



Data Science **PORTFOLIO**

By Muhammad Rajwa Athoriq

LinkedIn : www.linkedin.com/in/muhamad-rajwa-ath

GitHub : <https://github.com/rajwaAth>

IMPROVING EMPLOYEE RETENTION BY PREDICTING EMPLOYEE ATTRITION

| | |
|-----------|--|
| 03 | ABOUT ME |
| 04 | BUSINESS UNDERSTANDING |
| 06 | DATA UNDERSTANDING |
| 08 | DATA CLEANSING |
| 09 | FEATURE ENGINEERING |
| 10 | EXPLORATORY DATA ANALYSIS |
| 16 | DATA PREPROCESSING |
| 17 | MODELING & EVALUATION |
| 19 | FEATURE IMPORTANCE & BUSINESS RECOMENDATION |

introducing **ABOUT ME**

“I’m Rajwa, i dedicated data science student proficient in data analysis and machine learning techniques. Through rigorous training and academic exploration, I have honed my programming and data analysis skills. Experienced in optimizing marketing costs through data-driven insights. Committed to continuous learning and contributing to innovative projects in the field of information technology.”



 **Muhamad Rajwa AthorIQ**

BUSINESS UNDERSTANDING

Sebuah perusahaan teknologi menghadapi **masalah turnover** yang tinggi **di kalangan karyawan**, terutama dalam tim pengembang perangkat lunak. **Tingkat turnover ini menyebabkan biaya rekrutmen meningkat, penundaan proyek, dan menurunkan produktivitas.** Perusahaan ingin memahami faktor-faktor yang menyebabkan karyawan keluar dan mengambil langkah preventif untuk meningkatkan retensi.

[READ MORE](#)



PROBLEM

Turnover yang tinggi pada karyawan menyebabkan biaya rekrutmen meningkat, penundaan proyek, dan menurunkan produktivitas

OBJECTIVE

Menganalisis alasan utama terjadinya turnover & membangun model predictive karyawan dengan potensi resign tinggi

GOAL

Mengusulkan strategi preventif

Data UNDERSTANDING

Jumlah data yang digunakan berjumlah 287 baris dengan rincian sebagai berikut:

1. Imputasi missing value:

- Kolom numerik diimputasi dengan median.
- Kolom kategorikal diimputasi dengan modus.

2. Penghapusan kolom:

- Kolom IkutProgramLOP dan kolom kurang relevan seperti Username, PernahBekerja, EnterpriseID, NomorHP, dan Email dihapus.

| | column | type | null | number of unique value | sample unique value |
|----|------------------------------------|---------|------|------------------------|--|
| 0 | Username | object | 0 | 285 | [spiritedPorpoise3, jealousGelding2, pluckyMue... |
| 1 | EnterpriseID | int64 | 0 | 287 | [111065, 106080, 106452, 106325, 111171] |
| 2 | StatusPernikahan | object | 0 | 5 | [Belum_menikah, Menikah, Bercerai, Lainnya, -] |
| 3 | JenisKelamin | object | 0 | 2 | [Pria, Wanita] |
| 4 | StatusKepegawaian | object | 0 | 3 | [Outsource, FullTime, Internship] |
| 5 | Pekerjaan | object | 0 | 14 | [Software Engineer (Back End), Data Analyst, S... |
| 6 | JenjangKarir | object | 0 | 3 | [Freshgraduate_program, Senior_level, Mid_level] |
| 7 | PerformancePegawai | object | 0 | 5 | [Sangat_bagus, Sangat_kurang, Bagus, Biasa, Ku... |
| 8 | AsalDaerah | object | 0 | 5 | [Jakarta Timur, Jakarta Utara, Jakarta Pusat, ... |
| 9 | HiringPlatform | object | 0 | 9 | [Employee_Referral, Website, Indeed, LinkedIn, ... |
| 10 | SkorSurveyEngagement | int64 | 0 | 5 | [4, 3, 2, 1, 5] |
| 11 | SkorKepuasanPegawai | float64 | 5 | 5 | [4.0, 3.0, 5.0, nan, 2.0] |
| 12 | JumlahKeikutsertaanProjek | float64 | 3 | 9 | [0.0, 4.0, 6.0, nan, 7.0] |
| 13 | JumlahKeterlambatanSebulanTerakhir | float64 | 1 | 7 | [0.0, 4.0, 3.0, 5.0, 2.0] |
| 14 | JumlahKetidakhadiran | float64 | 6 | 22 | [9.0, 3.0, 11.0, 6.0, 10.0] |
| 15 | NomorHP | object | 0 | 287 | [+6282232522xxx, +6281270745xxx, +6281346215xx... |
| 16 | Email | object | 0 | 287 | [spiritedPorpoise3135@yahoo.com, jealousGeldin... |
| 17 | TingkatPendidikan | object | 0 | 3 | [Magister, Sarjana, Doktor] |
| 18 | PernahBekerja | object | 0 | 2 | [1, yes] |
| 19 | IkutProgramLOP | float64 | 258 | 2 | [1.0, 0.0, nan] |
| 20 | AlasanResign | object | 66 | 11 | [masih_bekerja, toxic_culture, jam_kerja, gant... |
| 21 | TanggalLahir | object | 0 | 284 | [1972-07-01, 1984-04-26, 1974-01-07, 1979-11-2... |
| 22 | TanggalHiring | object | 0 | 97 | [2011-01-10, 2014-01-06, 2014-2-17, 2013-11-11... |
| 23 | TanggalPenilaianKaryawan | object | 0 | 127 | [2016-2-15, 2020-1-17, 2016-01-10, 2020-02-04, ... |
| 24 | TanggalResign | object | 0 | 53 | [-, 2018-6-16, 2014-9-24, 2018-09-06, 2019-01-12] |


```
StatusPernikahan: ['Belum_menikah' 'Menikah' 'Bercerai' 'Lainnya' '-']

JenisKelamin: ['Pria' 'Wanita']

StatusKepegawaian: ['Outsource' 'FullTime' 'Internship']

Pekerjaan: ['Software Engineer (Back End)' 'Data Analyst'
'Software Engineer (Front End)' 'Product Manager'
'Software Engineer (Android)' 'Scrum Master'
'Product Design (UX Researcher)' 'Product Design (UI & UX)'
'Digital Product Manager' 'Data Engineer' 'Software Engineer (iOS)'
'DevOps Engineer' 'Software Architect' 'Machine Learning Engineer']

JenjangKarir: ['Freshgraduate_program' 'Senior_level' 'Mid_level']

PerformancePegawai: ['Sangat_bagus' 'Sangat_kurang' 'Bagus' 'Biasa' 'Kurang']

AsalDaerah: ['Jakarta Timur' 'Jakarta Utara' 'Jakarta Pusat' 'Jakarta Selatan'
'Jakarta Barat']

HiringPlatform: ['Employee_Referral' 'Website' 'Indeed' 'LinkedIn' 'CareerBuilder'
'Diversity_Job_Fair' 'Google_Search' 'On-line_Web_application' 'Other']

SkorSurveyEngagement: [4 3 2 1 5]

SkorKepuasanPegawai: [4. 3. 5. 2. 1.]

JumlahKeikutsertaanProjek: [0. 4. 6. 7. 3. 5. 1. 2. 8.]

JumlahKeterlambatanSebulanTerakhir: [0. 4. 3. 5. 2. 6. 1.]

TingkatPendidikan: ['Magister' 'Sarjana' 'Doktor']

AlasanResign: ['masih_bekerja' 'toxic_culture' 'jam_kerja' 'ganti_karir' 'tidak_bahagia'
'internal_conflict' 'Product Design (UI & UX)' 'kejelasan_karir'
'tidak_bisa_remote' 'apresiasi' 'leadership']
```

