Improving Employee Retention by Predicting Employee Attrition Using Machine Learning

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

Rakamin Academy

Created by:
Yehezkiel Novianto Aryasena
hezkyaryasena@gmail.com
www.linkedin.com/in/yehezkielnov

"**Saya Yehezkiel Novianto Aryasena**. Saya merupakan fresh graduate dari Institut Teknologi Sepuluh Nopember Surabaya. Saat menjalani masa perkuliahan, saya memiliki pengalaman organisasi dan kepanitiaan yang membuat saya mampu bekerja mandiri maupun dalam tim. Saya memiliki ketertarikan untuk mempelajari hal baru terutama pada bidang data science dan saat ini sedang mendalami pengetahuan saya dalam hal tersebut dengan mengikuti course yang diselenggarakan oleh Rakamin. "

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

- Tulislah proses data preprocessing yang kamu lakukan, dan jelaskan secara singkat bagaimana kamu melakukannya, dan alasan mengapa kamu melakukan proses tersebut.

- Source code yang sudah kamu buat, dapat ditampilkan dan berikan link untuk mengakses file tersebut. Contohnya seperti di pojok kanan bawah.

Untuk selengkapnya, dapat melihat jupyter notebook disini:
https://colab.research.google.com/drive/1ddra-xzafJ5sgl2WM1eWmMrSIuxHzbg_?usp=sharing

Data Pre-Processing I Missing Value

- Untuk kolom IkutProgramLOP dapat di drop karena persentase missing value melebihi 40%
- Untuk kolom berikut ini: JumlahKetidakhadiran, SkorKepuasaanPegawai, JumlahKeikutsertaanProjek, row yang berisi missing value akan di drop karena missing value diantara 1%-39%
- Untuk kolom JumlahKeterlambatanSebulanTerakhir akan diisi dengan nilai aggregat karena persentase missing value yang kecil
- Untuk kolom alasan resign akan diisi value 'masih_bekerja' karena merepresentasikan karyawan yang masih bekerja di perusahaan

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

Rakamin Academy

```
df.isna().mean().sort_values(ascending=False) * 100
```

```
IkutProgramLOP                    89.895470
AlasanResign                      22.996516
JumlahKetidakhadiran               2.090592
SkorKepuasanPegawai                1.742160
JumlahKeikutsertaanProjek          1.045296
JumlahKeterlambatanSebulanTerakhir 0.348432
TanggalPenilaianKaryawan           0.000000
TanggalHiring                      0.000000
TanggalLahir                       0.000000
PernahBekerja                      0.000000
TingkatPendidikan                  0.000000
Email                              0.000000
NomorHP                            0.000000
Username                           0.000000
EnterpriseID                       0.000000
SkorSurveyEngagement               0.000000
HiringPlatform                     0.000000
AsalDaerah                         0.000000
PerformancePegawai                 0.000000
JenjangKarir                       0.000000
Pekerjaan                          0.000000
StatusKepegawaian                  0.000000
JenisKelamin                       0.000000
StatusPernikahan                   0.000000
TanggalResign                      0.000000
dtype: float64
```

Data Pre-Processing II
Duplicated Value

- Tidak ditemukan data yang duplikat

```
[ ]  df.duplicated().sum()

     0
```

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

Data Pre-Processing III
Un-used Value

- Dapat dilihat pada tabel bahwa kolom PernahBekerja sebenarnya berisi 1 unique value saja karena disini saya mengasumsikan bahwa nilai '1' pada kolom tersebut sama valuenya dengan nilai 'Yes'. Sehingga dapat langsung kita drop saja kolom tersebut.

