	0	18.000000	primary
21	1	70	Junior_cycle_incomplete-vocational_school	0	43.000000	vocatic
22	1	114	Senior_cycle_terminal_leaver-secondary_school	1	38.934599	seconc
23	1	117	Senior_cycle_terminal_leaver-secondary_school	1	42.000000	seconc
24	1	112	Senior_cycle_terminal_leaver-secondary_school	1	43.000000	seconc

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	
25	1	88	3rd_level_complete	1	38.934599	seconc
26	1	106	3rd_level_complete	1	42.000000	seconc
27	1	125	3rd_level_complete	1	28.000000	seconc
28	1	94	Senior_cycle_terminal_leaver-secondary_school	1	38.934599	seconc
29	1	103	Junior_cycle_terminal_leaver-secondary_school	0	38.934599	seconc
...	...	...	...	...	...	...
470	1	129	Junior_cycle_terminal_leaver-secondary_school	0	18.000000	seconc
471	1	122	Senior_cycle_terminal_leaver-secondary_school	1	62.000000	seconc
472	1	121	3rd_level_incomplete	1	37.000000	seconc
473	1	129	3rd_level_incomplete	1	38.934599	seconc
474	1	122	Senior_cycle_terminal_leaver-secondary_school	1	40.000000	seconc
475	1	126	Junior_cycle_terminal_leaver-vocational_school	0	51.000000	vocatic
476	1	122	3rd_level_complete	1	35.000000	seconc
477	1	123	Senior_cycle_terminal_leaver-secondary_school	1	65.000000	seconc
478	1	119	3rd_level_complete	1	71.000000	seconc
479	0	120	3rd_level_complete	1	40.000000	seconc

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	
480	0	127	Junior_cycle_incomplete-vocational_school	0	35.000000	vocatic
481	0	127	Senior_cycle_terminal_leaver-secondary_school	1	62.000000	seconc
482	0	120	Senior_cycle_terminal_leaver-secondary_school	1	61.000000	seconc
483	0	127	Senior_cycle_terminal_leaver-secondary_school	1	58.000000	seconc
484	0	123	Senior_cycle_terminal_leaver-secondary_school	1	37.000000	seconc
485	0	120	3rd_level_complete	1	37.000000	seconc
486	0	123	Senior_cycle_terminal_leaver-secondary_school	1	37.000000	seconc
487	0	122	Junior_cycle_terminal_leaver-vocational_school	0	18.000000	vocatic
488	0	119	3rd_level_complete	1	37.000000	seconc
489	1	130	Senior_cycle_incomplete-vocational_school	0	38.934599	vocatic
490	1	134	3rd_level_incomplete	1	62.000000	seconc
491	1	136	3rd_level_complete	1	61.000000	seconc
492	1	135	3rd_level_complete	1	61.000000	seconc
493	1	140	3rd_level_complete	1	71.000000	seconc
494	1	131	Senior_cycle_terminal_leaver-secondary_school	1	30.000000	seconc
495	1	137	3rd_level_complete	1	62.000000	seconc

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	
496	1	136	3rd_level_complete	1	18.000000	seconc
497	1	132	3rd_level_complete	1	37.000000	seconc
498	0	135	3rd_level_complete	1	62.000000	seconc
499	0	134	3rd_level_complete	1	38.934599	seconc

494 rows × 6 columns

- *Encoding Categorical Features* pada Atribut *Educational Level* dan *Type School*

In [35]:
```
kategori3 = {'Educational_level': {'3rd_level_complete':10, '3rd_level_
incomplete':9, 'Senior_cycle_terminal_leaver-secondary_school':8,
                                   'Senior_cycle_incomplete-secondary_sch
ool':7, 'Senior_cycle_incomplete-vocational_school':6,
                                    'Junior_cycle_terminal_leaver-seconda
ry_school':5, 'Junior_cycle_terminal_leaver-vocational_school':4,
                                   'Junior_cycle_incomplete-secondary_sch
ool':3,'Junior_cycle_incomplete-vocational_school':2,
                                   'Primary_terminal_leaver':1},
                       'Type_school':{'secondary':1,'vocational':2,'primar
y_terminal_leaver':3}}
file_copy.replace(kategori3,inplace=True)
file_copy
```

Out[35]:

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	Type_school
0	1	113	3	0	28.000000	1
1	1	101	1	0	28.000000	3
2	1	110	8	1	69.000000	1
3	1	121	5	0	57.000000	1

