Improving Employee Retention by Predicting Employee Attrition Using Machine Learning

Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

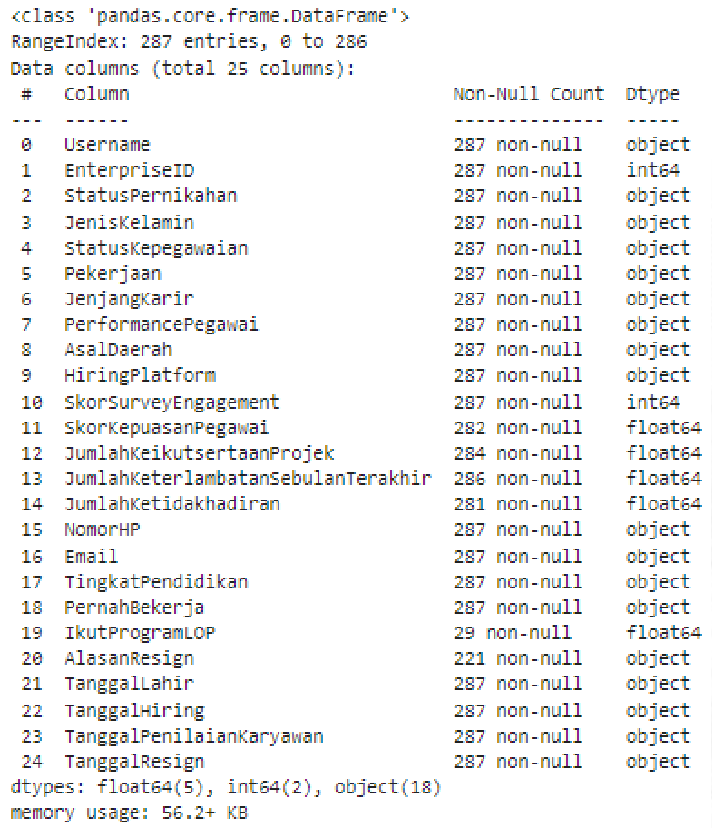
Created by:
Yasmin Fauziah
yasminfauziah63@gmail.com
https://www.linkedin.com/in/yasmin-fauziah-85b738239/

“Bachelor of Physics from Padjadjaran University> Someone who enjoys learning new things, has good analytical and planning skill. Enjoy to solve problem related to data analysis using Excel, SQL, Phython and Looker Studio. Have a high interest in a career in the data field.”

Overview

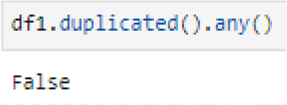
"Human resources (HR) is the main asset that needs to be managed properly by the company so that business goals can be achieved effectively and efficiently. On this occasion, we will face a problem about human resources in the company. Our focus is to find out how to keep employees to stay in the current company which can result in increased costs for employee recruitment and training for those who have just entered. By knowing the main factors that cause employee disengagement, companies can immediately address them by creating programs that are relevant to employee problems. "

Data Preprocessing



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 287 entries, 0 to 286
Data columns (total 25 columns):
 #   Column                              Non-Null Count  Dtype  
---  ------                              --------------  -----  
 0   Username                            287 non-null    object 
 1   EnterpriseID                        287 non-null    int64  
 2   StatusPernikahan                    287 non-null    object 
 3   JenisKelamin                        287 non-null    object 
 4   StatusKepegawaian                   287 non-null    object 
 5   Pekerjaan                           287 non-null    object 
 6   JenjangKarir                        287 non-null    object 
 7   PerformancePegawai                  287 non-null    object 
 8   AsalDaerah                          287 non-null    object 
 9   HiringPlatform                      287 non-null    object 
 10  SkorSurveyEngagement                287 non-null    int64  
 11  SkorKepuasanPegawai                 282 non-null    float64
 12  JumlahKeikutsertaanProjek           284 non-null    float64
 13  JumlahKeterlambatanSebulanTerakhir  286 non-null    float64
 14  JumlahKetidakhadiran                281 non-null    float64
 15  NomorHP                             287 non-null    object 
 16  Email                               287 non-null    object 
 17  TingkatPendidikan                   287 non-null    object 
 18  PernahBekerja                       287 non-null    object 
 19  IkutProgramLOP                      29 non-null     float64
 20  AlasanResign                        221 non-null    object 
 21  TanggalLahir                        287 non-null    object 
 22  TanggalHiring                       287 non-null    object 
 23  TanggalPenilaianKaryawan            287 non-null    object 
 24  TanggalResign                       287 non-null    object 
dtypes: float64(5), int64(2), object(18)
memory usage: 56.2+ KB
```

Handling Duplicate

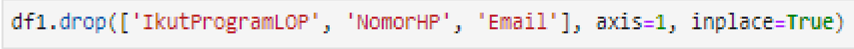


```
df1.duplicated().any()
```

```
False
```

No duplicate data

Drop Features



```
df1.drop(['IkutProgramLOP', 'NomorHP', 'Email'], axis=1, inplace=True)
```

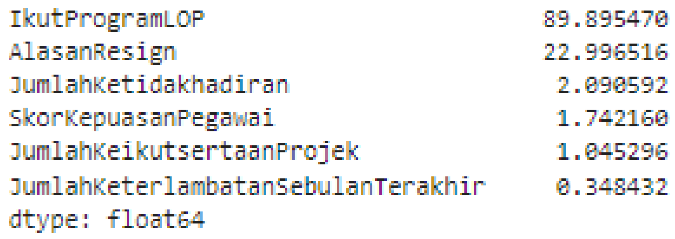
Drop “IkutProgramLOP” because most of the data is null and this feature is not a determining factor for employee attrition.

Drop “NomorHP” and “Email” because these features act as user identifiers and are not determinants of employee attrition.

for the details can access jupyter notebook here

Data Preprocessing

Handling Missing Value



