Improving Employee Retention by Predicting Employee Attrition Using Machine Learning

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

Created by:
Limatan Luviar
Limatan.junior@gmail.com
Limatan Luviar

Graduated from Universitas Sriwijaya majoring in Informatics Management in 2018 and interned at a government institution. Enrolling in and completing the Data Science Bootcamp at Rakamin Academy in 2024 was driven by my interest in data analysis. Currently, I am actively seeking job opportunities and have completed various data-driven projects at Rakamin. My focus is on Data Analyst roles, where I can utilize my analytical skills and business understanding to make a significant impact

Untuk selengkapnya, dapat melihat jupyter notebook disini

Overview

“Sumber daya manusia (SDM) adalah aset utama yang perlu dikelola dengan baik oleh perusahaan agar tujuan bisnis dapat tercapai dengan efektif dan efisien. Pada kesempatan kali ini, kita akan menghadapi sebuah permasalahan tentang sumber daya manusia yang ada di perusahaan. Fokus kita adalah untuk mengetahui bagaimana cara menjaga karyawan agar tetap bertahan di perusahaan yang ada saat ini yang dapat mengakibatkan bengkaknya biaya untuk rekrutmen karyawan serta pelatihan untuk mereka yang baru masuk. Dengan mengetahui faktor utama yang menyebabkan karyawan tidak merasa, perusahaan dapat segera menanggulanginya dengan membuat program-program yang relevan dengan permasalahan karyawan. “

Data Preprocessing

Rakamin Academy

Dataset ini terdiri dari 25 kolom dan 287 baris data dan berisi informasi tentang karyawan, termasuk data demografis (seperti jenis kelamin, status pernikahan, dan asal daerah), data pekerjaan (seperti jenis pekerjaan dan performa), dan data historis (seperti tanggal hiring, penilaian karyawan, dan resign). Dataset ini juga mencakup beberapa nilai null, terutama pada fitur "IkutProgramLOP" dan "AlasanResign".

	Feature	Data Type	Null	Null (%)	Unique	Unique Sample
0	Username	object	0	0.000000	285	[spiritedPorpoise3, jealousGelding2, pluckyMue...
1	EnterpriseID	int64	0	0.000000	287	[111065, 106080, 106452, 106325, 111171]
2	StatusPernikahan	object	0	0.000000	5	[Belum_menikah, Menikah, Bercerai, Lainnya, -]
3	JenisKelamin	object	0	0.000000	2	[Pria, Wanita]
4	StatusKepegawaian	object	0	0.000000	3	[Outsource, FullTime, Internship]
5	Pekerjaan	object	0	0.000000	14	[Software Engineer (Back End), Data Analyst, S...
6	JenjangKarir	object	0	0.000000	3	[Freshgraduate_program, Senior_level, Mid_level]
7	PerformancePegawai	object	0	0.000000	5	[Sangat_bagus, Sangat_kurang, Bagus, Biasa, Ku...
8	AsalDaerah	object	0	0.000000	5	[Jakarta Timur, Jakarta Utara, Jakarta Pusat, ...
9	HiringPlatform	object	0	0.000000	9	[Employee_Referral, Website, Indeed, LinkedIn,...
10	SkorSurveyEngagement	int64	0	0.000000	5	[4, 3, 2, 1, 5]
11	SkorKepuasanPegawai	float64	5	1.742160	5	[4.0, 3.0, 5.0, nan, 2.0]
12	JumlahKeikutsertaanProjek	float64	3	1.045296	9	[0.0, 4.0, 6.0, nan, 7.0]
13	JumlahKeterlambatanSebulanTerakhir	float64	1	0.348432	7	[0.0, 4.0, 3.0, 5.0, 2.0]
14	JumlahKetidakhadiran	float64	6	2.090592	22	[9.0, 3.0, 11.0, 6.0, 10.0]
15	NomorHP	object	0	0.000000	287	[+6282232522xxx, +6281270745xxx, +6281346215xx...
16	Email	object	0	0.000000	287	[spiritedPorpoise3135@yahoo.com, jealousGeldin...
17	TingkatPendidikan	object	0	0.000000	3	[Magister, Sarjana, Doktor]
18	PernahBekerja	object	0	0.000000	2	[1, yes]
19	IkutProgramLOP	float64	258	89.895470	2	[1.0, 0.0, nan]
20	AlasanResign	object	66	22.996516	11	[masih_bekerja, toxic_culture, jam_kerja, gant...
21	TanggalLahir	object	0	0.000000	284	[1972-07-01, 1984-04-26, 1974-01-07, 1979-11-2...
22	TanggalHiring	object	0	0.000000	97	[2011-01-10, 2014-01-06, 2014-2-17, 2013-11-11...
23	TanggalPenilaianKaryawan	object	0	0.000000	127	[2016-2-15, 2020-1-17, 2016-01-10, 2020-02-04,...
24	TanggalResign	object	0	0.000000	53	[-, 2018-6-16, 2014-9-24, 2018-09-06, 2019-01-12]

