Improving Employee Retention by Predicting Employee Attrition Using Machine Learning

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

Created by:
Devia Febyanti
devia.febyanti99@gmail.com
https://www.linkedin.com/in/devia-febyanti/

Experienced data operator with proven work history in the field of education. Currently motivated to become a Data Scientist to help driving financial inclusion through the use of big data and machine learning.

Bachelor of Geophysics
2018-2022

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

```
RangeIndex: 287 entries, 0 to 286
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
```

Shape

- 287 data rows
- 25 features

Data Type

- Float64 : 5 feature
- Int64 : 2 features
- Object : 18 features

Missing Value

- SkorKepuasanPegawai
- JumlahKeikutsertaanProjek
- JumlahKeterlambatanSebulanTerakhir
- JumlahKetidakhadiran
- IkutProgramLOP
- AlasanResign

Duplicated

None

Tidak dilakukan handling

- Fitur TanggalLahir, TanggalHiring, TanggalPenilaianKaryawan, dan TanggalResign akan diubah menjadi tipe data datetime.
- Variabel target yang merupakan AlasanResign adalah data kategorikal dan harus diubah menjadi data numerik.

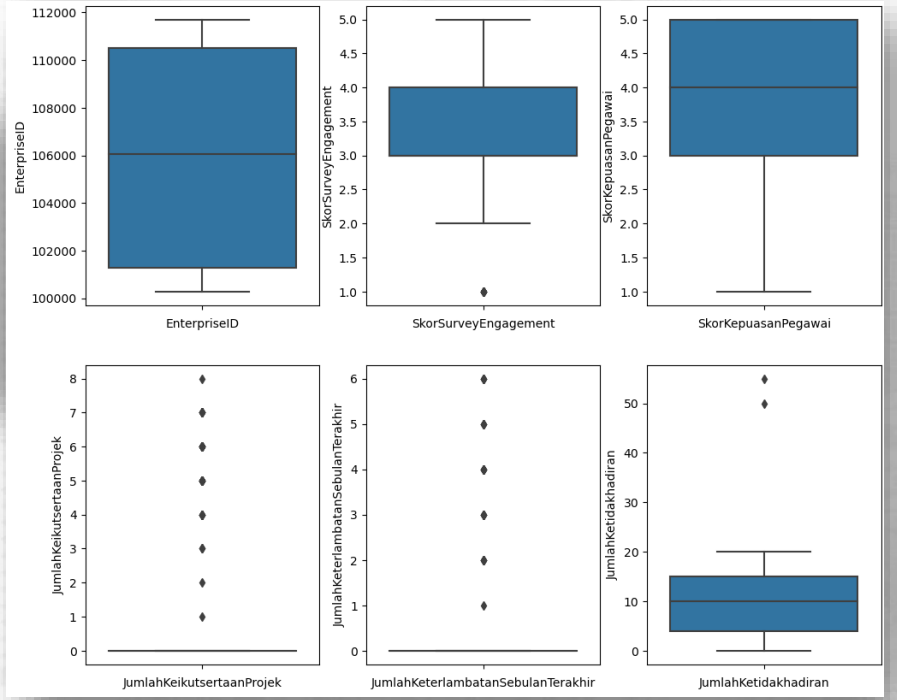
Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Preprocessing

EXPLORATORY DATA ANALYSIS (EDA)

Fitur Numerik

	EnterpriseID	SkorSurveyEngagement	SkorKepuasanPegawai	JumlahKeikutsertaanProjek	JumlahKeterlambatanSebulanTerakhir	JumlahKetidakhadiran
count	287.000000	287.000000	287.000000	287.000000	287.000000	287.000000
mean	105923.324042	3.101045	3.905923	1.167247	0.411150	10.229965
std	4044.977599	0.836388	0.905423	2.285537	1.273018	6.991709
min	100282.000000	1.000000	1.000000	0.000000	0.000000	0.000000
25%	101269.000000	3.000000	3.000000	0.000000	0.000000	4.000000
50%	106069.000000	3.000000	4.000000	0.000000	0.000000	10.000000
75%	110514.500000	4.000000	5.000000	0.000000	0.000000	15.000000
max	111703.000000	5.000000	5.000000	8.000000	6.000000	55.000000



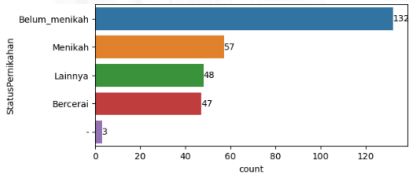
- Akan dicek apakah jumlah ketidakhadiran tersebut benar-benar memiliki outlier atau tidak.
- Nilai jumlah ketidakhadiran yang ekstrim pada feature tersebut wajar karena kedua karyawan telah bekerja selama 11 dan 13 tahun lamanya.

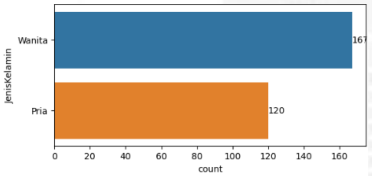
Untuk selengkapnya, dapat melihat jupyter notebook disini

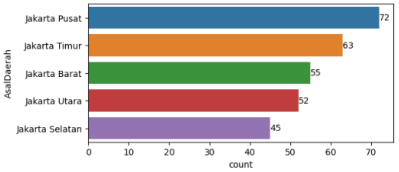
Data Preprocessing

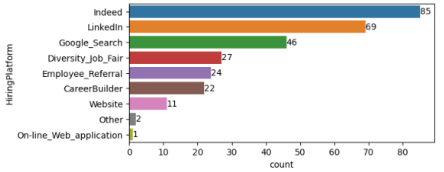
Fitur Kategori

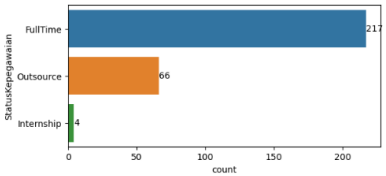
EXPLORATORY DATA ANALYSIS (EDA)

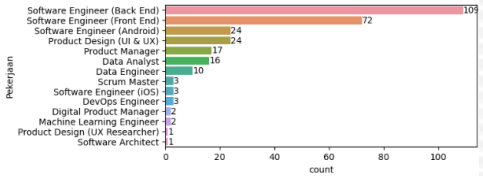


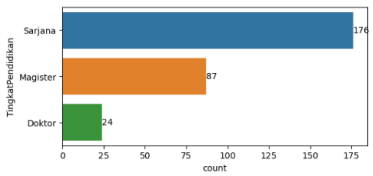


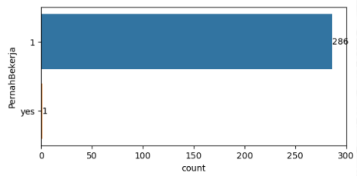


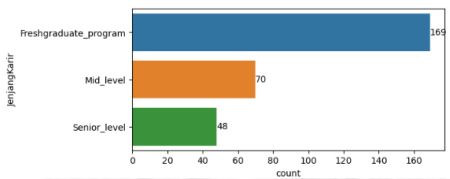


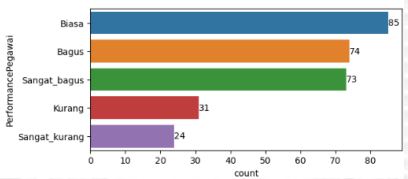


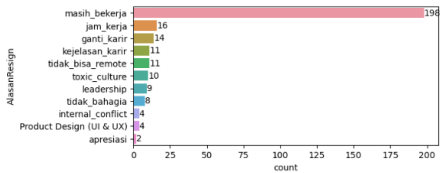












- Nilai "-" pada fitur **StatusPernikahan** akan diganti menjadi "Belum_menikah"
- Seluruh nilai pada fitur **PernahBekerja** adalah sama yaitu "1" dan "yes". Fitur ini dapat dihapus saja karena hanya mengandung 1 nilai unique.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Preprocessing

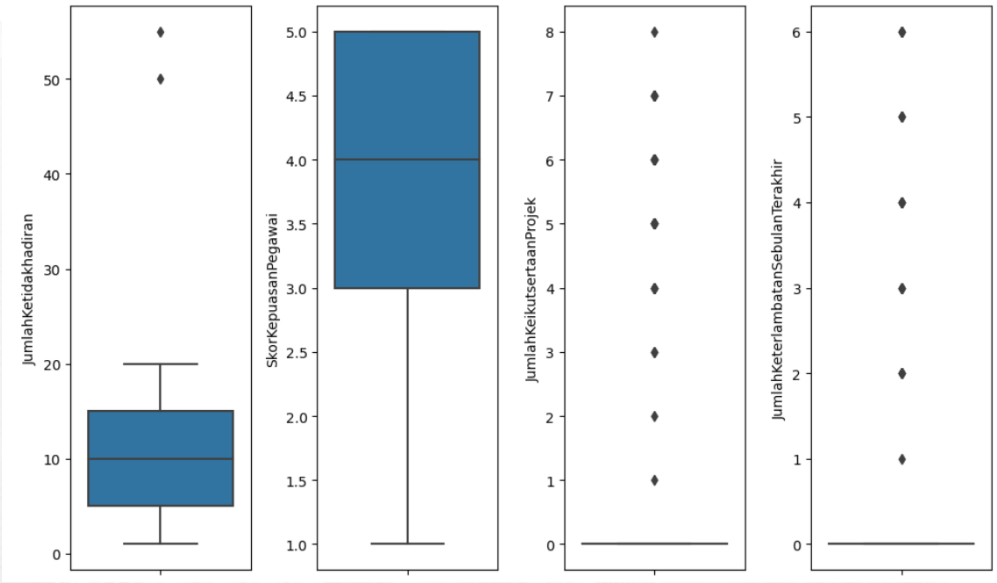
Rakamin Academy

HANDLING MISSING VALUE

The features that has missing value (percentage):

IkutProgramLOP	89.895470
AlasanResign	22.996516
JumlahKetidakhadiran	2.090592
SkorKepuasanPegawai	1.742160
JumlahKeikutsertaanProjek	1.045296
JumlahKeterlambatanSebulanTerakhir	0.348432

- Fitur IkutProgramLOP akan dihapus karena didominasi oleh missing value, yaitu hampir 90% dari dataset.
- Fitur lainnya akan dilakukan handling missing value.



