Rakamin Academy

Kelompok Data Lab

1. Oktavian Dwi Putra
2. Muhammad Ilham
3. Maulid Uly Mada
4. Seto Bhanu Adyatma
5. Mega Pangastuti
6. Rasyidah Maulida P. A
7. M. Hasbi Ashshiddiqi
8. Raihan Damar

1. Data Cleansing

1. Data Cleansing (50 poin)

A. Handle missing values

```
Jumlah nilai kosong pada masing-masing kolom :
Age                         0
Attrition                   0
BusinessTravel              0
DailyRate                   0
Department                  0
DistanceFromHome            0
Education                   0
EducationField              0
EmployeeCount               0
EmployeeNumber              0
EnvironmentSatisfaction     0
Gender                      0
HourlyRate                  0
JobInvolvement              0
JobLevel                    0
JobRole                     0
JobSatisfaction             0
MaritalStatus               0
MonthlyIncome               0
MonthlyRate                 0
NumCompaniesWorked          0
Over18                      0
OverTime                    0
PercentSalaryHike           0
PerformanceRating           0
RelationshipSatisfaction    0
StandardHours               0
StockOptionLevel            0
TotalWorkingYears           0
TrainingTimesLastYear       0
WorkLifeBalance             0
YearsAtCompany              0
YearsInCurrentRole          0
YearsSinceLastPromotion     0
YearsWithCurrManager        0
dtype: int64
```

Berdasarkan informasi di samping dapat disimpulkan bahwa **tidak terdapat missing value** pada semua kolom yang ada sehingga tidak diperlukan proses penghapusan *missing values*.

1. Data Cleansing (50 poin)

B. Handle data duplicated

```
# Menampilkan jumlah data duplicate
print("Jumlah data duplikat :")
df.duplicated().sum()
```

```
Jumlah data duplikat :
0
```

Berdasarkan informasi di samping dapat disimpulkan bahwa **tidak terdapat data yang terduplikasi** sehingga tidak diperlukan proses penghapusan *duplicated data*.

1. Data Cleansing (50 poin)

C. Handle outliers

```
# Menghapus outlier
from scipy import stats

print(f'Jumlah baris sebelum dilakukan penghapusan data outliers : {len(df)}')

filtered_entries = np.array([True] * len(df))

for col in ['MonthlyIncome', 'NumCompaniesWorked', 'TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany', 'YearsInCurrentRole', 'YearsSinceLastPromotion', 'YearsWithCurrManager']:
    zscore = abs(stats.zscore(df[col])) # hitung absolute z-scorenya
    filtered_entries = (zscore < 3) & filtered_entries # keep yang kurang dari 3 absolute z-scorenya

df = df[filtered_entries] # filter, cuma ambil yang z-scorenya dibawah 3

print(f'Jumlah baris setelah dilakukan penghapusan data outliers : {len(df)}')
```

```
Jumlah baris sebelum dilakukan penghapusan data outliers : 1470
Jumlah baris setelah dilakukan penghapusan data outliers : 1387
```

- Dilakukan proses penghapusan data outliers pada kolom **MonthlyIncome, NumCompaniesWorked, TotalWorkingYears, TrainingTimesLastYear, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion,** dan **YearsWithCurrManager.**
- Berdasarkan informasi di atas, didapatkan bahwa sebelum dilakukan penghapusan data *outliers* terdapat **1470** baris data dan setelah dilakukan penghapusan data *outliers* terdapat **1387** baris data.

1. Data Cleansing (50 poin)

D. Feature encoding

Melakukan feature encoding pada kolom **Attrition, BusinessTravel, Gender,** dan **Overtime** menjadi **numerical** dengan menggunakan metode label encoding sehingga model lebih mudah dipahami.

```
# Melakukan label encoding pada kolom bertipe kategorik
df['Attrition'] = LabelEncoder().fit_transform(df['Attrition'])
df['BusinessTravel'] = LabelEncoder().fit_transform(df['BusinessTravel'])
df['Gender'] = LabelEncoder().fit_transform(df['Gender'])
df['OverTime'] = LabelEncoder().fit_transform(df['OverTime'])

# Menampilkan data setelah dilakukan proses label encoding
df.head(3)
```

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	EnvironmentSatisfaction	Gender	HourlyRate	JobInvolvement	JobLevel	JobRole	JobSatisfaction	MaritalSta
0	41	1	2	1102	Sales	1	2	Life Sciences	1	1	2	0	94	3	2	Sales Executive	4	Si
1	49	0	1	279	Research & Development	8	1	Life Sciences	1	2	3	1	61	2	2	Research Scientist	2	Mar
2	37	1	2	1373	Research & Development	2	2	Other	1	4	4	1	92	2	1	Laboratory Technician	3	Si

1. Data Cleansing (50 poin)

D. Feature encoding - One hot encoding

Melakukan *feature encoding* pada kolom yang memiliki lebih dari dua tingkatan/kelas yakni kolom **Attrition, BusinessTravel, Gender, Overtime,** dan **Over18** dengan menggunakan metode **One hot encoding** sehingga memunculkan seluruh tingkatan/kelas menjadi feature yang baru.