Terdapat adanya kesalahan dalam value pada kolom **‘StatusPernikahan’** dan **‘AlasanResign’**, dikarenakan keduanya merupakan kolom kategori maka value yang salah tersebut akan direplace dengan value terbanyak pada kolom tersebut

Data CLEANSING

Handle Missing Values

```
def impute_missing_value(df):  
    for col in df.columns:  
        if pd.api.types.is_numeric_dtype(df[col]):  
            df[col] = df[col].fillna(df[col].median())  
        else:  
            df[col] = df[col].fillna(df[col].mode()[0])  
    return df
```

Melakukan imputasi dengan menggunakan nilai paling sering muncul (mode) untuk kolom dengan tipe kategori dan nilai tengah (median) untuk kolom bertipe numerik

```
# Feature StatusPernikahan  
df_cln['StatusPernikahan'].replace(  
    {'-': df_cln['StatusPernikahan'].mode()[0]},  
    inplace=True)  
  
# Feature AlasanResign  
df_cln['AlasanResign'].replace(  
    {'Product Design (UI & UX)': df_cln['AlasanResign'].mode()[0]},  
    inplace=True)
```

Karena kolom yang mengalami kesalahan berisi data kategorik, nilai yang salah akan diganti dengan nilai yang paling sering (mode) muncul pada kolom tersebut.