Data Preprocessing

- Dapat dilihat bahwa kolom "IkutProgramLOP" memiliki persentase missing value yang sangat tinggi, yaitu 89.90%, sementara kolom lainnya memiliki persentase missing value yang relatif rendah. Perlu dilakukan penanganan khusus terhadap kolom "IkutProgramLOP", seperti imputasi nilai atau penghapusan kolom jika tidak relevan untuk analisis yang akan dilakukan. Sedangkan untuk kolom lainnya, dapat dipertimbangkan untuk imputasi nilai atau penghapusan baris jika jumlah missing valuenya relatif sedikit.
- hasil tersebut juga menunjukkan bahwa tidak ada nilai yang hanya berupa spasi ganda (whitespace) untuk setiap kolom dalam dataset. Hal ini menunjukkan bahwa jika terdapat nilai kosong dalam dataset, nilai tersebut benar-benar kosong (null) dan bukan hanya spasi ganda.
- Dataset tidak memiliki nilai yang terduplikat, sehingga tidak perlu dilakukan penghapusan atau pengelolaan data duplikat.
- Drop Feature kolom 'IkutProgramLOP', 'NomorHP', dan 'Email' telah dihapus dari DataFrame df_prep. Hal ini dilakukan karena kolom-kolom tersebut tidak diperlukan dalam analisis atau pemodelan data yang akan dilakukan.

	Jumlah Missing Value	Persentase (%)
SkorKepuasanPegawai	5	1.74
JumlahKeikutsertaanProjek	3	1.05
JumlahKeterlambatanSebulanTerakhir	1	0.35
JumlahKetidakhadiran	6	2.09
IkutProgramLOP	258	89.90
AlasanResign	66	23.00

```
# Mencari Nilai Null dengan Whitespace dalam DataFrame
white_space = []
for col in df_prep.columns:
    for val in df_prep[col]:
        if isinstance(val, str) and '  ' in val:
            white_space.append(val)

# Output
print(white_space)
```

0.0s

```
[]
```

```
print('There is',df_prep.duplicated().sum(),'duplicated value')
```

0.0s

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

0.0s

Handle Missing Value

```
df_prep['JumlahKetidakhadiran'].fillna(df_prep['JumlahKetidakhadiran'].median(), inplace=True)
df_prep['AlasanResign'].fillna(df_prep['AlasanResign'].mode()[0], inplace=True)
df_prep['SkorKepuasanPegawai'].fillna(0, inplace=True)
df_prep['JumlahKeikutsertaanProjek'].fillna(df_prep['JumlahKeikutsertaanProjek'].median(), inplace=True)
df_prep['JumlahKeterlambatanSebulanTerakhir'].fillna(df_prep['JumlahKeterlambatanSebulanTerakhir'].median(), inplace=True)
```

- Syntax tersebut digunakan untuk mengisi nilai null pada beberapa kolom tertentu dalam DataFrame `df_prep` dengan nilai tertentu. Mengisi nilai null pada kolom 'JumlahKetidakhadiran' dengan nilai median dari kolom tersebut.
- Mengisi nilai null pada kolom 'AlasanResign' dengan nilai modus (nilai yang paling sering muncul) dari kolom tersebut.
- Mengisi nilai null pada kolom 'SkorKepuasanPegawai' dengan nilai 0.
- Mengisi nilai null pada kolom 'JumlahKeikutsertaanProjek' dengan nilai median dari kolom tersebut.
- Mengisi nilai null pada kolom 'JumlahKeterlambatanSebulanTerakhir' dengan nilai median dari kolom tersebut.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Handle Inconsistent Data

