Analisa untuk Memprediksi Karyawan yang Keluar (Churn)

Final Project

Dibimbing
Data Science Batch 23B

Rini Kustiah Rahmadiana

Latar Belakang

salah satu tantangan utama yang dihadapi perusahaan adalah mengelola turnover karyawan atau churn. Churn karyawan dapat menyebabkan kerugian finansial, kehilangan produktivitas, dan menurunkan moral di tempat kerja. Oleh karena itu, penting bagi perusahaan untuk dapat mengantisipasi dan mencegah churn karyawan sebanyak mungkin.

Dengan kemajuan dalam bidang machine learning (ML) dan analisis data, perusahaan memiliki kesempatan untuk menggunakan pendekatan prediktif untuk mengidentifikasi karyawan yang berisiko tinggi untuk churn di masa depan.

. Dengan memanfaatkan teknik ML seperti model klasifikasi, perusahaan dapat menganalisis data historis dan variabel-variabel penting untuk membangun model yang dapat memprediksi churn karyawan dengan tingkat akurasi yang tinggi.

”

Prediksi adalah kunci untuk membuka pintu masa depan. Dengan alat ML, kita dapat membaca isyarat yang tersembunyi dalam data untuk mengantisipasi langkah-langkah berikutnya

TOPIK

Penggunaan ML untuk memprediksi karyawan yang churn

Konten

Latar Belakang

Kesimpulan

Data Feature dan EDA

Insight

Machine Learning

Rekomendasi

Fokus Pembahasan

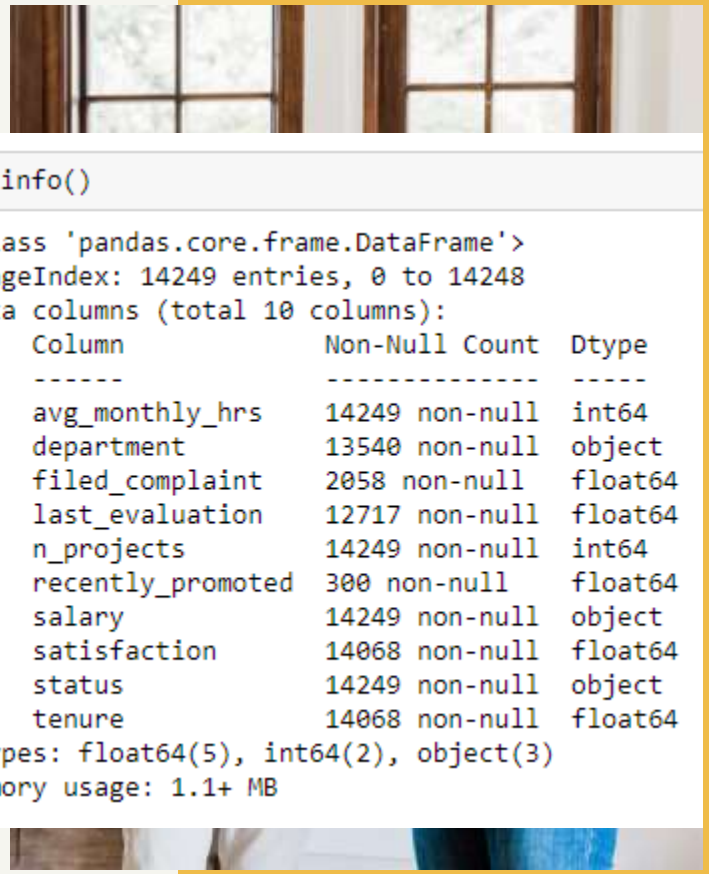
Dalam Dataset ini kami ingin mengetahui tentang 2 hal, yaitu:

1. Faktor apa saja yang paling berpengaruh yang membuat karyawan churn?
2. Membuat model Machine Learning tentang kasus ini.

Dataset (employee_churn.csv)

https://www.kaggle.com/datasets/ninopadilla13/employee-churn

Kolom	
1.	avg_monthly_hrs
2.	department
3.	filed_complaint
4.	last_evaluation
5.	n_projects
6.	recently_promoted
7.	salary
8.	satisfaction
9.	status
10.	tenure



```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14249 entries, 0 to 14248
Data columns (total 10 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   avg_monthly_hrs    14249 non-null  int64
 1   department         13540 non-null  object
 2   filed_complaint    2058 non-null   float64
 3   last_evaluation    12717 non-null  float64
 4   n_projects         14249 non-null  int64
 5   recently_promoted  300 non-null    float64
 6   salary             14249 non-null  object
 7   satisfaction       14068 non-null  float64
 8   status             14249 non-null  object
 9   tenure             14068 non-null  float64
dtypes: float64(5), int64(2), object(3)
memory usage: 1.1+ MB
```

```
# number of missing values per column
df.isna().sum()
```

```
avg_monthly_hrs            0
department               709
filed_complaint        12191
last_evaluation         1532
n_projects                 0
recently_promoted      13949
salary                     0
satisfaction             181
status                     0
tenure                   181
dtype: int64
```

```
df = df.dropna(subset=['tenure'])


df.filed_complaint.fillna(0, inplace=True)
df.recently_promoted.fillna(0, inplace=True)
df['last_evaluation_missing'] = df.last_evaluation.isnull().astype(int)
df.last_evaluation.fillna(0, inplace=True)

df['department'].fillna('Missing',inplace=True)

df.department.replace('information_technology', 'IT', inplace=True)
```

```
# Transformed Dataset
df.sample(10)
```

	avg_monthly_hrs	department	filed_complaint	last_evaluation	n_projects	recently_promoted	salary	satisfaction	status	tenure	last_evaluation_missing
3	188	support	0.0	0.707017	3	0.0	low	0.800805	Employed	2.0	0
8	155	management	0.0	1.000000	4	0.0	low	0.851027	Employed	3.0	0
4	255	engineering	1.0	0.889249	5	0.0	low	0.762451	Employed	4.0	0

```
df.isnull().sum()
```

```
avg_monthly_hrs            0
department                 0
filed_complaint            0
last_evaluation            0
n_projects                 0
recently_promoted          0
salary                     0
satisfaction               0
status                     0
tenure                     0
last_evaluation_missing    0
dtype: int64
```

