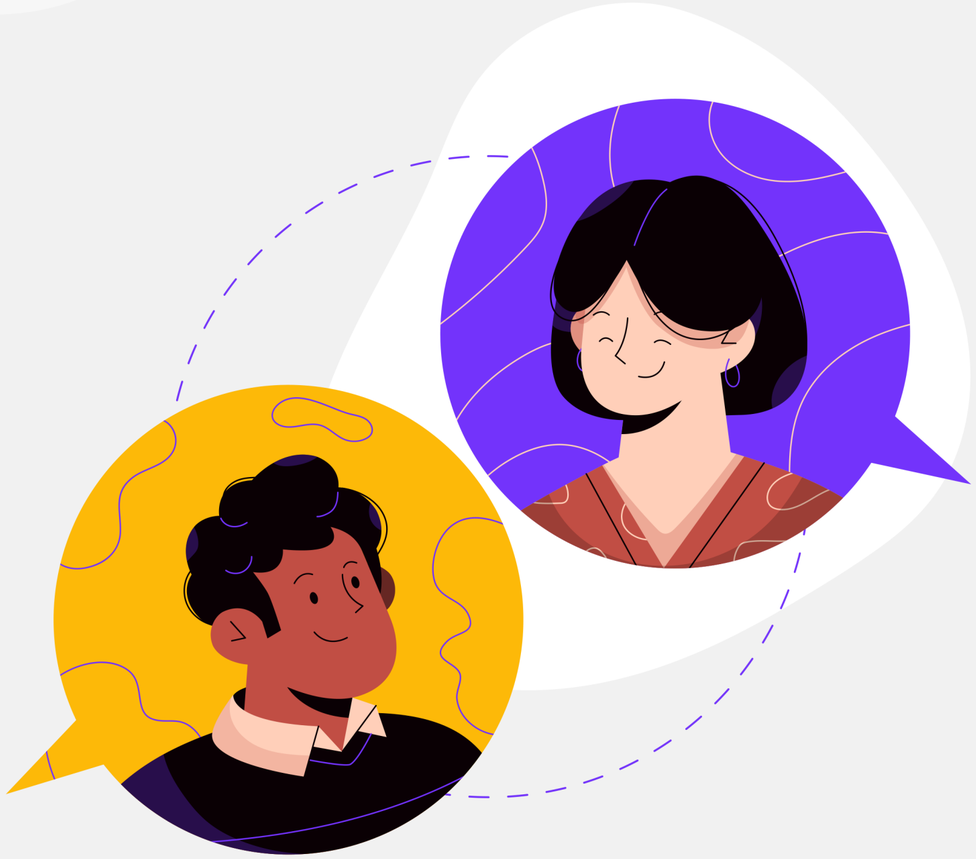
Improving Employee Retention by Predicting Employee Attrition Using Machine Learning

By: Joni Syofian

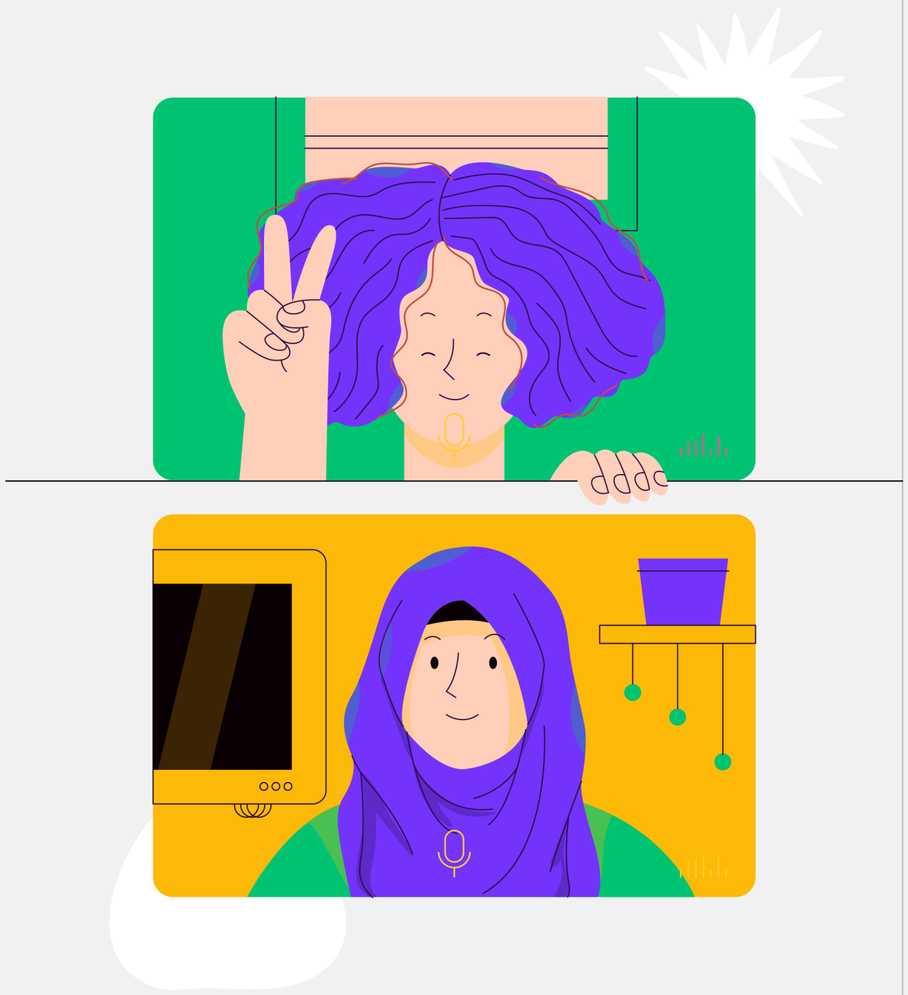


Supported by:
Rakamin Academy

ABOUT ME

Joni is a fresh graduate student from Bandung Institute of Technology. He is interested in data science, data analytics, and ocean issues. To improve his skills in the field of data, he took several courses and just completed a data science bootcamp with a good grade.