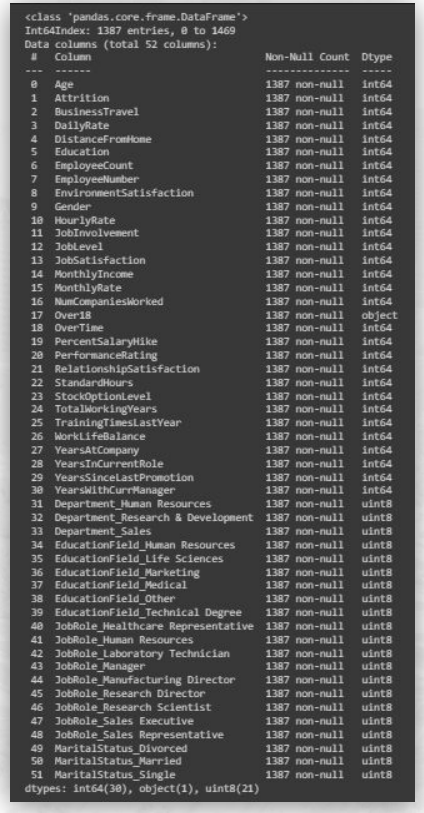
```
# Melakukan one-hot encoding kolom kategorik
for cat in categorical:
  if cat not in ['Attrition', 'BusinessTravel', 'Gender', 'OverTime', 'Over18']:
    df1 = pd.get_dummies(df[cat], prefix=cat)
    df  = df.drop(cat, axis = 1)
    df  = df.join(df1)

# Menampilkan data setelah dilakukan proses one-hot encoding
df.head(3)
```

	Age	Attrition	BusinessTravel	DailyRate	DistanceFromHome	Education	EmployeeCount	EmployeeNumber	EnvironmentSatisfaction	Gender	HourlyRate	JobInvolvement	JobLevel	JobSatisfaction	MonthlyIncome	MonthlyRate	NumCompaniesWorked	O
0	41	1	2	1102	1	2	1	1	2	0	94	3	2	4	5993	19479	8	
1	49	0	1	279	8	1	1	2	3	1	61	2	2	2	5130	24907	1	
2	37	1	2	1373	2	2	1	4	4	1	92	2	1	3	2090	2396	6	

1. Data Cleansing (50 poin)

D. Feature encoding

Setelah dilakukan proses *feature encoding* terdapat peningkatan jumlah kolom menjadi 51 kolom, yang sebelumnya berjumlah 34 kolom. Peningkatan jumlah kolom ini disebabkan oleh penggunaan metode One Hot Encoding yang nantinya kelas/tingkatan pada kolom tersebut menjadi kolom baru seperti kolom MaritalStatus_Divorced, MaritalStatus_Married, MaritalStatus_Single, dll.