Tidak ada lagi data yang missing

```
# drop duplicated rows
df = df.drop_duplicates()
```

```
# sanity check
df.duplicated().sum()
```

```
0
```

Data Encoder

```
# Transforming Categorical Data (One Hot Encoding)

OHE_Cols = ['salary']

for col in OHE_Cols:
 data1 = pd.get_dummies(df[[col]],dtype=int)
 data2 = df.drop(columns=col)


df = pd.concat([data1, data2], axis = 1)
```

```
# Transformed Dataset
df.sample(10)
```

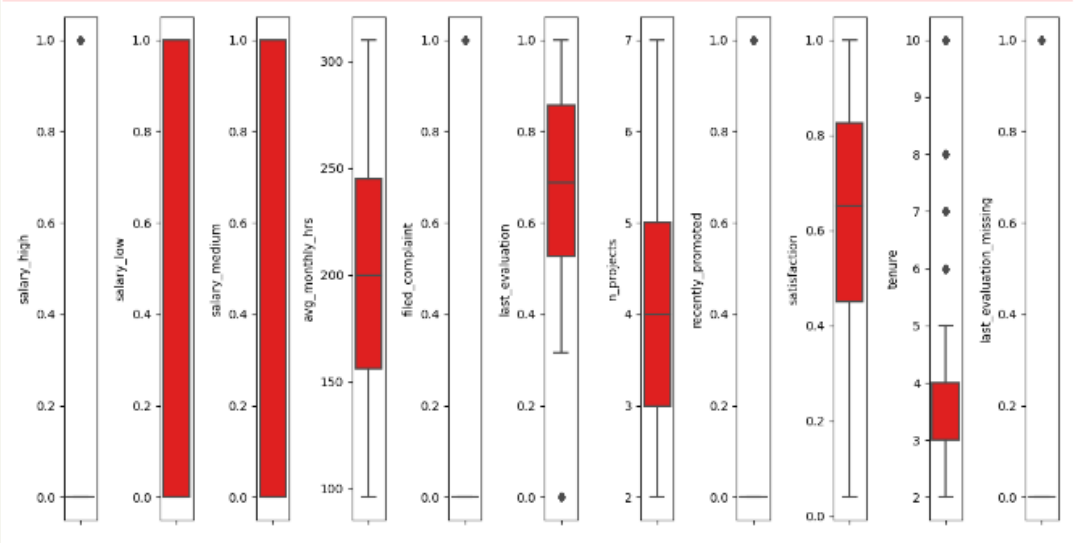
salary_high	salary_low	salary_medium	avg_monthly_hrs	department	filed_complaint	last_evaluation	n_projects	recently_promoted	satisfaction	status	ter
0	1	0	156	product	0.0	0.638501	4	0.0	0.779843	Employed	
0	0	1	131	support	1.0	0.432786	5	0.0	0.902914	Employed	
0	1	0	221	engineering	1.0	0.805053	4	0.0	0.932998	Employed	
0	0	1	198	sales	0.0	0.924947	4	0.0	0.625847	Employed	