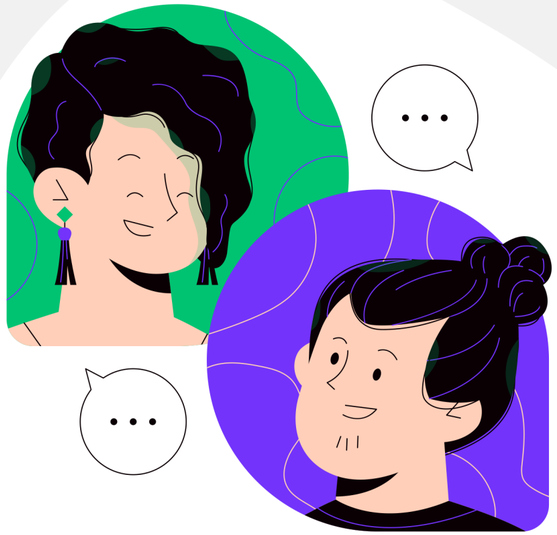
Table of Content



BUSINESS UNDERSTANDING

Background

Human resources (HR) are the main assets that need to be managed properly by a company so that business goals can be achieved effectively and efficiently. On this occasion, we will face a problem regarding human resources in the company. **Our focus is to find out how to keep employees afloat in an existing 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 employees to resign, companies can immediately overcome them by making programs that are relevant to employee problems.



Problem

Many employees have resigned from the company (31%)

Goals

- Reducing the rate of resignation from the company.
- Increase the retention of employees.

Objective

- Create a machine learning model that can predict employees who resign.
- Provide business recommendations based on the model that has been built.

Exploring Data & Data Preprocessing

Data Overview

#	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)

- **Existing data** is data that contains **related employee information whether related to reign or not**.
- There are **25 columns** and **287 rows** consisting of float64(5), int64(2), and object(18).
- **There are missing values** in the column (`SkorKepuasanPegawai`, `JumlahKeikutsertaanProjek`, `JumlahKeterlambatanSebulanTerakhir`, `JumlahKetidakhadiran`, `IkutProgramLOP` and`AlasanResign`)
- **There are no duplicate data**.
- **There are data types that do not match the data**.

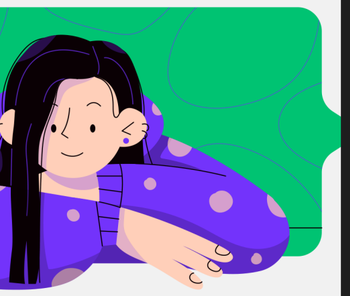
Stats | Bar Chart | Spectrum | Heat Map | Dendrogram

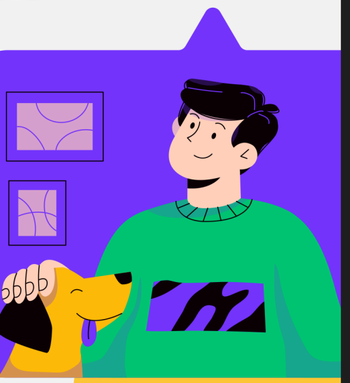
Missing Statistics

Missing Cells	339
Missing Cells (%)	4.7%
Missing Columns	6
Missing Rows	260
Avg Missing Cells per Column	13.56
Avg Missing Cells per Row	1.18

Exploring Data

Before data preprocessing is carried out, a missing value and an incorrect value are checked in the existing column. From the search results, it was found that:





```
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
```

From the information it is found that **the value that has the most is "masih_bekerja".** From this, it can be said that the data entered is not data from all employees who resigned but rather all employees. Assuming that, **when the data entry is "masih_bekerja" it means that the employee is still working,** while the other reason is the reason for resigning.
The existence of a missing value in this column can be traced further, namely by comparing it with the `TanggalResign` column to ascertain whether the missing value means that the employee is still working or not. When the `TanggalResign` column is empty or the like, it is assumed that the employee is still working.
The strange thing found here is the AlasanResign value, that is, with the **Product Design (UI & UX) value**, this value is the one that should work. So in this case, **this value will be replaced with "Lainnya".**

Exploring Data

- **JumlahKetidakhadiran Column**

```
4.0     20
2.0     20
15.0    19
7.0     17
16.0    17
6.0     16
14.0    16
3.0     16
19.0    15
13.0    15
17.0    14
1.0     14
20.0    13
9.0     12
11.0    12
10.0    10
8.0     10
5.0      9
12.0     7
18.0     7
50.0     1
55.0     1
Name: JumlahKetidakhadiran, dtype: int64
```

From the value in the `JumlahKetidakhadiran` column, there is no value of 0. Therefore, it is assumed that **the missing value in this column indicates that the employee is always present.** So that in the next stage the missing value in this column will **be filled with a value of 0.**

- **SkorKepuasanPegawai Column**

```
3.0    96
5.0    91
4.0    85
2.0     8
1.0     2
Name: SkorKepuasanPegawai, dtype: int64
```

```
1  df['SkorKepuasanPegawai'].isnull().sum()
```

```
5
```

From the information above, **it is found that 4 of the 5 missing values contained in the `SkorKepuasanPegawai` column are for employees who are still working.** So for the missing value handle in this column, **it is assumed that the Employee Satisfaction Score is at a value of 3, where this value of 3 is also the mode in the `SkorKepuasanPegawai` column.**

Exploring Data

- **JumlahKeikutsertaanProjek Column**

```
0.0    221
6.0     20
5.0     20
7.0      9
4.0      8
3.0      3
1.0      1
2.0      1
8.0      1
Name: JumlahKeikutsertaanProjek, dtype: int64
```

From the information above, it is found that **there is one dominant value,** namely the **value 0**, so this column can be filled with a value of 0.

