Rakamin Academy

Improving Employee Retention by Predicting Employee Attrition Using Machine Learning

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

Created by:
Saila Fikriyya
sailafs@gmail.com
linkedin.com/in/sailafs1203/

I'm a health psychology graduate turned data scientist, bridging the gap between human behavior and data-driven solutions. With a Master's in Health Psychology from the University of Stirling and ongoing training in data science, I combine expertise in research methods, data analysis, and understanding human behavior. My experience spans from studying post-traumatic growth in cancer survivors to applying data science techniques in digital healthcare and career development. I'm passionate about leveraging interdisciplinary skills to create evidence-based strategies that improve health outcomes and overall well-being, always prioritizing a human-centric approach in collaborative environments.

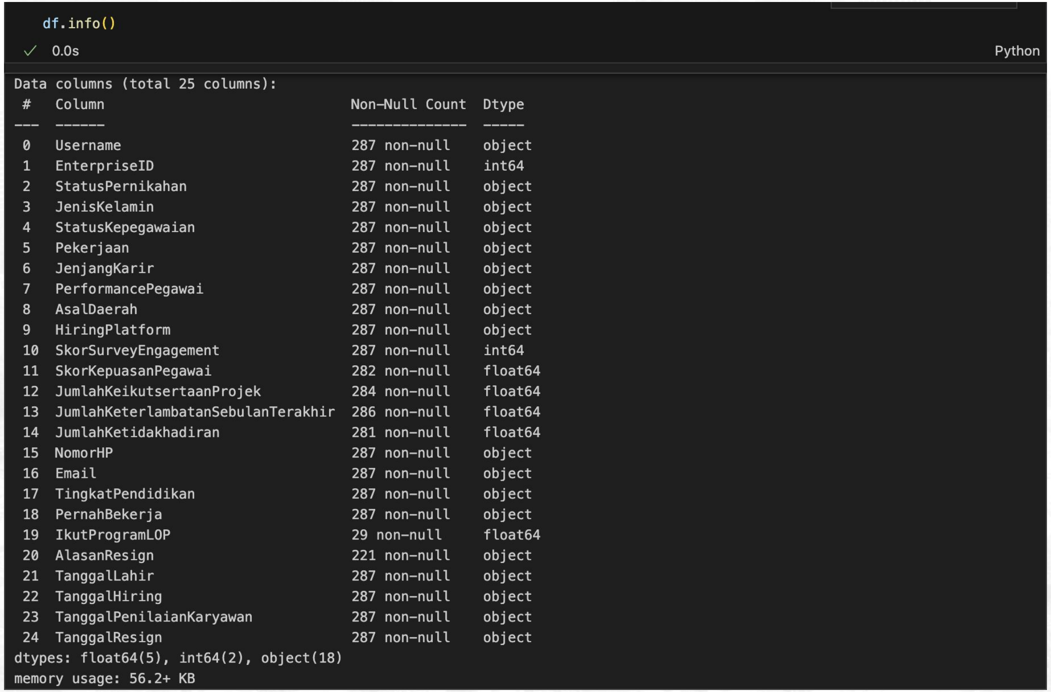
Overview

"Human resources (HR) are the main asset that needs to be managed well by the company in order for business objectives to be achieved effectively and efficiently." On this occasion, we will address an issue regarding human resources within the company. Our focus is to understand how to retain employees in the current company, which can lead to inflated costs for recruiting new staff and training for newcomers. By identifying the main factors that cause employees to feel disengaged, the company can promptly address these issues by creating programs that are relevant to employees' concerns."

Data Preprocessing

- Step 1: Quick Look of the data

  - There are 287 rows with 24 features, consists of int, float, and object.
  - Some of null data are detected from features as below:
    - ‘SkorKepuasanPegawai'
    - 'JumlahKeikutsertaanProjek'
    - 'JumlahKeterlambatanSebulanTerakhir'
    - 'JumlahKetidakhadiran'
    - 'IkutProgramLOP'
    - 'AlasanResign'