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1387 entries, 0 to 1469
Data columns (total 52 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Age                                   1387 non-null   int64
 1   Attrition                             1387 non-null   int64
 2   BusinessTravel                        1387 non-null   int64
 3   DailyRate                             1387 non-null   int64
 4   DistanceFromHome                      1387 non-null   int64
 5   Education                             1387 non-null   int64
 6   EmployeeCount                         1387 non-null   int64
 7   EmployeeNumber                        1387 non-null   int64
 8   EnvironmentSatisfaction               1387 non-null   int64
 9   Gender                                1387 non-null   int64
 10  HourlyRate                            1387 non-null   int64
 11  JobInvolvement                        1387 non-null   int64
 12  JobLevel                              1387 non-null   int64
 13  JobSatisfaction                       1387 non-null   int64
 14  MonthlyIncome                         1387 non-null   int64
 15  MonthlyRate                           1387 non-null   int64
 16  NumCompaniesWorked                    1387 non-null   int64
 17  Over18                                1387 non-null   object
 18  OverTime                              1387 non-null   int64
 19  PercentSalaryHike                     1387 non-null   int64
 20  PerformanceRating                     1387 non-null   int64
 21  RelationshipSatisfaction              1387 non-null   int64
 22  StandardHours                         1387 non-null   int64
 23  StockOptionLevel                      1387 non-null   int64
 24  TotalWorkingYears                     1387 non-null   int64
 25  TrainingTimesLastYear                 1387 non-null   int64
 26  WorkLifeBalance                       1387 non-null   int64
 27  YearsAtCompany                        1387 non-null   int64
 28  YearsInCurrentRole                    1387 non-null   int64
 29  YearsSinceLastPromotion               1387 non-null   int64
 30  YearsWithCurrManager                  1387 non-null   int64
 31  Department_Human Resources            1387 non-null   uint8
 32  Department_Research & Development     1387 non-null   uint8
 33  Department_Sales                      1387 non-null   uint8
 34  EducationField_Human Resources        1387 non-null   uint8
 35  EducationField_Life Sciences          1387 non-null   uint8
 36  EducationField_Marketing              1387 non-null   uint8
 37  EducationField_Medical                1387 non-null   uint8
 38  EducationField_Other                  1387 non-null   uint8
 39  EducationField_Technical Degree       1387 non-null   uint8
 40  JobRole_Healthcare Representative     1387 non-null   uint8
 41  JobRole_Human Resources               1387 non-null   uint8
 42  JobRole_Laboratory Technician         1387 non-null   uint8
 43  JobRole_Manager                       1387 non-null   uint8
 44  JobRole_Manufacturing Director        1387 non-null   uint8
 45  JobRole_Research Director             1387 non-null   uint8
 46  JobRole_Research Scientist            1387 non-null   uint8
 47  JobRole_Sales Executive               1387 non-null   uint8
 48  JobRole_Sales Representative          1387 non-null   uint8
 49  MaritalStatus_Divorced                1387 non-null   uint8
 50  MaritalStatus_Married                 1387 non-null   uint8
 51  MaritalStatus_Single                  1387 non-null   uint8
dtypes: int64(30), object(1), uint8(21)
```

2. Feature Engineering

2. Feature Engineering (35 poin)

A. Feature selection

