Rakamin Academy

# Improving Employee Retention by Predicting Employee Attrition Using Machine Learning

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

**Created by:**
**Muhammad Cikal Merdeka**
**Email :** mcikalmerdeka@gmail.com
**LinkedIn :** linkedin.com/in/mcikalmerdeka
**Github :** github.com/mcikalmerdeka

Dedicated entry-level data scientist with analytical and experimental background of Physics. My graduation 2023, a pivotal year marked by significant advancements in artificial intelligence with the introduction of GPT-4 and other generative AI models, has fueled my curiosity and excitement to delve into the field of data. I have comprehensive grasp of data science methodology from business understanding to modelling process with proficiency in **Python, SQL, Tableau, Power BI, Looker Studio and other tools** related to data analytics workflow from several coursework and bootcamps.

# Project Overview

Employee attrition poses a significant challenge to organizations, leading to substantial costs associated with hiring, training, and lost productivity. High turnover rates can disrupt business operations, lower morale, and decrease organizational efficiency. Traditional methods of predicting and mitigating employee turnover often rely on retrospective analyses and generalized strategies that fail to address individual employee needs and circumstances.

This project aims to develop a machine learning model to predict employee attrition accurately and provide actionable insights to improve employee retention strategies. By leveraging historical employee data, the model will identify patterns and factors contributing to employee turnover, enabling HR departments to implement proactive measures to retain valuable talent.

For more details, you can view the Jupyter notebook here.

# Dataset Overview

The original dataframe has 287 rows and 25 columns. The columns of our dataset are :

- □ Username : Username of the employee account
- □ EnterpriseID : ID of the employee in the company
- □ StatusPernikahan : Marital status of the employee
- □ JenisKelamin : Gender of the employee
- □ StatusKepegawaian : Employment status of the employee
- □ Pekerjaan : Role of the employee
- □ JenjangKarir : Level of experience of the employee
- □ PerformancePegawai : Employee performance category score
- □ AsalDaerah : Employee region of origin
- □ HiringPlatform : Platform the employee application is accepted
- □ SkorSurveyEngagement : Level of employee engagement within the organization
- □ SkorKepuasanPegawai : Level of how satisfied employees are with their job and the workplace
- □ JumlahKeikutsertaanProjek : Number of times the employee join a project
- □ JumlahKeterlambatanSebulanTerakhir : Number of times the employee is late
- □ JumlahKetidakhadiran : Number of times the employee is absent
- □ NomorHP : Handphone number of the employee
- □ Email : Personal email of the employee
- □ TingkatPendidikan : Education level Handphone number of the employee
- □ PernahBekerja : Whether the employee have previous work experience or not
- □ IkutProgramLOP : Whether the employee join LOP Program or not
- □ AlasanResign : Reason for resignation of the employee
- □ TanggalLahir : Birth date of the employee
- □ TanggalHiring : Hiring date of the employee
- □ TanggalPenilaianKaryawan : Scoring date of the employee
- □ TanggalResign : Resignation date of the employee

# Data Cleaning & Preprocessing

## Drop Unnecessary Columns and Handling Inconsistent Values Format

- PernahBekerja and IkutProgramLOP columns in the dataset is initially dropped because they don't provide valuable information for analysis or modeling process.

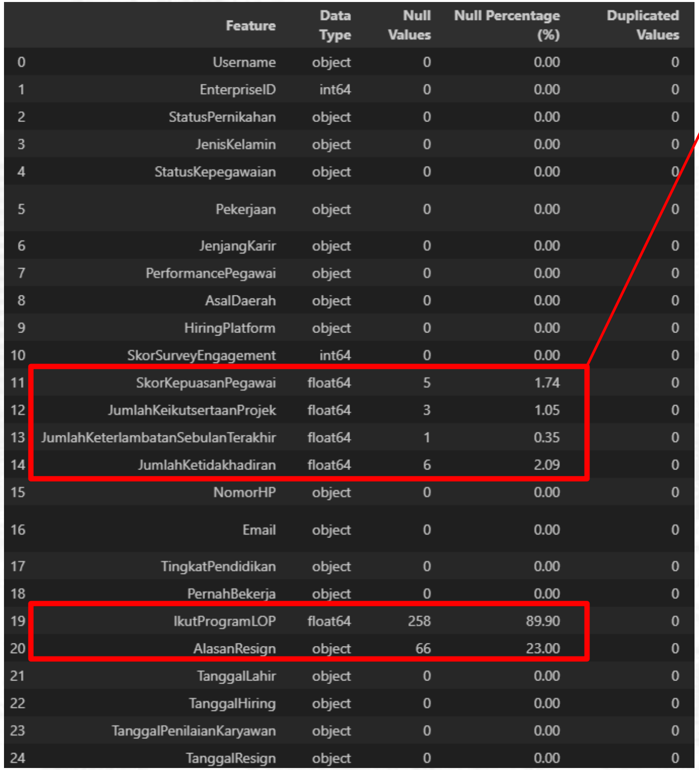
- Values in some columns are renamed maintain format similarity for the entire dataframe and also add more information/context to them. The values will be in the format of title case (ex : Data Analyst). Columns that have their values renamed are : (column name – example of their original values format)