- **PernahBekerja Column**

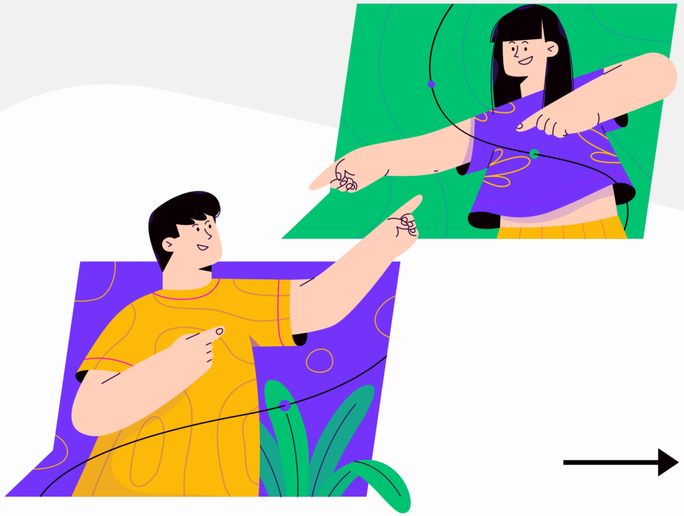
```
1      286
yes      1
Name: PernahBekerja, dtype: int64
```

In the `PernahBekerja` column, there **are two unique values, namely 'yes' and '1'** which actually **have the same meaning**, so this column will be dropped.

- **JumlahKeterlambatanSebulanTerakhir Column**

```
0.0    255
4.0      8
3.0      6
2.0      6
5.0      5
6.0      5
1.0      1
Name: JumlahKeterlambatanSebulanTerakhir, dtype: int64
```

From the information above, it is found that **there is one dominant value**, namely the **value 0**, so this column can be filled with a value of 0 for handling missing values.



Data Preprocessing

- **Change data types**

```
# change the data type to datetime
df2['EnterpriseID']= df2['EnterpriseID'].astype(str)
df2['TanggalLahir']=pd.to_datetime(df2['TanggalLahir'])
df2['TanggalHiring']=pd.to_datetime(df2['TanggalHiring'])
df2['TanggalPenilaianKaryawan']=pd.to_datetime(df2['TanggalPenilaianKaryawan'])
df2['TanggalResign']=pd.to_datetime(df2['TanggalResign'],errors='coerce')
```

- **Handling Missing Values & Incorrect Values**

```
# Handle missing value
df_drop['AlasanResign'].fillna('masih_bekerja', inplace=True)
df_drop['JumlahKetidakhadiran'].fillna(0, inplace=True)
df_drop['SkorKepuasanPegawai'].fillna(df['SkorKepuasanPegawai'].mode()[0], inplace=True)
df_drop['JumlahKeikutsertaanProjek'].fillna(df_drop['JumlahKeikutsertaanProjek'].mode()[0], inplace=True)
df_drop['JumlahKeterlambatanSebulanTerakhir'].fillna(df_drop['JumlahKeterlambatanSebulanTerakhir'].mode()[0], inplace=True)
df_drop.isnull().sum()
```

```
# replace the value '-' in the StatusPernikahan column with 'Lainnya'.
df_drop['StatusPernikahan'] = df_drop['StatusPernikahan'].replace({'-':'Lainnya'})
```

```
# change the value 'Product Design (UI & UX)' in the `AlasanResign` column to 'Lainnya'
df_drop['AlasanResign'] = df_drop['AlasanResign'].replace({'Product Design (UI & UX)':'Lainnya'})
```

- **Handling Outliers**

```
df_drop = df_drop.drop(df_drop[df_drop.JumlahKetidakhadiran>20].index)
```

There are two values can be dropped in this coulumn because they have a high `JumlahKetidakhadiran` value compared to other employees.

- **Drop Column**

Column drop is performed on a column that **has many missing values**, in this case the `IkutProgramLOP` column, and column drop is performed on **features that contain identities** related to employees, such as the `Username`, `EnterpriseID`, `NomorHP`, and `Email` columns. Then, also drop the `TanggalPenilaianKaryawan` column because it is **not related to the employee** but rather to the person assessing the employee. In addition, this is also done in the `PernahBekerja` column because it **only has one value**.

```
ls_drop = [ # missing value > 80%
    'IkutProgramLOP',
    # identity
    'Username',
    'EnterpriseID',
    'NomorHP',
    'Email',
    # unrelated to employees
    'TanggalPenilaianKaryawan',
    # has only one value
    'PernahBekerja'
]
```

```
df_drop = df2.drop(ls_drop, axis=1)
df_drop.info()
```