Handle Missing Values

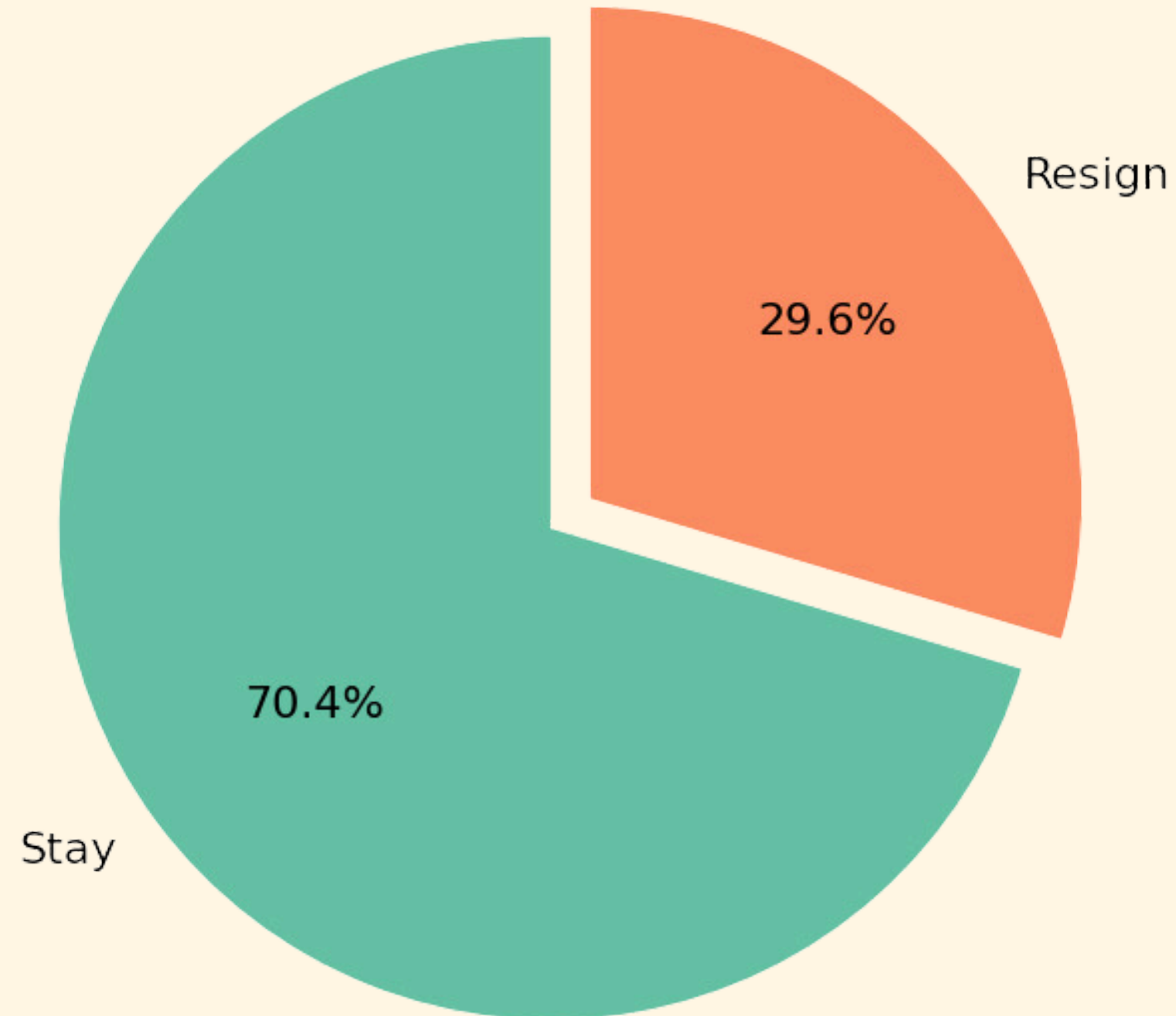
Feature ENGINEERING

Kesimpulan:

1. **Ekstrak tahun** dari kolom tanggal; nilai '-' dipertahankan jika tidak valid.
2. Hitung **AgeAtResign** (tahun resign - tahun lahir) dan **LengthWorked** (tahun resign - tahun hiring); default 0 jika tahun resign kosong.
3. **Kolom Resign**: 'Stay' untuk "masih_bekerja", 'Resign' untuk lainnya.
4. Hapus kolom tanggal asli.

```
df_cln['YearBirth'] = df_cln['TanggalLahir'].map(  
    lambda x: int(x[:4]) if x != '-' else '-')  
df_cln['YearHiring'] = df_cln['TanggalHiring'].map(  
    lambda x: int(x[:4]) if x != '-' else '-')  
df_cln['YearPenilaian'] = df_cln['TanggalPenilaianKaryawan'].map(  
    lambda x: int(x[:4]) if x != '-' else '-')  
df_cln['YearResign'] = df_cln['TanggalResign'].map(  
    lambda x: int(x[:4]) if x != '-' else '-')  
df_cln['AgeAtResign'] = df_cln.apply(  
    lambda row: row['YearResign'] - row['YearBirth']  
    if row['YearResign'] != '-' else 0, axis=1)  
df_cln['LengthWorked'] = df_cln.apply(  
    lambda row: row['YearResign'] - row['YearHiring']  
    if row['YearResign'] != '-' else 0, axis=1).astype(int)  
df_cln['LengthWorked'] = df_cln['LengthWorked'].map(  
    lambda x: 0 if x < 0 else x)  
df_cln['Resign'] = df_cln['AlasanResign'].apply(  
    lambda x: 'Stay' if x == 'masih_bekerja' else 'Resign')  
  
# # Drop columns  
df_cln.drop(['TanggalLahir', 'TanggalHiring',  
            'TanggalPenilaianKaryawan', 'TanggalResign'],  
            axis=1, inplace=True)
```

Percentage of Employees Resign or Not



Data Status

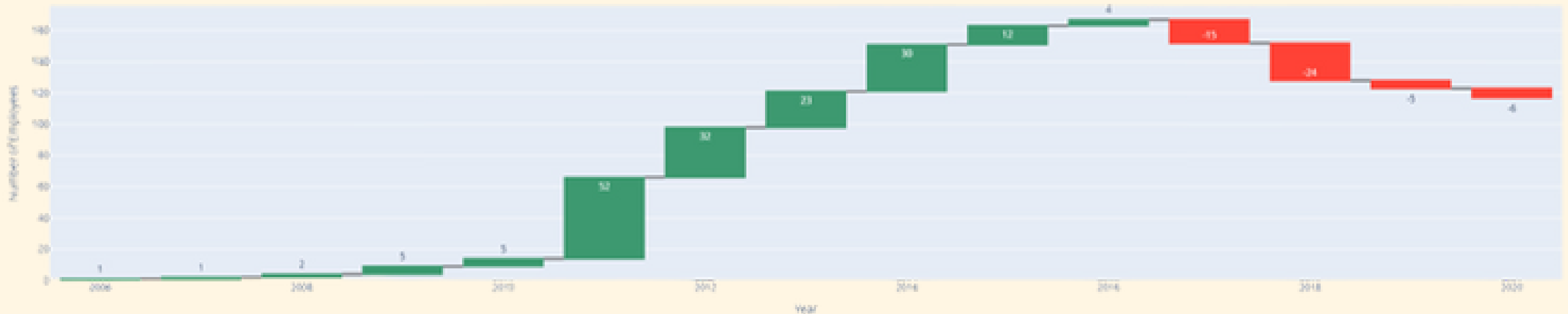
Imbalance

Interpretasi

Tingkat turnover 29,6% menunjukkan perlunya perhatian pada alasan karyawan resign agar perusahaan dapat menciptakan lingkungan kerja yang mendukung dan meningkatkan retensi.