- Untuk feature **IkutProgramLOP** akan dihapus karena memiliki nilai null yang terlalu banyak, yaitu hampir 90%.
- Untuk nilai null pada feature **AlasanResign** akan diganti menjadi *masih_bekerja* karena belum ada nilai pada feature TanggalResign yang artinya karyawan tersebut memang belum resign.
- Untuk nilai null pada feature **JumlahKetidakhadiran** akan diisi dengan 0 dimana diasumsikan bahwa nilai null tersebut berarti karyawan belum pernah tidak hadir. Asumsi ini didukung dengan tidak adanya nilai 0 pada feature tersebut.
- Untuk nilai null pada feature **SkorKepuasanPegawai**, **JumlahKeikutsertaanProjek** dan **JumlahKeterlambatanSebulanTerakhir** akan diisi dengan nilai median dari masing-masing kolom karena mengandung outlier.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Preprocessing

Rakamin Academy

HANDLING DUPLICATED DATA

- None

HANDLING INCOMPATIBLE DATA

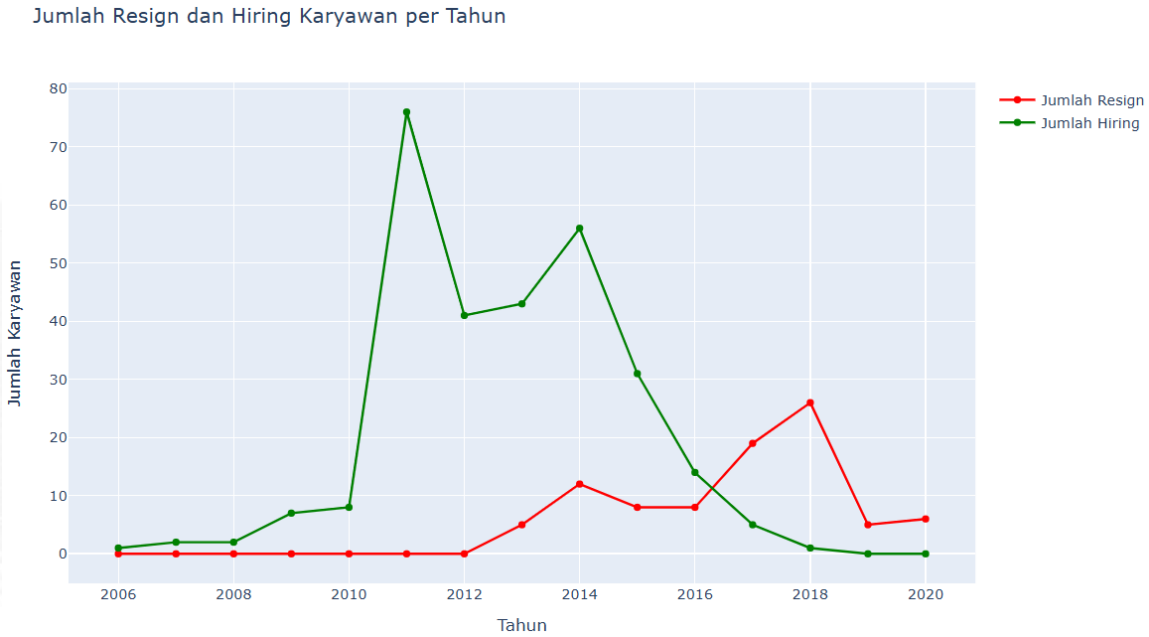
- Nilai "-" pada fitur StatusPernikahan akan diganti menjadi "Belum_menikah"

DELETE UNNECESSARY FEATURE

- Fitur **NomorHP, Email,** dan **EnterpriseID** didrop karena hanya berperan sebagai identitas user dan bukan merupakan faktor penentu "employee attrition“
- Seluruh nilai pada fitur **PernahBekerja** adalah sama yaitu "1" dan "yes". Fitur ini dapat dihapus saja karena hanya mengandung 1 nilai unique.
- Fitur **IkutProgramLOP** dihapus karena memiliki nilai null yang terlalu banyak, yaitu hampir 90%.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Annual Report on Employee Number Changes

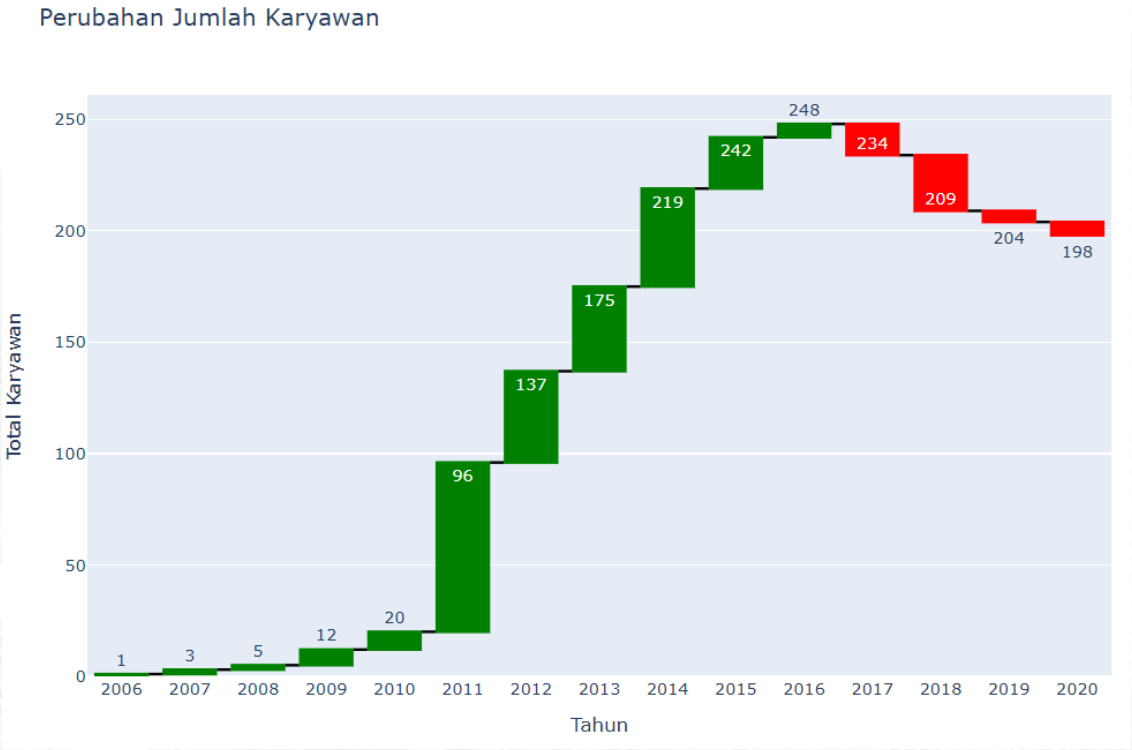


- Periode 2006-2010: Jumlah karyawan yang di-hiring meningkat secara bertahap. Tidak ada karyawan yang resign selama periode ini. Perusahaan berada dalam fase pertumbuhan awal dengan penambahan karyawan secara konsisten dan tanpa kehilangan karyawan.
- Periode 2011-2014: Terdapat peningkatan signifikan pada tahun 2011. Mulai ada karyawan yang resign pada tahun 2013 dan meningkat pada 2014. Perusahaan mengalami ekspansi besar-besaran, terutama pada 2011. Namun, mulai ada tanda-tanda pengunduran diri karyawan, yang bisa menjadi indikasi berbagai hal seperti perubahan manajemen, budaya kerja, atau kondisi pasar.
- Periode 2015-2016: Jumlah karyawan baru mulai menurun pada 2015. Jumlah karyawan yang resign tetap stabil. Perusahaan mungkin memasuki fase stabilisasi atau restrukturisasi setelah periode pertumbuhan pesat. Penurunan jumlah hiring bisa menunjukkan upaya pengendalian biaya atau mencapai kapasitas optimal. Jumlah resign yang stabil bisa menunjukkan kondisi kerja yang relatif stabil.