- Mengganti nilai '-' pada kolom 'StatusPernikahan' dengan nilai 'Belum_menikah'. Penggantian dilakukan secara langsung pada DataFrame `df_prep`.
- Mengganti nilai 1 pada kolom 'PernahBekerja' dengan nilai 'yes'. Hasil penggantian disimpan kembali ke kolom 'PernahBekerja' dalam DataFrame `df_prep`.
- Menambahkan kolom baru 'Resigned' ke DataFrame, di mana nilai 0 akan diberikan jika nilai dalam kolom 'TanggalResign' adalah '-', dan nilai 1 akan diberikan jika tidak.

```
df_prep['StatusPernikahan'].replace(['-'], 'Belum_menikah', inplace=True)
df_prep['PernahBekerja'] = df_prep['PernahBekerja'].replace(1,'yes')
```

✓ 0.0s

```
df_prep['Resigned'] = np.where(df_prep['TanggalResign']=='-', 0, 1)
```

✓ 0.0s

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Preprocessing

Target
terdapat 198 karyawan yang belum mengundurkan diri dan 89 karyawan yang telah mengundurkan diri.

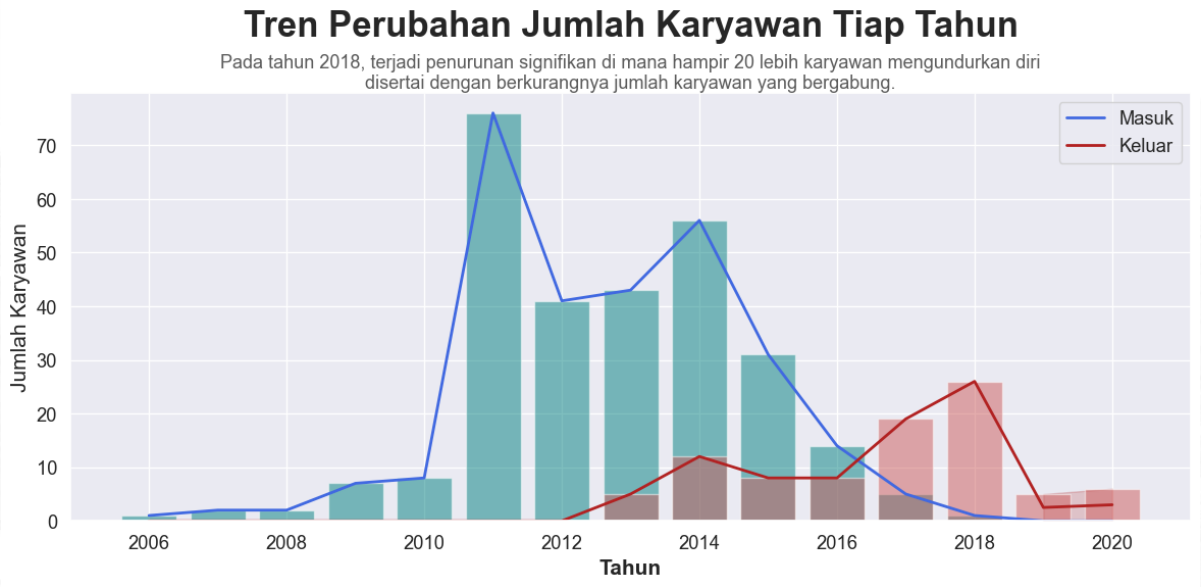
```
df_prep['Resigned'] = np.where(df_prep['TanggalResign']=='-', 0, 1)
```

```
df_prep['Resigned'].value_counts()
```

```
0    198
1     89
Name: Resigned, dtype: int64
```