INSIGHT

Annual Report on Employee Number Changes

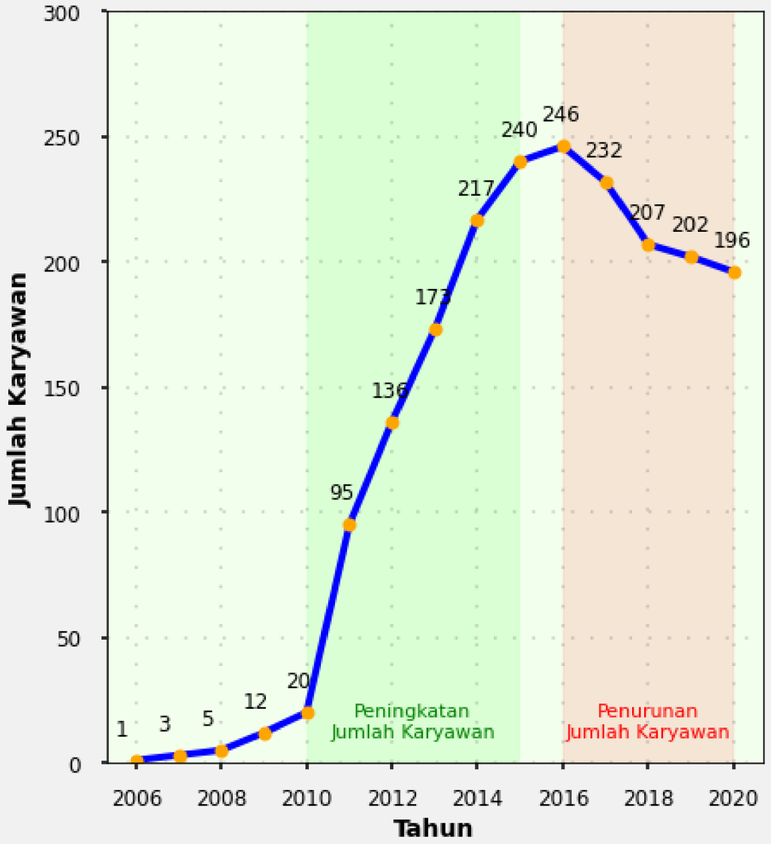
Resign Reason Analysis for Employee Attrition Management Strategy

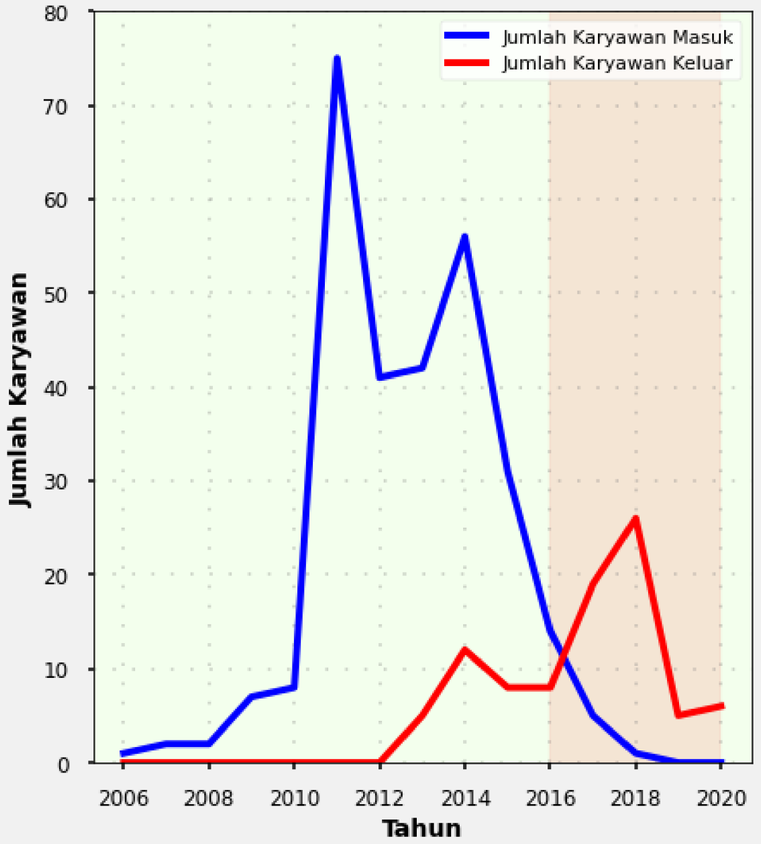
Resign Reason Analysis for Employee Attrition Management Strategy

Annual Report on Employee Number Changes

Terjadi peningkatan jumlah karyawan yang signifikan dari tahun 2010 hingga 2015
Namun, pada tahun 2016 hingga 2020 perusahaan mengalami penurunan jumlah karyawan

Penurunan jumlah karyawan ini ditandai dengan banyaknya karyawan yang keluar dibandingkan yang masuk.
Hal ini perlu dibenahi agar karyawan lain tetap bertahan sehingga rancangan strategi jangka panjang tetap dapat dilaksanakan tanpa kendala.