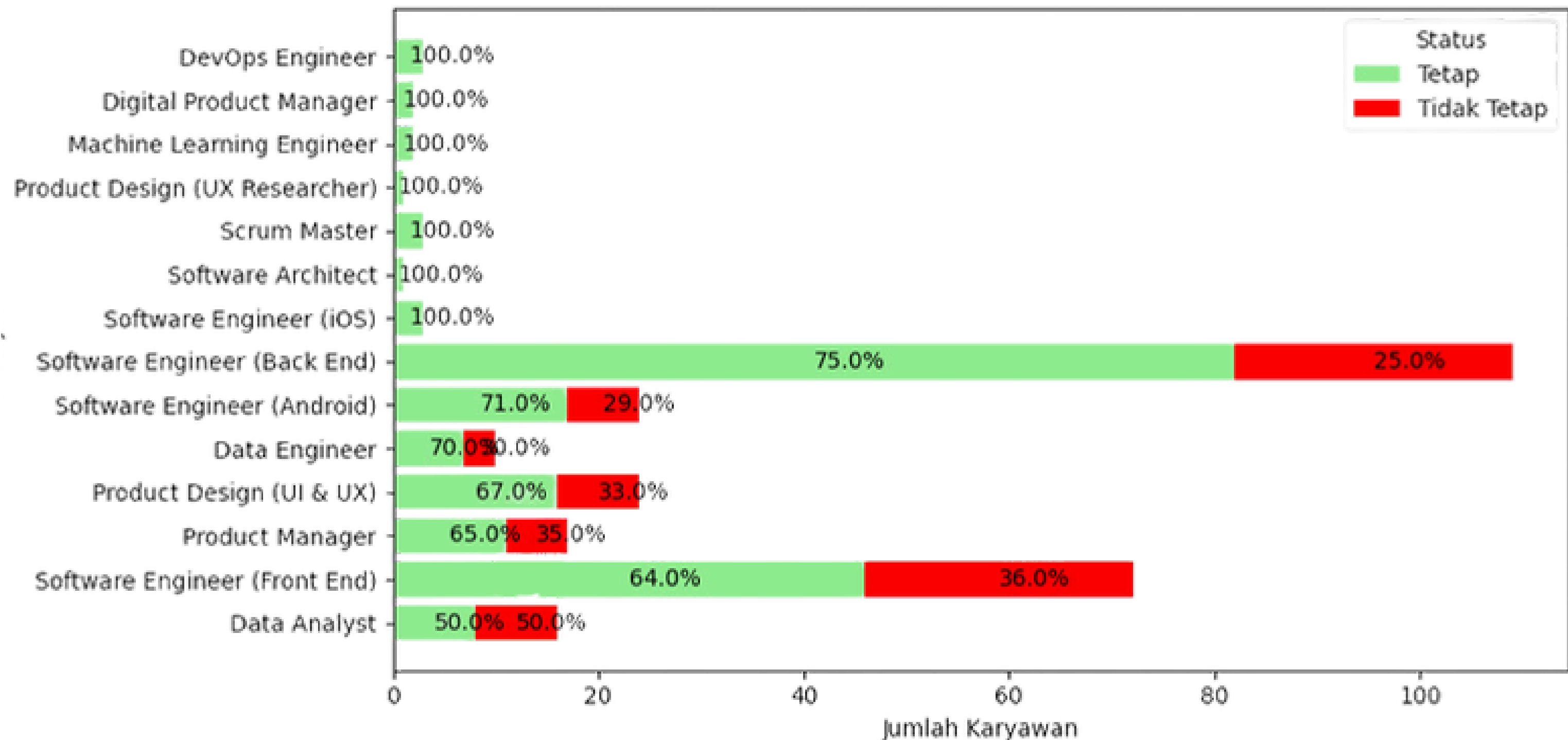
Number of Ups and DOWNS OF EMPLOYEES

Perusahaan tumbuh signifikan dari 2006 hingga 2014, dengan lonjakan pada 2011. Namun, terjadi penurunan karyawan signifikan pada 2017-2018 (-15, -24) dan tren menurun berlanjut hingga 2020, mengindikasikan tantangan seperti turnover tinggi atau restrukturisasi yang perlu segera diatasi.