column	unique_value
Username	[spiritedPorpoise3, jealousGelding2, pluckyMue...
EnterpriseID	[111065, 106080, 106452, 106325, 111171, 10641...
StatusPernikahan	[Belum_menikah, Menikah, Bercerai, Lainnya, -]
JenisKelamin	[Pria, Wanita]
StatusKepegawaian	[Outsource, FullTime, Internship]
Pekerjaan	[Software Engineer (Back End), Data Analyst, S...
JenjangKarir	[Freshgraduate_program, Senior_level, Mid_level]
PerformancePegawai	[Sangat_bagus, Sangat_kurang, Bagus, Biasa, Ku...
AsalDaerah	[Jakarta Timur, Jakarta Utara, Jakarta Pusat, ...
HiringPlatform	[Employee_Referral, Website, Indeed, LinkedIn,...
SkorSurveyEngagement	[4, 3, 2, 1, 5]
SkorKepuasanPegawai	[4.0, 3.0, 5.0, 2.0, 1.0]
JumlahKeikutsertaanProjek	[0.0, 4.0, 6.0, 5.0, 7.0, 3.0, 1.0, 2.0, 8.0]
JumlahKeterlambatanSebulanTerakhir	[0.0, 4.0, 3.0, 5.0, 2.0, 6.0, 1.0]
JumlahKetidakhadiran	[9.0, 3.0, 11.0, 6.0, 10.0, 19.0, 4.0, 2.0, 50...
NomorHP	[+6282232522xxx, +6281270745xxx, +6281346215xx...
Email	[spiritedPorpoise3135@yahoo.com, jealousGeldin...
TingkatPendidikan	[Magister, Sarjana, Doktor]
PernahBekerja	[1, yes]
AlasanResign	[masih_bekerja, toxic_culture, jam_kerja, gant...
TanggalLahir	[1972-07-01, 1984-04-26, 1974-01-07, 1979-11-2...
TanggalHiring	[2011-01-10, 2014-01-06, 2014-2-17, 2013-11-11...
TanggalPenilaianKaryawan	[2016-2-15, 2020-1-17, 2016-01-10, 2020-02-04,...
TanggalResign	[-, 2018-6-16, 2014-9-24, 2018-09-06, 2019-01-...

Data Pre-Processing IV
Adjust Value 1

- Terdapat value '-' yang tidak jelas merefer ke status pernikahan yang seperti apa. Sehingga treatment yang perlu dilakukan adalah mengubah value '-' menjadi 'Lainnya'

```
[ ]  df['StatusPernikahan'].unique()

     array(['Belum_menikah', 'Menikah', 'Bercerai', 'Lainnya', '-'],
           dtype=object)
```

```
df['StatusPernikahan'] = df['StatusPernikahan'].replace(['-'], 'Lainnya')
df['StatusPernikahan'].unique()
```

```
array(['Belum_menikah', 'Menikah', 'Bercerai', 'Lainnya'], dtype=object)
```

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

Data Pre-Processing IV
Adjust Value 2

- Terdapat value 'Product Design (UI & UX)' yang seharusnya merupakan value dari kolom 'Pekerjaan'. Hal ini mungkin merupakan kesalahan input sehingga lebih baik value diganti dengan value 'tidak_disebutkan' . Sehingga treatment yang perlu dilakukan adalah mengubah value 'Product Design (UI & UX)' menjadi 'tidak_disebutkan' saja.

```
df['AlasanResign'].unique()
```

```
array(['masih_bekerja', 'toxic_culture', 'jam_kerja', 'ganti_karir',
       'tidak_bahagia', 'internal_conflict', 'Product Design (UI & UX)',
       'kejelasan_karir', 'tidak_bisa_remote', 'apresiasi', 'leadership'],
      dtype=object)
```

```
df['AlasanResign'] = df['AlasanResign'].replace(['Product Design (UI & UX)'], 'tidak_disebutkan')
df['AlasanResign'].unique()
```

```
array(['masih_bekerja', 'toxic_culture', 'jam_kerja', 'ganti_karir',
       'tidak_bahagia', 'internal_conflict', 'tidak_disebutkan',
       'kejelasan_karir', 'tidak_bisa_remote', 'apresiasi', 'leadership'],
      dtype=object)
```

Data Pre-Processing IV
Adjust Value 2

- Terdapat value 'Product Design (UI & UX)' yang seharusnya merupakan value dari kolom 'Pekerjaan'. Hal ini mungkin merupakan kesalahan input sehingga lebih baik value diganti dengan value 'tidak_disebutkan' . Sehingga treatment yang perlu dilakukan adalah mengubah value 'Product Design (UI & UX)' menjadi 'tidak_disebutkan' saja.

```
df['AlasanResign'].unique()
```

```
array(['masih_bekerja', 'toxic_culture', 'jam_kerja', 'ganti_karir',
       'tidak_bahagia', 'internal_conflict', 'Product Design (UI & UX)',
       'kejelasan_karir', 'tidak_bisa_remote', 'apresiasi', 'leadership'],
      dtype=object)
```

```
df['AlasanResign'] = df['AlasanResign'].replace(['Product Design (UI & UX)'], 'tidak_disebutkan')
df['AlasanResign'].unique()
```

```
array(['masih_bekerja', 'toxic_culture', 'jam_kerja', 'ganti_karir',
       'tidak_bahagia', 'internal_conflict', 'tidak_disebutkan',
       'kejelasan_karir', 'tidak_bisa_remote', 'apresiasi', 'leadership'],
      dtype=object)
```