```
IkutProgramLOP                      89.895470
AlasanResign                        22.996516
JumlahKetidakhadiran                 2.090592
SkorKepuasanPegawai                  1.742160
JumlahKeikutsertaanProjek            1.045296
JumlahKeterlambatanSebulanTerakhir   0.348432
dtype: float64
```

There are 4 features that have null value. "IkutProgramLOP" dropped

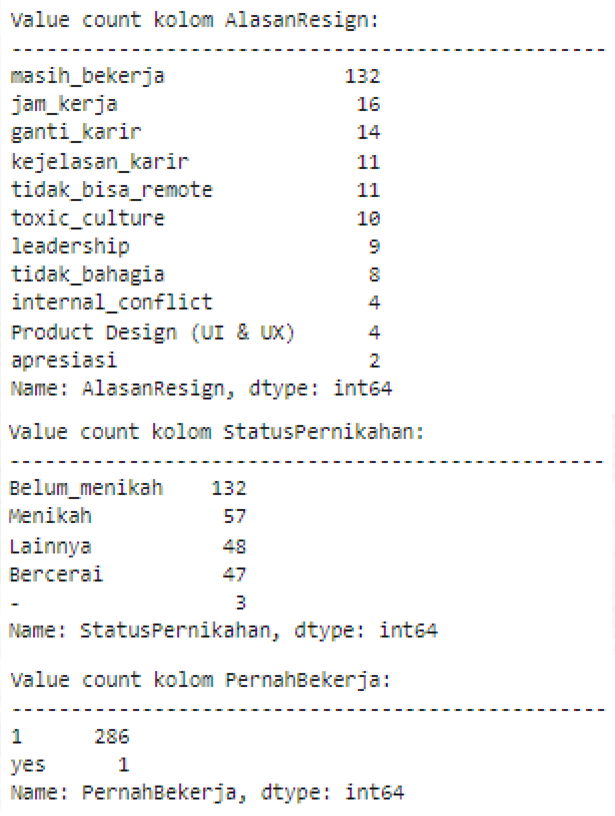
"AlasanResign" is fill with the mode value

"JumlahKetidakhadiran" is fill with the median value

"SkorKepuasanPegawai" is fill with the zero value

"JumlahKeikutsertaanProjek" is fill with the median value

"JumlahKeterlambatanSebulanTerakhir" is fill with the median value



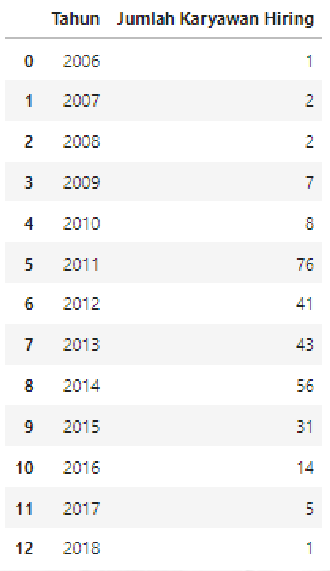
```
Value count kolom AlasanResign:
--------------------------------------------------
masih_bekerja            132
jam_kerja                 16
ganti_karir               14
kejelasan_karir           11
tidak_bisa_remote         11
toxic_culture             10
leadership                 9
tidak_bahagia              8
internal_conflict          4
Product Design (UI & UX)   4
apresiasi                  2
Name: AlasanResign, dtype: int64

Value count kolom StatusPernikahan:
--------------------------------------------------
Belum_menikah    132
Menikah           57
Lainnya           48
Bercerai          47
-                  3
Name: StatusPernikahan, dtype: int64