Percentage employee

STAY AND RESIGN

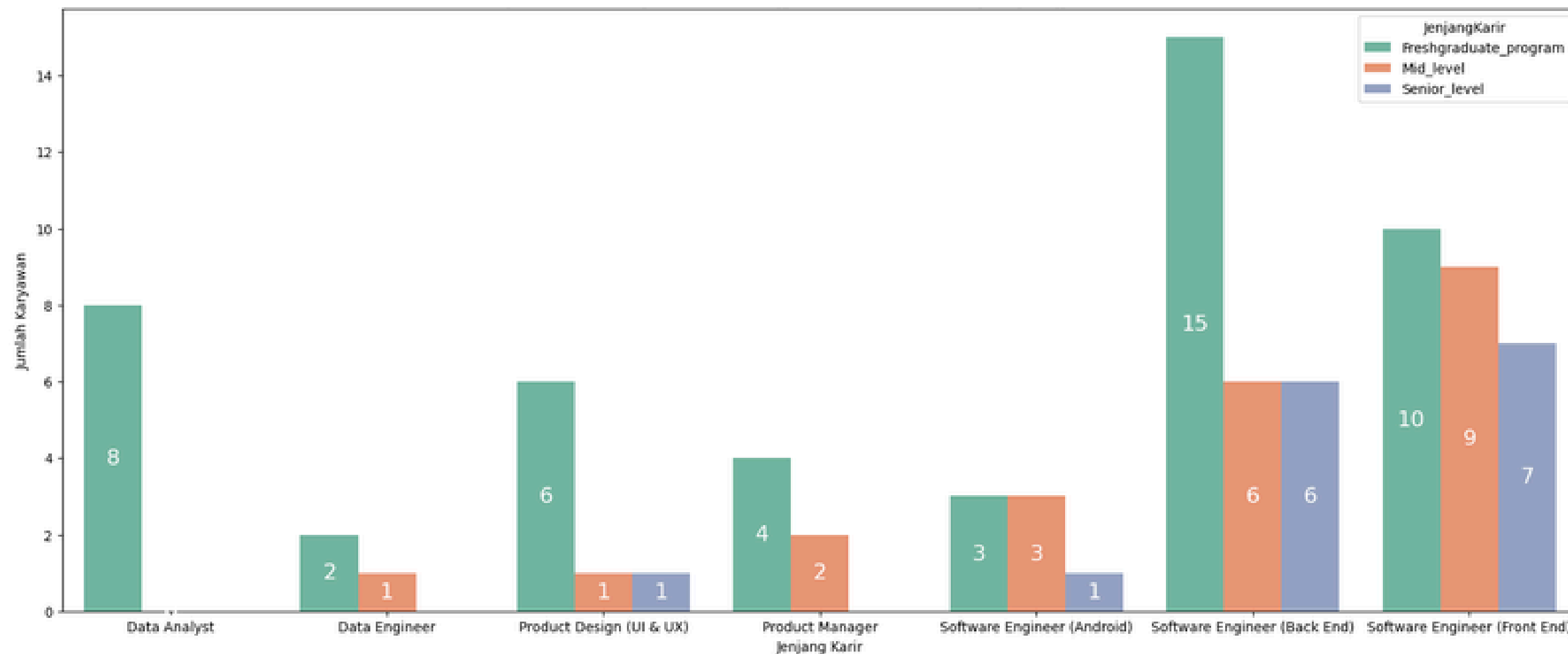


Percentage OF REASONS FOR RESIGN BY DIVISION

1. Tidak Bisa Remote: Dominan pada Software Engineer (Front End) (72.73%) dan Data Analyst (27.27%).
2. Tidak Bahagia: Menonjol pada Software Engineer (Android) (62.50%).
3. Jam Kerja: Berdampak pada Software Engineer (Back End) (37.50%).
4. Toxic Culture: Signifikan pada Data Analyst (60.00%).
5. Kejelasan Karir: Memengaruhi Product Manager (36.36% dan 33.33%).
6. Leadership: 4 dari 6 divisi mengalami permasalahan pada sisi leadership