- Periode **2017-2020**: Jumlah karyawan baru menurun drastis, serta tidak ada karyawan baru di tahun 2019 dan 2020. Jumlah karyawan yang resign meningkat signifikan pada tahun 2017 dan 2018, kemudian menurun di tahun 2019 dan 2020. Periode ini menandakan adanya masalah serius dalam perusahaan. Penurunan drastis dalam hiring dan peningkatan jumlah resign menunjukkan **kondisi yang mengkhawatirkan**. Mungkin ada masalah internal seperti ketidakpuasan karyawan, perubahan manajemen, atau kondisi keuangan yang buruk. Penurunan jumlah resign pada tahun 2019 dan 2020 mungkin disebabkan oleh stabilisasi internal atau kurangnya karyawan yang tersisa untuk resign.

Untuk selengkapnya, dapat melihat jupyter notebook disini

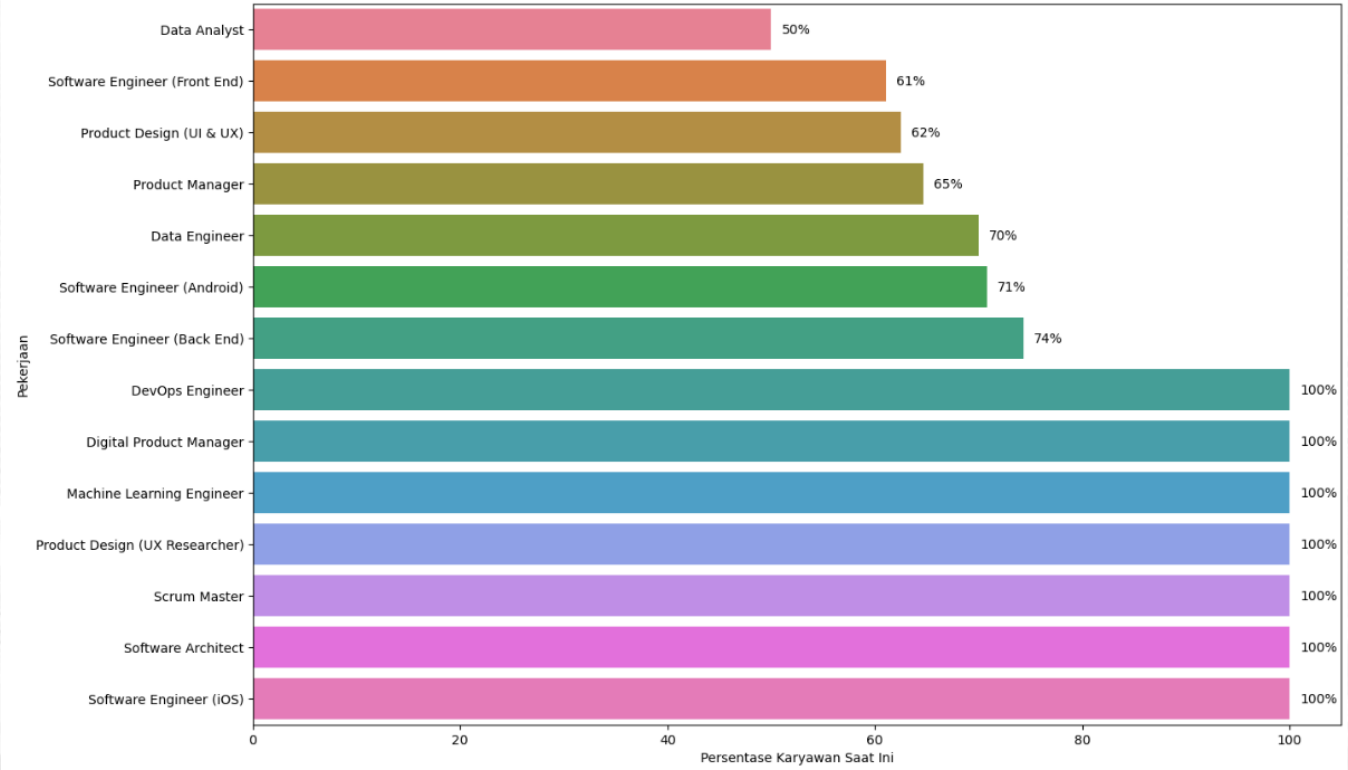
Annual Report on Employee Number Changes



- 2006-2010: Terdapat peningkatan jumlah karyawan secara bertahap dari 1 karyawan di tahun 2006 hingga 20 karyawan di tahun 2010.
- 2011: Terjadi lonjakan besar dalam jumlah karyawan, meningkat menjadi 96 karyawan.
- 2012-2014: Jumlah karyawan terus meningkat dengan lonjakan besar di tahun 2012 dan 2014. Sejak 2006 hingga 2014 perusahaan sedang berkembang dan mungkin mengalami pertumbuhan bisnis yang pesat selama periode ini.
- 2015-2016: Pertumbuhan lebih stabil dengan peningkatan jumlah karyawan yang lebih kecil. Pada periode ini, pertumbuhan jumlah karyawan masih positif, meskipun dengan laju yang lebih lambat. Ini mungkin menunjukkan bahwa perusahaan telah mencapai titik stabil di mana ekspansi tidak secepat sebelumnya tetapi masih dalam kondisi yang baik.
- 2017-2020: Terjadi penurunan jumlah karyawan secara bertahap, dengan penurunan terbesar pada tahun 2017 dan 2018. Penurunan jumlah karyawan selama periode ini bisa menjadi tanda kekhawatiran. Alasan penurunan perlu dianalisis lebih lanjut, apakah disebabkan oleh faktor internal seperti restrukturisasi atau efisiensi, atau faktor eksternal seperti kondisi pasar yang menurun atau persaingan yang meningkat.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Resign Reason Analysis for Employee Attrition Management Strategy

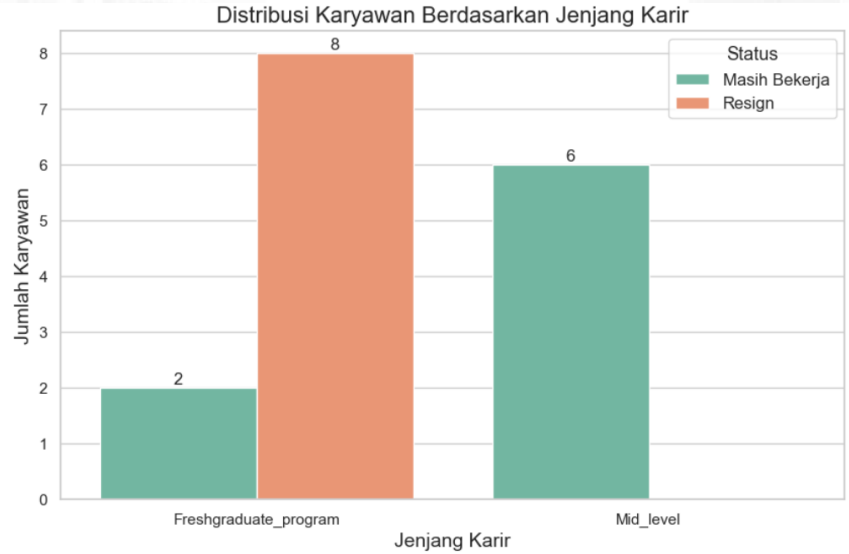


Data Analyst adalah pekerjaan/divisi yang memiliki tingkat resign tertinggi.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Resign Reason Analysis for Employee Attrition Management Strategy

Distribusi Karyawan Berdasarkan Jenjang Karir



Interpretasi

1. **Freshgraduate Program** Tingkat resign yang sangat tinggi di program ini menunjukkan bahwa banyak karyawan fresh graduate yang meninggalkan perusahaan. Ini bisa disebabkan oleh beberapa faktor seperti ekspektasi yang tidak terpenuhi, kurangnya peluang pengembangan karir, atau ketidakpuasan dengan budaya kerja.
2. **Mid Level** Jumlah karyawan yang masih bekerja di tingkat mid-level relatif tinggi. Ini bisa menunjukkan bahwa karyawan di level ini merasa lebih puas dengan kondisi kerja mereka, atau mungkin mereka melihat peluang karir yang lebih baik dibandingkan dengan fresh graduates.