```
# Melakukan penghapusan kolom dengan nilai unik 1 dan nilai unik yang sama dengan jumlah baris
df_new = df.drop(columns = ['EmployeeNumber', 'EmployeeCount', 'StandardHours', 'Over18'])
df_new.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1387 entries, 0 to 1469
Data columns (total 48 columns):
 #   Column                             Non-Null Count  Dtype
---  ------                             --------------  -----
 0   Age                                1387 non-null   int64
 1   Attrition                          1387 non-null   int64
 2   BusinessTravel                     1387 non-null   int64
 3   DailyRate                          1387 non-null   int64
 4   DistanceFromHome                   1387 non-null   int64
 5   Education                          1387 non-null   int64
 6   EnvironmentSatisfaction            1387 non-null   int64
 7   Gender                             1387 non-null   int64
 8   HourlyRate                         1387 non-null   int64
 9   JobInvolvement                     1387 non-null   int64
 10  JobLevel                           1387 non-null   int64
 11  JobSatisfaction                    1387 non-null   int64
 12  MonthlyIncome                      1387 non-null   int64
 13  MonthlyRate                        1387 non-null   int64
 14  NumCompaniesWorked                 1387 non-null   int64
 15  OverTime                           1387 non-null   int64
 16  PercentSalaryHike                  1387 non-null   int64
 17  PerformanceRating                  1387 non-null   int64
 18  RelationshipSatisfaction           1387 non-null   int64
 19  StockOptionLevel                   1387 non-null   int64
 20  TotalWorkingYears                  1387 non-null   int64
 21  TrainingTimesLastYear              1387 non-null   int64
 22  WorkLifeBalance                    1387 non-null   int64
 23  YearsAtCompany                     1387 non-null   int64
 24  YearsInCurrentRole                 1387 non-null   int64
 25  YearsSinceLastPromotion            1387 non-null   int64
 26  YearsWithCurrManager               1387 non-null   int64
 27  Department_Human Resources         1387 non-null   uint8
 28  Department_Research & Development  1387 non-null   uint8
 29  Department_Sales                   1387 non-null   uint8
 30  EducationField_Human Resources     1387 non-null   uint8
 31  EducationField_Life Sciences       1387 non-null   uint8
 32  EducationField_Marketing           1387 non-null   uint8
 33  EducationField_Medical             1387 non-null   uint8
 34  EducationField_Other               1387 non-null   uint8
 35  EducationField_Technical Degree    1387 non-null   uint8
 36  JobRole_Healthcare Representative  1387 non-null   uint8
 37  JobRole_Human Resources            1387 non-null   uint8
 38  JobRole_Laboratory Technician      1387 non-null   uint8
 39  JobRole_Manager                    1387 non-null   uint8
 40  JobRole_Manufacturing Director     1387 non-null   uint8
 41  JobRole_Research Director          1387 non-null   uint8
 42  JobRole_Research Scientist         1387 non-null   uint8
 43  JobRole_Sales Executive            1387 non-null   uint8
 44  JobRole_Sales Representative       1387 non-null   uint8
 45  MaritalStatus_Divorced             1387 non-null   uint8
 46  MaritalStatus_Married              1387 non-null   uint8
 47  MaritalStatus_Single               1387 non-null   uint8
dtypes: int64(27), uint8(21)
memory usage: 364.1 KB
```

Setelah dilakukan proses *feature encoding*, selanjutnya dilakukan penghapusan kolom yang memiliki nilai unik sebanyak 1 dan kolom yang memiliki nilai unik sebanyak jumlah baris. Kemudian terjadi penurunan jumlah kolom menjadi 48 kolom.

2. Feature Engineering (35 poin) - Ilham

A. Feature selection

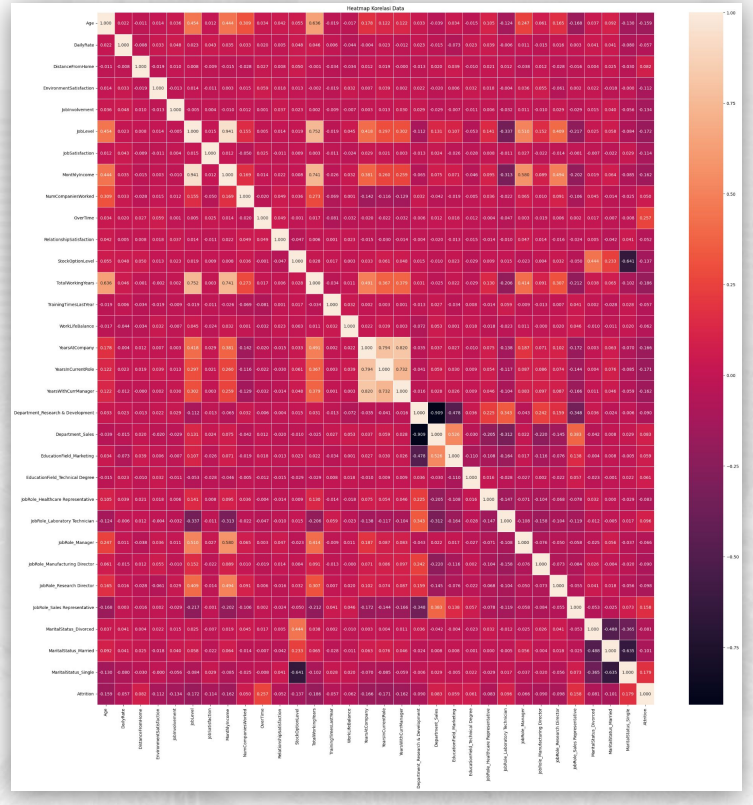
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1387 entries, 0 to 1469
Data columns (total 32 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Age                                   1387 non-null   int64
 1   Attrition                             1387 non-null   int64
 2   DailyRate                             1387 non-null   int64
 3   DistanceFromHome                      1387 non-null   int64
 4   EnvironmentSatisfaction               1387 non-null   int64
 5   JobInvolvement                        1387 non-null   int64
 6   JobLevel                              1387 non-null   int64
 7   JobSatisfaction                       1387 non-null   int64
 8   MonthlyIncome                         1387 non-null   int64
 9   NumCompaniesWorked                    1387 non-null   int64
 10  OverTime                              1387 non-null   int64
 11  RelationshipSatisfaction              1387 non-null   int64
 12  StockOptionLevel                      1387 non-null   int64
 13  TotalWorkingYears                     1387 non-null   int64
 14  TrainingTimesLastYear                 1387 non-null   int64
 15  WorkLifeBalance                       1387 non-null   int64
 16  YearsAtCompany                        1387 non-null   int64
 17  YearsInCurrentRole                    1387 non-null   int64
 18  YearsWithCurrManager                  1387 non-null   int64
 19  Department_Research & Development     1387 non-null   uint8
 20  Department_Sales                      1387 non-null   uint8
 21  EducationField_Marketing              1387 non-null   uint8
 22  EducationField_Technical Degree       1387 non-null   uint8
 23  JobRole_Healthcare Representative     1387 non-null   uint8
 24  JobRole_Laboratory Technician         1387 non-null   uint8
 25  JobRole_Manager                       1387 non-null   uint8
 26  JobRole_Manufacturing Director        1387 non-null   uint8
 27  JobRole_Research Director             1387 non-null   uint8
 28  JobRole_Sales Representative          1387 non-null   uint8
 29  MaritalStatus_Divorced                1387 non-null   uint8
 30  MaritalStatus_Married                 1387 non-null   uint8
 31  MaritalStatus_Single                  1387 non-null   uint8
dtypes: int64(19), uint8(13)
memory usage: 266.6 KB
```

```python
# Menghilangkan kolom yang memiliki nilai korelasi rendah terhadap target (korelasi dibawah 0.05)
# Kolom target yang akan digunakan sebagai referensi
target_column = 'Attrition'

# Hitung korelasi terhadap kolom target
correlation_with_target = df_new.corr()[target_column].drop(target_column)

# Hilangkan kolom-kolom dengan korelasi di bawah 0.05 terhadap kolom target
threshold = 0.05
columns_to_drop = correlation_with_target[abs(correlation_with_target) < threshold].index
data_filtered = df_new.drop(columns=columns_to_drop)
```