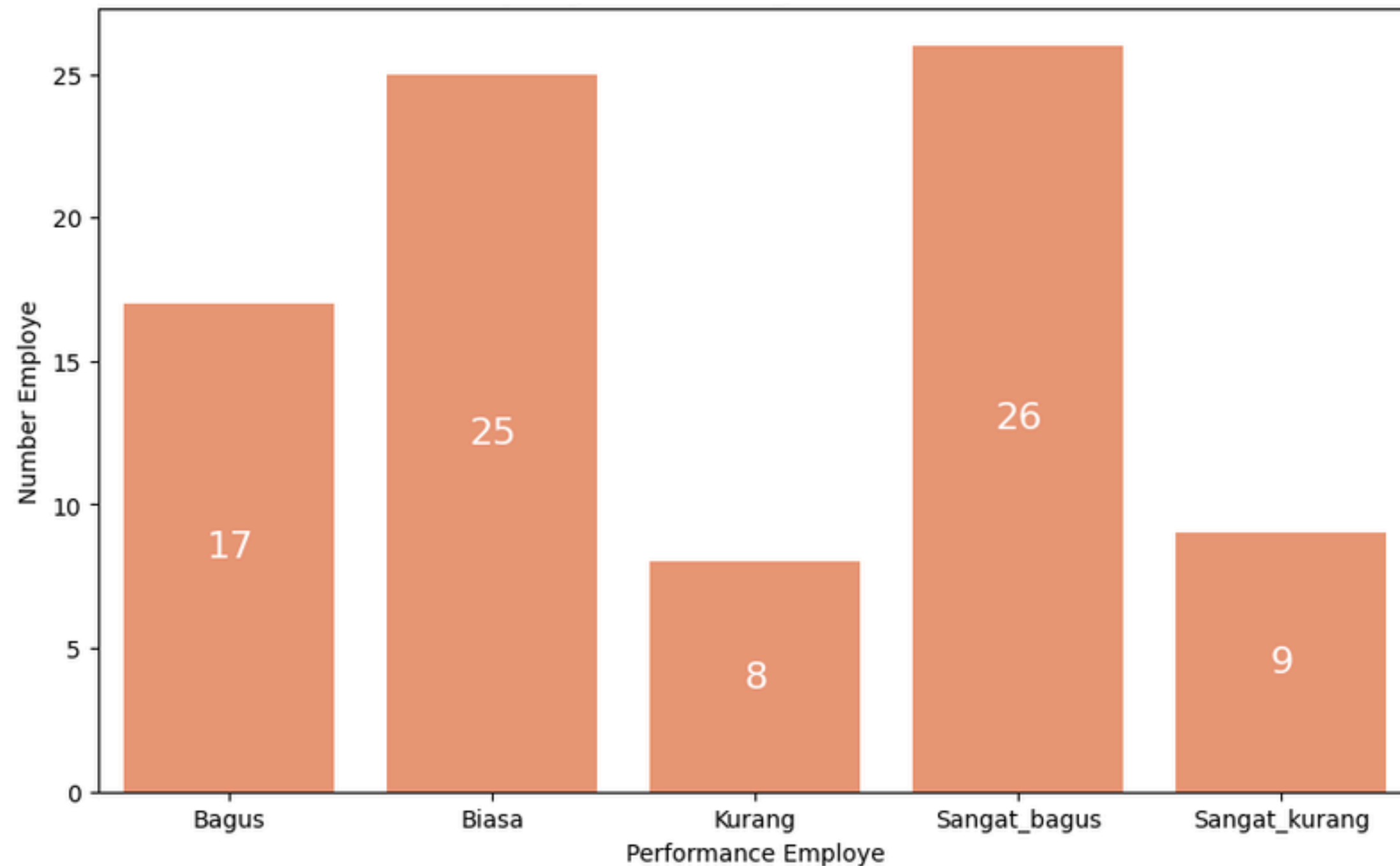


Number of Resigned EMPLOYEES BY CAREER LEVEL



Karyawan dengan jenjang karier fresh graduate memiliki potensi terbesar untuk resign, terlihat dari semua divisi yang menunjukkan jumlah resign terbanyak berasal dari kategori fresh graduate.

Number of Employees RESIGN BASED ON PERFORMANCE



01 Sangat bagus

Resignnya karyawan berperforma sangat bagus adalah masalah serius yang dapat mengancam stabilitas operasional dan menurunkan daya saing perusahaan di pasar.

02 Biasa

Diikuti oleh karyawan dengan kinerja yang tergolong standar. Meskipun demikian, mereka memiliki potensi untuk berkembang menjadi sumber daya unggul apabila diberikan pelatihan yang tepat

03 Bagus

Karyawan berperforma baik merupakan kelompok ketiga terbanyak yang resign, mengindikasikan masalah dalam kepuasan kerja.

Data

REPROCESSING

ENCODING FEATURE

Dilakukan label encoding untuk kolom dengan nilai ordinal dan one-hot encoding untuk kolom dengan nilai non-ordinal.

SPLITTING DATA

Membagi dataset menjadi data train dan test dengan perbandingan 70:20, 70% pada train dan 20% pada test

SCALING VALUES

Melakukan scaling nilai menggunakan StandardScaler untuk menyeragamkan skala nilai pada setiap kolom.

MODELING

| | model | acc_train | acc_test | prec_train | prec_test | rec_train | rec_test | f1_train | f1_test | roc_train | roc_test |
|---|---------------------|-----------|----------|------------|-----------|-----------|----------|----------|---------|-----------|----------|
| 0 | Logistic Regression | 0.99 | 0.98 | 0.99 | 0.95 | 1.00 | 1.00 | 0.99 | 0.97 | 0.99 | 0.99 |
| 1 | Decision Tree | 1.00 | 0.95 | 1.00 | 0.94 | 1.00 | 0.89 | 1.00 | 0.91 | 1.00 | 0.93 |
| 2 | Random Forest | 1.00 | 0.98 | 1.00 | 0.95 | 1.00 | 1.00 | 1.00 | 0.97 | 1.00 | 0.99 |
| 3 | Ada Boost | 1.00 | 0.98 | 1.00 | 0.95 | 1.00 | 1.00 | 1.00 | 0.97 | 1.00 | 0.99 |
| 4 | SVC | 0.99 | 0.95 | 0.99 | 0.94 | 1.00 | 0.89 | 0.99 | 0.91 | 0.99 | 0.93 |
| 5 | KNeighbors | 0.94 | 0.81 | 0.92 | 0.68 | 0.97 | 0.72 | 0.94 | 0.70 | 0.94 | 0.79 |
| 6 | Gaussian NB | 0.54 | 0.31 | 0.52 | 0.31 | 1.00 | 1.00 | 0.69 | 0.47 | 0.54 | 0.50 |
| 7 | XGBoost | 1.00 | 0.97 | 1.00 | 0.94 | 1.00 | 0.94 | 1.00 | 0.94 | 1.00 | 0.96 |

Logistic Regression menunjukkan risiko overfitting paling rendah dibandingkan AdaBoost dan Random Forest, dengan skor train di bawah 100% dan gap train-test terkecil meskipun kinerja test serupa.

Focused Matrix

F1 - Score

Reason

Memilih F1-Score menyeimbangkan deteksi resign (Recall) dan tindakan tepat sasaran (Precision), meningkatkan efisiensi dan mengurangi kerugian.

Model EVALUATION

Setelah dilakukan hyperparameter tuning, skor menunjukkan tidak ada perubahan dari pengaturan default, yang menandakan bahwa model sudah memiliki skor yang baik sejak awal.