Rekomendasi

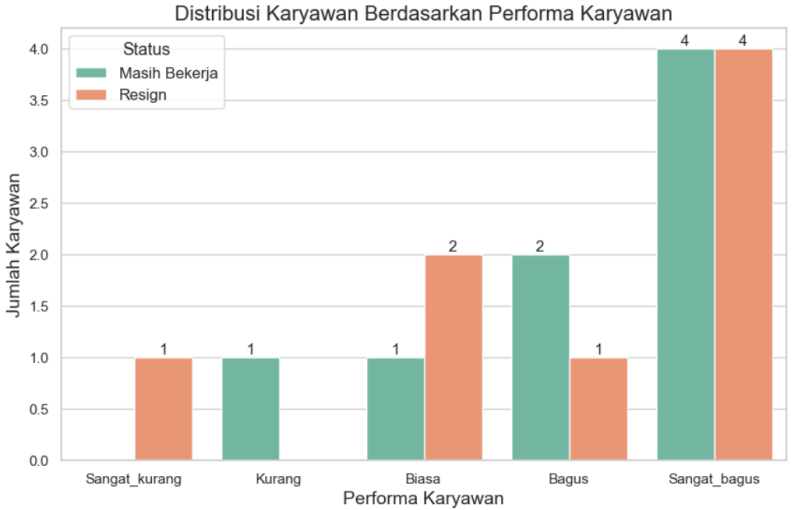
Untuk mengurangi tingkat resign yang tinggi pada program freshgraduate dan meningkatkan retensi karyawan, manajemen dapat mempertimbangkan langkah-langkah berikut:

1. Tinjau kembali program onboarding dan pelatihan untuk fresh graduates untuk memastikan bahwa mereka diberikan sumber daya dan dukungan yang diperlukan untuk sukses di peran mereka.
2. Meningkatkan komunikasi antara manajemen dan karyawan fresh graduate untuk memahami kebutuhan dan kekhawatiran mereka. Ini bisa dilakukan melalui survei rutin atau sesi umpan balik.
3. Menerapkan program pengembangan karir dan mentorship yang lebih baik. Mentor yang berpengalaman dapat membantu fresh graduates menavigasi lingkungan kerja dan mengembangkan keterampilan yang diperlukan.
4. Menawarkan insentif dan penghargaan untuk karyawan yang menunjukkan kinerja baik. Ini bisa berupa bonus, penghargaan, atau peluang pengembangan profesional.
5. Menciptakan budaya kerja yang positif dan inklusif di mana karyawan merasa dihargai dan didukung.
6. Memberikan peluang rotasi kerja sehingga karyawan dapat merasakan berbagai aspek pekerjaan dan menemukan bidang yang paling sesuai dengan minat dan keterampilan mereka.
7. Tinjau dan sesuaikan skala gaji untuk memastikan bahwa kompensasi yang ditawarkan kompetitif dengan pasar. Ini penting untuk menarik dan mempertahankan talenta terbaik.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Resign Reason Analysis for Employee Attrition Management Strategy

Distribusi Karyawan Berdasarkan Performa Karyawan



Rekomendasi

1. Karyawan dengan performa "Sangat Bagus" memiliki jumlah yang sama antara yang masih bekerja dan yang resign. Manajemen perlu mengidentifikasi alasan mengapa karyawan berperforma tinggi ini resign. Fokus pada meningkatkan kepuasan kerja mereka melalui insentif, kesempatan pengembangan karir, dan lingkungan kerja yang kondusif.
2. Karyawan dengan performa "Biasa" menunjukkan jumlah resign yang lebih tinggi daripada performa "Bagus". Evaluasi kondisi kerja mereka dan tawarkan dukungan atau pelatihan yang dapat membantu mereka meningkatkan performa dan merasa lebih dihargai di perusahaan.
3. Meskipun jumlah karyawan dengan performa "Kurang" dan "Sangat Kurang" yang masih bekerja tidak tinggi, penting untuk memahami dan menangani penyebab kinerja rendah. Berikan pelatihan tambahan atau peluang untuk pindah ke posisi yang mungkin lebih sesuai dengan keahlian mereka.
4. Perusahaan dapat mempertimbangkan untuk mengimplementasikan program pengembangan karir yang terstruktur untuk semua tingkat performa. Program ini bisa meliputi pelatihan, mentoring, dan jalur promosi yang jelas.
5. Rutin melakukan survei kepuasan karyawan untuk memahami kebutuhan dan kekhawatiran mereka. Gunakan data ini untuk membuat perubahan yang dapat meningkatkan retensi dan kepuasan kerja.
6. Membuat budaya kerja yang mendukung komunikasi terbuka antara manajemen dan karyawan sehingga karyawan merasa nyaman untuk menyampaikan masukan dan keluhan mereka.

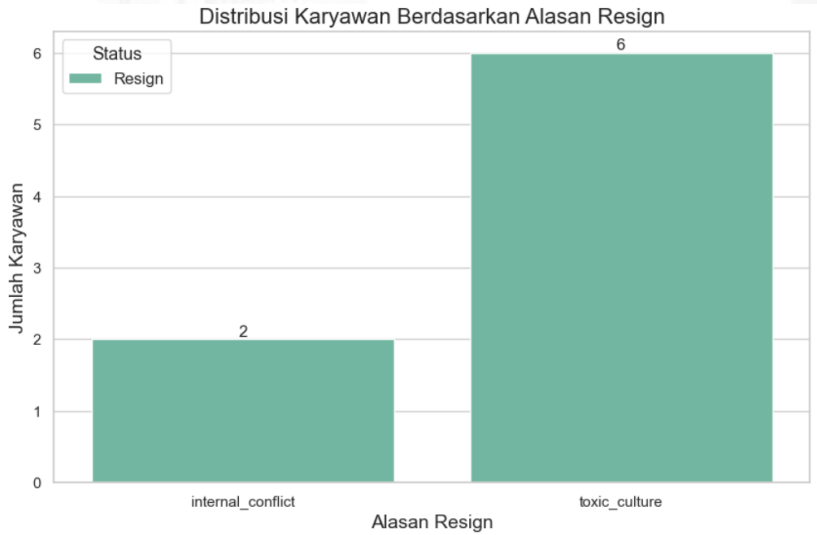
Interpretasi

- Jumlah karyawan dengan performa "Sangat Bagus" yang masih bekerja dan yang resign sama-sama tinggi, yaitu 4 orang.
- Jumlah karyawan dengan performa "Bagus" yang masih bekerja lebih banyak daripada yang resign, dengan perbandingan 2 berbanding 1.
- Jumlah karyawan dengan performa "Biasa" yang masih bekerja lebih sedikit daripada yang resign, dengan perbandingan 1 berbanding 2.
- Jumlah karyawan dengan performa "Kurang" yang masih bekerja adalah 1 orang, sementara tidak ada yang resign.
- Jumlah karyawan dengan performa "Sangat Kurang" yang resign adalah 1 orang, sementara tidak ada yang masih bekerja.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Resign Reason Analysis for Employee Attrition Management Strategy

Distribusi Karyawan Berdasarkan Alasan Resign



Interpretasi

- Sebanyak 6 karyawan resign karena alasan "toxic culture". Ini menunjukkan bahwa lingkungan kerja yang tidak sehat adalah masalah utama yang menyebabkan karyawan meninggalkan perusahaan.
- Sebanyak 2 karyawan resign karena "internal conflict". Ini menunjukkan adanya konflik internal di dalam perusahaan yang juga berkontribusi terhadap keputusan karyawan untuk resign.

Rekomendasi

- Perusahaan harus melakukan upaya serius untuk mengatasi masalah "**toxic culture**". Hal ini bisa dimulai dengan mengadakan workshop tentang perilaku yang diinginkan di tempat kerja, menyusun kode etik yang jelas, dan memastikan bahwa nilai-nilai perusahaan dijalankan dengan baik. Manajemen juga harus lebih proaktif dalam mengidentifikasi dan menangani perilaku atau praktik yang merusak lingkungan kerja.
- Untuk mengatasi masalah "**internal conflict**", perusahaan harus meningkatkan komunikasi antar departemen dan antar individu. Pelatihan manajemen konflik bisa diberikan kepada manajer dan karyawan untuk membantu mereka mengelola dan menyelesaikan konflik dengan lebih efektif. Perusahaan juga bisa mengadakan sesi mediasi atau konseling bagi karyawan yang mengalami konflik.
- Melakukan survei kepuasan karyawan secara rutin untuk mengidentifikasi masalah sejak dini. Survei ini bisa membantu perusahaan memahami perasaan dan persepsi karyawan terhadap lingkungan kerja. Berdasarkan hasil survei, perusahaan dapat membuat rencana aksi yang tepat untuk meningkatkan kepuasan kerja dan mengurangi tingkat resign.
- Menyediakan program pengembangan karir dan pelatihan untuk karyawan agar mereka merasa dihargai dan memiliki kesempatan untuk berkembang dalam perusahaan. Hal ini dapat meningkatkan motivasi dan loyalitas karyawan.
- Menerapkan kebijakan "zero tolerance" terhadap perilaku negatif yang berkontribusi terhadap "toxic culture". Ini termasuk tindakan bullying, diskriminasi, dan pelecehan di tempat kerja.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Build an Automated Resignation Behavior Prediction using Machine Learning