Data Pre-Processing V
Adjust Data Types

- Terlihat bahwa kolom yang berbau tanggal masih bertipe object. Sehingga treatment yang perlu dilakukan adalah mengubah tipe data tersebut menjadi datetime.

```
df['TanggalLahir']=pd.to_datetime(df['TanggalLahir'])
df['TanggalHiring']=pd.to_datetime(df['TanggalHiring'])
df['TanggalPenilaianKaryawan']=pd.to_datetime(df['TanggalPenilaianKaryawan'])
df['TanggalResign']=pd.to_datetime(df['TanggalResign'],errors='coerce')
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 276 entries, 0 to 286
Data columns (total 23 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Username                    276 non-null    object
 1   EnterpriseID                276 non-null    int64
 2   StatusPernikahan            276 non-null    object
 3   JenisKelamin                276 non-null    object
 4   StatusKepegawaian           276 non-null    object
 5   Pekerjaan                   276 non-null    object
 6   JenjangKarir                276 non-null    object
 7   PerformancePegawai          276 non-null    object
 8   AsalDaerah                  276 non-null    object
 9   HiringPlatform              276 non-null    object
 10  SkorSurveyEngagement        276 non-null    int64
 11  SkorKepuasanPegawai         276 non-null    float64
 12  JumlahKeikutsertaanProjek   276 non-null    float64
 13  JumlahKeterlambatanSebulanTerakhir  276 non-null  float64
 14  JumlahKetidakhadiran        276 non-null    float64
 15  NomorHP                     276 non-null    object
 16  Email                       276 non-null    object
 17  TingkatPendidikan           276 non-null    object
 18  AlasanResign                276 non-null    object
 19  TanggalLahir                276 non-null    object
 20  TanggalHiring               276 non-null    object
 21  TanggalPenilaianKaryawan    276 non-null    object
 22  TanggalResign               276 non-null    object
dtypes: float64(4), int64(2), object(17)
memory usage: 51.8+ KB
```

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

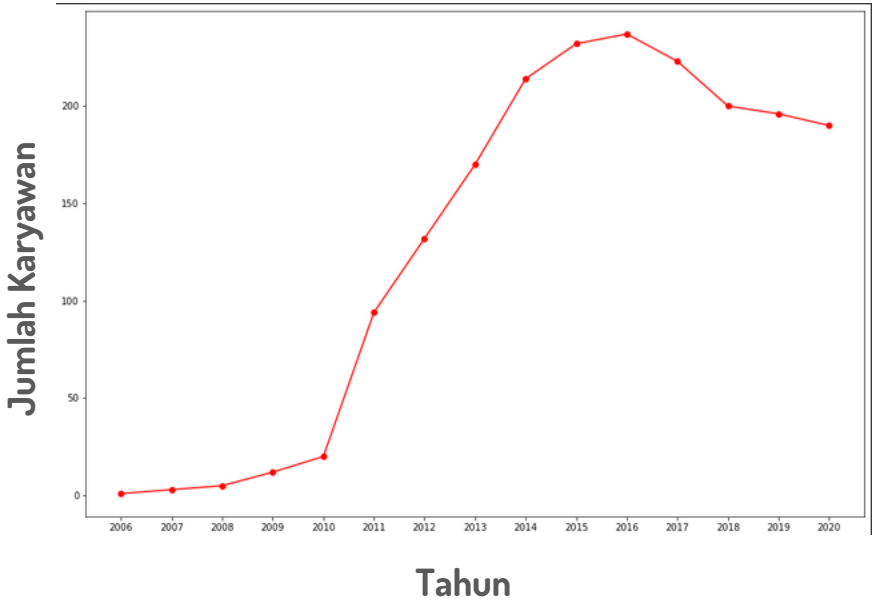
Annual Report on Employee Number Changes

Masukkan grafik visualisasi pada tugas ini, kemudian tuliskan pula hasil analisismu, insight apa saja yang kamu dapatkan.

EDA (I)
Annual Report on Employee Number Changes

Insight:

- Perusahaan mengalami kenaikan jumlah karyawan yang signifikan dari tahun 2010 dengan jumlah karyawan terbanyak pada tahun 2016
- Hal ini perlu diperhatikan departemen perekrutan karyawan untuk mempertimbangkan apakah alokasi sumber daya sudah siap untuk dapat mencukupi karyawan karyawan yang baru ini