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 14068 entries, 0 to 14248
Data columns (total 13 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   salary_high              14068 non-null  int32  
 1   salary_low               14068 non-null  int32  
 2   salary_medium            14068 non-null  int32  
 3   avg_monthly_hrs          14068 non-null  int64  
 4   department               14068 non-null  object 
 5   filed_complaint          14068 non-null  float64
 6   last_evaluation          14068 non-null  float64
 7   n_projects               14068 non-null  int64  
 8   recently_promoted        14068 non-null  float64
 9   satisfaction             14068 non-null  float64
 10  status                   14068 non-null  object 
 11  tenure                   14068 non-null  float64
 12  last_evaluation_missing  14068 non-null  int32  
dtypes: float64(5), int32(4), int64(2), object(2)
memory usage: 1.3+ MB
```

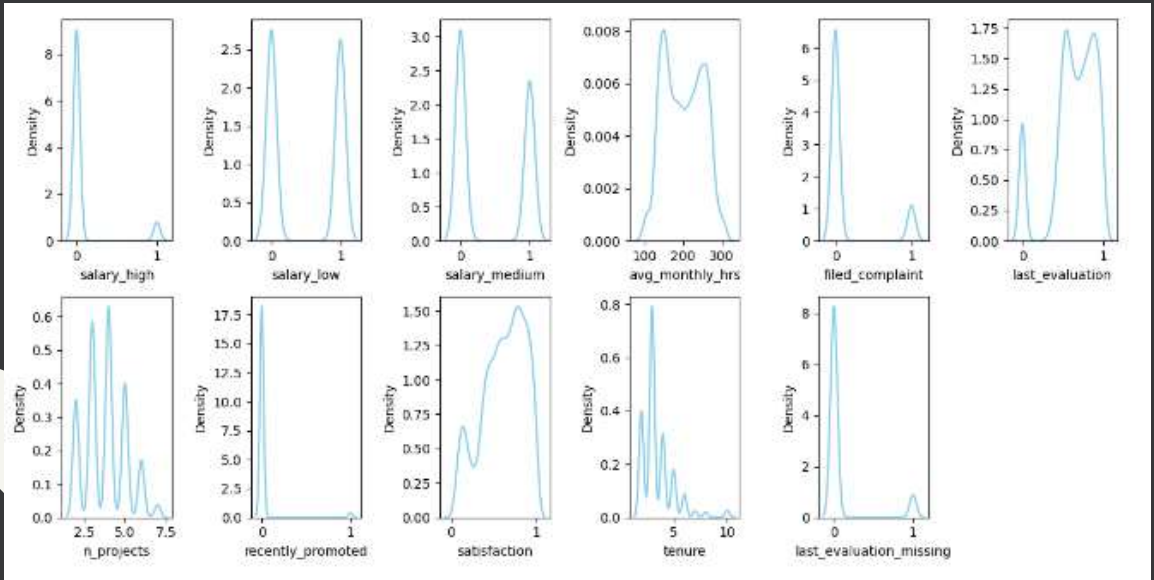
EXPLORATORY DATA ANALYSIS

Univariate Analysis



Dari Boxplot terlihat adanya outlier di salary_high,last_evaluation, recently_promoted,tenure dan last_evaluation_missing, namun bisa diabaikan karena tidak terlalu berpengaruh.

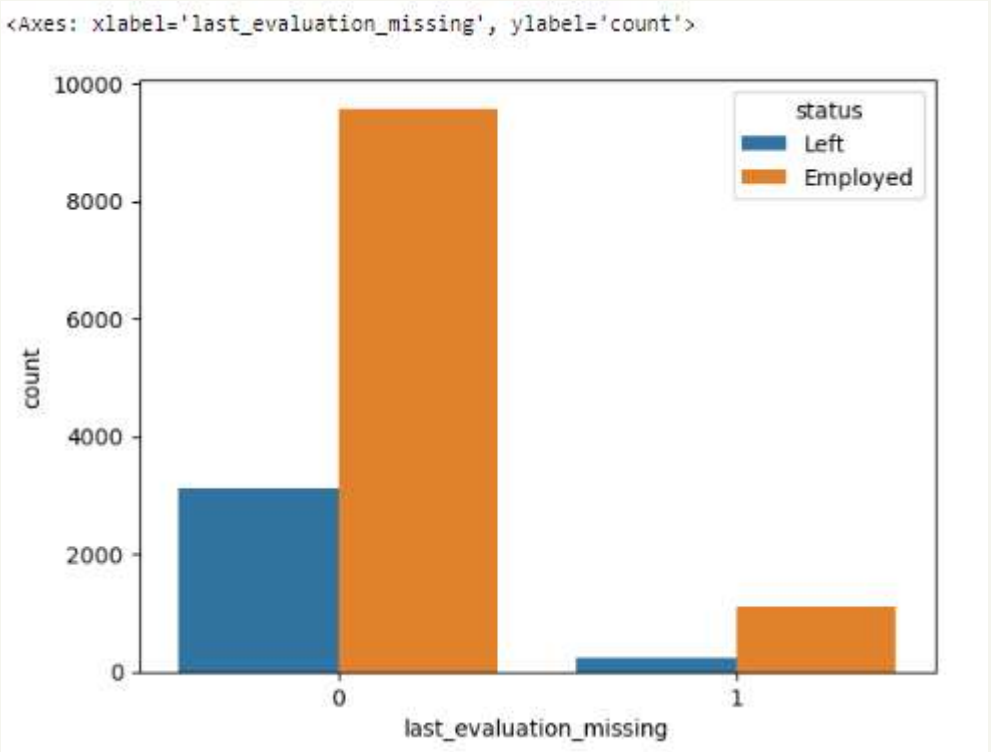
KDE plot for knowing the distribution form



1. avg_monthly_hrs sebarannya adalah normal skew