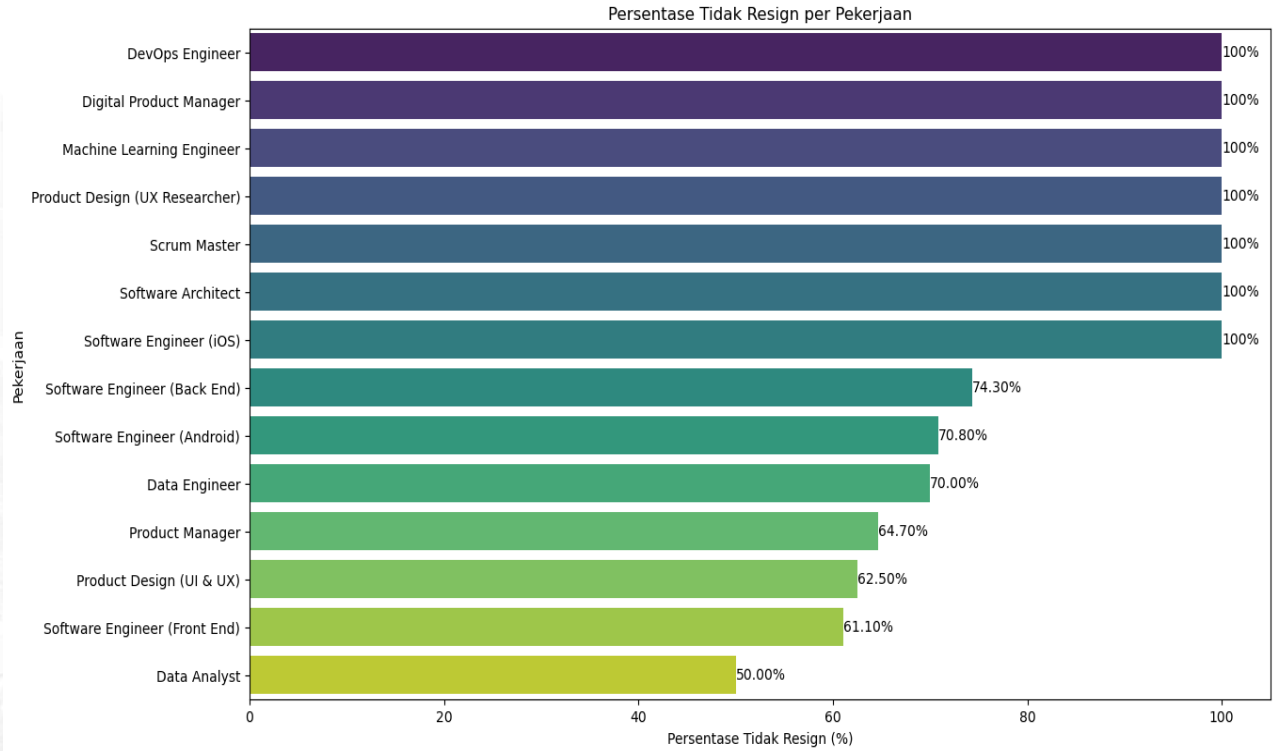
Annual Report on Employee Number Changes

Perusahaan mengalami peningkatan jumlah hiring dari tahun ke tahun, namun juga mengalami peningkatan jumlah resign. Hal ini menyebabkan perubahan jumlah karyawan yang signifikan setiap tahunnya. Meskipun terjadi peningkatan total karyawan secara umum, terdapat tahun-tahun tertentu di mana terjadi penurunan jumlah karyawan akibat tingginya jumlah resign.