Value count kolom PernahBekerja:
--------------------------------------------------
1      286
yes      1
Name: PernahBekerja, dtype: int64
```

Handling Inconsistent

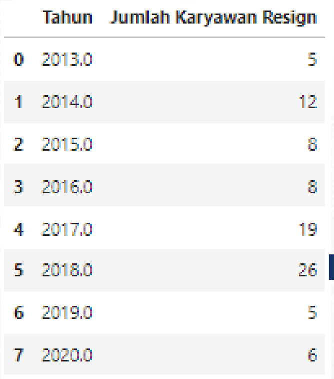
- "Product Design (UI & UX)" value in "AlasanResign" feature is replaced with "DII"
- "-" value in "StatusPernikahan" is replaced with "Belum_menikah"
- "yes" value in "PernahBekerja" is replaced with "1"

Annual Report on Employee Number Changes

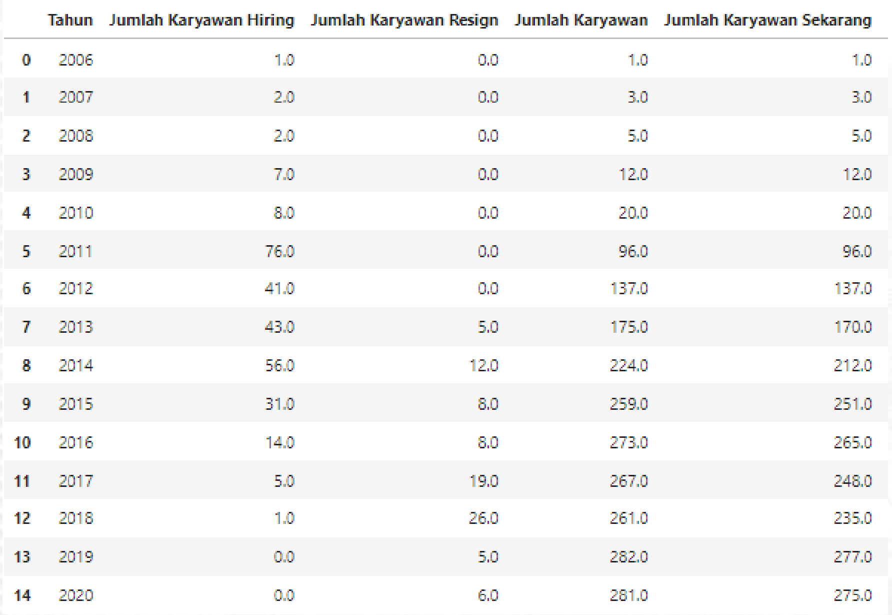
Grouping Total Employee Hiring and Resign by Year



	Tahun	Jumlah Karyawan Hiring
0	2006	1
1	2007	2
2	2008	2
3	2009	7
4	2010	8
5	2011	76
6	2012	41
7	2013	43
8	2014	56
9	2015	31
10	2016	14
11	2017	5
12	2018	1



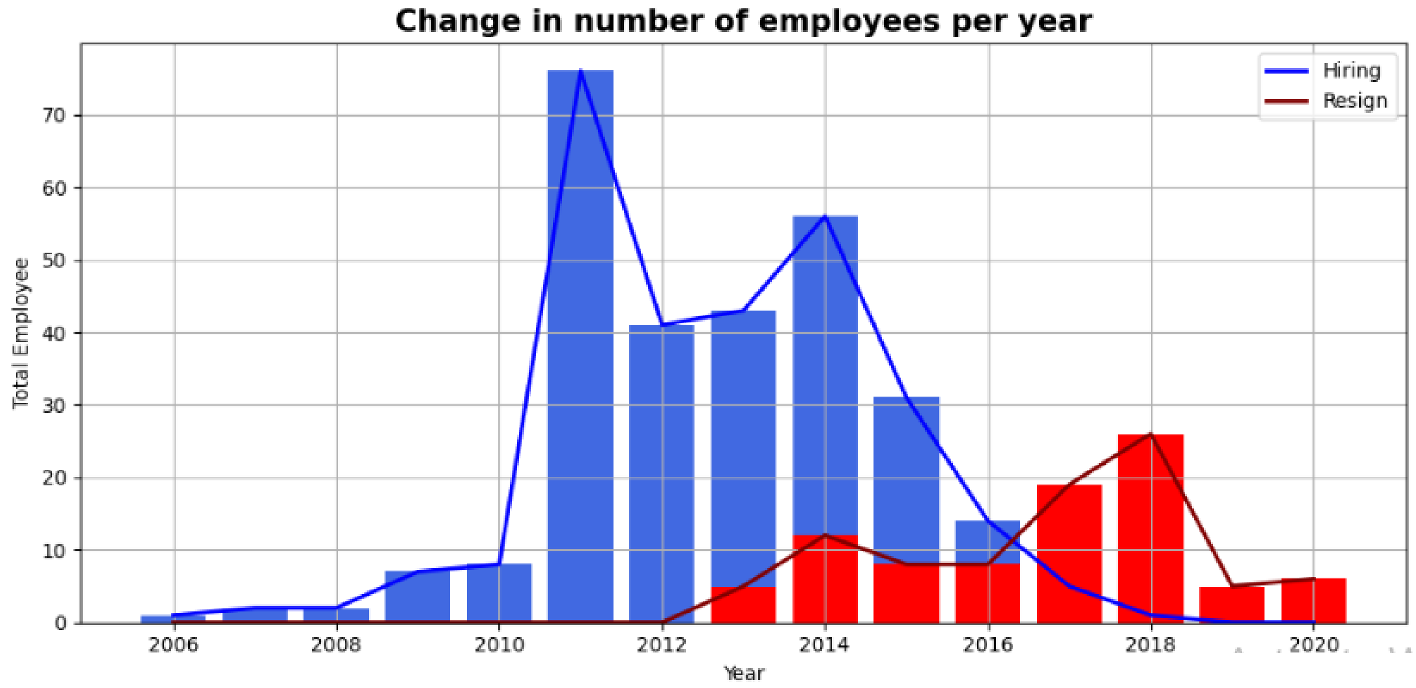
	Tahun	Jumlah Karyawan Resign
0	2013.0	5
1	2014.0	12
2	2015.0	8
3	2016.0	8
4	2017.0	19
5	2018.0	26
6	2019.0	5
7	2020.0	6



	Tahun	Jumlah Karyawan Hiring	Jumlah Karyawan Resign	Jumlah Karyawan	Jumlah Karyawan Sekarang
0	2006	1.0	0.0	1.0	1.0
1	2007	2.0	0.0	3.0	3.0
2	2008	2.0	0.0	5.0	5.0
3	2009	7.0	0.0	12.0	12.0
4	2010	8.0	0.0	20.0	20.0
5	2011	76.0	0.0	96.0	96.0
6	2012	41.0	0.0	137.0	137.0
7	2013	43.0	5.0	175.0	170.0
8	2014	56.0	12.0	224.0	212.0
9	2015	31.0	8.0	259.0	251.0
10	2016	14.0	8.0	273.0	265.0
11	2017	5.0	19.0	267.0	248.0
12	2018	1.0	26.0	261.0	235.0
13	2019	0.0	5.0	282.0	277.0
14	2020	0.0	6.0	281.0	275.0

for the details can access jupyter notebook here

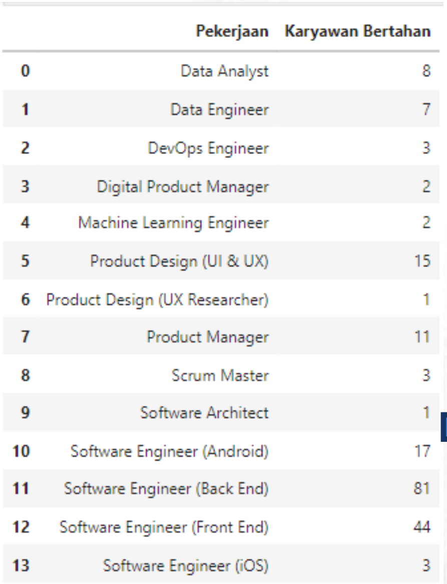
Annual Report on Employee Number Changes



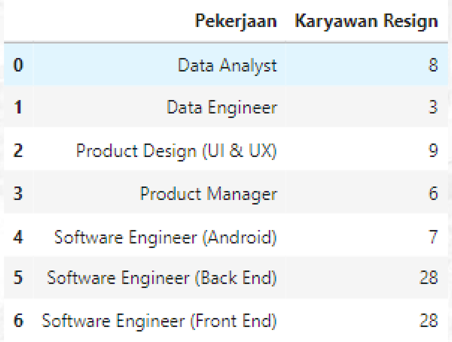
Interpretation:

The growth in the number of employees occurred in the range of 2006 - 2018. In the 2013-2020 period, it appears that the company's condition is worrying because the number of employees continues to decrease until the peak occurred in 2018. This could be a sign that the company may be experiencing internal problems such as lack of growth opportunities, poor working environment, or financial problems.

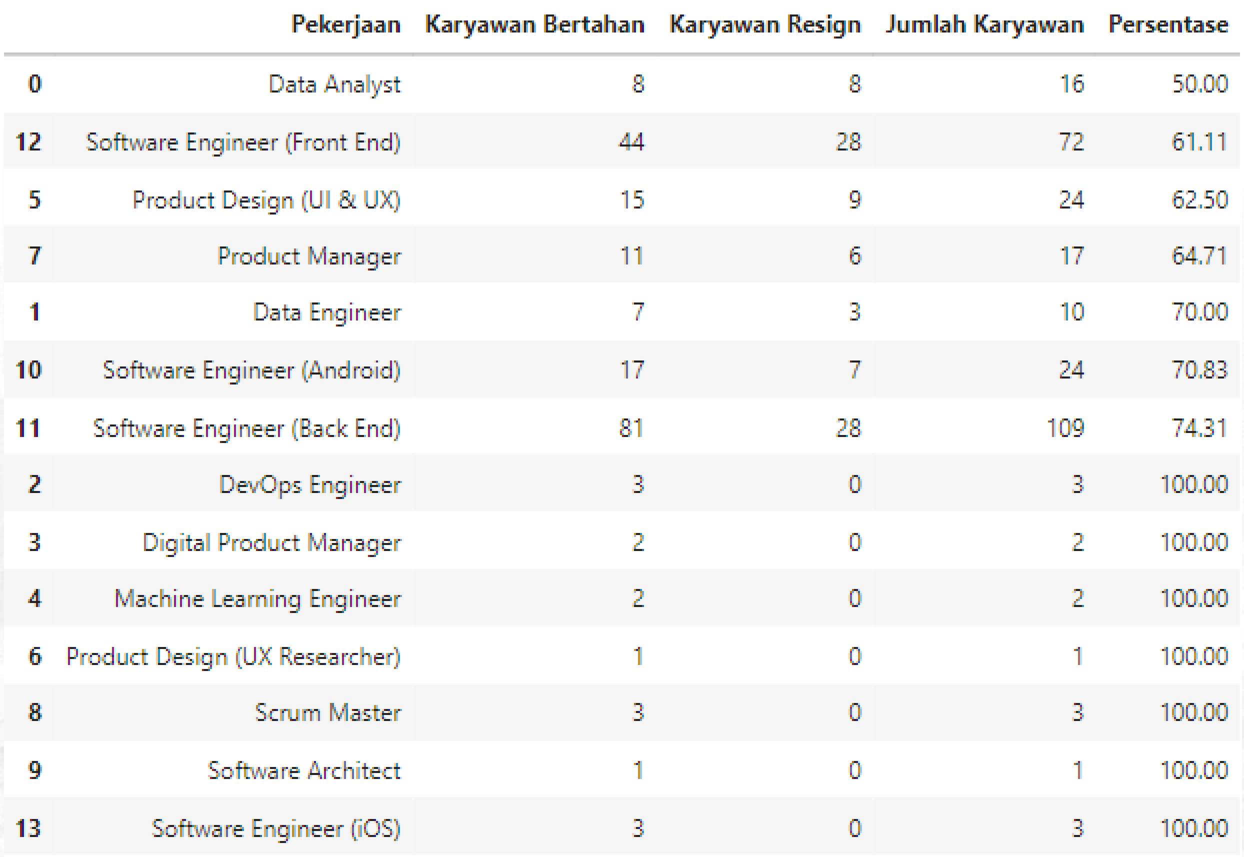
Resign Reason Analysis for Employee Attrition Management Strategy



	Pekerjaan	Karyawan Bertahan
0	Data Analyst	8
1	Data Engineer	7
2	DevOps Engineer	3
3	Digital Product Manager	2
4	Machine Learning Engineer	2
5	Product Design (UI & UX)	15
6	Product Design (UX Researcher)	1
7	Product Manager	11
8	Scrum Master	3
9	Software Architect	1
10	Software Engineer (Android)	17
11	Software Engineer (Back End)	81
12	Software Engineer (Front End)	44
13	Software Engineer (iOS)	3



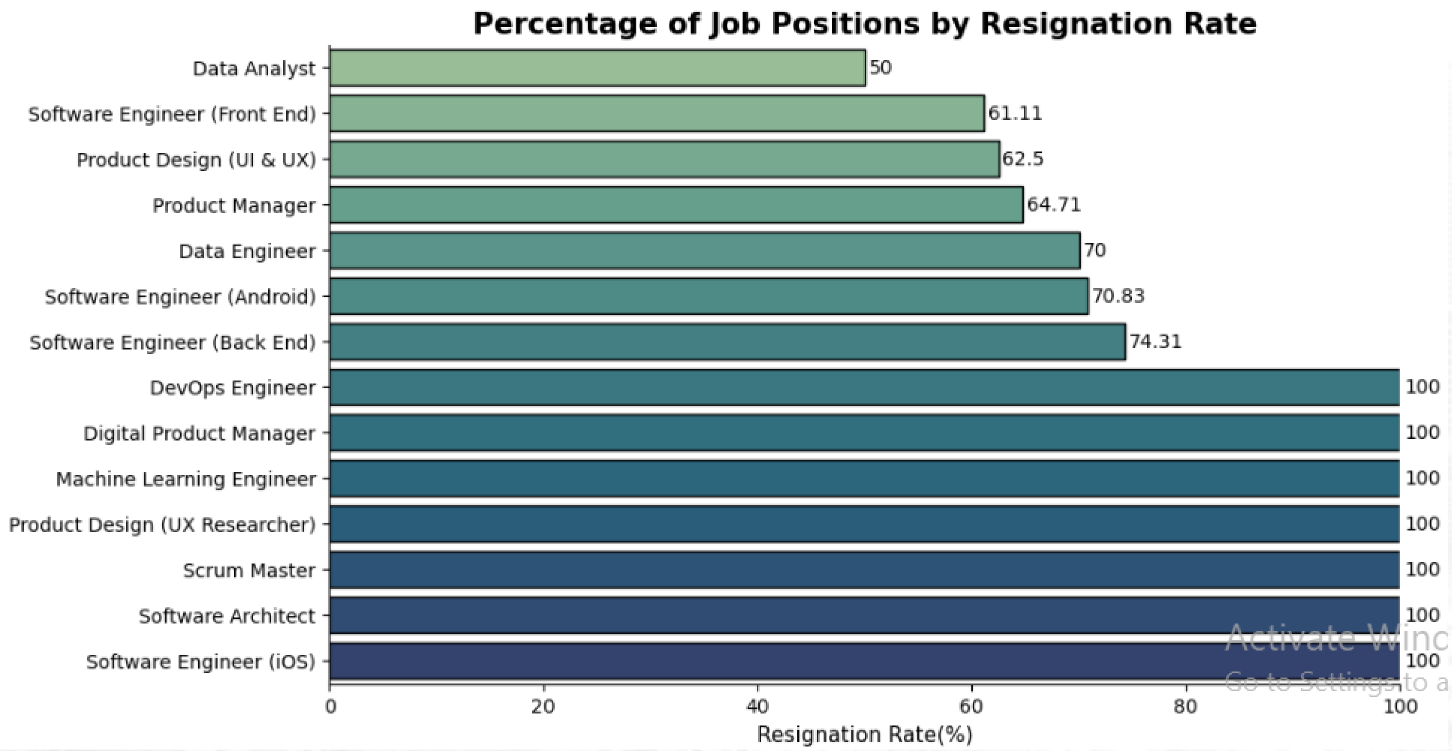
	Pekerjaan	Karyawan Resign
0	Data Analyst	8
1	Data Engineer	3
2	Product Design (UI & UX)	9
3	Product Manager	6
4	Software Engineer (Android)	7
5	Software Engineer (Back End)	28
6	Software Engineer (Front End)	28



	Pekerjaan	Karyawan Bertahan	Karyawan Resign	Jumlah Karyawan	Persentase
0	Data Analyst	8	8	16	50.00
12	Software Engineer (Front End)	44	28	72	61.11
5	Product Design (UI & UX)	15	9	24	62.50
7	Product Manager	11	6	17	64.71