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	Type_school
4	1	82	4	0	18.000000	2
5	1	85	4	0	28.000000	2
6	1	84	1	0	38.934599	3
7	1	98	2	0	43.000000	2
8	1	92	4	0	33.000000	2
9	1	90	1	0	18.000000	3
10	1	88	4	0	28.000000	2
11	1	84	4	0	18.000000	2
12	1	114	10	1	18.000000	1
13	1	70	4	0	57.000000	2
14	1	109	8	1	40.000000	1
15	1	83	4	0	43.000000	2
16	1	108	4	0	69.000000	2
17	1	95	3	0	43.000000	1
18	1	90	4	0	28.000000	2
19	1	107	8	1	28.000000	1
20	1	86	1	0	18.000000	3
21	1	70	2	0	43.000000	2
22	1	114	8	1	38.934599	1
23	1	117	8	1	42.000000	1
24	1	112	8	1	43.000000	1
25	1	88	10	1	38.934599	1

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	Type_school
26	1	106	10	1	42.000000	1
27	1	125	10	1	28.000000	1
28	1	94	8	1	38.934599	1
29	1	103	5	0	38.934599	1
...	...	...	...	...	...	...
470	1	129	5	0	18.000000	1
471	1	122	8	1	62.000000	1
472	1	121	9	1	37.000000	1
473	1	129	9	1	38.934599	1
474	1	122	8	1	40.000000	1
475	1	126	4	0	51.000000	2
476	1	122	10	1	35.000000	1
477	1	123	8	1	65.000000	1
478	1	119	10	1	71.000000	1
479	0	120	10	1	40.000000	1
480	0	127	2	0	35.000000	2
481	0	127	8	1	62.000000	1
482	0	120	8	1	61.000000	1
483	0	127	8	1	58.000000	1
484	0	123	8	1	37.000000	1
485	0	120	10	1	37.000000	1
486	0	123	8	1	37.000000	1

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	Type_school
487	0	122	4	0	18.000000	2
488	0	119	10	1	37.000000	1
489	1	130	6	0	38.934599	2
490	1	134	9	1	62.000000	1
491	1	136	10	1	61.000000	1
492	1	135	10	1	61.000000	1
493	1	140	10	1	71.000000	1
494	1	131	8	1	30.000000	1
495	1	137	10	1	62.000000	1
496	1	136	10	1	18.000000	1
497	1	132	10	1	37.000000	1
498	0	135	10	1	62.000000	1
499	0	134	10	1	38.934599	1

494 rows × 6 columns

- *Feature Creation* sebagai Atribut Baru yang menggambarkan Tingkat Kemampuan

Atribut baru ini dinamakan dengan atribut **Count_Indeks** dimana *values* pada atribut indeks akan diperoleh dari rumus sebagai berikut :

$$CountIndeks = (\frac{DVRT+Prestige}{maxValueOf(DVRT)+maxValueOf(Prestige)}) * 100$$

In [36]:
```
count3 = 0
count3_indeks = 0
for i in file_ps:
    for j in file_dvrt:
```

```
        count3 = file_ps[i]+file_dvrt[j]
        count3_indeks = (count3/215)*100
    print(count3_indeks)
```

```
0      65.581395
1      60.000000
2      83.255814
3      82.790698
4      46.511628
5      52.558140
6      57.178883
7      65.581395
8      58.139535
9      50.232558
10     53.953488
11     47.441860
12     61.395349
13     59.069767
14     69.302326
15     58.604651
16     82.325581
17     64.186047
18     54.883721
19     62.790698
20     48.372093
21     52.558140
22     71.132372
23     73.953488
24     72.093023
25     59.039348
26     68.837209
27     71.162791
28     61.830046
29     66.016093
         ...
470    68.372093
471    85.581395
472    73.488372
473    78.109116

474    75.348837
```

```
474    75.348837
475    82.325581
476    73.023256
477    87.441860
478    88.372093
479    74.418605
480    75.348837
481    87.906977
482    84.186047
483    86.046512
484    74.418605
485    73.023256
486    74.418605
487    65.116279
488    72.558140
489    78.574232
490    91.162791
491    91.627907
492    91.162791
493    98.139535
494    74.883721
495    92.558140
496    71.627907
497    78.604651
498    91.627907
499    80.434697
Length: 494, dtype: float64
```

In [37]:

```
raw_data3={'Sex': file_copy['Sex'].values,
           'DVRT': file_copy['DVRT'].values,
           'Educational_level': file_copy['Educational_level'].values,
           'Leaving_Certificate' : file_copy['Leaving_Certificate'].valu
es ,
           'Prestige_score' : file_copy['Prestige_score'].values,
             'Type_school': file_copy['Type_school'].values,
           'Count_Indeks': count3_indeks}
df3 = pd.DataFrame(raw_data3, columns=['Sex','DVRT','Educational_level'
,'Leaving_Certificate','Prestige_score','Type_school','Count_Indeks'])
df3
```

Out[37]:

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	Type_school	Co
0	1	113	3	0	28.000000	1	65.
1	1	101	1	0	28.000000	3	60.
2	1	110	8	1	69.000000	1	83.
3	1	121	5	0	57.000000	1	82.
4	1	82	4	0	18.000000	2	46.
5	1	85	4	0	28.000000	2	52.
6	1	84	1	0	38.934599	3	57.
7	1	98	2	0	43.000000	2	65.
8	1	92	4	0	33.000000	2	58.
9	1	90	1	0	18.000000	3	50.
10	1	88	4	0	28.000000	2	53.
11	1	84	4	0	18.000000	2	47.
12	1	114	10	1	18.000000	1	61.
13	1	70	4	0	57.000000	2	59.
14	1	109	8	1	40.000000	1	69.
15	1	83	4	0	43.000000	2	58.
16	1	108	4	0	69.000000	2	82.
17	1	95	3	0	43.000000	1	64.
18	1	90	4	0	28.000000	2	54.
19	1	107	8	1	28.000000	1	62.
20	1	86	1	0	18.000000	3	48.
21	1	70	2	0	43.000000	2	52.

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	Type_school	Co
22	1	114	8	1	38.934599	1	71.
23	1	117	8	1	42.000000	1	73.
24	1	112	8	1	43.000000	1	72.
25	1	88	10	1	38.934599	1	59.
26	1	106	10	1	42.000000	1	68.
27	1	125	10	1	28.000000	1	71.
28	1	94	8	1	38.934599	1	61.
29	1	103	5	0	38.934599	1	66.
...	...	...	...	...	...	...	...
470	1	129	5	0	18.000000	1	68.
471	1	122	8	1	62.000000	1	85.
472	1	121	9	1	37.000000	1	73.
473	1	129	9	1	38.934599	1	78.
474	1	122	8	1	40.000000	1	75.
475	1	126	4	0	51.000000	2	82.
476	1	122	10	1	35.000000	1	73.
477	1	123	8	1	65.000000	1	87.
478	1	119	10	1	71.000000	1	88.
479	0	120	10	1	40.000000	1	74.
480	0	127	2	0	35.000000	2	75.
481	0	127	8	1	62.000000	1	87.
482	0	120	8	1	61.000000	1	84.

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	Type_school	Co
483	0	127	8	1	58.000000	1	86.
484	0	123	8	1	37.000000	1	74.
485	0	120	10	1	37.000000	1	73.
486	0	123	8	1	37.000000	1	74.
487	0	122	4	0	18.000000	2	65.
488	0	119	10	1	37.000000	1	72.
489	1	130	6	0	38.934599	2	78.
490	1	134	9	1	62.000000	1	91.
491	1	136	10	1	61.000000	1	91.
492	1	135	10	1	61.000000	1	91.
493	1	140	10	1	71.000000	1	98.
494	1	131	8	1	30.000000	1	74.
495	1	137	10	1	62.000000	1	92.
496	1	136	10	1	18.000000	1	71.
497	1	132	10	1	37.000000	1	78.
498	0	135	10	1	62.000000	1	91.
499	0	134	10	1	38.934599	1	80.

494 rows × 7 columns

- Diskretisasi dengan menggunakan *Equal Interval*

In [38]:
```
bins3 = [0,20,40,60,80,100]
group3_names=['E','D','C','B','A']
```

```
df3['Index']=pd.cut(df3['Count_Indeks'], bins3, labels=group3_names)
df3.to_csv('Irish Preprocessing Skenario 3.csv')
df3
```

Out[38]:

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	Type_school	Co
0	1	113	3	0	28.000000	1	65.
1	1	101	1	0	28.000000	3	60.
2	1	110	8	1	69.000000	1	83.
3	1	121	5	0	57.000000	1	82.
4	1	82	4	0	18.000000	2	46.
5	1	85	4	0	28.000000	2	52.
6	1	84	1	0	38.934599	3	57.
7	1	98	2	0	43.000000	2	65.
8	1	92	4	0	33.000000	2	58.
9	1	90	1	0	18.000000	3	50.
10	1	88	4	0	28.000000	2	53.
11	1	84	4	0	18.000000	2	47.
12	1	114	10	1	18.000000	1	61.
13	1	70	4	0	57.000000	2	59.
14	1	109	8	1	40.000000	1	69.
15	1	83	4	0	43.000000	2	58.
16	1	108	4	0	69.000000	2	82.
17	1	95	3	0	43.000000	1	64.
18	1	90	4	0	28.000000	2	54.