Rakamin Academy

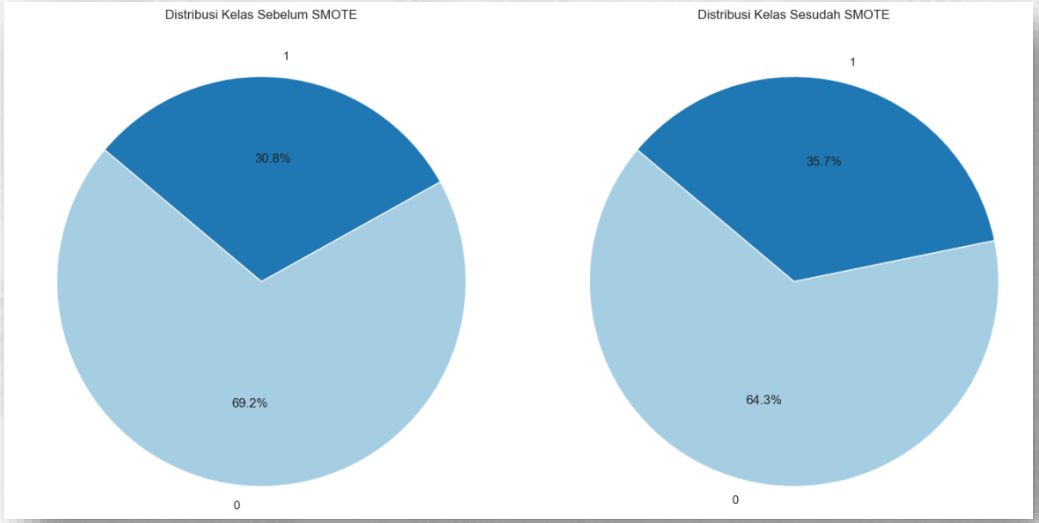
Feature Extraction	• Menambahkan fitur target “Resign” • Menambahkan fitur “Lama Bekerja”
Feature Selection	Beberapa fitur yang akan diremove: • **Username** karena merupakan identitas (unique) • **StatusPernikahan, JenisKelamin** dan **AsalDaerah** untuk menghindari diskriminasi • **StatusKepegawaian** karena pada dasarnya hanya pegawai fulltime yang benar-benar bekerja untuk perusahaan sementara outsource dan internship tidak berstatus pegawai perusahaan • **Pekerjaan** dan **HiringPlatform** karena terlalu banyak nilai unique • **AlasanResign, TanggalLahir, TanggalPenilaianKaryawan, TanggalResign** dan **TahunResign** karena fitur tidak relevan untuk memprediksi resign • **TanggalHiring** dan **TahunHiring** karena sudah diconvert ke fitur LamaBekerja
Feature Encoding	• Label encoding: **PerformancePegawai, TingkatPendidikan** • One hot encoding: **JenjangKarir**
Handling Missing Value	None
Handling Duplicated Data	None
Handling Outlier	Fitur: **JumlahKetidakhadiran** → Menghapus baris yang memiliki JumlahKetidakhadiran > 30

Untuk selengkapnya, dapat melihat jupyter notebook disini

Build an Automated Resignation Behavior Prediction using Machine Learning

Rakamin Academy

Split Data Test & Data Train	• 70% Data Train • 30% Data Test
Feature Transformation	Normalisasi menggunakan MinMaxScaler
Handling Class Imbalance	SMOTE • Jumlah minoritas: 20% total data • Jumlah mayoritas: 80% total data



1: Resign (Positif)
0: Tidak Resign (Negatif)

- Setelah penerapan SMOTE, jumlah karyawan yang resign (Class 1) meningkat dari 61 menjadi 76, menghasilkan distribusi yang lebih seimbang.
- Proporsi karyawan yang resign meningkat dari 30.8% menjadi 35.7%, membuat model lebih mampu mengenali pola resign.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Build an Automated Resignation Behavior Prediction using Machine Learning

Rakamin Academy

DATA MODELING

	Model	Accuracy (Train)	Accuracy (Test)	Precision (Train)	Precision (Test)	Recall (Train)	Recall (Test)	F1-Score (Train)	F1-Score (Test)	ROC AUC (Train)	ROC AUC (Test)
0	LogisticRegression	0.765258	0.764706	0.644444	0.588235	0.763158	0.769231	0.698795	0.666667	0.855167	0.846806
1	DecisionTreeClassifier	1.000000	0.741176	1.000000	0.562500	1.000000	0.692308	1.000000	0.620690	1.000000	0.727510
2	RandomForestClassifier	1.000000	0.835294	1.000000	0.800000	1.000000	0.615385	1.000000	0.695652	1.000000	0.891460
3	XGBClassifier	1.000000	0.823529	1.000000	0.703704	1.000000	0.730769	1.000000	0.716981	1.000000	0.862451

- **RandomForestClassifier** dan **XGBClassifier** menunjukkan kinerja yang sangat baik pada data pelatihan, dengan akurasi 100% pada data pelatihan. Namun, kita harus memperhatikan performa pada data uji untuk menentukan model terbaik.
- RandomForestClassifier memiliki akurasi uji tertinggi (0.835), diikuti oleh XGBClassifier (0.824).
- XGBClassifier menunjukkan performa yang cukup baik pada data uji, dengan hasil yang mendekati RandomForestClassifier.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Build an Automated Resignation Behavior Prediction using Machine Learning

Tuning Hyperparameter untuk XGBoost

Classification Report:

	precision	recall	f1-score	support
0	0.83	0.98	0.90	59
1	0.93	0.54	0.68	26
accuracy			0.85	85
macro avg	0.88	0.76	0.79	85
weighted avg	0.86	0.85	0.83	85

1. Kelas 0 (Negatif)

- Precision (0.83): Dari semua prediksi kelas 0, 83% benar-benar kelas 0. Ini menunjukkan bahwa model cukup akurat dalam memprediksi kelas 0.
- Recall (0.98): Dari semua data sebenarnya kelas 0, model berhasil mendeteksi 98%. Ini menunjukkan bahwa model sangat baik dalam menemukan kelas 0.
- F1-Score (0.90): F1-Score adalah rata-rata harmonis dari precision dan recall. Skor 0.90 menunjukkan keseimbangan yang baik antara precision dan recall untuk kelas 0.

2. Kelas 1 (Positif)

- Precision (0.93): Dari semua prediksi kelas 1, 93% benar-benar kelas 1. Ini menunjukkan bahwa model sangat akurat dalam memprediksi kelas 1.
- Recall (0.54): Dari semua data sebenarnya kelas 1, model hanya berhasil mendeteksi 54%. Ini menunjukkan bahwa model mungkin melewatkan banyak kasus positif.
- F1-Score (0.68): F1-Score untuk kelas 1 adalah 0.68, yang menunjukkan adanya trade-off antara precision dan recall. Model mungkin memiliki kesulitan dalam mendeteksi semua kasus positif dengan baik.

Kesimpulan

Model memiliki akurasi yang baik secara keseluruhan, tetapi ada perbedaan signifikan antara precision dan recall untuk kelas 0 dan 1. Ini berarti model mungkin lebih baik dalam mendeteksi kelas 0 dibandingkan dengan kelas 1.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Build an Automated Resignation Behavior Prediction using Machine Learning

Rakamin Academy

Validasi Akhir Model

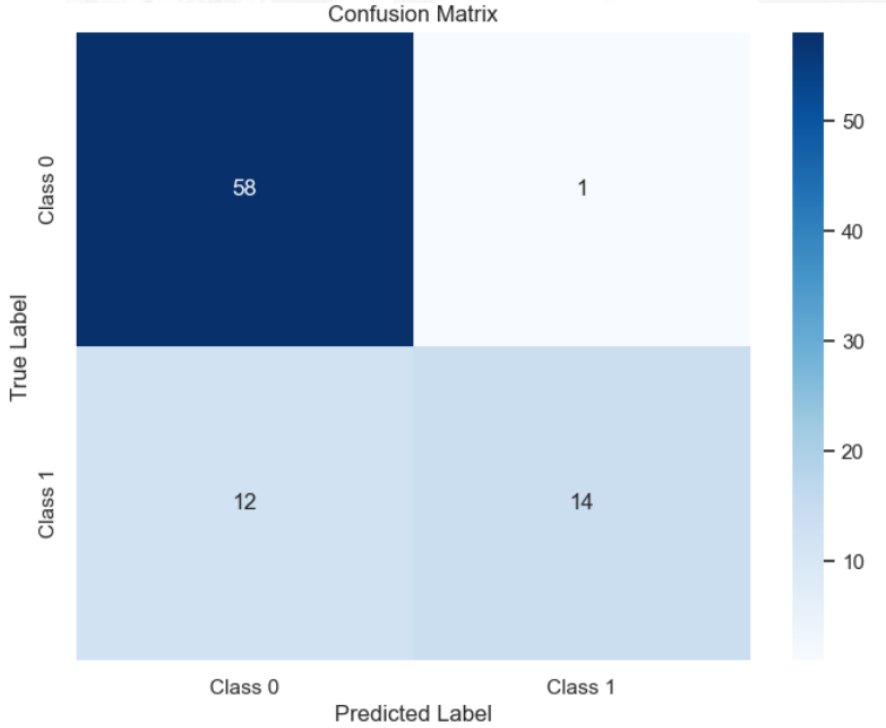
```
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.98      0.90        59
           1       0.93      0.54      0.68        26

    accuracy                           0.85        85
   macro avg       0.88      0.76      0.79        85
weighted avg       0.86      0.85      0.83        85

Confusion Matrix:
[[58  1]
 [12 14]]
ROC AUC Score: 0.89
```