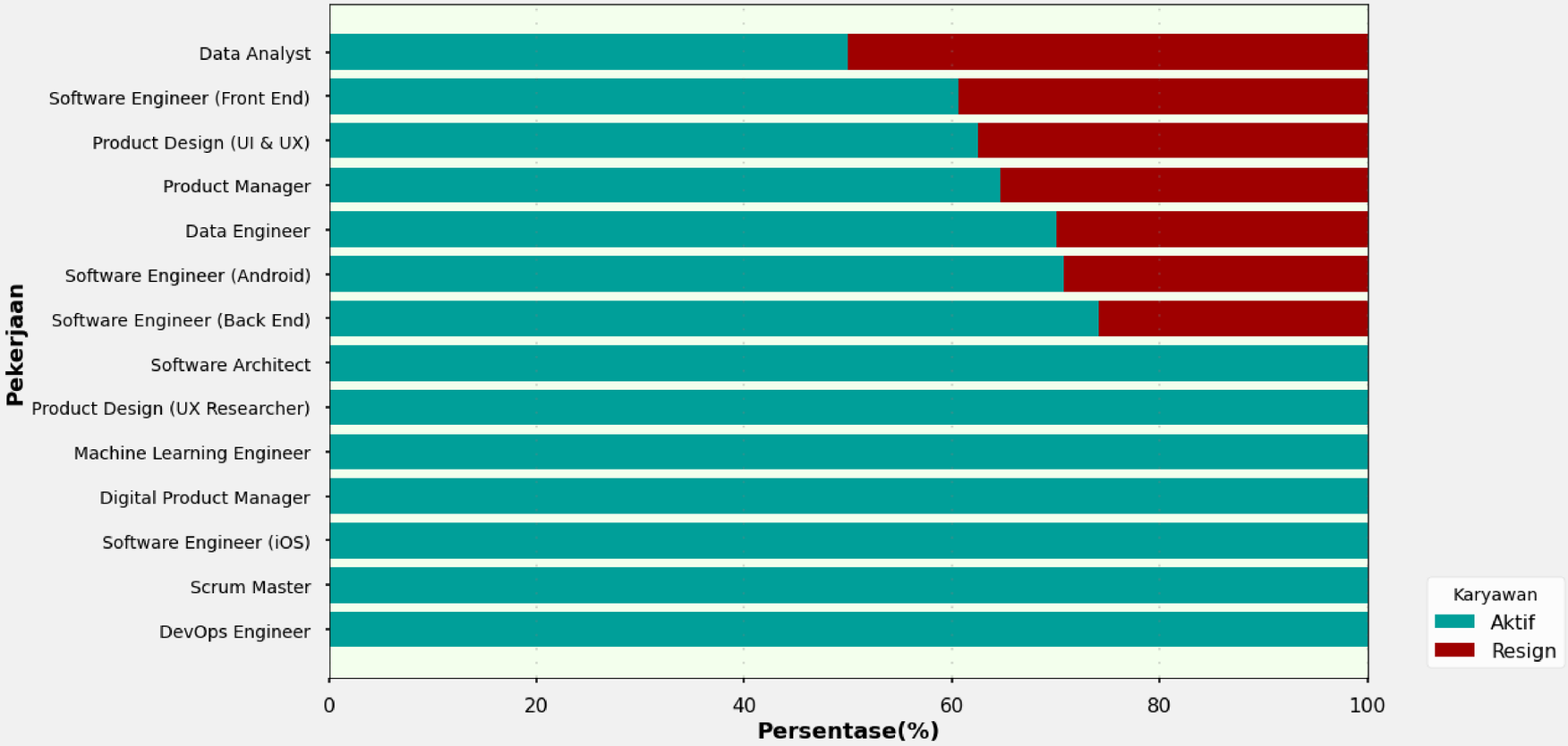


There was a **significant increase in the number of employees from 2010 to 2015.** However, from **2016 to 2020**, the company experienced a **decrease in the number of employees**. The decrease in the number of employees was marked by the number of employees leaving compared to those entering. This needs to be addressed so that other employees can survive, and long-term strategic plans can still be implemented without problems.

Resign Reason Analysis for Employee Attrition Management Strategy

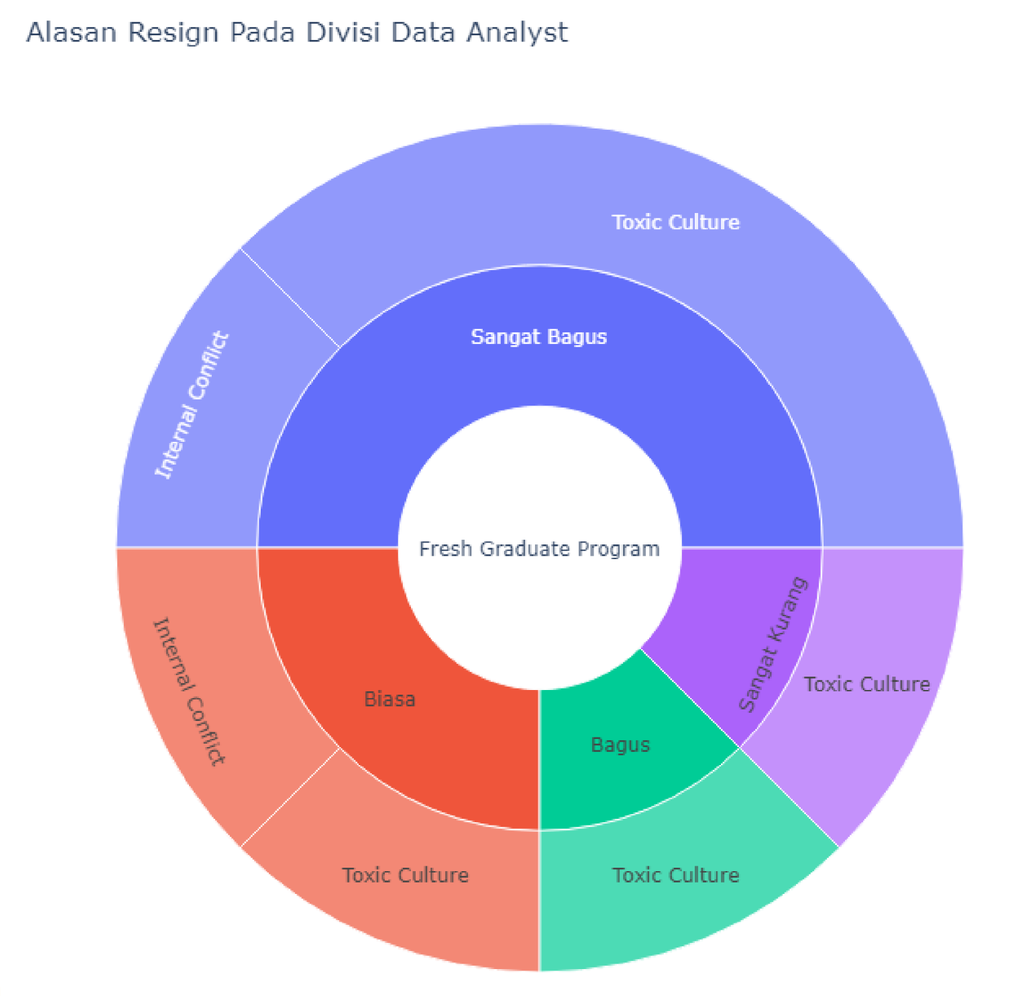
50% Karyawan dari divisi data analyst resign