2. filed_complaint, sebaran distribusi '0' (tdk komplain) sangat tinggi dibanding distribusi'1'(komplain)
3. last_evaluation, sebaran distribusi ke arah skew negatif, dengan ekor /melandai ke kiri
4. n_project, sebaran distribusi ke arah skew positif, dengan ekor /melandai ke kanan
5. recently_promoted,sebaran distribusi '0' (tdk promosi) sangat tinggi dibanding distribusi'1'(promosi)
6. satisfaction,sebarannya adalah negatif skew
7. tenure,sebaran distribusi ke arah skew positif, dengan ekor /melandai ke kanan
8. last_evaluation_missing,sebarannya adalah normal skew
9. salary_high, sebarannya adalah normal skew
10. salary_low, sebarannya adalah normal skew
11. salary_medium, sebarannya adalah normal skew

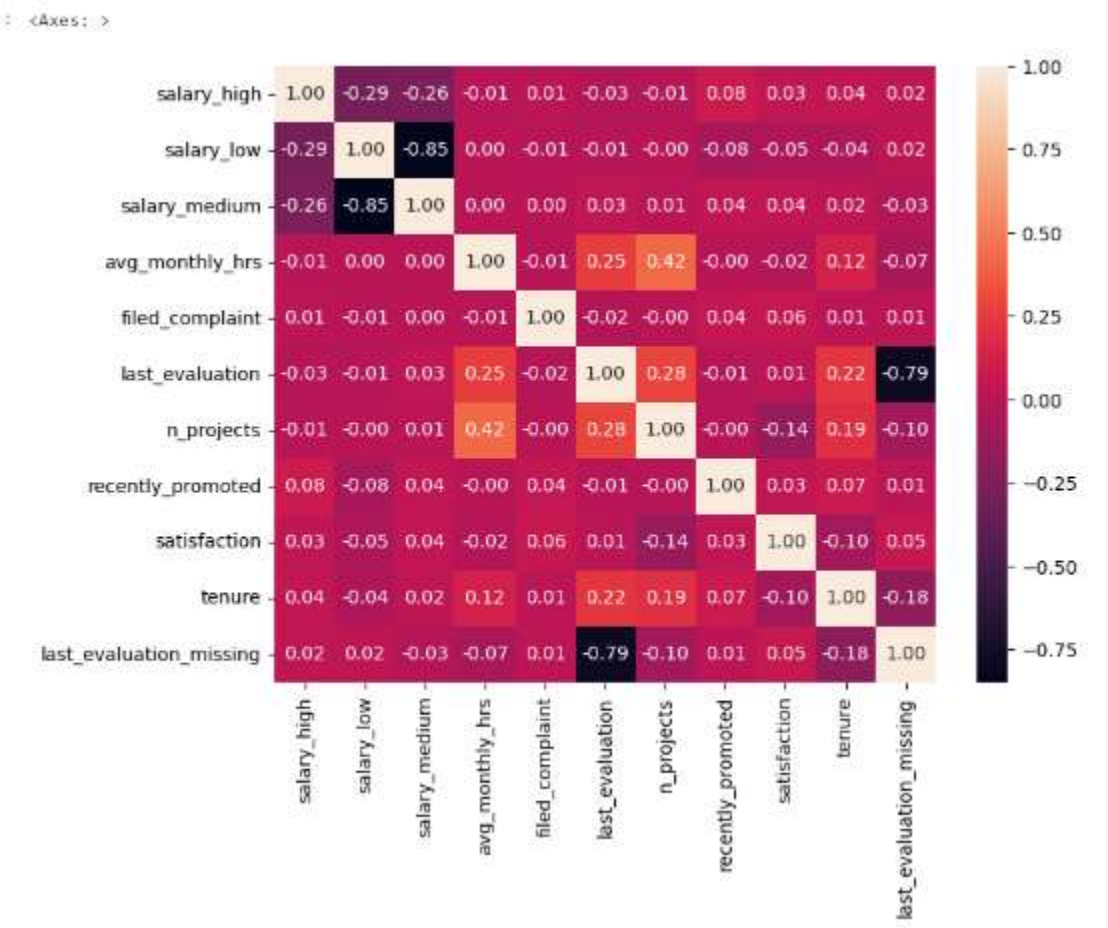
Bivariate Analysis

status = 1 berarti karyawan keluar (churn), dan status = 0 berarti sebaliknya Dapat dikatakan bahwa pada last_evaluation_missing yang banyak jumlahnya adalah yang 0 (tidak churn).



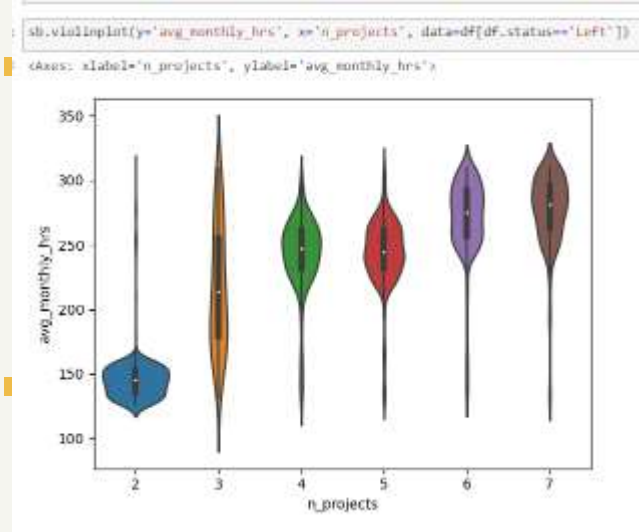
HEATMAP

Dari heatmap diatas dapat dilihat bahwa warna terang menunjukkan nilai - nilai yang berkolerasi dari satu kolom ke kolom lainnya. dengan kata lain kolom 'avg_monthly_hrs','filed_complaint','last_evaluation', 'n_projects','recently_promoted', 'satisfaction', 'status','tenure','last_evaluation_missing' tidak terlalu memiliki korelasi yang kuat karena < 0.5.



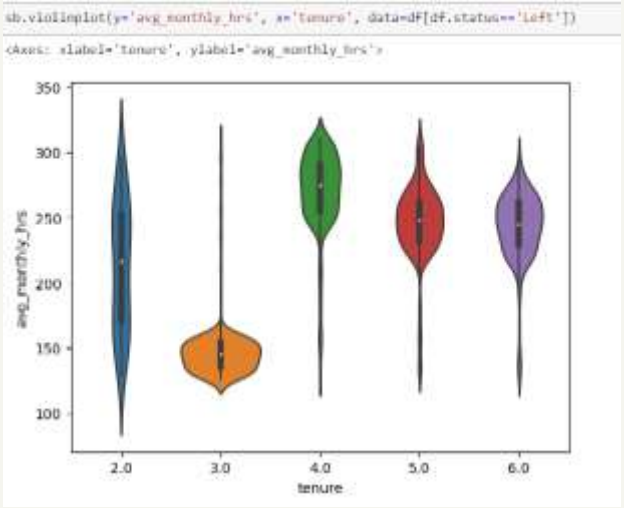