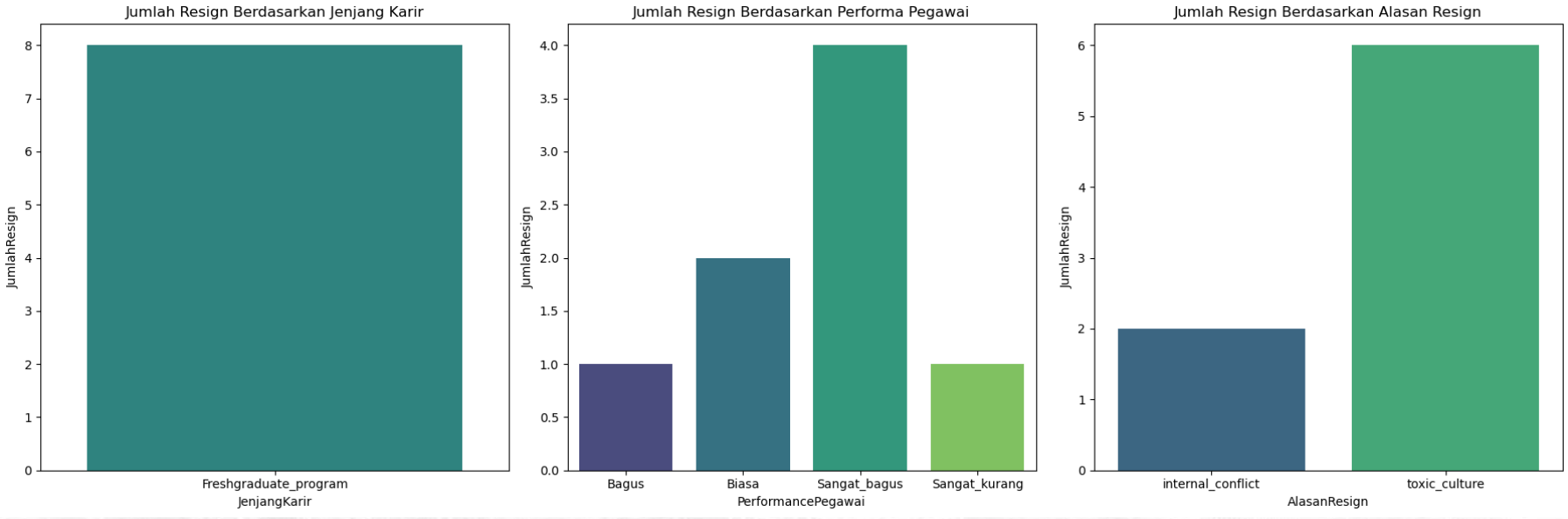
Resign Reason Analysis for Employee Attrition Management Strategy

divisi yang memiliki tingkat resign tertinggi adalah "Software Engineer (Back End) dan Software Engineer (Front End)" dengan persentase 74.3% & 61.1 % akan tetapi Data Analyst Memiliki Persentase tidak resign paling rendah dari total karyawan di divisi tersebut yang telah resign.



Untuk selengkapnya, dapat melihat jupyter notebook disini

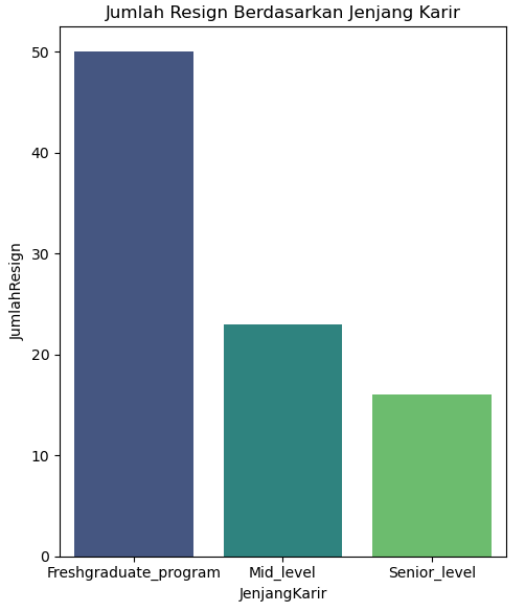
Resign Reason Analysis for Employee Attrition Management Strategy