Ini merupakan divisi dengan persentase resign yang terbesar
dengan posisi kedua dan ketiga adalah Software Engineer (Front End)(38.89%) dan Product Design (UI & UX)(37.50%)



From the picture, it is clear that **data analysts** are the **division with the highest percentage** of employees, namely **50%**. As a result, an investigation into the reasons why employees leave this division will be conducted.

Resign Reason Analysis for Employee Attrition Management Strategy

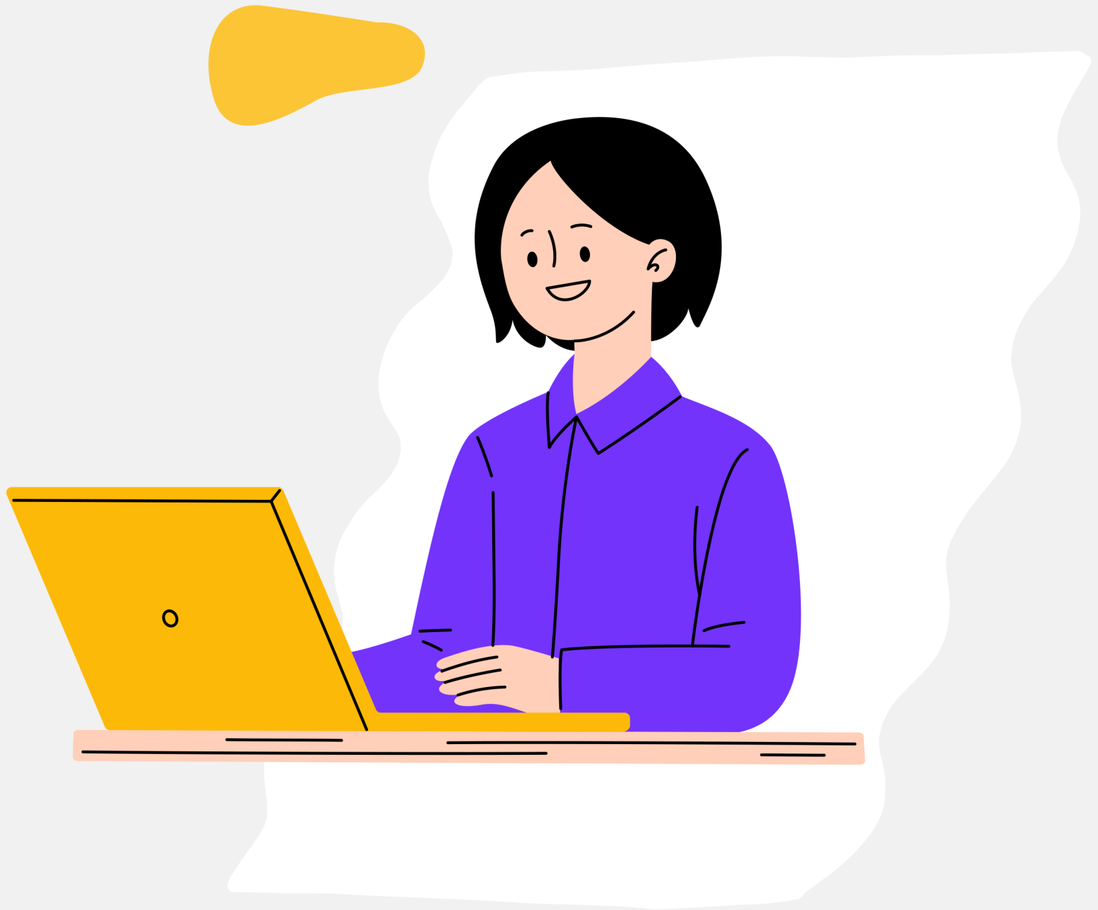


	JenjangKarir	PerformancePegawai	AlasanResign	JumlahKaryawan
0	Fresh Graduate Program	Sangat Bagus	Toxic Culture	3
1	Fresh Graduate Program	Bagus	Toxic Culture	1
2	Fresh Graduate Program	Biasa	Internal Conflict	1
3	Fresh Graduate Program	Biasa	Toxic Culture	1
4	Fresh Graduate Program	Sangat Bagus	Internal Conflict	1
5	Fresh Graduate Program	Sangat Kurang	Toxic Culture	1

From the information above, it is found that the data analyst division is **the division with the highest percentage of resignations, of which 50% are fresh graduates**. Three out of eight fresh graduates resigned for **reasons of toxic_culter and internal_conflict** and it's very **unfortunate that these three people are employees with very good performance**. This needs to be a lesson for the company's self-improvement so that something like this does not happen again.

The thing that must be done by the company is to **improve the internals of the company so that employees remain comfortable while working**. As a result, **businesses can engage in a variety of activities that foster a sense of kinship (internally)**. Moreover, many of the resigning employees are fresh graduates who feel that the environment is toxic and that there are also internal problems.