1	Data Engineer	7	3	10	70.00
10	Software Engineer (Android)	17	7	24	70.83
11	Software Engineer (Back End)	81	28	109	74.31
2	DevOps Engineer	3	0	3	100.00
3	Digital Product Manager	2	0	2	100.00
4	Machine Learning Engineer	2	0	2	100.00
6	Product Design (UX Researcher)	1	0	1	100.00
8	Scrum Master	3	0	3	100.00
9	Software Architect	1	0	1	100.00
13	Software Engineer (iOS)	3	0	3	100.00

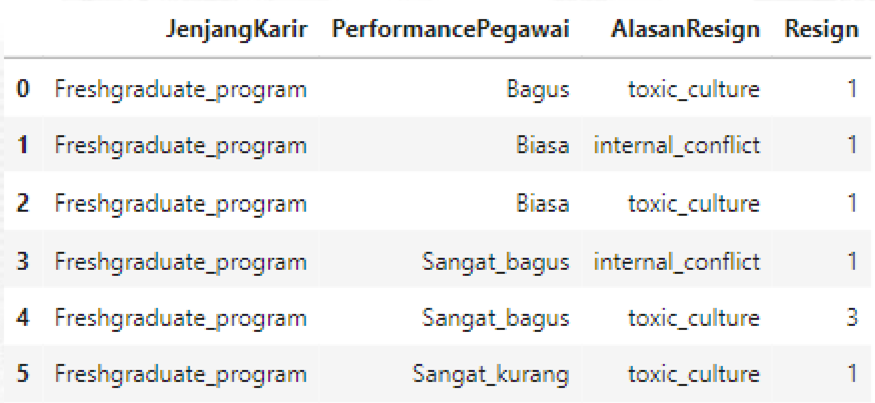
for the details can access jupyter notebook here

Resign Reason Analysis for Employee Attrition Management Strategy



Based on job position, Data Analyst has highest resignation rate (50%).

Resign Reason Analysis for Employee Attrition Management Strategy



	JenjangKarir	PerformancePegawai	AlasanResign	Resign
0	Freshgraduate_program	Bagus	toxic_culture	1
1	Freshgraduate_program	Biasa	internal_conflict	1
2	Freshgraduate_program	Biasa	toxic_culture	1
3	Freshgraduate_program	Sangat_bagus	internal_conflict	1
4	Freshgraduate_program	Sangat_bagus	toxic_culture	3
5	Freshgraduate_program	Sangat_kurang	toxic_culture	1

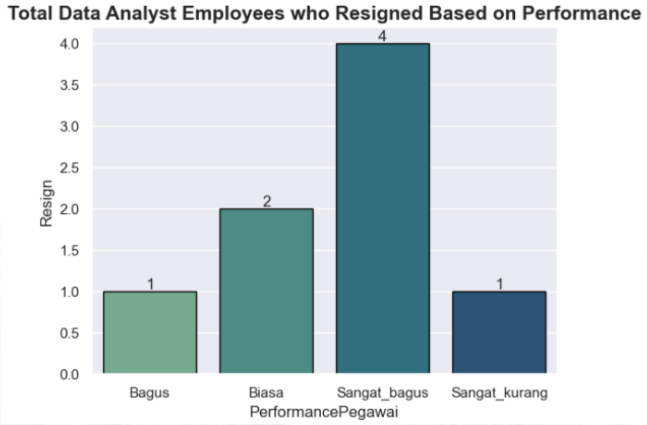
Interpretation:

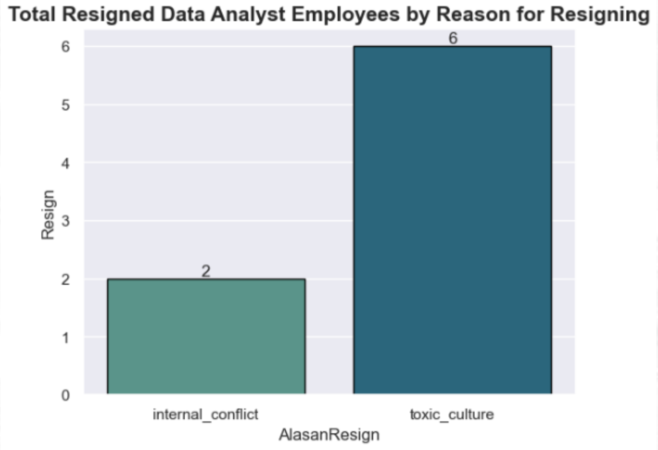
All employees who resigned in the Data Analyst position were Fresh Graduate Program.

Recommendation:

The company can offer fresh graduate program employees more competitive benefits, conduct training, better self-development opportunities and create a more supportive work environment.

Resign Reason Analysis for Employee Attrition Management Strategy





Interpretation:

Of the 8 Data Analyst employees who resigned, 2 of them had good performance and the other 4 were very good. This is certainly very detrimental to the company because the majority of employees who resign are employees with good performance.

Recommendation:

The company can offer better salary, benefits and work-life balance to employees with good performance. In addition, the company is expected to offer good career paths and self-development to employees with good performance so that these employees feel valued and feel they will have a good career path in the company.

Interpretation:

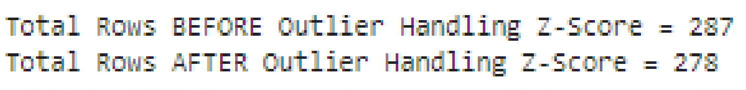
Of the 8 employees, 6 Data Analyst employees resigned due to toxic culture and 2 resigned due to internal conflict. Both reasons illustrate that there are unfavorable factors from the internal position of the company's own Data Analyst.

Recommendation:

The company can create an effective feedback system so that employees feel they can give input and receive constructive feedback. The company should also be able to resolve internal conflicts that occur between employees by facilitating meetings between employees to resolve problems. In addition, the company should re-evaluate the work culture and ensure that the culture is positive and motivates employees.

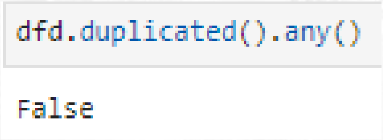
Build an Automated Resignation Behavior Prediction using Machine Learning

Handling Outlier