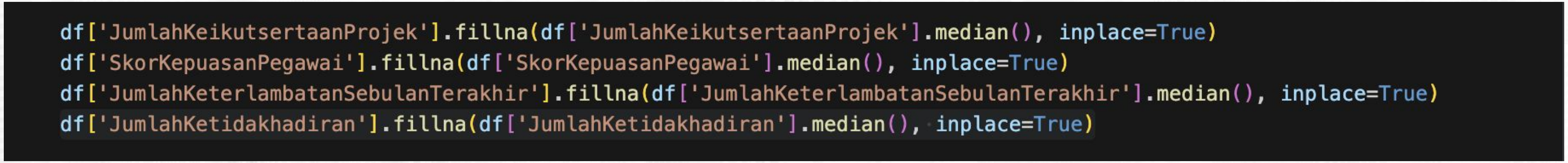
```
df.info()
```

```
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
```

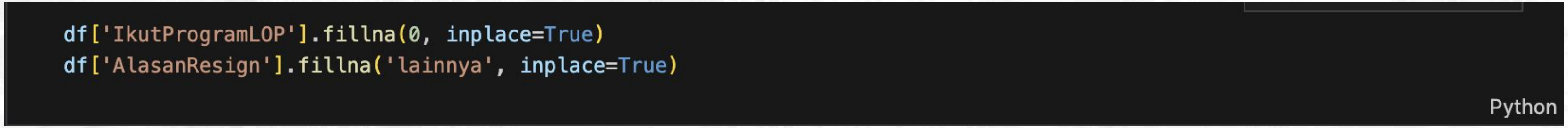
For more information, see jupyter notebook here

Data Preprocessing

- Step 2: Fill in Missing Data
  - All missing data is filled by median since mean would resulted in decimal.
  - For 'JumlahKetidakhadiran', missing data were filled with 0 indicating did not participate.
  - For ‘AlasanResign’, that has NA values, a new value 'lainnya' was created.



```python
df['JumlahKeikutsertaanProjek'].fillna(df['JumlahKeikutsertaanProjek'].median(), inplace=True)
df['SkorKepuasanPegawai'].fillna(df['SkorKepuasanPegawai'].median(), inplace=True)
df['JumlahKeterlambatanSebulanTerakhir'].fillna(df['JumlahKeterlambatanSebulanTerakhir'].median(), inplace=True)
df['JumlahKetidakhadiran'].fillna(df['JumlahKetidakhadiran'].median(), inplace=True)
```

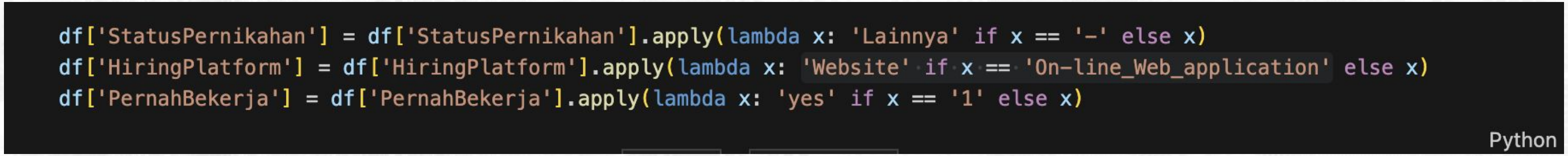


```python
df['IkutProgramLOP'].fillna(0, inplace=True)
df['AlasanResign'].fillna('lainnya', inplace=True)
```

For more information, see jupyter notebook here

Data Preprocessing

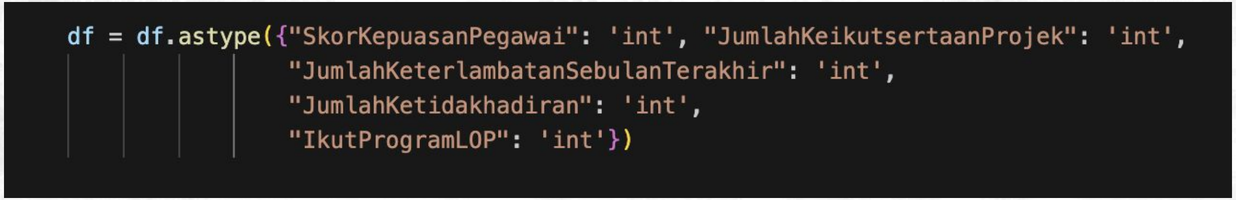
- Step 3: Change Irrelevant Value
  - While checking value counts for categorical data, some irrelevant values were detected.
    - For ‘StatusPernikahan’, there were ‘-’ and ‘Lainnya’ which might indicate the same thing. Both values than merged to be ‘Lainnya’.
    - For ‘HiringPlatform’, the same thing was observed between values 'Website' and 'On-line_Web_application' which then merged to be ‘Website’
    - For ‘PernahBekerja’, there were 2 unique values → ‘1’ and ‘yes, which might indicate the same thing. Values than merged into ‘yes’



```python
df['StatusPernikahan'] = df['StatusPernikahan'].apply(lambda x: 'Lainnya' if x == '-' else x)
df['HiringPlatform'] = df['HiringPlatform'].apply(lambda x: 'Website' if x == 'On-line_Web_application' else x)
df['PernahBekerja'] = df['PernahBekerja'].apply(lambda x: 'yes' if x == '1' else x)
```

Python

  - Float data types also replace to be integer for uniformity purposes.