Selain itu, hasil cross-validation menunjukkan bahwa model memiliki skor yang optimal pada setiap fold-nya.

| model | f1_train | f1_test |
|------------------------------|----------|---------|
| Logistic Regression(Default) | 0.99 | 0.97 |
| Logistic Regression (Tuning) | 0.99 | 0.97 |

```
Best parameters: {'solver': 'saga', 'penalty': 'l1',  
                  'max_iter': 1000, 'l1_ratio': 1.0, 'C': 1.0}
```

Cross-validation scores for each fold:

Fold 1: 0.9846153846153847

Fold 2: 0.9846153846153847

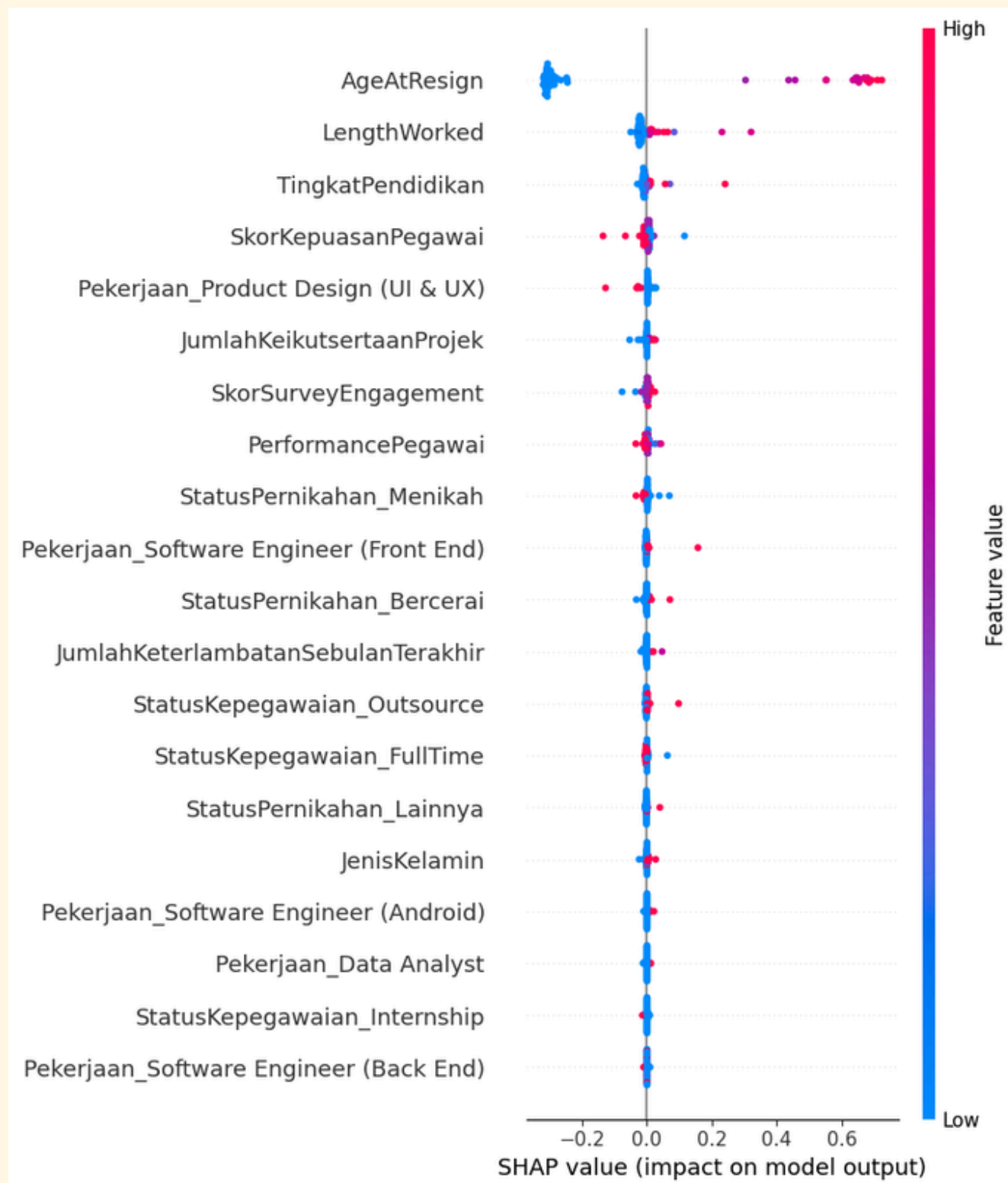
Fold 3: 0.9850746268656716

Fold 4: 1.0

Fold 5: 1.0

Mean cross-validation score: 0.9908610792192881

Feature IMPORTANCE



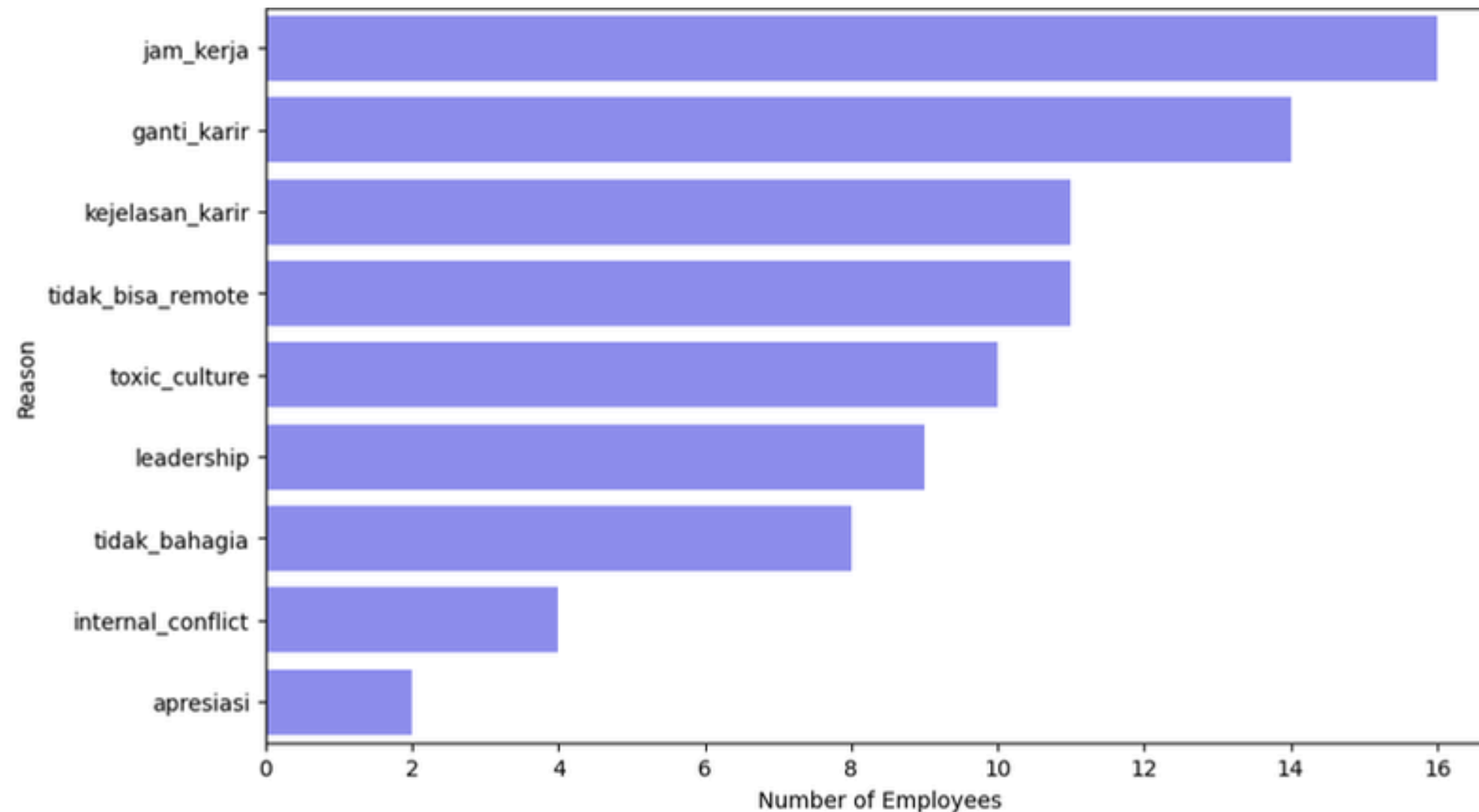
Kesimpulan:

- **Usia Saat Resign:** Karyawan yang sudah tua cenderung lebih sering resign.
- **Lama Bekerja:** Semakin lama bekerja, semakin kecil kemungkinan resign.
- **Tingkat Pendidikan:** Karyawan dengan tingkat pendidikan lebih tinggi cenderung resign lebih sering.
- **Skor Kepuasan Karyawan:** Karyawan dengan kepuasan rendah lebih berisiko resign.

Rekomendasi Bisnis:


- **Usia Saat Resign:** Untuk karyawan yang lebih tua, tawarkan fleksibilitas kerja, program pensiun, atau jalur pengembangan yang sesuai dengan tahap karir mereka untuk mempertahankan mereka lebih lama.