```
Total Rows BEFORE Outlier Handling Z-Score = 287
Total Rows AFTER Outlier Handling Z-Score = 278
```

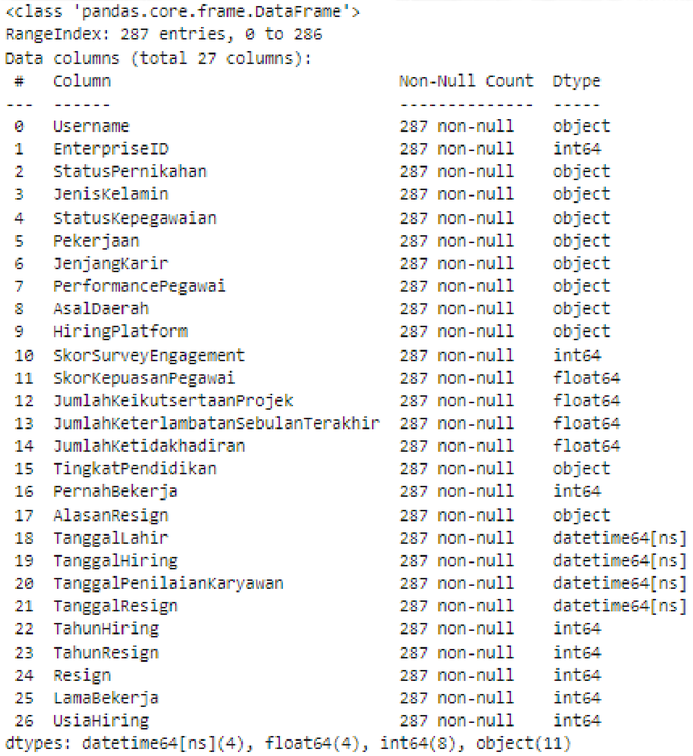
Data Duplicate