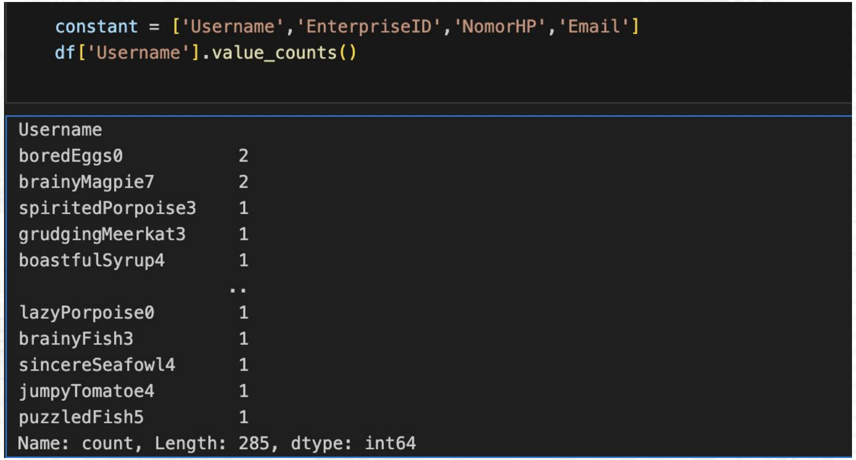


```python
df = df.astype({"SkorKepuasanPegawai": 'int', "JumlahKeikutsertaanProjek": 'int',
                "JumlahKeterlambatanSebulanTerakhir": 'int',
                "JumlahKetidakhadiran": 'int',
                "IkutProgramLOP": 'int'})
```

For more information, see jupyter notebook here

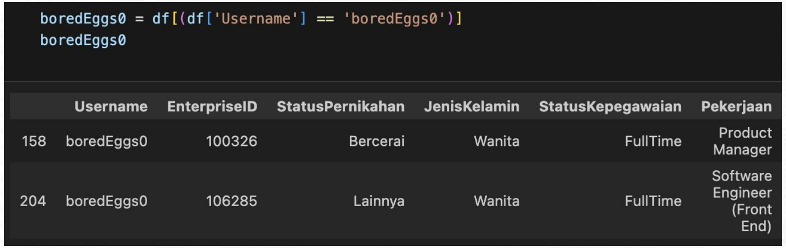
Data Preprocessing

- Step 4a: Drop Constant
    - For unique columns, supposed that 'constant' is a list of unique features as below. Upon checking, there are 2 username which were used twice with different people. As such, no need to change as the column will be dropped.



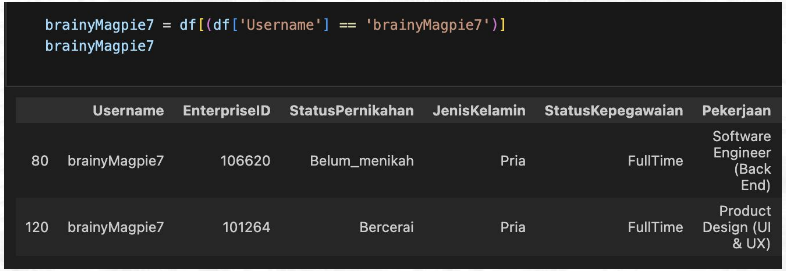
```python
constant = ['Username','EnterpriseID','NomorHP','Email']
df['Username'].value_counts()
```

```
Username
boredEggs0          2
brainyMagpie7       2
spiritedPorpoise3   1
grudgingMeerkat3    1
boastfulSyrup4      1
                   ..
lazyPorpoise0       1
brainyFish3         1
sincereSeafowl4     1
jumpyTomatoe4       1
puzzledFish5        1
Name: count, Length: 285, dtype: int64
```



```python
boredEggs0 = df[(df['Username'] == 'boredEggs0')]
boredEggs0
```