Selanjutnya dilakukan penghapusan kolom yang memiliki nilai korelasi yang rendah terhadap target (korelasi dibawah 0.05). Kemudian terjadi penurunan jumlah kolom menjadi 32 kolom.

2. Feature Engineering (35 poin)

A. Feature selection



Dilakukan proses analisa menggunakan heatmap untuk melihat korelasi antar feature. Kemudian karena terdapat **multicollinearity** (korelasi yang memiliki nilai diatas 0.7) antara kolom **“JobLevel”, “MonthlyIncome”,** dan **“TotalWorkingYears”** serta kolom **“YearsAtCompany”, “YearsInCurrentRole”,** dan **“YearsWithCurrManager”** sehingga dipilih salah satu kolom saja.

2. Feature Engineering (35 poin)

A. Feature selection

```
data_filtered = data_filtered.drop(['JobLevel', 'MonthlyIncome', 'YearsAtCompany', 'YearsWithCurrManager'], axis = 1)
data_filtered.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1387 entries, 0 to 1469
Data columns (total 28 columns):
 #   Column                             Non-Null Count  Dtype
---  ------                             --------------  -----
 0   Age                                1387 non-null   int64
 1   DailyRate                          1387 non-null   int64
 2   DistanceFromHome                   1387 non-null   int64
 3   EnvironmentSatisfaction            1387 non-null   int64
 4   JobInvolvement                     1387 non-null   int64
 5   JobSatisfaction                    1387 non-null   int64
 6   NumCompaniesWorked                 1387 non-null   int64
 7   OverTime                           1387 non-null   int64
 8   RelationshipSatisfaction           1387 non-null   int64
 9   StockOptionLevel                   1387 non-null   int64
 10  TotalWorkingYears                  1387 non-null   int64
 11  TrainingTimesLastYear              1387 non-null   int64
 12  WorkLifeBalance                    1387 non-null   int64
 13  YearsInCurrentRole                 1387 non-null   int64
 14  Department_Research & Development  1387 non-null   uint8
 15  Department_Sales                   1387 non-null   uint8
 16  EducationField_Marketing           1387 non-null   uint8
 17  EducationField_Technical Degree    1387 non-null   uint8
 18  JobRole_Healthcare Representative  1387 non-null   uint8
 19  JobRole_Laboratory Technician      1387 non-null   uint8
 20  JobRole_Manager                    1387 non-null   uint8
 21  JobRole_Manufacturing Director     1387 non-null   uint8
 22  JobRole_Research Director          1387 non-null   uint8
 23  JobRole_Sales Representative       1387 non-null   uint8
 24  MaritalStatus_Divorced             1387 non-null   uint8
 25  MaritalStatus_Married              1387 non-null   uint8
 26  MaritalStatus_Single               1387 non-null   uint8
 27  Attrition                          1387 non-null   int64
dtypes: int64(15), uint8(13)
memory usage: 223.3 KB
```

Karena terdapat **multicollinearity**, dilakukan penghapusan kolom yang memiliki korelasi lebih rendah terhadap target yaitu **“JobLevel”** dan **“MonthlyIncome”** serta kolom **“YearsAtCompany”** dan **“YearsWithCurrManager”**.