```
dfd.duplicated().any()
```

```
False
```

Missing Value

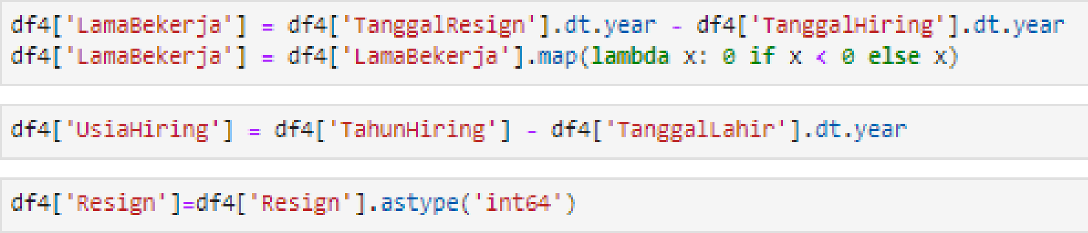


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 287 entries, 0 to 286
Data columns (total 27 columns):
 #   Column                               Non-Null Count  Dtype
---  ------                               --------------  -----
 0   Username                             287 non-null    object
 1   EnterpriseID                         287 non-null    int64
 2   StatusPernikahan                     287 non-null    object
 3   JenisKelamin                         287 non-null    object
 4   StatusKepegawaian                    287 non-null    object
 5   Pekerjaan                            287 non-null    object
 6   JenjangKarir                         287 non-null    object
 7   PerformancePegawai                   287 non-null    object
 8   AsalDaerah                           287 non-null    object
 9   HiringPlatform                       287 non-null    object
 10  SkorSurveyEngagement                 287 non-null    int64
 11  SkorKepuasanPegawai                  287 non-null    float64
 12  JumlahKeikutsertaanProjek            287 non-null    float64
 13  JumlahKeterlambatanSebulanTerakhir   287 non-null    float64
 14  JumlahKetidakhadiran                 287 non-null    float64
 15  TingkatPendidikan                    287 non-null    object
 16  PernahBekerja                        287 non-null    int64
 17  AlasanResign                         287 non-null    object
 18  TanggalLahir                         287 non-null    datetime64[ns]
 19  TanggalHiring                        287 non-null    datetime64[ns]
 20  TanggalPenilaianKaryawan             287 non-null    datetime64[ns]
 21  TanggalResign                        287 non-null    datetime64[ns]
 22  TahunHiring                          287 non-null    int64
 23  TahunResign                          287 non-null    int64
 24  Resign                               287 non-null    int64
 25  LamaBekerja                          287 non-null    int64
 26  UsiaHiring                           287 non-null    int64
dtypes: datetime64[ns](4), float64(4), int64(8), object(11)
```

for the details can access jupyter notebook here

Build an Automated Resignation Behavior Prediction using Machine Learning

Feature Engineering



```
df4['LamaBekerja'] = df4['TanggalResign'].dt.year - df4['TanggalHiring'].dt.year
df4['LamaBekerja'] = df4['LamaBekerja'].map(lambda x: 0 if x < 0 else x)
```

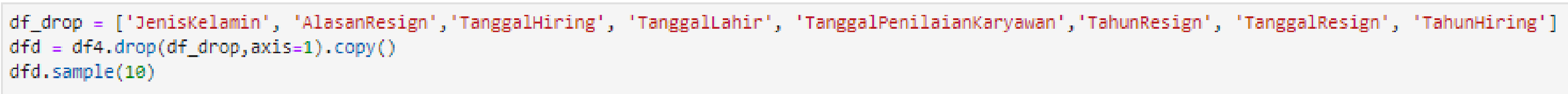
```
df4['UsiaHiring'] = df4['TahunHiring'] - df4['TanggalLahir'].dt.year
```

```
df4['Resign']=df4['Resign'].astype('int64')
```

Create 3 new features:

- LamaBekerja
- UsiaHiring
- Resign

Feature Selection



```
df_drop = ['JenisKelamin', 'AlasanResign','TanggalHiring', 'TanggalLahir', 'TanggalPenilaianKaryawan','TahunResign', 'TanggalResign', 'TahunHiring']
dfd = df4.drop(df_drop,axis=1).copy()
dfd.sample(10)
```

Features that were removed:

JenisKelamin, avoid discrimination.

AlasanResign, irrelevant feature to predict resignation

TanggalPenilaianKaryawan, TanggalResign and TahunResign, features are not relevant to predict resignation

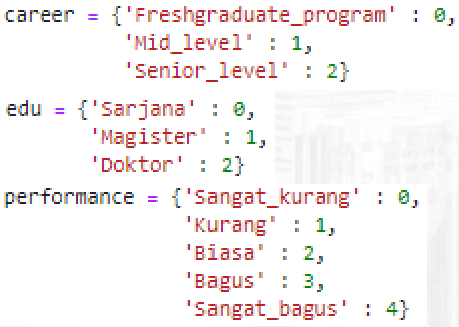
TanggalLahir, TanggalHiring and TahunHiring, already converted to LamaBekerja and UsiaHiring

HiringPlatform -> too many unique values

Build an Automated Resignation Behavior Prediction using Machine Learning

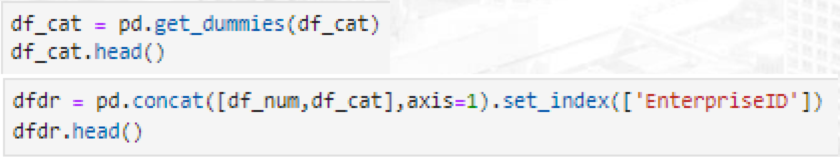
Feature Encoding

Label Encoding



```
career = {'Freshgraduate_program' : 0,
          'Mid_level' : 1,
          'Senior_level' : 2}
edu = {'Sarjana' : 0,
       'Magister' : 1,
       'Doktor' : 2}
performance = {'Sangat_kurang' : 0,
               'Kurang' : 1,
               'Biasa' : 2,
               'Bagus' : 3,
               'Sangat_bagus' : 4}
```