- Accuracy (0.85): Model memiliki akurasi 85%, yang menunjukkan bahwa model secara keseluruhan memprediksi dengan benar 85% dari data.
- ROC AUC Score: 0.89
  Skor ini menunjukkan bahwa model memiliki kemampuan yang sangat baik dalam membedakan antara kelas positif dan negatif. Nilai mendekati 1 menandakan bahwa model cukup baik dalam memisahkan kelas-kelas ini secara keseluruhan.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Build an Automated Resignation Behavior Prediction using Machine Learning

Confusion Matrix



- True Positives (TP): 14 (Kelas 1 yang benar-benar diklasifikasikan sebagai kelas 1)
- False Positives (FP): 1 (Kelas 0 yang salah diklasifikasikan sebagai kelas 1)
- False Negatives (FN): 12 (Kelas 1 yang salah diklasifikasikan sebagai kelas 0)
- True Negatives (TN): 58 (Kelas 0 yang benar-benar diklasifikasikan sebagai kelas 0)

Untuk selengkapnya, dapat melihat jupyter notebook disini

Build an Automated Resignation Behavior Prediction using Machine Learning

Rakamin Academy

Result

```
Akurasi: 0.85 (85%)
Precision untuk kelas 1: 0.93 (93%)
Recall untuk kelas 1: 0.54 (54%)
Precision untuk kelas 0: 0.83 (83%)
Recall untuk kelas 0: 0.98 (98%)
```

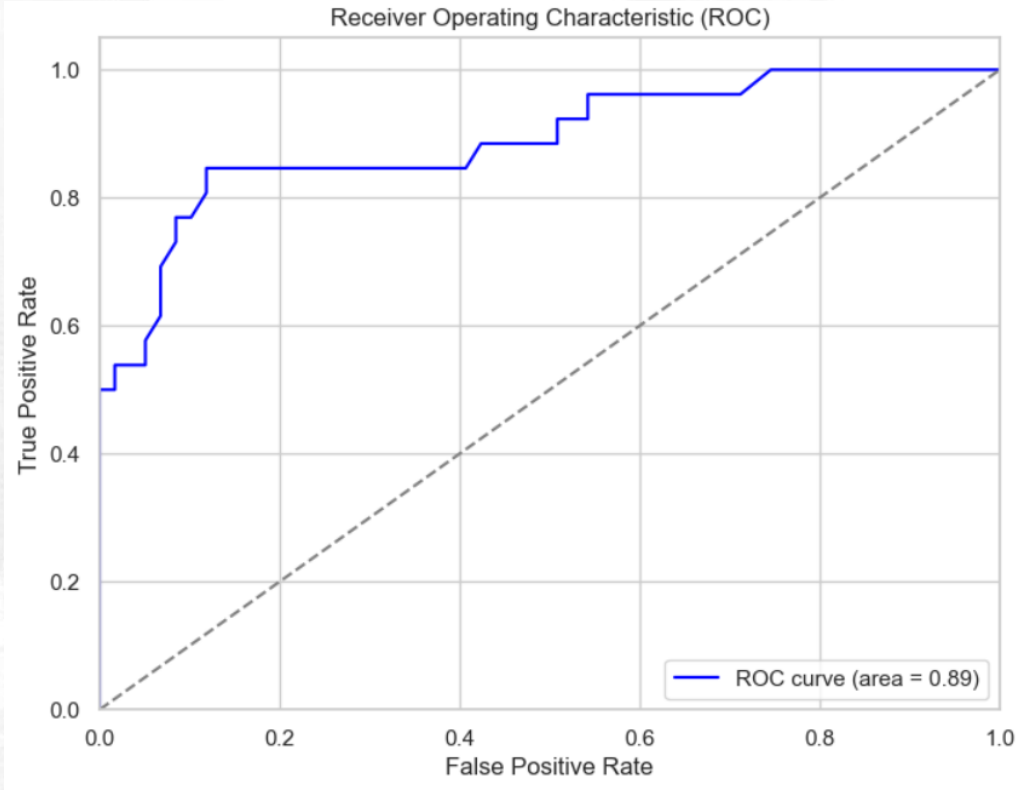
- Precision yang sangat tinggi untuk kelas 1 (93%) menunjukkan bahwa model **sangat baik** dalam memastikan bahwa prediksi positifnya benar, yang penting dalam konteks di mana false positives mahal.
- Recall yang sangat tinggi untuk kelas 0 (98%) menunjukkan bahwa model **hampir sempurna** dalam mengidentifikasi instance negatif, mengurangi false negatives.
- Recall yang rendah untuk kelas 1 (54%) menunjukkan bahwa model **masih melewatkan** sejumlah besar instance positif, yang berarti banyak positif yang tidak terdeteksi.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Build an Automated Resignation Behavior Prediction using Machine Learning

Receiver Operating Characteristic (ROC)

- Kurva ROC jauh dari garis diagonal sehingga model lebih baik daripada tebakan acak.
- AUC (Area Under the Curve) bernilai 0.89, yang menunjukkan bahwa model memiliki kinerja yang **sangat baik**.
- Kurva ROC ini menunjukkan bahwa model memiliki tingkat True Positive Rate yang tinggi dan False Positive Rate yang rendah, yang berarti model ini cukup efektif dalam membedakan antara kelas positif dan negatif.

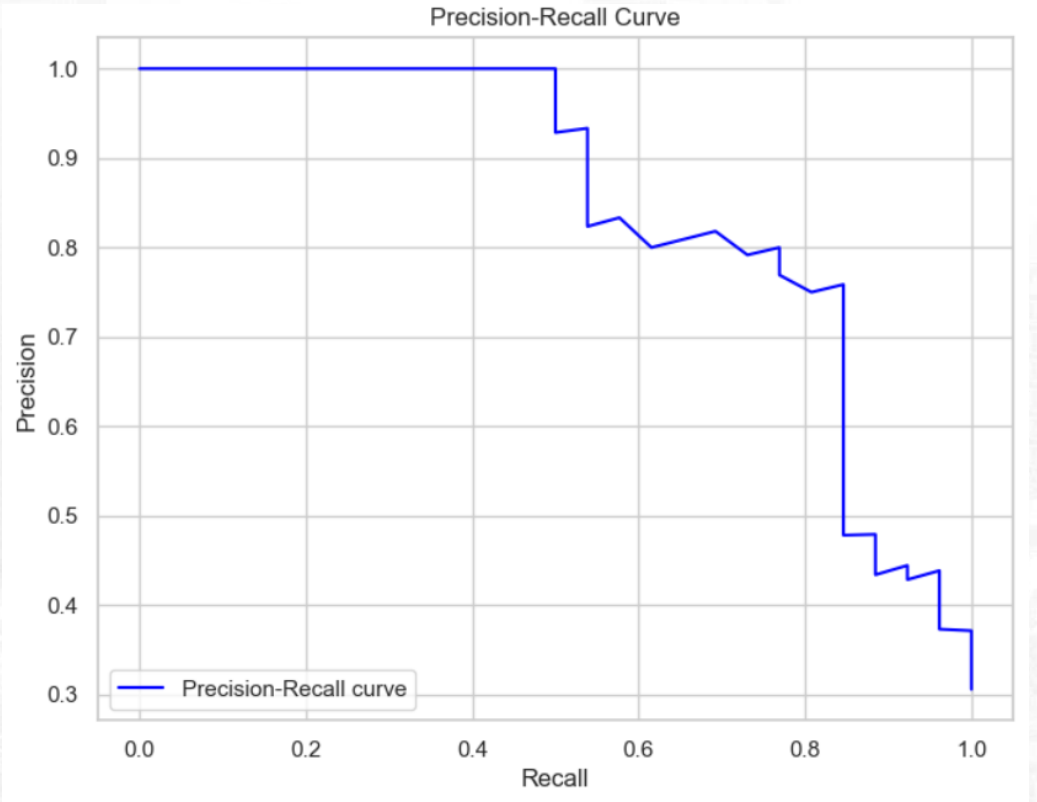


Untuk selengkapnya, dapat melihat jupyter notebook disini

Build an Automated Resignation Behavior Prediction using Machine Learning

Precision-Recall Curve

- Kurva menunjukkan precision yang tinggi ketika recall rendah, yang berarti ketika model mengklasifikasikan suatu instance sebagai positif, kemungkinan besar itu benar-benar positif. Ini adalah tanda bahwa model efektif dalam mengidentifikasi kasus positif, meskipun mungkin tidak menangkap semua kasus positif.
- Precision menurun seiring dengan meningkatnya recall, menunjukkan bahwa saat model mencoba menangkap lebih banyak instance positif, jumlah false positive juga meningkat. Ini adalah pola yang umum dan dapat diterima, tetapi kita harus mempertimbangkan seberapa cepat precision menurun.
- Jika precision tetap cukup tinggi (misalnya, di atas 0.7) bahkan saat recall mendekati 1.0, ini menunjukkan bahwa **model cukup baik** dalam menyeimbangkan antara menangkap banyak instance positif dan menjaga kesalahan positif rendah.