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	Type_school	Co
19	1	107	8	1	28.000000	1	62.
20	1	86	1	0	18.000000	3	48.
21	1	70	2	0	43.000000	2	52.
22	1	114	8	1	38.934599	1	71.
23	1	117	8	1	42.000000	1	73.
24	1	112	8	1	43.000000	1	72.
25	1	88	10	1	38.934599	1	59.
26	1	106	10	1	42.000000	1	68.
27	1	125	10	1	28.000000	1	71.
28	1	94	8	1	38.934599	1	61.
29	1	103	5	0	38.934599	1	66.
...	...	...	...	...	...	...	...
470	1	129	5	0	18.000000	1	68.
471	1	122	8	1	62.000000	1	85.
472	1	121	9	1	37.000000	1	73.
473	1	129	9	1	38.934599	1	78.
474	1	122	8	1	40.000000	1	75.
475	1	126	4	0	51.000000	2	82.
476	1	122	10	1	35.000000	1	73.
477	1	123	8	1	65.000000	1	87.
478	1	119	10	1	71.000000	1	88.
479	0	120	10	1	40.000000	1	74.

	Sex	DVRT	Educational_level	Leaving_Certificate	Prestige_score	Type_school	Co
480	0	127	2	0	35.000000	2	75.
481	0	127	8	1	62.000000	1	87.
482	0	120	8	1	61.000000	1	84.
483	0	127	8	1	58.000000	1	86.
484	0	123	8	1	37.000000	1	74.
485	0	120	10	1	37.000000	1	73.
486	0	123	8	1	37.000000	1	74.
487	0	122	4	0	18.000000	2	65.
488	0	119	10	1	37.000000	1	72.
489	1	130	6	0	38.934599	2	78.
490	1	134	9	1	62.000000	1	91.
491	1	136	10	1	61.000000	1	91.
492	1	135	10	1	61.000000	1	91.
493	1	140	10	1	71.000000	1	98.
494	1	131	8	1	30.000000	1	74.
495	1	137	10	1	62.000000	1	92.
496	1	136	10	1	18.000000	1	71.
497	1	132	10	1	37.000000	1	78.
498	0	135	10	1	62.000000	1	91.
499	0	134	10	1	38.934599	1	80.

494 rows × 8 columns

HASIL IMPLEMENTASI

Dari 3 skenario diatas, maka hasil implementasi preprocessing dapat dilihat perbedaannya dengan melihat perbedaan tiap indeks pada tiap record yang ada pada tiap skenario, seperti berikut.

In [39]:
```
skenario1 = pd.read_csv('Irish Preprocessing Skenario 1.csv')
skenario2 = pd.read_csv('Irish Preprocessing Skenario 2.csv')
skenario3 = pd.read_csv('Irish Preprocessing Skenario 3.csv')
```

In [40]:
```
indeks1 = skenario1['Indeks'].values
indeks2 = skenario2['Indeks'].values
indeks3 = skenario3['Index'].values
```

In [41]:
```
nilai12 = 0
diff12 = 0
for i in range(len(indeks1)):
            if indeks1[i] == indeks2[i] :
                nilai12 += 1
            else:
                diff12+= 1
print('Jumlah indeks sama pada skenario 1 dan 2: ',nilai12)
print('Jumlah indeks berbeda pada skenario 1 dan 2: ',diff12)
```

```
Jumlah indeks sama pada skenario 1 dan 2:  489
Jumlah indeks berbeda pada skenario 1 dan 2:  11
```

In [42]:
```
nilai13 = 0
nilai23 = 0
diff13 = 0
diff23=0
for i in range(len(indeks3)):
            if indeks1[i] == indeks3[i] :
                nilai13 += 1
```

```
            elif indeks1[i] != indeks3[i]:
                diff13 += 1
            if indeks2[i] == indeks3[i]:
                nilai23 += 1
            elif indeks2[i] != indeks3[i]:
                diff23 += 1
print('Jumlah indeks sama pada skenario 1 dan 3: ',nilai13)
print('Jumlah indeks berbeda pada skenario 1 dan 3: ',diff13)
print()
print('Jumlah indeks sama pada skenario 2 dan 3: ',nilai23)
print('Jumlah indeks berbeda pada skenario 2 dan 3: ',diff23)
```

```
Jumlah indeks sama pada skenario 1 dan 3:  273
Jumlah indeks berbeda pada skenario 1 dan 3:  221

Jumlah indeks sama pada skenario 2 dan 3:  269
Jumlah indeks berbeda pada skenario 2 dan 3:  225
```

KESIMPULAN

Dari hasil perhitungan diatas, maka dapat dilihat bahwa skenario 1 dan 2 memiliki hasil indeks yang cukup mirip, dengan perbedaan hanya terdapat pada 11 record. Sedangkan untuk perbandingan kedua skenario dengan skenario 3, skenario 1 dan 2 memiliki hasil indeks yang cukup berbeda dengan skenario 3. Sehingga, dapat dikatakan bahwa dengan metode preprocessing yang dilakukan akan memiliki hasil yang tidak terlalu berbeda apabila diterapkan skenario mean dan modus pada atribut, tanpa melakukan ignore pada missing values seperti pada skenario ke-3.