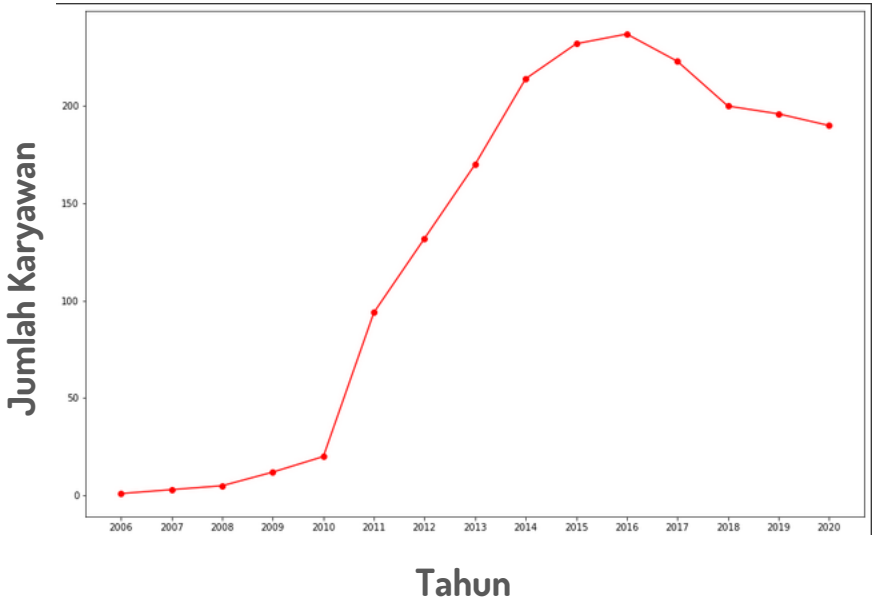


Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

EDA (II)
Annual Report on Employee Number Changes

Insight:

- Pada tahun 2016 sampai tahun 2020, setiap tahunnya terjadi penurunan jumlah karyawan
- Hal ini menandakan jumlah karyawan yang melakukan *resign* melebihi jumlah karyawan baru. Sehingga perusahaan perlu melakukan sesuatu untuk mempertahankan karyawan yang ada
- Hal yang dapat dilakukan perusahaan yaitu dapat memprediksi karyawan yang akan resign sehingga dapat mengantisipasi dan mengurangi jumlah karyawan resign pada tahun tersebut



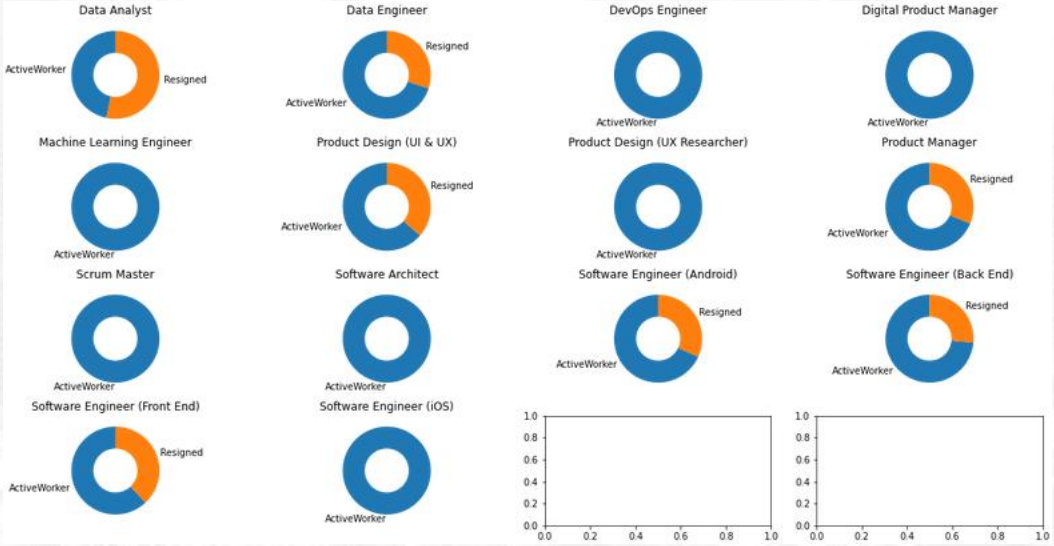
Resign Reason Analysis for Employee Attrition Management Strategy

Masukkan grafik visualisasi pada tugas ini, kemudian tuliskan pula hasil analisismu, insight apa saja yang kamu dapatkan.

Resign Reason Analysis for Employee Attrition Management Strategy

Insight:

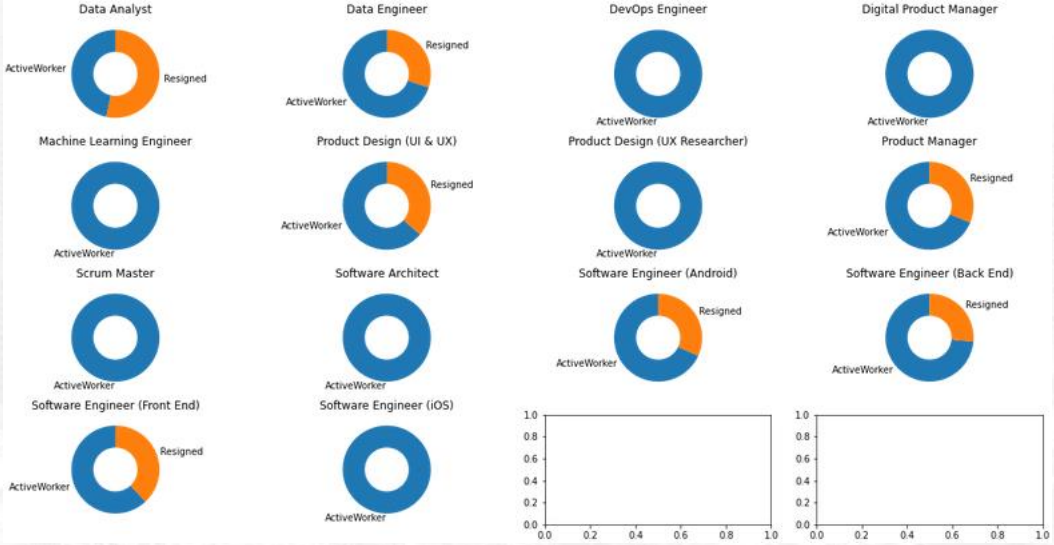
1. Pada visualisasi pie chart disamping diketahui bahwa jenis pekerjaan yang persentase karyawan resign paling tinggi berada pada karyawan dengan posisi pekerjaan sebagai Data Analyst
2. Beberapa posisi pekerjaan dalam perusahaan dinilai cukup baik dalam memberikan lingkungan kerja yang baik kepada karyawannya. Hal ini ditandai dengan tidak/belum adanya karyawan yang resign pada posisi pekerjaan tersebut. Beberapa posisi pekerjaan yang dimaksud seperti DevOps Engineer, Digital product Manager, ML Engineer, dst.



Resign Reason Analysis for Employee Attrition Management Strategy

Insight:

3. Hal yang dapat dilakukan perusahaan untuk mengurangi karyawan yang resign adalah dengan mempertahankan lingkungan kerja yang sudah baik pada posisi pekerjaan yang tidak/belum ada karyawan yang resign serta memperbaiki lingkungan kerja pada posisi pekerjaan masih memiliki karyawan yang resign dilihat dari alasan mereka resign



Build an Automated Resignation Behavior Prediction using Machine Learning

- Check kembali hasil preprocessing apakah outlier, duplikasi dan missing data telah ter-handle dengan benar
- Melakukan feature transformation dan feature engineering agar data menjadi siap untuk dilakukan modelling
- Melakukan Split data train dan testing lalu dilakukan modelling dengan menggunakan metode machine learning yang berbeda
- Lakukan evaluation dan bandingkan metode machine learning mana yang memiliki hasil terbaik dengan membuat tabel perbandingan

Data Pre-Processing (Part 2) Missing Value

Disini masih ada nilai nan/null pada TanggalResign karena value yang awalnya '-' akan diganti menjadi nan yang disebabkan oleh pd.to_datetime(df['TanggalResign'],errors='coerce') . Namun pada model nantinya juga kita tidak memerlukan feature yang berhubungan dengan tanggal maka semua feature tanggal akan di drop.

```
df_new=df.copy()
df_new.isna().sum()
```

```
Username                          0
EnterpriseID                      0
StatusPernikahan                  0
JenisKelamin                      0
StatusKepegawaian                 0
Pekerjaan                         0
JenjangKarir                      0
PerformancePegawai                0
AsalDaerah                        0
HiringPlatform                    0
SkorSurveyEngagement              0
SkorKepuasanPegawai               0
JumlahKeikutsertaanProjek         0
JumlahKeterlambatanSebulanTerakhir 0
JumlahKetidakhadiran              0
NomorHP                           0
Email                             0
TingkatPendidikan                 0
AlasanResign                      0
TanggalLahir                      0
TanggalHiring                     0
TanggalPenilaianKaryawan          0
TanggalResign                   190
dtype: int64
```

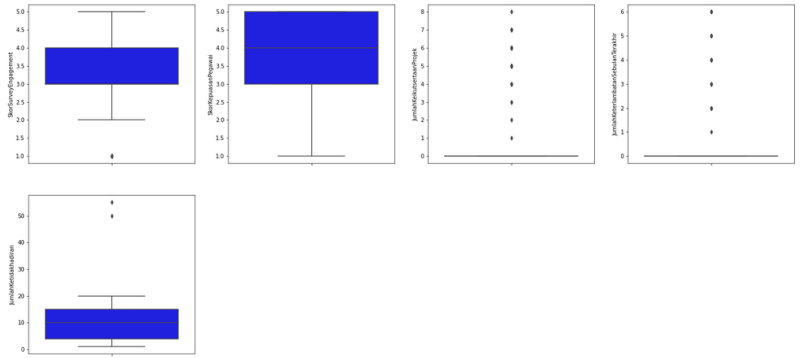
Data Pre-Processing (Part 2) Irrelevant Feature