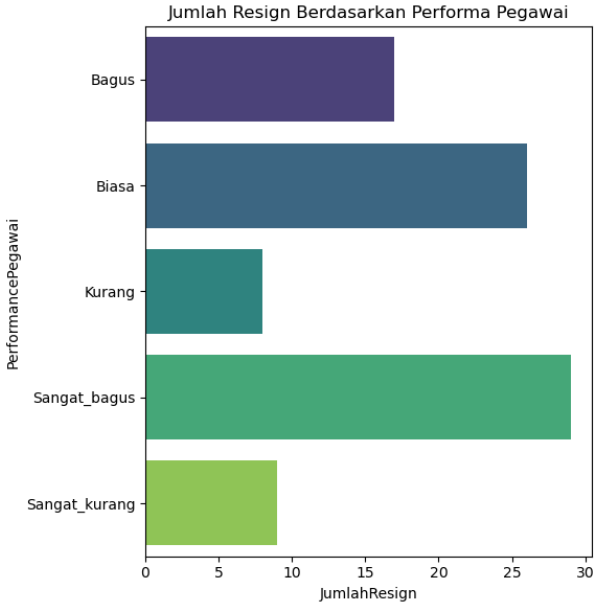


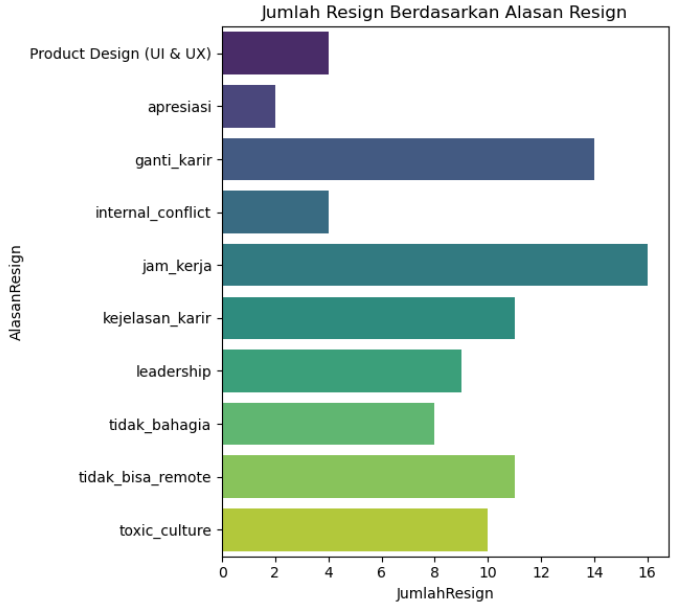
Berdasarkan chart di atas, divisi Data Analyst mengalami tingkat resign yang cukup tinggi. Banyak yang mengundurkan diri dari program fresh graduate, meskipun ada yang memiliki performa baik dan sangat baik. Alasan resign yang paling dominan adalah budaya kerja yang tidak sehat Hal ini terlihat dari jumlah yang signifikan dari karyawan yang mengundurkan diri karena alasan tersebut, yaitu 6 orang dari total 8 orang yang mengundurkan diri dari program fresh graduate.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Resign Reason Analysis for Employee Attrition Management Strategy







Berdasarkan Chart diatas, Tingkat resign di perusahaan terjadi di berbagai jenjang karir, dengan mayoritas terjadi pada program fresh graduate. Performa pegawai juga tidak menjadi faktor utama, karena terdapat resign dari berbagai tingkat performa. Alasan resign yang dominan adalah masalah jam kerja, keinginan untuk ganti karir, dan kurangnya kejelasan dalam karir. Selain itu, budaya kerja yang tidak sehat juga menjadi faktor penting dalam keputusan resign karyawan.

Build an Automated Resignation Behavior Prediction using Machine Learning

Rakamin Academy

Data terdiri dari 25 fitur dengan berbagai tipe data. Tidak ada nilai null kecuali pada kolom TanggalResign dan TahunResign. Kolom EnterpriseID memiliki 287 nilai unik, sementara kolom lainnya memiliki nilai unik yang bervariasi. Kolom PernahBekerja hanya memiliki satu nilai ('yes'). Tidak ada baris duplikat dalam dataset.