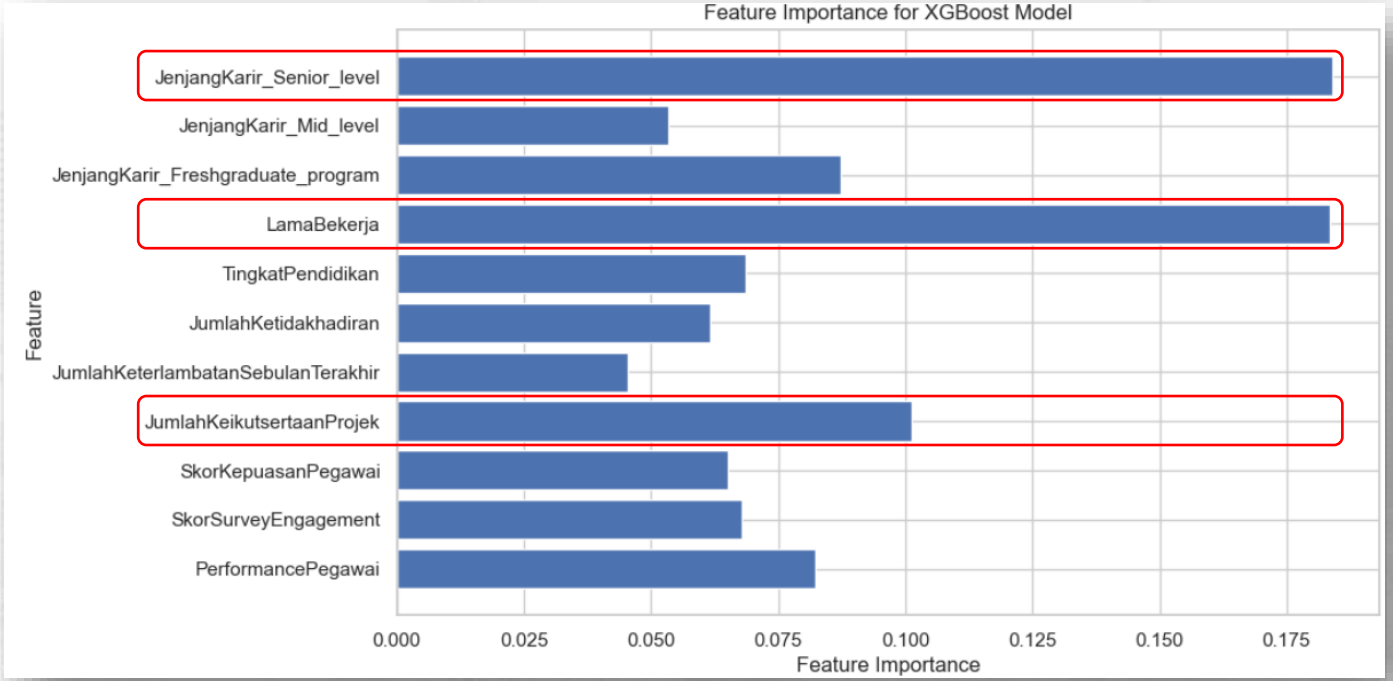


Untuk selengkapnya, dapat melihat jupyter notebook disini

Presenting Machine Learning Products to the Business Users

Rakamin Academy

FEATURE IMPORTANCE

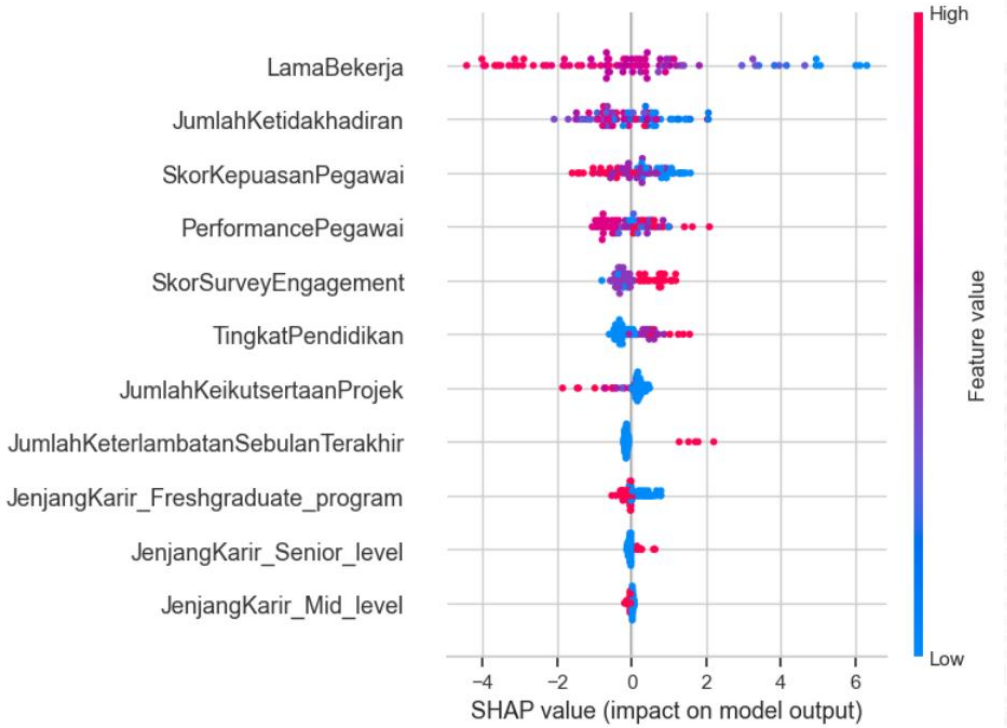


Jenjang karir, lama bekerja, dan jumlah keikutsertaan proyek merupakan faktor-faktor utama yang mempengaruhi target prediksi dalam model XGBoost ini. Hal ini bisa memberikan wawasan berharga dalam pengelolaan karyawan dan pengambilan keputusan strategis di perusahaan.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Presenting Machine Learning Products to the Business Users

SHAP Values



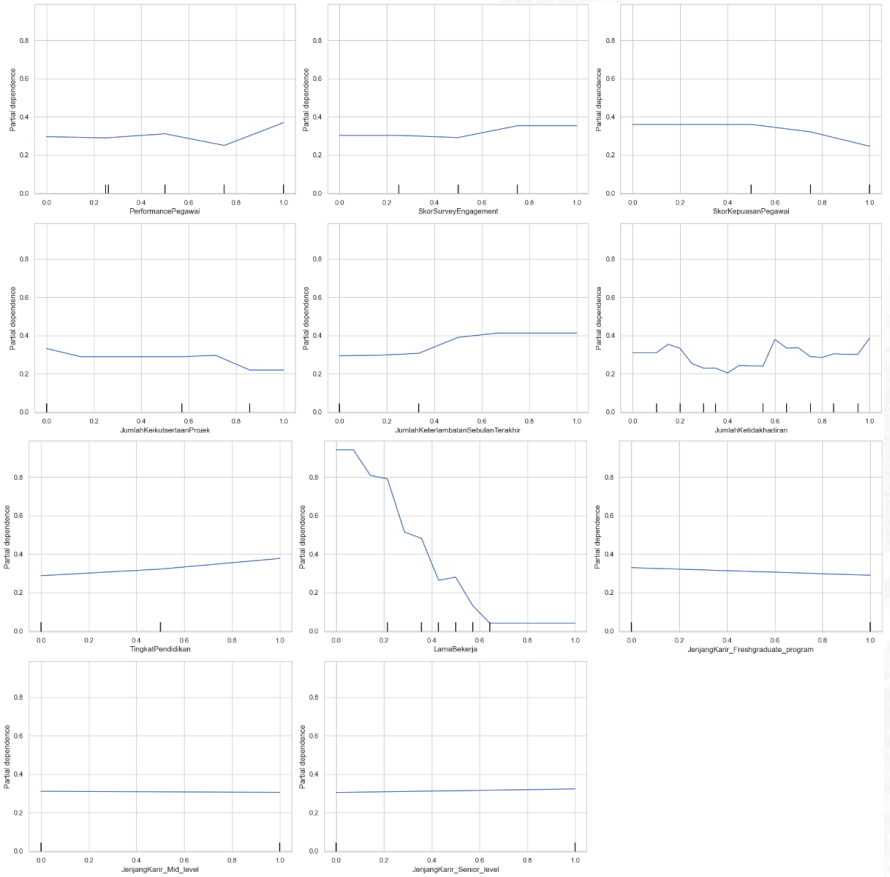
Fitur-fitur seperti **LamaBekerja, JumlahKetidakhadiran, SkorKepuasanPegawai, dan PerformancePegawai** memiliki pengaruh terbesar pada prediksi model XGBoost.