  □ StatusPernikahan – Belum_menikah
  □ StatusKepegawaian – FullTime
  □ JenjangKarir – Senior_level
  □ PerformancePegawai – Sangat_bagus
  □ HiringPlatform – Employee_Referral
  □ AlasanResign – toxic_culture

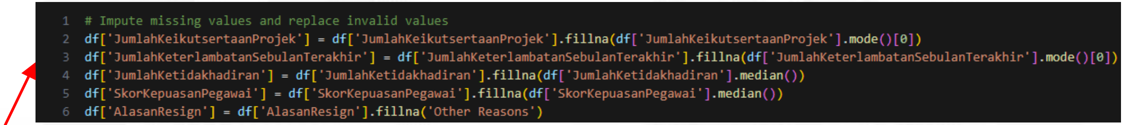
For more details, you can view the Jupyter notebook here.

# Data Cleaning & Preprocessing

## Handling Duplicated and Missing Values



| | Feature | Data Type | Null Values | Null Percentage (%) | Duplicated Values |
|---|---|---|---|---|---|
| 0 | Username | object | 0 | 0.00 | 0 |
| 1 | EnterpriseID | int64 | 0 | 0.00 | 0 |
| 2 | StatusPernikahan | object | 0 | 0.00 | 0 |
| 3 | JenisKelamin | object | 0 | 0.00 | 0 |
| 4 | StatusKepegawaian | object | 0 | 0.00 | 0 |
| 5 | Pekerjaan | object | 0 | 0.00 | 0 |
| 6 | JenjangKarir | object | 0 | 0.00 | 0 |
| 7 | PerformancePegawai | object | 0 | 0.00 | 0 |
| 8 | AsalDaerah | object | 0 | 0.00 | 0 |
| 9 | HiringPlatform | object | 0 | 0.00 | 0 |
| 10 | SkorSurveyEngagement | int64 | 0 | 0.00 | 0 |
| 11 | SkorKepuasanPegawai | float64 | 5 | 1.74 | 0 |
| 12 | JumlahKeikutsertaanProjek | float64 | 3 | 1.05 | 0 |
| 13 | JumlahKeterlambatanSebulanTerakhir | float64 | 1 | 0.35 | 0 |
| 14 | JumlahKetidakhadiran | float64 | 6 | 2.09 | 0 |
| 15 | NomorHP | object | 0 | 0.00 | 0 |
| 16 | Email | object | 0 | 0.00 | 0 |
| 17 | TingkatPendidikan | object | 0 | 0.00 | 0 |
| 18 | PernahBekerja | object | 0 | 0.00 | 0 |
| 19 | IkutProgramLOP | float64 | 258 | 89.90 | 0 |
| 20 | AlasanResign | object | 66 | 23.00 | 0 |
| 21 | TanggalLahir | object | 0 | 0.00 | 0 |
| 22 | TanggalHiring | object | 0 | 0.00 | 0 |
| 23 | TanggalPenilaianKaryawan | object | 0 | 0.00 | 0 |
| 24 | TanggalResign | object | 0 | 0.00 | 0 |



```python
# Impute missing values and replace invalid values
df['JumlahKeikutsertaanProjek'] = df['JumlahKeikutsertaanProjek'].fillna(df['JumlahKeikutsertaanProjek'].mode()[0])
df['JumlahKeterlambatanSebulanTerakhir'] = df['JumlahKeterlambatanSebulanTerakhir'].fillna(df['JumlahKeterlambatanSebulanTerakhir'].mode()[0])
df['JumlahKetidakhadiran'] = df['JumlahKetidakhadiran'].fillna(df['JumlahKetidakhadiran'].median())
df['SkorKepuasanPegawai'] = df['SkorKepuasanPegawai'].fillna(df['SkorKepuasanPegawai'].median())
df['AlasanResign'] = df['AlasanResign'].fillna('Other Reasons')
```

- No duplicated values in this dataset.

- There are 6 columns that missing values which are : SkorLepuasanPegawai, JumlahKeikutsertaanProjek, JumlahKeterlambatanSebulanTerakhir, JumlahKetidakhadiran, IkutProgramLOP, and AlasanResign.

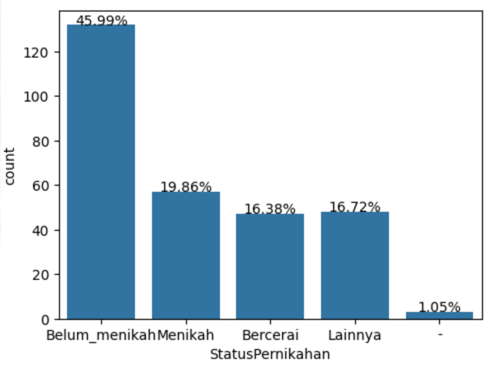
- We will handle them by **imputation with median and mode values** considering the distribution of the values and context (more explanation in code).