Kemudian terjadi penurunan jumlah kolom menjadi 28 kolom.

2. Feature Engineering (35 poin)

B. Feature extraction

```
# Membuat fitur baru
data_filtered['SatisfactionMean'] = ( df_new['EnvironmentSatisfaction'] + df_new['JobSatisfaction'] + df_new['RelationshipSatisfaction'] ) / 3
data_filtered['IncomeTotalYears'] = df_new['MonthlyIncome']*df_new['TotalWorkingYears']
data_filtered['JobLevelSatisfaction'] = df_new['JobLevel'] * df_new['JobSatisfaction']
data_filtered['YearsInCurrentRole_JobSatisfaction'] = df_new['YearsInCurrentRole'] * df_new['JobSatisfaction']
data_filtered['StockOption_JobSatisfaction'] = df_new['StockOptionLevel'] * df_new['JobSatisfaction']
```

Selanjutnya dilakukan pembuatan feature baru dengan rincian sebagai berikut :

1. **“SatisfactionMean”** yaitu hasil rataan dari kolom **satisfaction.**
2. **“IncomeTotalYears”** yaitu hasil kali antara kolom **“MonthlyIncome”** dengan **“TotalWorkingYears”.**
3. **“JobLevelSatisfaction”** yaitu hasil kali antara kolom **“JobLevel”** dengan **“JobSatisfaction”.**
4. **“YearsInCurrentRole_JobSatisfaction”** yaitu hasil kali antara kolom **“YearsInCurrentRole”** dengan **“JobSatisfaction”.**
5. **“StockOption_JobSatisfaction”** yaitu hasil kali antara **“StockOptionLevel”** dengan **“JobSatisfaction”.**

2. Feature Engineering (35 poin)

B. Feature extraction

Setelah proses pembuatan *feature* baru, dilakukan pengecekan korelasi feature baru terhadap target dengan menggunakan heatmap. Hasil korelasi yang didapatkan dari feature baru berkisar antara 0.13 sampai 0.18. Kemudin, terjadi peningkatan jumlah kolom menjadi 33 kolom.

2. Feature Engineering (35 poin)

B. Feature extraction

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1387 entries, 0 to 1469
Data columns (total 27 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Age                                   1387 non-null   int64
 1   DailyRate                             1387 non-null   int64
 2   DistanceFromHome                      1387 non-null   int64
 3   JobInvolvement                        1387 non-null   int64
 4   NumCompaniesWorked                    1387 non-null   int64
 5   OverTime                              1387 non-null   int64
 6   StockOptionLevel                      1387 non-null   int64
 7   TotalWorkingYears                     1387 non-null   int64
 8   TrainingTimesLastYear                 1387 non-null   int64
 9   WorkLifeBalance                       1387 non-null   int64
 10  YearsInCurrentRole                    1387 non-null   int64
 11  Department_Research & Development     1387 non-null   uint8
 12  Department_Sales                      1387 non-null   uint8
 13  EducationField_Marketing              1387 non-null   uint8
 14  EducationField_Technical Degree       1387 non-null   uint8
 15  JobRole_Healthcare Representative     1387 non-null   uint8
 16  JobRole_Laboratory Technician         1387 non-null   uint8
 17  JobRole_Manager                       1387 non-null   uint8
 18  JobRole_Manufacturing Director        1387 non-null   uint8
 19  JobRole_Research Director             1387 non-null   uint8
 20  JobRole_Sales Representative          1387 non-null   uint8
 21  MaritalStatus_Divorced                1387 non-null   uint8
 22  MaritalStatus_Married                 1387 non-null   uint8
 23  MaritalStatus_Single                  1387 non-null   uint8
 24  SatisfactionMean                      1387 non-null   float64
 25  JobLevelSatisfaction                  1387 non-null   int64
 26  Attrition                             1387 non-null   int64
dtypes: float64(1), int64(13), uint8(13)
memory usage: 212.4 KB
```