Violinplot

Distribusi 'avg_monthly_hrs' untuk setiap nilai 'n_projects' di status karyawan yang churn /Left.



Terlihat pada project 2 :skewnya positif, project 3 : skew normal, project 4,5,6,7 : skew negatif. rentang jam kerja terbanyak saat project ke 3

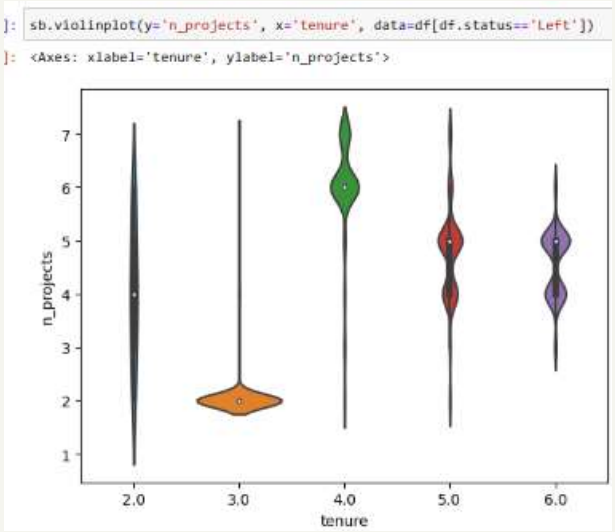
Distribusi 'avg_monthly_hrs' untuk setiap nilai 'tenure' di status karyawan yang churn /Left



Terlihat rata-rata tenure,skewnya normal hanya dimasa kerja 2 tahun. sedang di tahun ke 3 skewnya positif, ditahun ke 4, 5 dan 6 tahun masa kerja skewnya negatif.

Distribusi 'n_projects' untuk setiap nilai 'tenure' di status karyawan yang churn /Left.

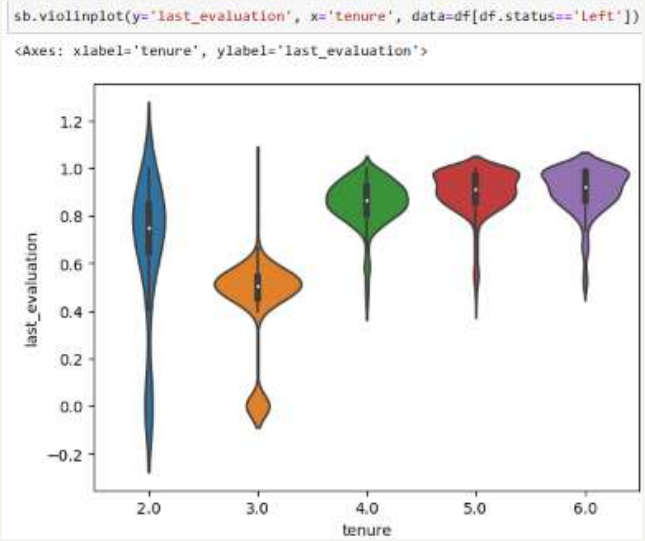
```
]: sb.violinplot(y='n_projects', x='tenure', data=df[df.status=='Left'])
]: <Axes: xlabel='tenure', ylabel='n_projects'>
```



Terlihat ditenure 3,skewnya positif dan di tenure 2,5,6 tahun skewnya normal dan di tenure 4 skewnya negatif.

Distribusi 'last_evaluation' untuk setiap nilai 'tenure' di status karyawan yang churn /Left

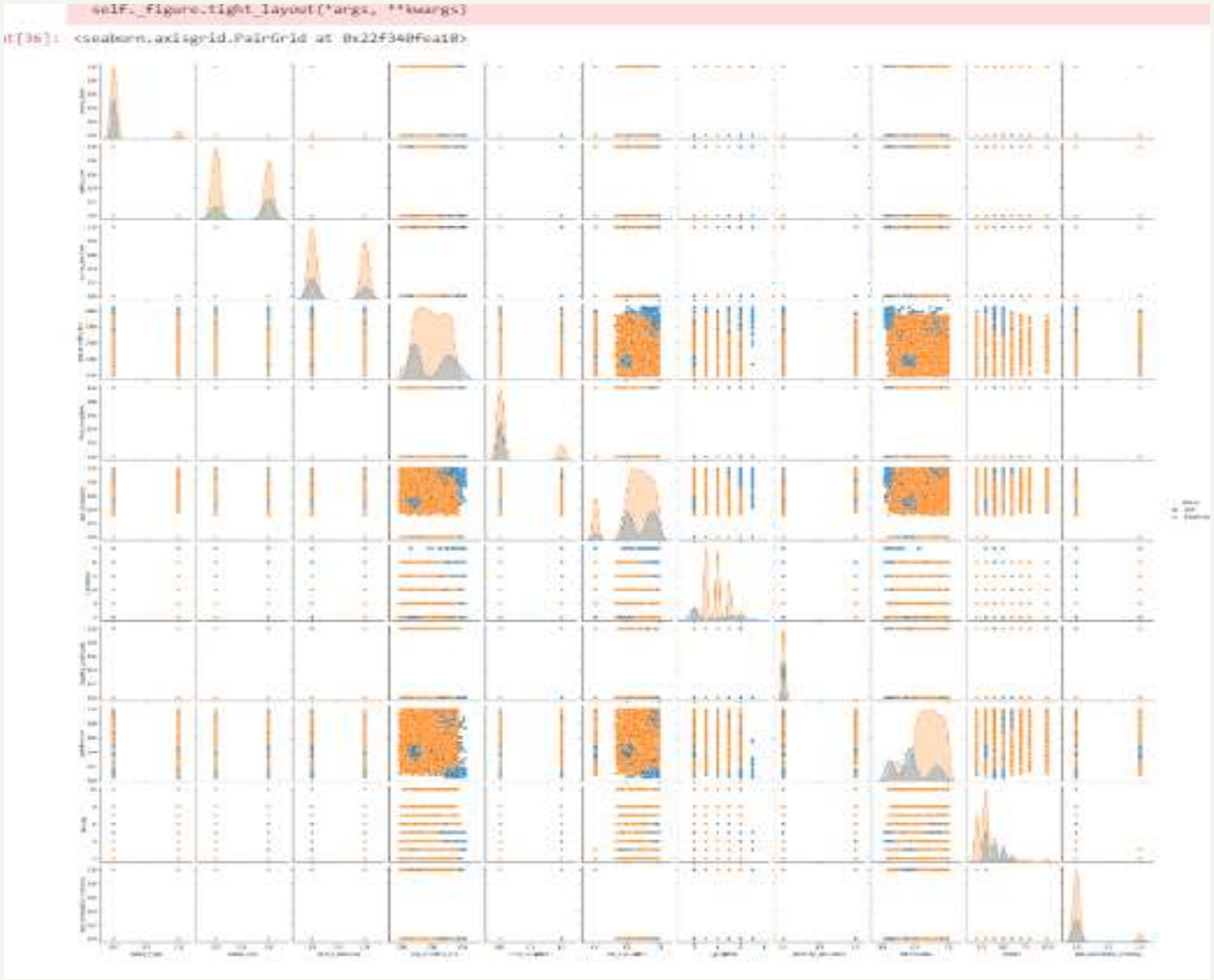
```
sb.violinplot(y='last_evaluation', x='tenure', data=df[df.status=='Left'])
<Axes: xlabel='tenure', ylabel='last_evaluation'>
```



Terlihat rata-rata tenure,skewnya negatif.

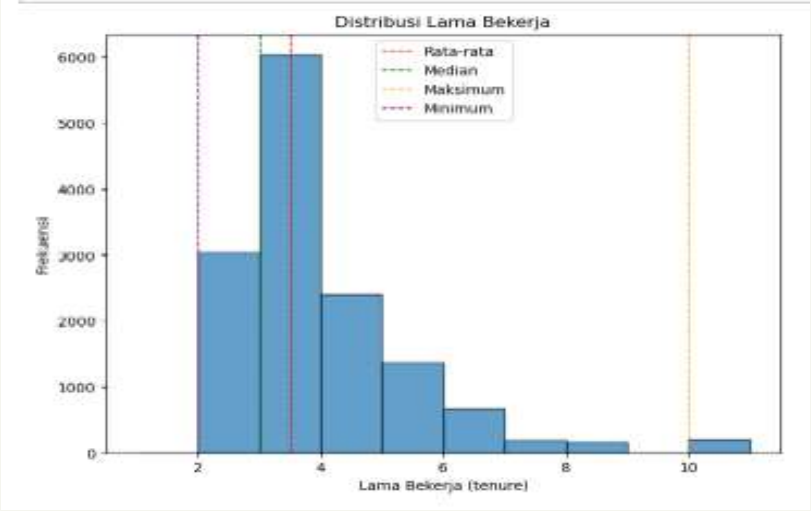
Pairplot