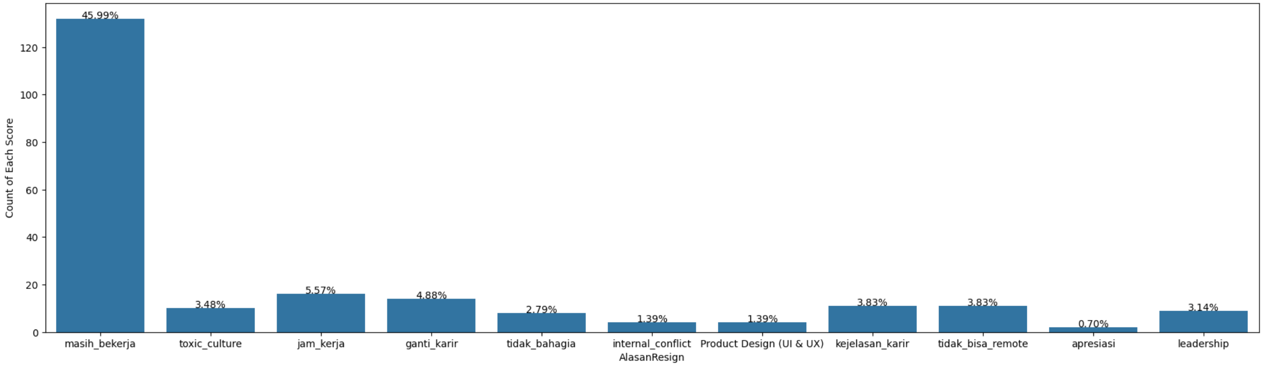
# Data Cleaning & Preprocessing

## Handling Out of Context Values

Some of the values in the dataset is out of context to what their columns should be. Columns that will be handled for this special case and the solutions are :

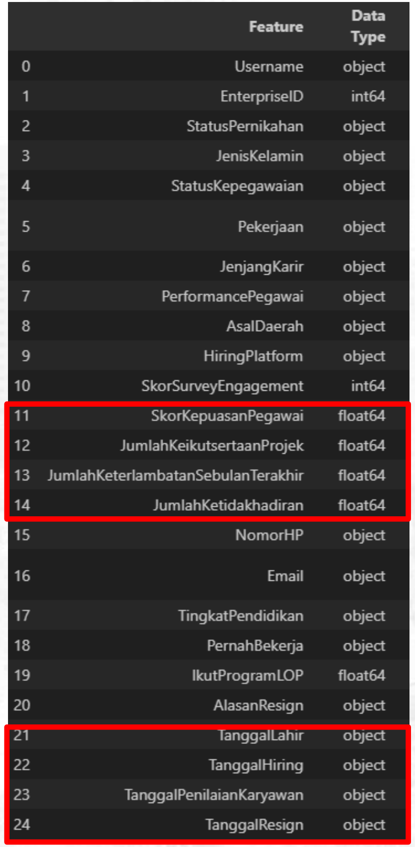
- ❑ StatusPernikahan : Replace the value of '-' to the mode.
- ❑ AlasanResign : Replace the value of 'Product Design (UI & UX)' to 'Oher Reasons'





# Data Cleaning & Preprocessing

## Data Types Correction



| | Feature | Data Type |
|---|---|---|
| 0 | Username | object |
| 1 | EnterpriseID | int64 |
| 2 | StatusPernikahan | object |
| 3 | JenisKelamin | object |
| 4 | StatusKepegawaian | object |
| 5 | Pekerjaan | object |
| 6 | JenjangKarir | object |
| 7 | PerformancePegawai | object |
| 8 | AsalDaerah | object |
| 9 | HiringPlatform | object |
| 10 | SkorSurveyEngagement | int64 |
| 11 | SkorKepuasanPegawai | float64 |
| 12 | JumlahKeikutsertaanProjek | float64 |
| 13 | JumlahKeterlambatanSebulanTerakhir | float64 |
| 14 | JumlahKetidakhadiran | float64 |
| 15 | NomorHP | object |
| 16 | Email | object |
| 17 | TingkatPendidikan | object |
| 18 | PernahBekerja | object |
| 19 | IkutProgramLOP | float64 |
| 20 | AlasanResign | object |
| 21 | TanggalLahir | object |
| 22 | TanggalHiring | object |
| 23 | TanggalPenilaianKaryawan | object |
| 24 | TanggalResign | object |

- Some columns have incorrect original datatypes which will be changed for better and correct result in further analysis.

- Float to integer : SkorKepuasanPegawai, JumlahKeikutsertaanProjek, JumlahKeterlambatanSebulanTerakhir, JumlahKetidakhadiran

**The values of these columns are discrete, so having float datatype might not actually make the code error, but it's just give better context in analysis and visualization later.**

- String to datetime : TanggalLahir, TanggalHiring, TanggalPenilaianKaryawan, TanggalResign

**Since we will need to extract datetime components, it's better to convert the datatypesof these columns to datetime.**