```
# Menghapus kolom yang memiliki korelasi lebih rendah dengan target
data_baru = data_filtered.drop(columns = ['EnvironmentSatisfaction', 'JobSatisfaction',
                                          'RelationshipSatisfaction', 'IncomeTotalYears',
                                          'YearsInCurrentRole_JobSatisfaction', 'StockOption_JobSatisfaction'])
```

Selanjutnya dilakukan penghapusan kolom yang dipergunakan untuk membuat *feature* baru dimana *feature* tersebut memiliki korelasi yang lebih rendah jika dibandingkan dengan korelasi feature baru terhadap target.

Kemudian, juga dilakukan penghapusan terhadap *feature* baru yang memiliki korelasi yang lebih rendah dibandingkan dengan korelasi kolom yang digunakan untuk membuat feature baru tersebut terhadap target.

Selanjutnya terjadi penurunan jumlah kolom menjadi 27 kolom.

2. Feature Engineering (35 poin)

C. Feature transformation

	Age	DailyRate	DistanceFromHome	JobInvolvement	NumCompaniesWorked	OverTime	StockOptionLevel	TotalWorkingYears	TrainingTimesLastYear	WorkLifeBalance	YearsInCurrentRole
count	1.387000e+03	1.387000e+03	1.387000e+03	1.387000e+03	1.387000e+03	1387.000000	1.387000e+03	1.387000e+03	1.387000e+03	1.387000e+03	1.387000e+03
mean	-3.201797e-16	-3.073725e-17	-7.684312e-18	-1.511248e-16	1.018171e-16	0.284787	4.098300e-17	-8.452743e-17	1.793006e-17	1.562477e-16	5.122875e-18
std	1.000361e+00	1.000361e+00	1.000361e+00	1.000361e+00	1.000361e+00	0.451476	1.000361e+00	1.000361e+00	1.000361e+00	1.000361e+00	1.000361e+00
min	-2.066430e+00	-1.746955e+00	-1.011249e+00	-2.444483e+00	-1.072199e+00	0.000000	-9.294273e-01	-1.538219e+00	-2.171420e+00	-2.501172e+00	-1.192419e+00
25%	-7.082852e-01	-8.400474e-01	-8.875735e-01	-1.034126e+00	-6.724300e-01	0.000000	-9.294273e-01	-6.423508e-01	-6.225595e-01	-1.082307e+00	-5.757580e-01
50%	-1.423914e-01	2.347108e-03	-2.691967e-01	3.762308e-01	-2.726613e-01	0.000000	2.382499e-01	-1.944169e-01	1.518710e-01	3.365585e-01	-2.674274e-01
75%	6.498599e-01	8.782389e-01	5.965307e-01	3.762308e-01	5.268762e-01	1.000000	2.382499e-01	4.028283e-01	1.518710e-01	3.365585e-01	9.658952e-01
max	2.687077e+00	1.716911e+00	2.451661e+00	1.786588e+00	2.525720e+00	1.000000	2.573604e+00	3.538366e+00	2.475162e+00	1.755424e+00	3.432540e+00

Dilakukan transformasi pada setiap *feature* dengan menggunakan **metode standarisasi**. Setelah dilakukan transformasi seluruh *feature* memiliki **simpangan baku = 1** dan **rata-rata mendekati 0**.

2. Feature Engineering (35 poin)

D. Class imbalance

```
# Membagi data menjadi target dan fitur
x = data_baru[[col for col in data_baru.columns if col not in ['Attrition']]]
y = data_baru['Attrition'].values

print("Jumlah perbandingan antara kelas attrition dan tidak attrition sebelum class imbalance :")
print(pd.Series(y).value_counts())
```

```
Jumlah perbandingan antara kelas attrition dan tidak attrition sebelum class imbalance :
0    1158
1     229
dtype: int64
```

```
# Melakukan handling class imbalance
from imblearn import over_sampling
x_over, y_over = over_sampling.SMOTE(sampling_strategy = 1).fit_resample(x, y)

print("Jumlah perbandingan antara kelas attrition dan tidak attrition setelah class imbalance :")
print(pd.Series(y_over).value_counts())
```