Dari hasil analisa terdapat kesimpulan : Karyawan yang churn rata-rata salarynya rendah, rata-rata jam kerjanya diatas 200 jam/bulan, rata-rata tidak puas pada perusahaan, rata-rata masa kerjanya diatas 3 tahun.



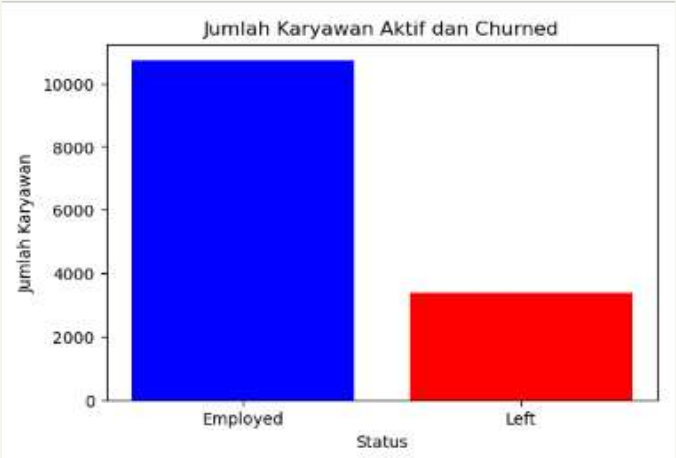
Deep Dive EDA

Berapa rata-rata, median, maksimum, minimum lama bekerja dan Bagaimana distribusi lama bekerja karyawan?



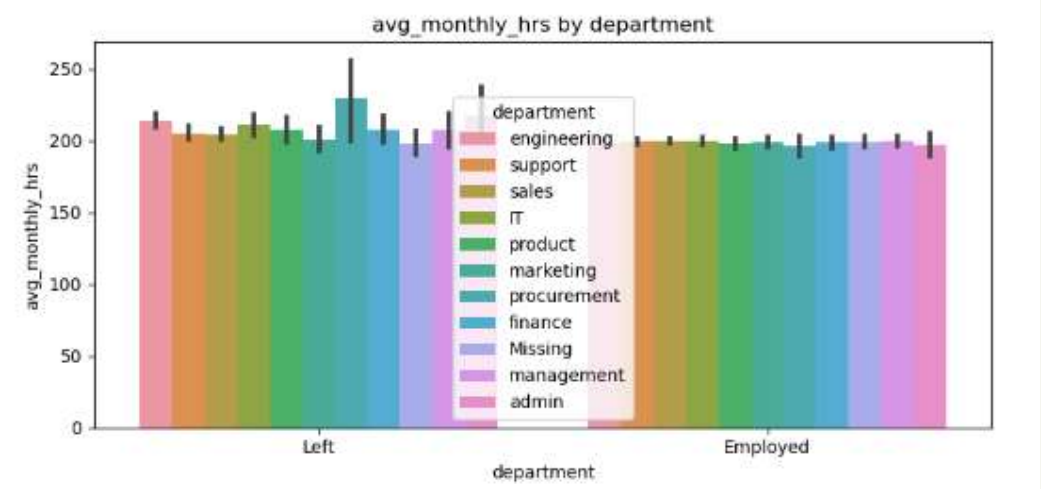
Karyawan terbanyak bekerja selama 3 tahun sebanyak sekitar 6000 orang.

Berapa jumlah karyawan yang masih aktif dan berhenti (churned)?



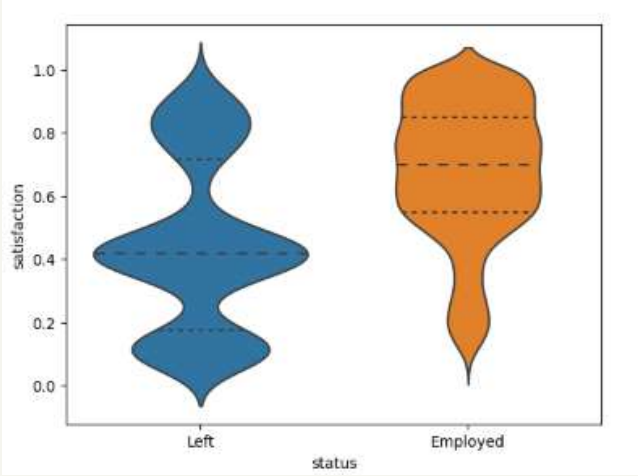
Dari barplot terlihat bahwa jumlah karyawan yang aktif sekitar 10,000 orang dan karyawan yang churn sekitar 3000 orang

Bagaimana distribusi jam kerja bulanan untuk karyawan yang churn/left dan employed berdasarkan departemennya?



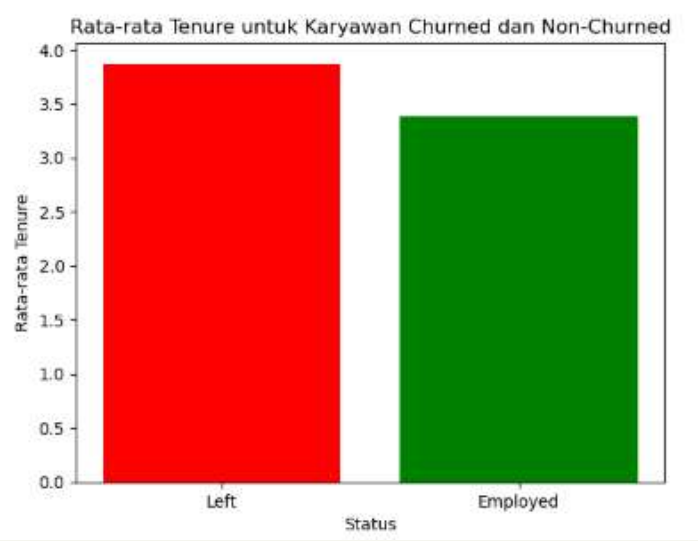
Berdasar Barplot diatas terlihat jam kerja karyawan yang churn/left lebih lama dibanding karyawan yang masih aktif yaitu diatas 200 jam/bulan

Bagaimana tingkat kepuasan karyawan yang left/churn dan yang employed?



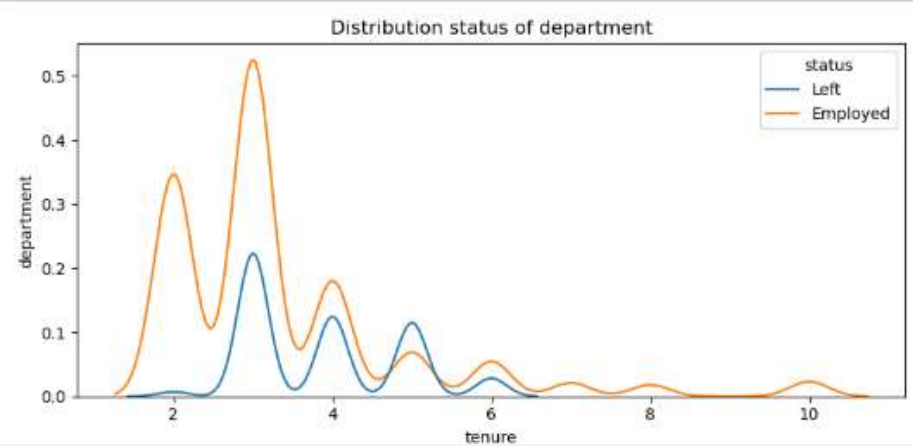
terlihat bahwa tingkat kepuasan karyawan yang churn/left lebih rendah dari yang employed yaitu rata-rata 40%

mencari rata-rata tenure dari employee yang berhenti(churn) dan dan tidak churn



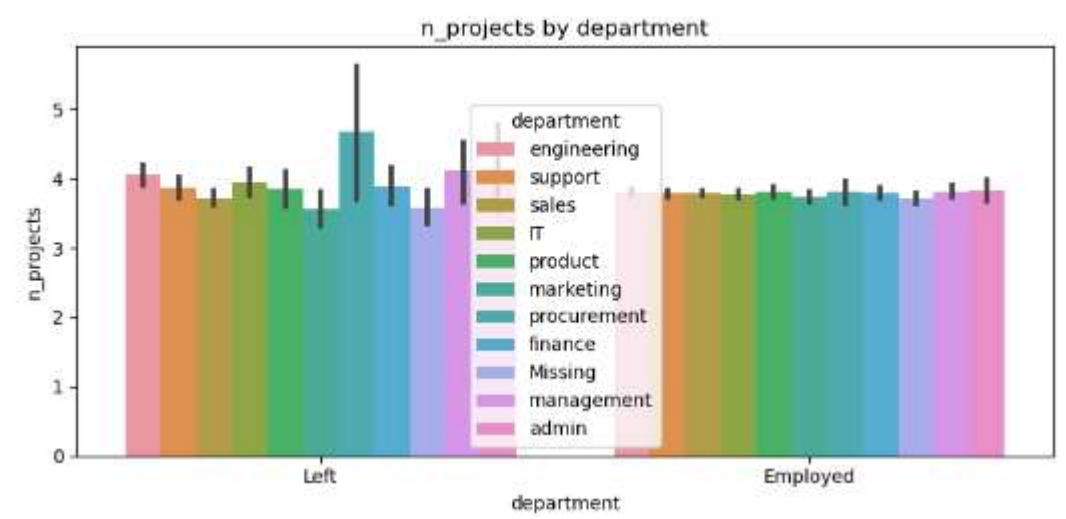
Rata-rata tenure (masa kerja) untuk yang churn lebih lama dari yang employed