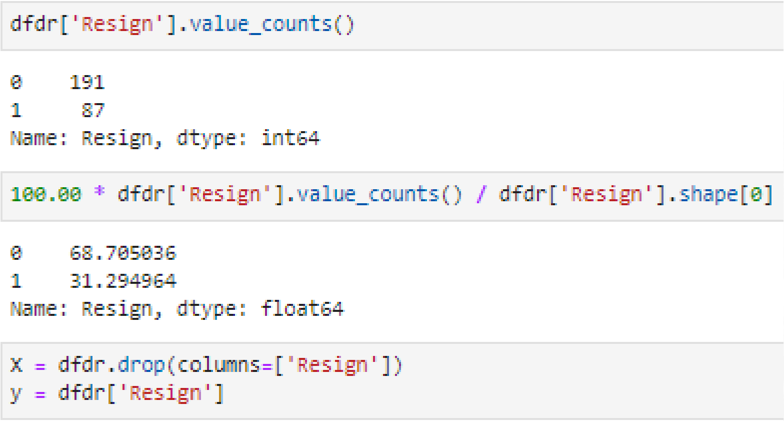
Onehot Encoding



```
df_cat = pd.get_dummies(df_cat)
df_cat.head()
```

```
dfdr = pd.concat([df_num,df_cat],axis=1).set_index(['EnterpriseID'])
dfdr.head()
```

Handling Class Imbalance



```
dfdr['Resign'].value_counts()
```

```
0    191
1     87
Name: Resign, dtype: int64
```

```
100.00 * dfdr['Resign'].value_counts() / dfdr['Resign'].shape[0]
```

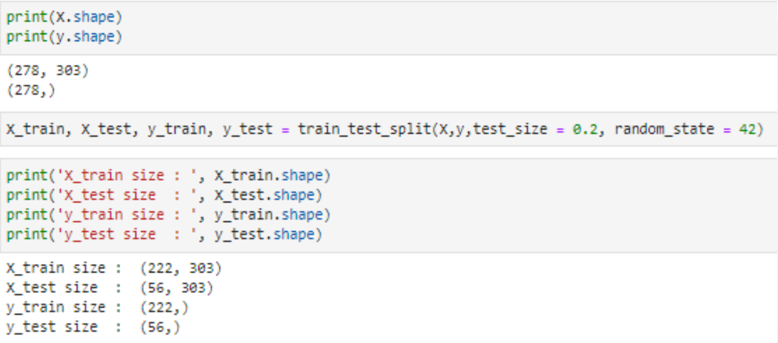
```
0    68.705036
1    31.294964
Name: Resign, dtype: float64
```

```
X = dfdr.drop(columns=['Resign'])
y = dfdr['Resign']
```

Build an Automated Resignation Behavior Prediction using Machine Learning

Rakamin Academy

Data Train/Test Split



```
print(X.shape)
print(y.shape)
```

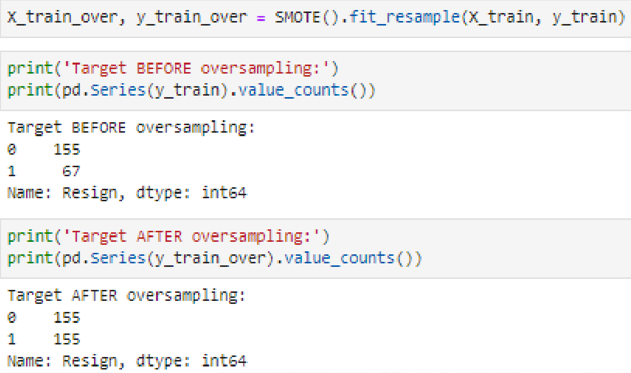
```
(278, 303)
(278,)
```

```
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size = 0.2, random_state = 42)
```

```
print('X_train size : ', X_train.shape)
print('X_test size  : ', X_test.shape)
print('y_train size : ', y_train.shape)
print('y_test size  : ', y_test.shape)
```

```
X_train size :  (222, 303)
X_test size  :  (56, 303)
y_train size :  (222,)
y_test size  :  (56,)
```

SMOTE



```
X_train_over, y_train_over = SMOTE().fit_resample(X_train, y_train)
```

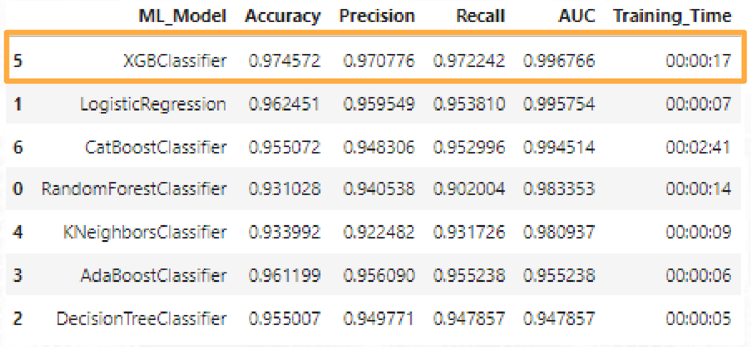
```
print('Target BEFORE oversampling:')
print(pd.Series(y_train).value_counts())
```

```
Target BEFORE oversampling:
0    155
1     67
Name: Resign, dtype: int64
```

```
print('Target AFTER oversampling:')
print(pd.Series(y_train_over).value_counts())
```

```
Target AFTER oversampling:
0    155
1    155
Name: Resign, dtype: int64
```