```
Jumlah perbandingan antara kelas attrition dan tidak attrition setelah class imbalance :
1    1158
0    1158
dtype: int64
```

Dilakukan pembagian data menjadi target (y) dan fitur (x) dalam suatu pemodelan machine learning. Pada proses ini, diasumsikan kolom 'Attrition' merupakan variabel target yang ingin diprediksi, dan kolom-kolom lainnya merupakan fitur-fitur yang digunakan untuk melakukan prediksi. Kemudian, didapatkan bahwa perbandingan antara kelas *attrition* dan tidak *attrition* pada kolom **AttritionNum** mendekati **1:5**, sehingga perlu dilakukan proses ***handling class imbalance***.

Dilakukan proses *handling class imbalance* menggunakan **metode SMOTE (Synthetic Minority Over-sampling Technique)**. Kelas 1 *(attrition)* memiliki jumlah sampel sebanyak 1158 dan kelas 0 (tidak *attrition*) juga memiliki jumlah sampel sebanyak 1158. Sehingga, perbandingan jumlah sampel antara kelas *'attrition'* dan 'tidak *attrition*' menjadi **1:1**, dengan kata lain setiap kelas memiliki jumlah sampel yang sama. Hal ini menunjukkan bahwa metode SMOTE telah berhasil menambahkan sampel sintetis pada kelas minoritas sehingga mengatasi **class imbalance** pada data.

2. Feature Engineering (35 poin)

D. Class imbalance

```
# Menggabungkan kembali data setelah dilakukan handling class imbalance
data_final = x_over
data_final['Attrition'] = y_over
data_final.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2316 entries, 0 to 2315
Data columns (total 27 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Age                                   2316 non-null   float64
 1   DailyRate                             2316 non-null   float64
 2   DistanceFromHome                      2316 non-null   float64
 3   JobInvolvement                        2316 non-null   float64
 4   NumCompaniesWorked                    2316 non-null   float64
 5   OverTime                              2316 non-null   int64
 6   StockOptionLevel                      2316 non-null   float64
 7   TotalWorkingYears                     2316 non-null   float64
 8   TrainingTimesLastYear                 2316 non-null   float64
 9   WorkLifeBalance                       2316 non-null   float64
 10  YearsInCurrentRole                    2316 non-null   float64
 11  Department_Research & Development     2316 non-null   uint8
 12  Department_Sales                      2316 non-null   uint8
 13  EducationField_Marketing              2316 non-null   uint8
 14  EducationField_Technical Degree       2316 non-null   uint8
 15  JobRole_Healthcare Representative     2316 non-null   uint8
 16  JobRole_Laboratory Technician         2316 non-null   uint8
 17  JobRole_Manager                       2316 non-null   uint8
 18  JobRole_Manufacturing Director        2316 non-null   uint8
 19  JobRole_Research Director             2316 non-null   uint8
 20  JobRole_Sales Representative          2316 non-null   uint8
 21  MaritalStatus_Divorced                2316 non-null   uint8
 22  MaritalStatus_Married                 2316 non-null   uint8
 23  MaritalStatus_Single                  2316 non-null   uint8
 24  SatisfactionMean                      2316 non-null   float64
 25  JobLevelSatisfaction                  2316 non-null   float64
 26  Attrition                             2316 non-null   int64
dtypes: float64(12), int64(2), uint8(13)
memory usage: 282.8 KB
```

Dilakukan penggabungan kembali feature yang telah dilakukan *handling class imbalance* menggunakan metode SMOTE pada **data_final.** Kemudian, data siap digunakan untuk proses pemodelan atau analisis selanjutnya.

2. Feature Engineering (35 poin)

E. Feature tambahan

Beberapa fitur tambahan yang mungkin akan sangat membantu untuk meningkatkan performa model :

- **WorkingHours (Integer)**: Jumlah jam kerja karyawan dalam satu bulan.
  Berdasarkan UU Ketenagakerjaan jumlah jam kerja karyawan dalam 1 bulan rata-rata adalah 173 jam. Fitur ini bisa digunakan untuk menganalisa durasi maksimal seorang karyawan bekerja melebihi jam kerja reguler yang akan membuat karyawan cenderung untuk bertahan di perusahaan.
- **Absence (Integer)**: Jumlah hari karyawan tidak masuk kerja dalam 1 bulan.
  Karyawan yang ingin keluar dari perusahaan cenderung untuk lebih banyak absen dibandingkan yang tidak.
- **ManagerSatisfaction (Integer)**: Tingkat kepuasan karyawan terhadap *manager*.
  Semakin tinggi tingkat kepuasan seorang karyawan terhadap *manager*, maka semakin rendah juga kemungkinan mereka untuk keluar dari perusahaan.
- **EmployeeEngagement (Integer)**: Tingkat keterlibatan karyawan dengan *event* yang diadakan oleh perusahaan dan hubungan dengan karyawan lainnya.
  Semakin tinggi tingkat keterlibatan seorang karyawan, maka semakin rendah juga kemungkinan mereka untuk keluar dari perusahaan.

Rakamin Academy

3. Git

3. Git (15 Poin)