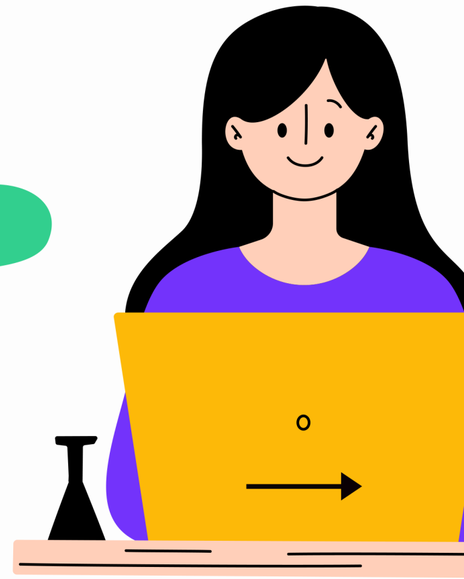
MODELING



MODELING

The following are the things to do at this stage:

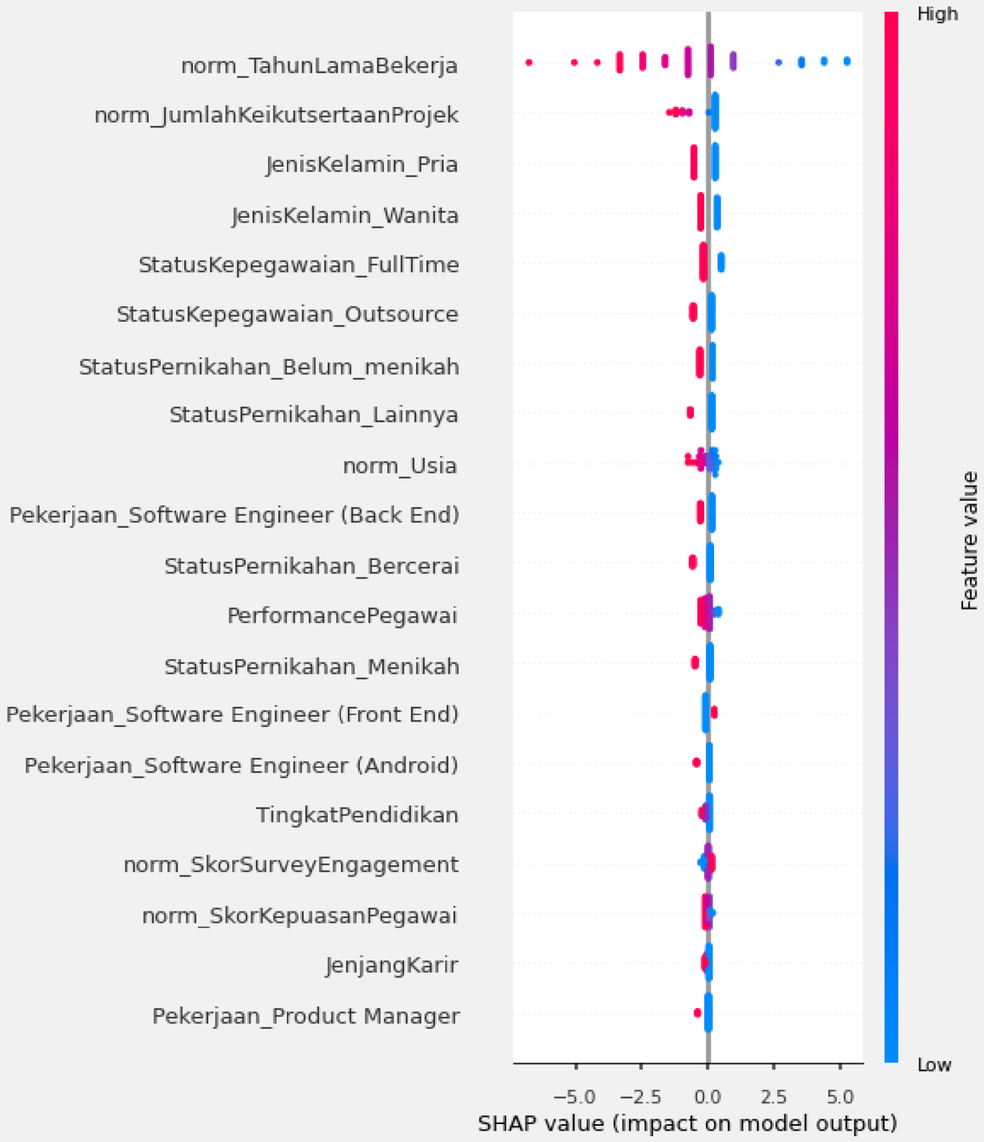
- **Split Train Test**
  Train : Test = 80 :20
- **Feature Scaling**
  The feature scaling used is Standardization
- **Oversampling**
  Oversampling is done on the train data set due to data imbalance.
- **Modeling**
  The algorithms used are LogisticRegression, KNN, DecisionTree, RandomForest, GradientBoostingClassifier, and XGBClassifier.
- **Tuning Hyperparameter**
  Performed on all algorithms to prevent overfitting.