Modeling

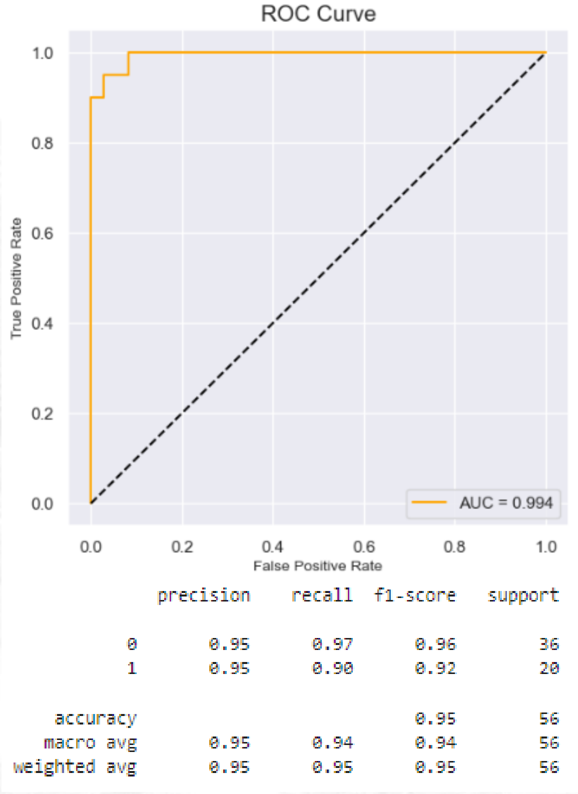


	ML_Model	Accuracy	Precision	Recall	AUC	Training_Time
5	XGBClassifier	0.974572	0.970776	0.972242	0.996766	00:00:17
1	LogisticRegression	0.962451	0.959549	0.953810	0.995754	00:00:07
6	CatBoostClassifier	0.955072	0.948306	0.952996	0.994514	00:02:41
0	RandomForestClassifier	0.931028	0.940538	0.902004	0.983353	00:00:14
4	KNeighborsClassifier	0.933992	0.922482	0.931726	0.980937	00:00:09
3	AdaBoostClassifier	0.961199	0.956090	0.955238	0.955238	00:00:06
2	DecisionTreeClassifier	0.955007	0.949771	0.947857	0.947857	00:00:05

Build an Automated Resignation Behavior Prediction using Machine Learning

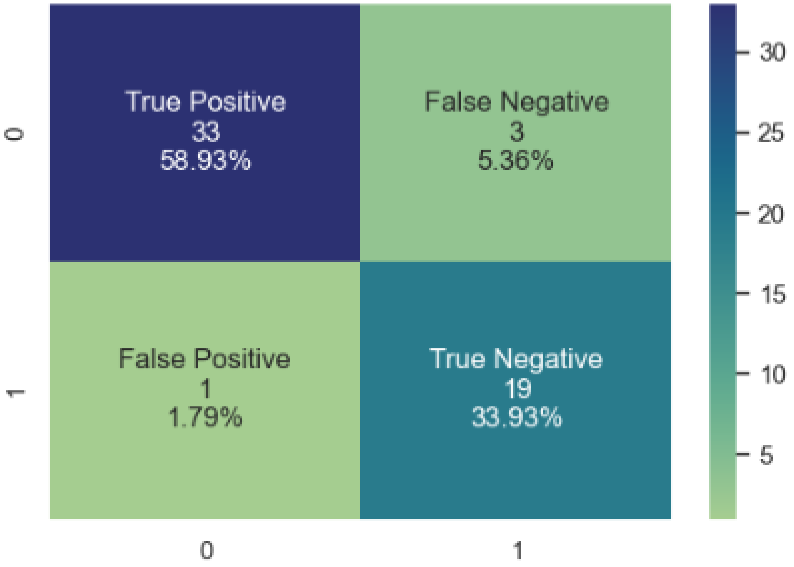
Rakamin Academy

ROC AUC

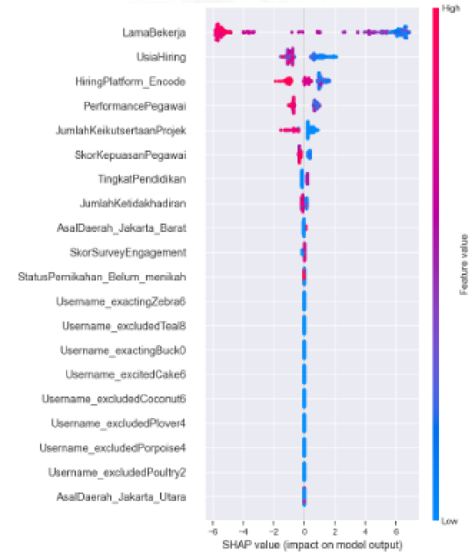


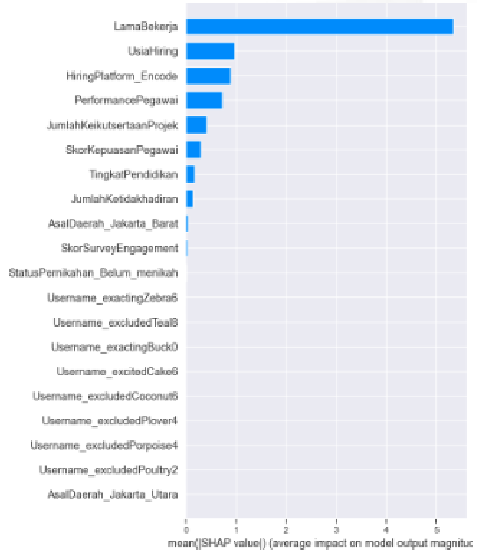
```
              precision    recall  f1-score   support

           0       0.95      0.97      0.96        36
           1       0.95      0.90      0.92        20

    accuracy                           0.95        56
   macro avg       0.95      0.94      0.94        56
weighted avg       0.95      0.95      0.95        56
```

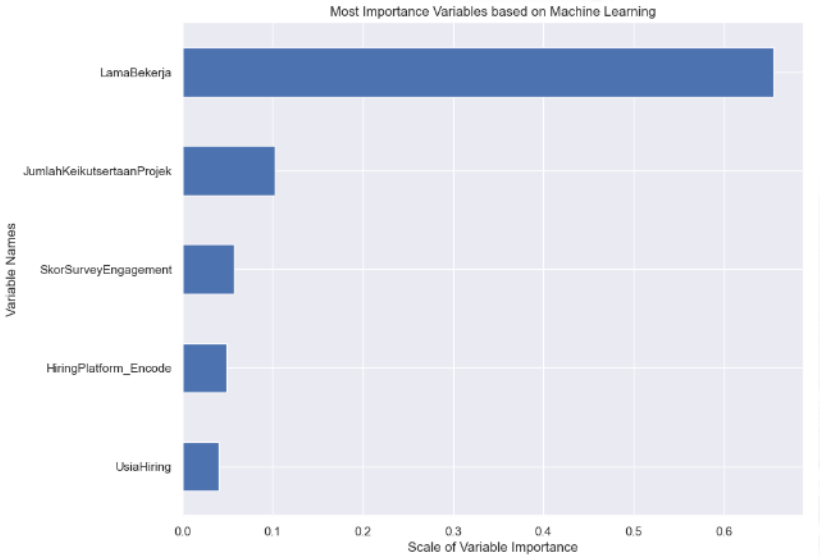
Confussion Matrix



Presenting Machine Learning Products to the Business Users







Interpretation:

It can be seen that "LamaBekerja" feature is the most important feature and is very dominant compared to other features in predicting the possibility of resigning from an employee. The SHAP value data shows that the smaller the length of service of an employee, the more likely the employee is to resign.

Recommendation:

The company can review the existing corporate culture so as not to create a toxic work environment and hold a career development program to maintain employees, especially new employees who have good self-development potential. In addition, the company can also conduct surveys and ask for feedback from employees to understand their needs.