	Feature	Data Type	Null	Null (%)	Unique	Unique Sample
0	Username	object	0	0.000000	285	[spiritedPorpoise3, jealousGelding2, pluckyMue...
1	EnterpriseID	int64	0	0.000000	287	[111065, 106080, 106452, 106325, 111171]
2	StatusPernikahan	object	0	0.000000	4	[Belum_menikah, Menikah, Bercerai, Lainnya]
3	JenisKelamin	object	0	0.000000	2	[Pria, Wanita]
4	StatusKepegawaian	object	0	0.000000	3	[Outsource, FullTime, Internship]
5	Pekerjaan	object	0	0.000000	14	[Software Engineer (Back End), Data Analyst, S...
6	JenjangKarir	object	0	0.000000	3	[Freshgraduate_program, Senior_level, Mid_level]
7	PerformancePegawai	object	0	0.000000	5	[Sangat_bagus, Sangat_kurang, Bagus, Biasa, Ku...
8	AsalDaerah	object	0	0.000000	5	[Jakarta Timur, Jakarta Utara, Jakarta Pusat, ...
9	HiringPlatform	object	0	0.000000	9	[Employee_Referral, Website, Indeed, LinkedIn,...
10	SkorSurveyEngagement	int64	0	0.000000	5	[4, 3, 2, 1, 5]
11	SkorKepuasanPegawai	float64	0	0.000000	6	[4.0, 3.0, 5.0, 0.0, 2.0]
12	JumlahKeikutsertaanProjek	float64	0	0.000000	9	[0.0, 4.0, 6.0, 7.0, 3.0]
13	JumlahKeterlambatanSebulanTerakhir	float64	0	0.000000	7	[0.0, 4.0, 3.0, 5.0, 2.0]
14	JumlahKetidakhadiran	float64	0	0.000000	22	[9.0, 3.0, 11.0, 6.0, 10.0]
15	TingkatPendidikan	object	0	0.000000	3	[Magister, Sarjana, Doktor]
16	PernahBekerja	object	0	0.000000	1	[yes]
17	AlasanResign	object	0	0.000000	11	[masih_bekerja, toxic_culture, jam_kerja, gant...
18	TanggalLahir	datetime64[ns]	0	0.000000	284	[1972-07-01T00:00:00.000000000, 1984-04-26T00:...
19	TanggalHiring	datetime64[ns]	0	0.000000	97	[2011-01-10T00:00:00.000000000, 2014-01-06T00:...
20	TanggalPenilaianKaryawan	datetime64[ns]	0	0.000000	127	[2016-02-15T00:00:00.000000000, 2020-01-17T00:...
21	TanggalResign	datetime64[ns]	198	68.989547	52	[NaT, 2018-06-16T00:00:00.000000000, 2014-09-2...
22	Resigned	int32	0	0.000000	2	[0, 1]
23	TahunHiring	int64	0	0.000000	13	[2011, 2014, 2013, 2016, 2015]
24	TahunResign	float64	198	68.989547	8	[nan, 2018.0, 2014.0, 2019.0, 2017.0]
25	Duplicate Rows	-	0	0.000000	-	-

Untuk selengkapnya, dapat melihat jupyter notebook disini

Build an Automated Resignation Behavior Prediction using Machine Learning

Rakamin Academy

Feature Encoding

- One-hot encoding dilakukan pada kolom-kolom yang memiliki beberapa kategori unik, seperti 'StatusPernikahan', 'JenisKelamin', 'StatusKepegawaian', 'JenjangKarir', 'AsalDaerah', 'HiringPlatform', dan 'TingkatPendidikan'.
- Label encoding dilakukan pada kolom-kolom seperti 'Pekerjaan', 'PernahBekerja', 'AlasanResign', 'Resigned', dan 'PerformancePegawai'.
- Beberapa kolom yang dianggap tidak diperlukan untuk analisis dihapus Kolom-kolom yang dihapus termasuk 'Username', 'EnterpriseID', 'TanggalLahir', 'TanggalHiring', 'TanggalPenilaianKaryawan', 'TanggalResign', dan 'TahunResign'.

Feature Transformation

Normalisasi ini penting karena beberapa model machine learning sensitif terhadap skala fitur. Dengan normalisasi Min-Max, nilai-nilai fitur akan diubah ke dalam rentang antara 0 dan 1, menjaga proporsi relatif antara nilai-nilai tersebut.

Feature Selection

memilih 15 fitur terbaik dari DataFrame menggunakan skor chi-squared. Fitur-fitur ini diharapkan memiliki pengaruh yang signifikan terhadap prediksi apakah seorang karyawan akan mengundurkan diri (Resigned).

Untuk selengkapnya, dapat melihat jupyter notebook disini

Build an Automated Resignation Behavior Prediction using Machine Learning

Rakamin Academy

Split Train & Test Data

Untuk menangani ketidakseimbangan kelas pada data target. Dalam kasus ini, setelah penerapan SMOTE, jumlah sampel untuk setiap kelas menjadi seimbang (143 sampel untuk setiap kelas, total 286 sampel).

Oversampling

Membagi dataset menjadi data latih dan data uji dengan proporsi yang sesuai.