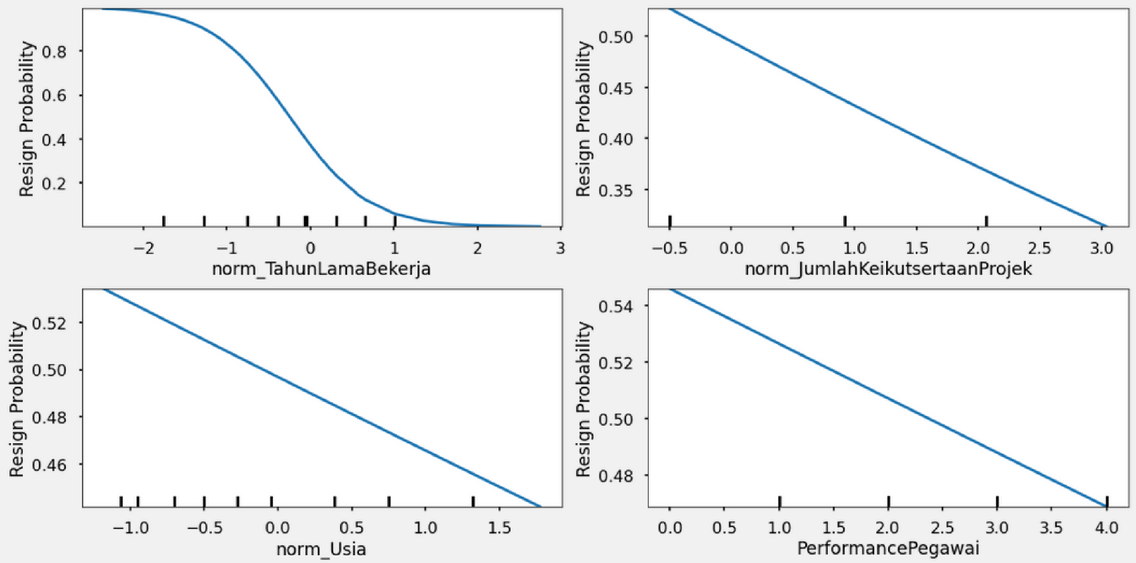


Model Comparison

	Accuracy	Precission	Recall	F1-Score	ROC-AUC (Test Prob)	ROC-AUC (Train Prob)
Tuned LogisticRegression	90.30%	68.40%	68.40%	68.40%	90.30%	93.10%
Tuned KNN	75.60%	52.20%	63.20%	57.10%	75.60%	100.00%
Tuned DecisionTree	82.30%	63.20%	63.20%	63.20%	82.30%	92.30%
Tunde RandomForest	77.30%	57.70%	78.90%	66.70%	73.00%	84.40%
Tuned GradientBoostingClassifier	87.70%	70.00%	73.70%	71.80%	87.50%	96.40%
Tuned XGBClassifier	87.30%	60.90%	73.70%	66.70%	87.30%	96.70%

From the results of several hyperparameter tunings, **the model chosen is a model with a logistic regression algorithm** because it **has good performance compared to other models**. It can be seen that the difference between the AUC train and test is not far, so this **model is not overfitting or underfitting** while the other models are still overfitting. Apart from that, evaluation metrics such as **accuracy, precision, recall, and f1-score also have better values than the other models.**





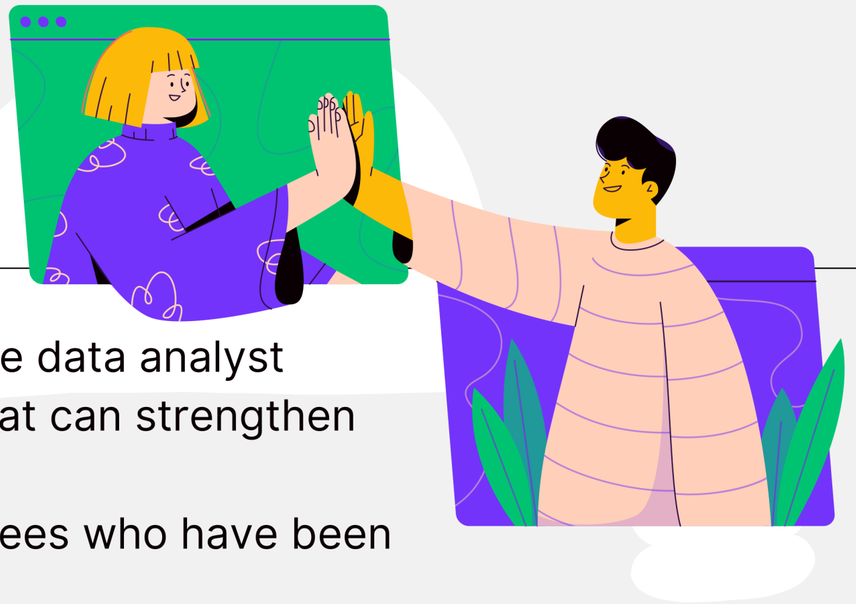
From the **importance of features** and **partial dependence**, it can be seen that:

- The **TahunLamaBekerja feature is the most important feature**. From this, it can be seen that there is a tendency where employees who resign are employees who have worked for a long time with little value, which means **that employees who tend to resign are employees who work for a short time**.
- Number of **JumlahKeikutsertaanProjek is the second important feature** where employees who tend to resign are employees **who have never participated in a project.**
- **Young employees have a high chance of resigning** compared to old age.
- **Employees** who have **low performance have a higher chance of resigning** compared to employees who have high performance.

Business Recommendation

Business Recommendation

As is well known, employees who tend to resign are those who have worked for a short time and have never participated in a project. Therefore, what companies can do to prevent employees from resigning is to make employees feel at home working with the company. For this reason, companies can do the following:



- **Improve internally**, as it is known that 50% of those who resign in the data analyst division have reasons related to internal, by carrying out activities that can strengthen the company's internal (employees with the company).
- **Give appreciation** in the form of loyalty on a regular basis to employees who have been working for a long time (> 10 years).
- **Provide projects to employees evenly (at least once**) so that all employees can feel they have contributed to the project.
- **Evaluate the work system** that is implemented so that employees are comfortable at work.
- Companies should **focus on retaining young employees** because t**hey are more likely to resign** than older employees.

THANKS

Email
jonisyofian14@gmail.com

LinkedIn
https://www.linkedin.com/in/jonisyofian/