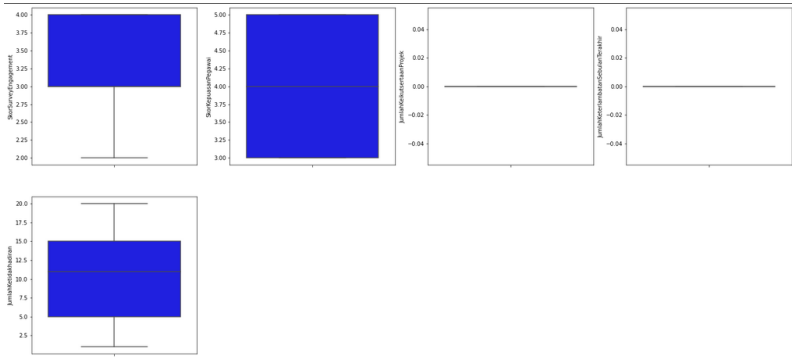
Menghapus beberapa feature lainnya yang tidak begitu berpengaruh kedalam model seperti: Username, EnterpriseID,NomorHP, dan Email

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 276 entries, 0 to 286
Data columns (total 19 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Username                        276 non-null    object
 1   EnterpriseID                    276 non-null    int64
 2   StatusPernikahan                276 non-null    object
 3   JenisKelamin                    276 non-null    object
 4   StatusKepegawaian               276 non-null    object
 5   Pekerjaan                       276 non-null    object
 6   JenjangKarir                    276 non-null    object
 7   PerformancePegawai              276 non-null    object
 8   AsalDaerah                      276 non-null    object
 9   HiringPlatform                  276 non-null    object
 10  SkorSurveyEngagement            276 non-null    int64
 11  SkorKepuasanPegawai             276 non-null    float64
 12  JumlahKeikutsertaanProjek       276 non-null    float64
 13  JumlahKeterlambatanSebulanTerakhir  276 non-null    float64
 14  JumlahKetidakhadiran            276 non-null    float64
 15  NomorHP                         276 non-null    object
 16  Email                           276 non-null    object
 17  TingkatPendidikan               276 non-null    object
 18  AlasanResign                    276 non-null    object
dtypes: float64(4), int64(2), object(13)
memory usage: 43.1+ KB
```

Data Pre-Processing (Part 2) Outliers

Masih terdapat outlier pada dataset sehingga dilakukan treatment dengan metode IQR

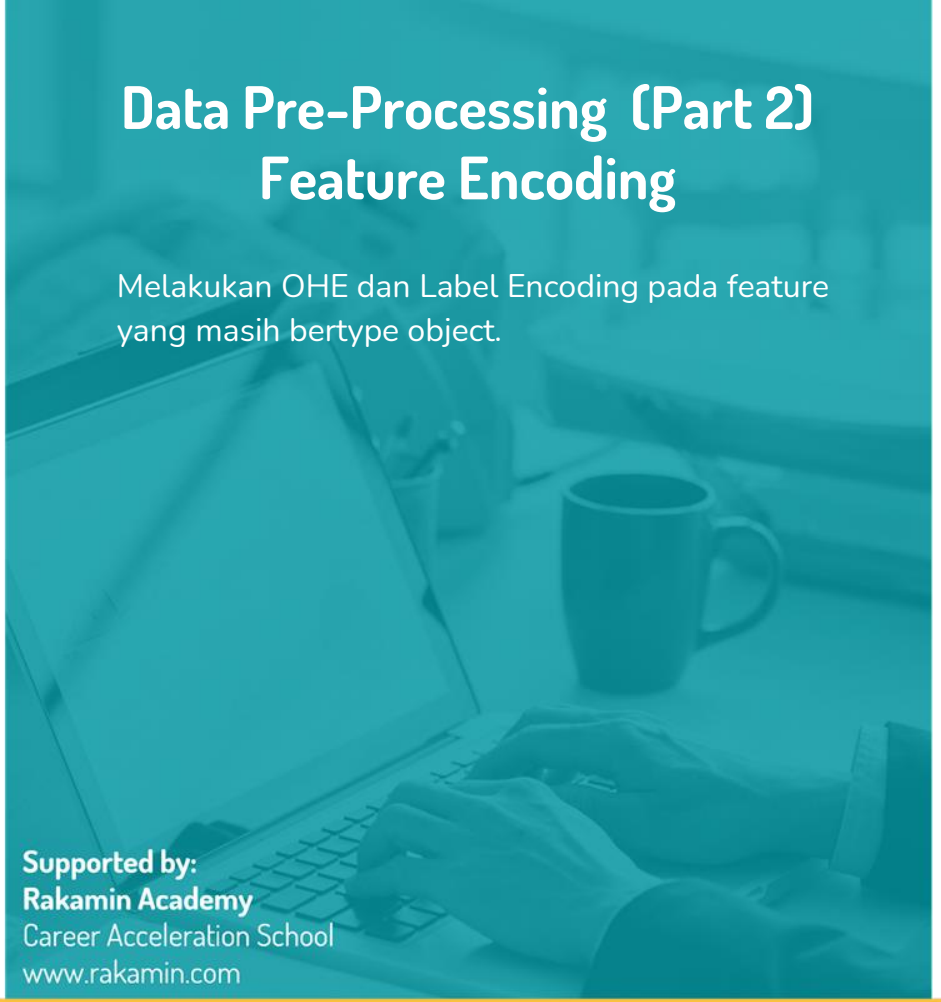




Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

Data Pre-Processing (Part 2) Feature Encoding

Melakukan OHE dan Label Encoding pada feature yang masih bertype object.



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

Rakamin Academy