- **Lama Bekerja:** Berikan penghargaan dan peluang pengembangan karir bagi karyawan yang sudah lama bekerja agar mereka tetap merasa dihargai dan termotivasi.
- **Tingkat Pendidikan:** Tawarkan peluang untuk pengembangan lebih lanjut, proyek yang menantang, dan jalur karir yang jelas bagi karyawan dengan pendidikan tinggi untuk menjaga mereka tetap terlibat.
- **Kepuasan Karyawan:** Tingkatkan faktor-faktor yang mempengaruhi kepuasan kerja, seperti lingkungan kerja, kompensasi, pengakuan, dan keseimbangan kerja-hidup, melalui survei dan umpan balik rutin.

The Most **COMMON REASON EMPLOYEE RESIGN**



- **Fleksibilitas Jam Kerja:** Berikan opsi jam kerja fleksibel atau kebijakan kerja jarak jauh untuk meningkatkan keseimbangan kerja-hidup.
- **Kejelasan Jalur Karir:** Sediakan kesempatan pengembangan dan pelatihan yang jelas untuk mendukung karir karyawan.
- **Perbaiki Budaya Kerja:** Ciptakan lingkungan kerja sehat dengan mengurangi budaya toksik dan meningkatkan komunikasi.
- **Kepemimpinan yang Lebih Baik:** Latih pemimpin untuk mengelola tim dengan lebih efektif dan menginspirasi karyawan.
- **Penghargaan dan Apresiasi:** Tingkatkan pengakuan terhadap karyawan berprestasi melalui sistem penghargaan yang jelas.

**THANKS
FOR
WATCHING**



thorIQ004@gmail.com