- Total sampel: 287
- Sampel latih: 200 (69.69%)
- Sampel uji: 87 (30.31%)

Modeling

melatih beberapa model machine learning dan mengevaluasi performanya menggunakan metrik seperti akurasi, presisi, recall, F1 score, dan ROC AUC. Model yang dilatih meliputi Logistic Regression, Support Vector Machine, Random Forest, dan K-Neighbors.

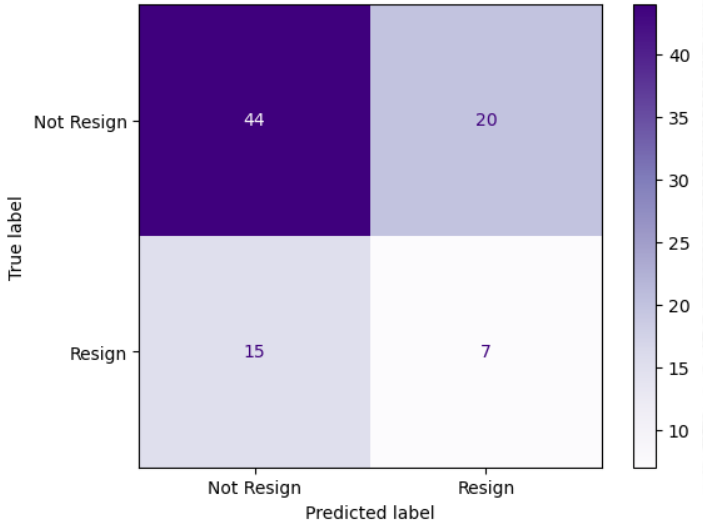
Untuk selengkapnya, dapat melihat jupyter notebook disini

Build an Automated Resignation Behavior Prediction using Machine Learning

Rakamin Academy

	Model	Accuracy Train	Accuracy Test	Precision Train	Precision Test	Recall Train	Recall Test	F1 Train	F1 Test	AUC Train	AUC Test
0	LogisticRegression	62.88	43.02	63.28	21.28	61.36	45.45	62.31	28.99	62.88	43.39
1	LogisticRegressionCV	62.12	44.19	62.12	24.00	62.12	54.55	62.12	33.33	62.15	42.26
2	DecisionTree	98.86	62.79	100.00	29.17	97.73	31.82	98.85	30.43	99.97	51.63
3	RandomForest	98.86	59.30	99.24	25.93	98.48	31.82	98.86	28.57	99.97	48.47
4	KNeighbors	81.44	59.30	76.10	31.43	91.67	50.00	83.16	38.60	91.14	50.82
5	AdaBoost	75.38	48.84	76.38	15.62	73.48	22.73	74.90	18.52	85.02	37.68
6	XGB	98.86	53.49	100.00	17.86	97.73	22.73	98.85	20.00	99.97	45.95

- Kesimpulannya, dari semua model yang diuji, RandomForest memiliki performa yang paling baik secara keseluruhan dengan nilai akurasi, precision, recall, F1-score, dan AUC yang relatif tinggi dan diikuti dengan Decision Tree.
- Model cenderung lebih baik dalam memprediksi data yang benar-benar negatif (TN) daripada yang benar-benar positif (TP).
- Model memiliki kecenderungan untuk salah memprediksi data yang sebenarnya negatif (FP) menjadi positif (resign) lebih sering daripada salah memprediksi data yang sebenarnya positif (FN) menjadi negatif (tidak resign).



Presenting Machine Learning Products to the Business Users

Rakamin Academy

Dari hasil feature importance, kita dapat melihat bahwa jumlah ketidakhadiran (absensi) merupakan fitur yang paling penting dalam memprediksi tingkat resignasi karyawan, diikuti oleh performa pegawai, skor kepuasan pegawai, dan skor survei engagement. Fitur-fitur ini memiliki pengaruh yang signifikan dalam model untuk memprediksi apakah seorang karyawan akan resign atau tidak. Fitur-fitur lain seperti jumlah keikutsertaan dalam proyek, jenjang karir, dan jumlah keterlambatan juga memiliki pengaruh, meskipun tidak sebesar fitur-fitur utama tersebut. Adanya program LOP juga mempengaruhi, meskipun dengan pengaruh yang lebih kecil.