```
pegawai=pd.get_dummies(df_clean['StatusKepegawaian'], prefix='status_kepegawaian')
df_clean=pd.concat([df_clean,pegawai], axis=1)

kerja=pd.get_dummies(df_clean['Pekerjaan'], prefix='pekerjaan')
df_clean=pd.concat([df_clean,kerja], axis=1)

karir=pd.get_dummies(df_clean['JenjangKarir'], prefix='jenjang_karir')
df_clean=pd.concat([df_clean,karir], axis=1)

platform=pd.get_dummies(df_clean['HiringPlatform'], prefix='platform')
df_clean=pd.concat([df_clean,platform], axis=1)

alasan=pd.get_dummies(df_clean['AlasanResign'], prefix='alasan')
df_clean=pd.concat([df_clean,alasan], axis=1)
```

```
mapping_performance={
    'Sangat_kurang':0,
    'Kurang':1,
    'Biasa':2,
    'Bagus':3,
    'Sangat_bagus':4,
}
df_clean['PerformancePegawai']=df_clean['PerformancePegawai'].map(mapping_performance)
```

Modeling + Evaluation
Define Target and CM Score Used

Target = karena model dibuat untuk memprediksi karyawan yang akan melakukan resign maka target adalah hasil Feature Encoding dari feature AlasanResign yaitu **alasan_masih_bekerja** (berarti **karyawan masih bekerja (1)** dan **karyawan resign (0)**)

CM Score Used = **Recall**, hal ini dikarenakan perusahaan perlu mengurangi kesalahan dalam memprediksi karyawan yang tidak resign padahal ingin resign. Dengan mengantisipasi hal ini terjadi maka perusahaan dapat melakukan retensi karyawan semaksimal mungkin. **False Negative**

Modeling + Evaluation

Dengan menggunakan 3 algoritma yaitu Logistic Regression, K-Nearest Neighborhood, dan KNN. Berikut ini merupakan hasil Confusion Matrix setiap Model.

Insight:

Nilai Recall tertinggi disini dipegang oleh ketiga model sehingga model yang dipakai bisa model apa saja dari ketiga ini.

	Log Res	KNN	Decision Tree
Accuracy	0.894737	0.649123	1.0
Recall	1.000000	1.000000	1.0