1. **LamaBekerja** memiliki dampak besar pada prediksi model, dengan variasi yang cukup besar pada nilai SHAP. Semakin lama seorang karyawan bekerja, semakin besar dampaknya terhadap hasil prediksi. Warna menunjukkan bahwa nilai yang lebih tinggi (merah) memiliki pengaruh besar, baik positif maupun negatif.
2. **JumlahKetidakhadiran** juga memiliki variasi besar dalam nilai SHAP, menunjukkan pengaruh signifikan pada prediksi. Tingkat ketidakhadiran yang lebih tinggi (merah) umumnya berdampak negatif pada prediksi model.
3. **SkorKepuasanPegawai** mempengaruhi model secara signifikan. Skor kepuasan yang lebih tinggi (merah) cenderung berdampak positif pada hasil prediksi.
4. **PerformancePegawai** memiliki dampak besar pada prediksi. Kinerja pegawai yang lebih tinggi (merah) cenderung meningkatkan hasil prediksi model.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Presenting Machine Learning Products to the Business Users

Partial Dependence Plot (PDP)

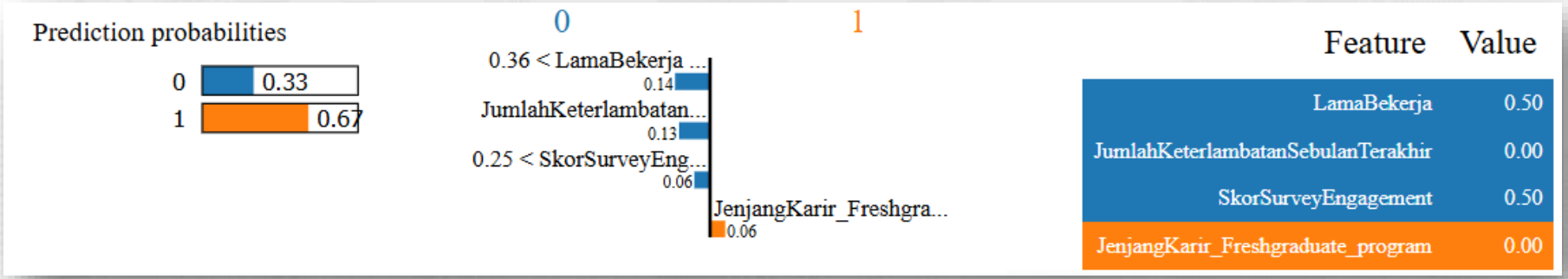


- Fitur seperti **JumlahKetidakhadiran** menunjukkan pengaruh yang cukup besar pada prediksi model, dengan perubahan dalam prediksi seiring dengan perubahan dalam fitur tersebut.
- **TingkatPendidikan** memiliki hubungan positif yang lemah dengan probabilitas resign. Semakin tinggi tingkat pendidikan, semakin tinggi probabilitas resign.
- **LamaBekerja** memiliki pengaruh yang sangat signifikan terhadap probabilitas resign. Karyawan dengan masa kerja yang lebih pendek (kurang dari 0.6) memiliki probabilitas resign yang jauh lebih tinggi.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Presenting Machine Learning Products to the Business Users

LIME Visualization



1. Model memprediksi bahwa karyawan ini memiliki kemungkinan 67% untuk resign.
2. LIME menunjukkan fitur-fitur yang paling berpengaruh terhadap prediksi tersebut, beserta nilai kontribusi masing-masing fitur:

- **LamaBekerja** (0.5): Fitur ini memberikan kontribusi terbesar terhadap prediksi resign. Karyawan dengan masa kerja lebih pendek cenderung memiliki risiko resign lebih tinggi.
- **SkorSurveyEngagement** (0.5): Karyawan dengan skor survey engagement yang rendah menunjukkan kecenderungan untuk resign lebih tinggi.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Preprocessing

Rakamin Academy

BUSINESS INSIGHT

Variabel Paling Berpengaruh	• LamaBekerja: Karyawan dengan masa kerja yang lebih pendek cenderung memiliki probabilitas resign yang lebih tinggi. • JumlahKetidakhadiran: Ketidakhadiran yang tinggi berhubungan dengan probabilitas resign yang lebih tinggi. • SkorKepuasanPegawai: Skor kepuasan yang rendah meningkatkan probabilitas resign.
Threshold untuk Resign	• LamaBekerja: Bisa ditemukan bahwa karyawan yang bekerja kurang dari X tahun memiliki risiko resign lebih tinggi. Misalnya, karyawan yang bekerja kurang dari 2 tahun memiliki probabilitas resign 30% lebih tinggi dibanding yang lain. • JumlahKetidakhadiran: Karyawan dengan lebih dari Y hari ketidakhadiran dalam setahun memiliki probabilitas resign yang lebih tinggi. • SkorKepuasanPegawai: Skor kepuasan di bawah Z (misalnya 70) mungkin menunjukkan risiko resign yang lebih tinggi.
Pengaruh Predictive Power Antara Variabel	Dengan SHAP values, kita bisa melihat berapa besar perbedaan pengaruh predictive power antara variabel seperti 'LamaBekerja' dan 'JumlahKetidakhadiran'. Misalnya, 'LamaBekerja' mungkin memiliki SHAP value yang dua kali lebih besar dibanding 'JumlahKetidakhadiran', menunjukkan bahwa 'LamaBekerja' lebih berpengaruh terhadap keputusan resign.
Individu dengan Risiko Tinggi	Dengan LIME, kita bisa mengidentifikasi individu-individu yang memiliki risiko resign tinggi dan melihat fitur spesifik yang mempengaruhi keputusan tersebut. Ini membantu dalam intervensi proaktif untuk mengurangi risiko resign.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Presenting Machine Learning Products to the Business Users

Rakamin Academy

Storytelling: Analisis dan Prediksi Resign Karyawan Menggunakan Model Machine Learning

Latar Belakang

Perusahaan XYZ menghadapi tantangan signifikan dengan tingkat resign karyawan yang tinggi, terutama di antara karyawan baru. Tingginya tingkat resign ini berdampak langsung pada produktivitas, biaya pelatihan, dan moral tim. Dengan data historis kepegawaian yang tersedia, perusahaan memutuskan untuk menggunakan model machine learning untuk memprediksi kemungkinan resign karyawan dan mengidentifikasi faktor-faktor utama yang mempengaruhi keputusan resign.

Pendekatan dan Metodologi

Saya menggunakan model XGBoost untuk membangun prediksi resign karyawan. Model ini dikenal karena kinerjanya yang kuat dalam berbagai masalah klasifikasi. Kami juga memanfaatkan alat-alat Interpretable/Explainable AI seperti SHAP dan LIME untuk mendapatkan wawasan mendalam tentang faktor-faktor yang berkontribusi terhadap prediksi model.

Hasil dan Temuan

1. Feature Importance

- LamaBekerja: Karyawan dengan masa kerja yang lebih pendek memiliki risiko resign lebih tinggi.
- JumlahKetidakhadiran: Ketidakhadiran yang tinggi berhubungan dengan probabilitas resign yang lebih tinggi.
- SkorKepuasanPegawai: Karyawan dengan skor kepuasan yang rendah cenderung lebih mungkin untuk resign.

2. Analisis SHAP

- LamaBekerja: Karyawan yang bekerja kurang dari 2 tahun memiliki probabilitas resign 30% lebih tinggi.
- JumlahKetidakhadiran: Karyawan dengan lebih dari 10 hari ketidakhadiran dalam setahun memiliki risiko resign yang meningkat.
- SkorKepuasanPegawai: Skor kepuasan di bawah 70 menunjukkan risiko resign yang signifikan.

3. Analisis LIME

Kombinasi dari rendahnya skor kepuasan dan tingginya jumlah ketidakhadiran secara signifikan meningkatkan probabilitas resign.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Presenting Machine Learning Products to the Business Users

Storytelling: Analisis dan Prediksi Resign Karyawan Menggunakan Model Machine Learning

Makna dari Temuan

- Identifikasi Risiko: Perusahaan dapat mengidentifikasi karyawan yang berisiko tinggi untuk resign dan mengambil tindakan pencegahan.
- Intervensi yang Tepat: Dengan memahami faktor-faktor utama yang mempengaruhi resign, perusahaan dapat merancang program retensi yang lebih efektif, seperti peningkatan kepuasan karyawan, program pelatihan untuk karyawan baru, dan kebijakan ketidakhadiran yang lebih baik.
- Pengurangan Biaya: Mengurangi tingkat resign akan menurunkan biaya pelatihan dan rekrutmen, serta meningkatkan produktivitas dan moral tim.

Rekomendasi

- Program Peningkatan Kepuasan Karyawan: Fokus pada inisiatif yang meningkatkan skor kepuasan, seperti pengembangan karir, pengakuan karyawan, dan keseimbangan kerja-hidup.
- Manajemen Ketidakhadiran: Implementasi kebijakan yang lebih ketat dan dukungan untuk mengurangi ketidakhadiran, seperti program kesehatan dan kebugaran.
- Pelatihan dan Pengembangan: Memberikan pelatihan tambahan dan dukungan bagi karyawan baru untuk meningkatkan masa kerja mereka dan mengurangi risiko resign awal.
- Analisis Berkelanjutan: Melakukan analisis berkelanjutan menggunakan model machine learning untuk memantau risiko resign dan mengidentifikasi tren baru.

Untuk selengkapnya, dapat melihat jupyter notebook disini