	Username	EnterpriseID	StatusPernikahan	JenisKelamin	StatusKepegawaian	Pekerjaan
158	boredEggs0	100326	Bercerai	Wanita	FullTime	Product Manager
204	boredEggs0	106285	Lainnya	Wanita	FullTime	Software Engineer (Front End)



```python
brainyMagpie7 = df[(df['Username'] == 'brainyMagpie7')]
brainyMagpie7
```

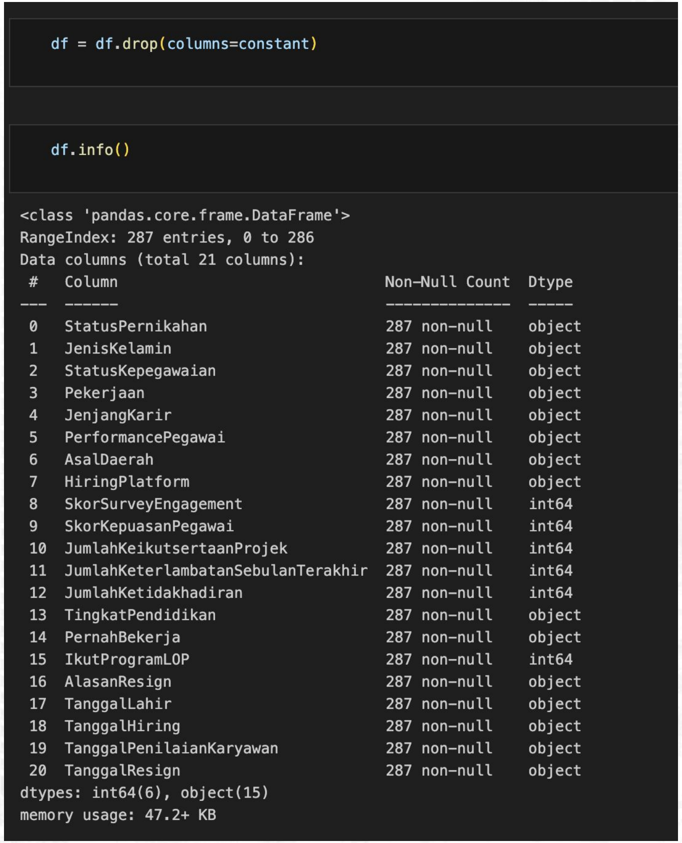
	Username	EnterpriseID	StatusPernikahan	JenisKelamin	StatusKepegawaian	Pekerjaan
80	brainyMagpie7	106620	Belum_menikah	Pria	FullTime	Software Engineer (Back End)
120	brainyMagpie7	101264	Bercerai	Pria	FullTime	Product Design (UI & UX)

For more information, see jupyter notebook here

Data Preprocessing

- Step 4b: Drop Constant

  - Columns that dropped using the code below and checking all to ensure no missing data



```
df = df.drop(columns=constant)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 287 entries, 0 to 286
Data columns (total 21 columns):
 #   Column                              Non-Null Count  Dtype 
---  ------                              --------------  ----- 
 0   StatusPernikahan                    287 non-null    object
 1   JenisKelamin                        287 non-null    object
 2   StatusKepegawaian                   287 non-null    object
 3   Pekerjaan                           287 non-null    object
 4   JenjangKarir                        287 non-null    object
 5   PerformancePegawai                  287 non-null    object
 6   AsalDaerah                          287 non-null    object
 7   HiringPlatform                      287 non-null    object
 8   SkorSurveyEngagement                287 non-null    int64 
 9   SkorKepuasanPegawai                 287 non-null    int64 
 10  JumlahKeikutsertaanProjek           287 non-null    int64 
 11  JumlahKeterlambatanSebulanTerakhir  287 non-null    int64 
 12  JumlahKetidakhadiran                287 non-null    int64 
 13  TingkatPendidikan                   287 non-null    object
 14  PernahBekerja                       287 non-null    object
 15  IkutProgramLOP                      287 non-null    int64 
 16  AlasanResign                        287 non-null    object
 17  TanggalLahir                        287 non-null    object
 18  TanggalHiring                       287 non-null    object
 19  TanggalPenilaianKaryawan            287 non-null    object
 20  TanggalResign                       287 non-null    object
dtypes: int64(6), object(15)
memory usage: 47.2+ KB
```

For more information, see jupyter notebook here