Precision	0.860465	0.649123	1.0
F1 Score	0.925000	0.787234	1.0

Modeling + Evaluation HyperParameter Tuning

Dengan menggunakan 3 algoritma yaitu Logistic Regression, K-Nearest Neighborhood, dan KNN. Berikut ini merupakan hasil Confusion Matrix setiap Model.

Insight:

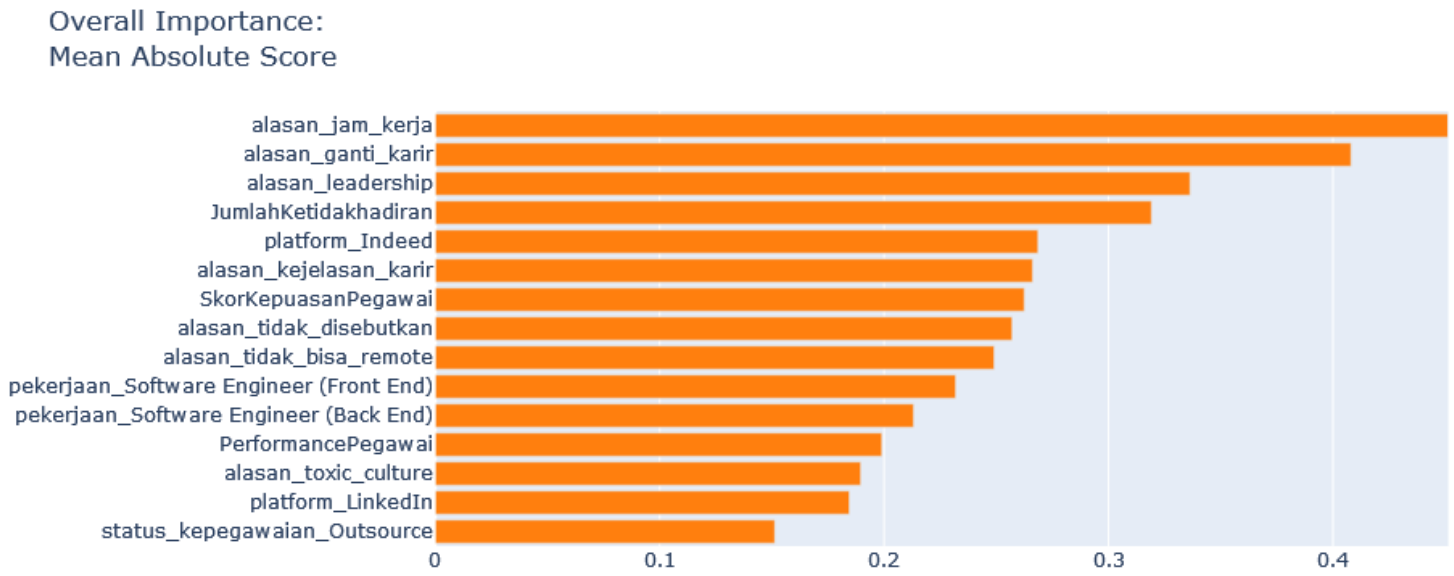
Setelah melakukan Hyperparameter Tuning, nilai Recall tertinggi disini hasilnya masih sama dengan model sebelumnya yaitu masih dipegang oleh ketiga model sehingga model yang dapat dipakai bisa model apa saja dari ketiga model ini.

	Logistic Regression Tuned	KNN Tuned	Decision Tree Tuned
Accuracy	1.0	0.649123	0.649123
Recall	1.0	1.000000	1.000000
Precision	1.0	0.649123	0.649123
F1 Score	1.0	0.787234	0.787234

Presenting Machine Learning Products to the Business Users

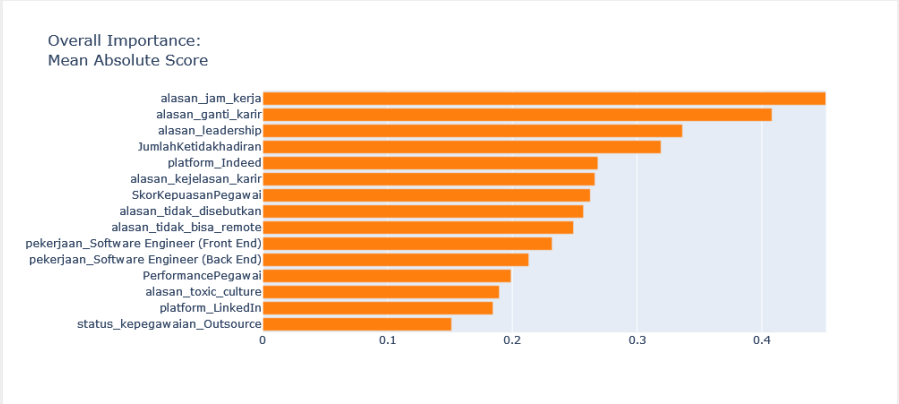
Buatlah story telling, makna, dan rekomendasi dari model machine learning yang kamu buat bisa sangat bermanfaat untuk perusahaan dalam perspektif bisnis atau kebutuhan penyelesaian masalah yang ada.

Presenting Machine Learning Products to the Business Users



Insight & Bussiness Recommendation

1. Berikut ini merupakan visualisasi dari feature importance yang ada dalam dataset
2. Terlihat bahwa kebanyakan karyawan resign karena 3 alasan utama yakni jam kerja, ganti karir, dan leadership
3. Hal yang dapat dilakukan perusahaan yaitu mengatur ulang jam kerja karyawan agar dapat memaksimalkan kinerja karyawan namun tetap memperhatikan perasaan karyawan serta mempromosikan posisi pekerjaan kepada karyawan yang sudah mencapai target tertentu
4. Alasan resign lainnya seperti kejelasan karir dapat ditangani dengan menetapkan kontrak diawal agar karyawan juga tidak merasa dirugikan oleh perusahaan.



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