Bagaimana distribusi karyawan yang churn/left dan employed berdasarkan tenure di departemen?



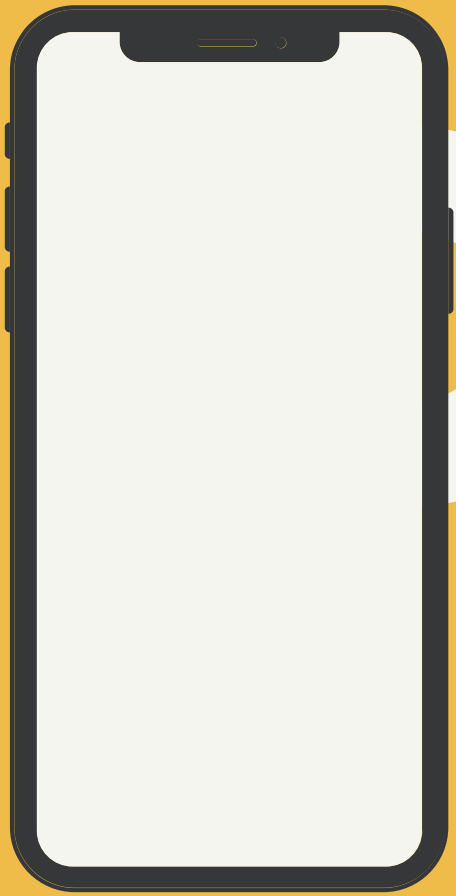
terlihat bahwa karyawan yang churn terbanyak di masa kerja 3 tahun keatas

Bagaimana distribusi karyawan yang churn/left dan employed di departemen berdasarkan proyek?



Dari plot diatas,berdasarkan proyek terlihat bahwa dari divisi Procurement/ manajemen jumlah yang churn adalah yang terbanyak

Machine Learning



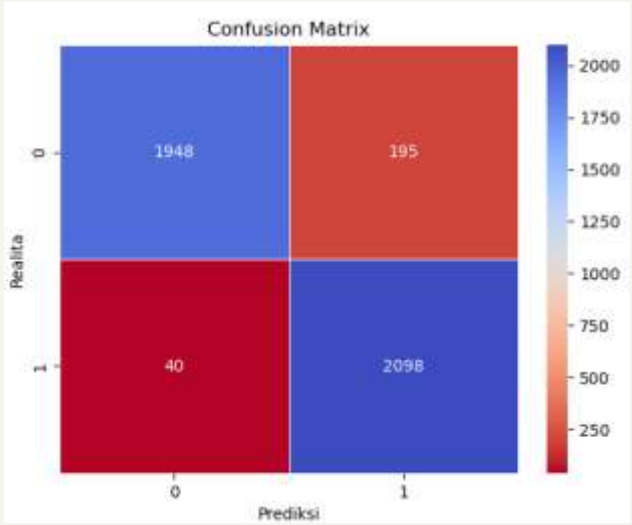
Model Comparison - Classifier

	Algoritm	ROC_AUC_Mean	ROC_AUC_Std	Recall_Mean	Recall_Std	Precision_Mean	Precision_Std	F1_score_Mean	F1_score_Std
1	Random Forest	97.47	0.10	96.46	0.58	90.51	0.33	93.31	0.31
3	KNN	93.31	0.29	92.53	0.14	82.68	0.40	87.32	0.21
4	Decision Tree Classifier	91.59	0.25	95.13	0.28	88.79	0.09	91.83	0.24
2	SVM	88.99	0.40	85.35	0.84	78.75	0.32	81.92	0.56
0	Logistic Regression	76.70	0.60	68.88	0.07	71.04	0.89	69.94	0.40
5	Gausian NB	75.56	0.62	82.16	0.42	61.13	0.31	70.10	0.23

Random Forest: Algoritma Random Forest memiliki kinerja yang paling tinggi di antara semua algoritma yang dievaluasi. Ini ditunjukkan dengan nilai ROC AUC_Mean sebesar 97.44, yang menandakan bahwa model ini memiliki kemampuan yang sangat baik dalam membedakan antara kelas positif dan negatif. Selain itu, F1-score_Mean yang mencapai 93.31 menunjukkan bahwa model ini mencapai keseimbangan yang baik antara recall dan precision.

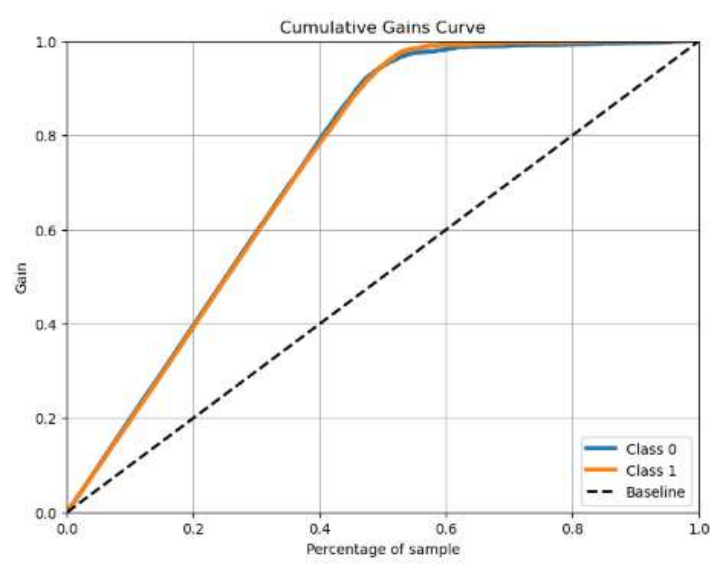
Evaluation For Data Test

	precision	recall	f1-score	support
0	0.98	0.91	0.94	2143
1	0.91	0.98	0.95	2138
accuracy			0.95	4281
macro avg	0.95	0.95	0.95	4281
weighted avg	0.95	0.95	0.95	4281

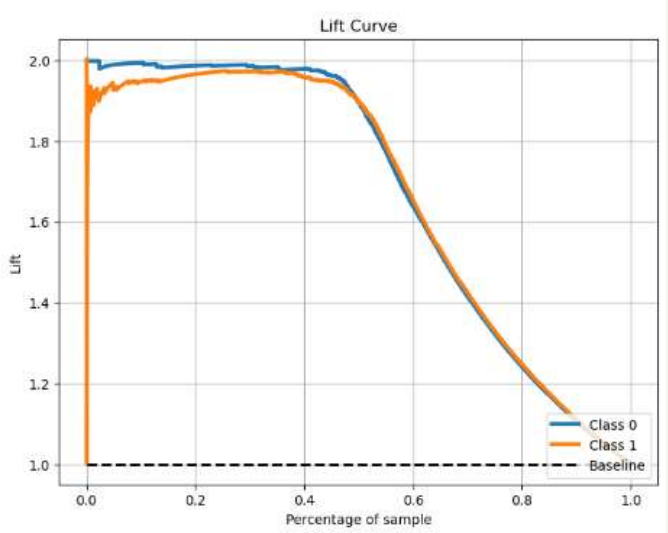


True Positive (TP) adalah 1943, yang merupakan jumlah karyawan yang diprediksi aktif dan kenyataannya masih aktif. True Negative (TN) adalah 2100, yang merupakan jumlah karyawan yang diprediksi berhenti (churn) dan kenyataannya churn . False Positive (FP) adalah 200, yang merupakan jumlah karyawan yang diprediksi aktif, tetapi sebenarnya tidak aktif (churn). False Negative (FN) adalah 38, yang merupakan jumlah karyawan yang diprediksi tidak aktif (churn),kenyataannya aktif.

Lift Curve



Berdasarkan analisis cumulative gains, model menunjukkan kinerja yang baik karena berhasil mengidentifikasi karyawan sebesar 98% yang akan churn (berhenti) saat memfokuskan pada 20% dari populasi dengan probabilitas tertinggi.



Grafik ini menunjukkan efektivitas model prediksi dalam membedakan antara karyawan yang aktif (warna biru) dan karyawan yang churn (warna oranye). Lift adalah perbandingan antara akurasi model dengan baseline (random chance). Jika lift > 1, model lebih baik daripada baseline. Pada awal persentase sampel, kurva biru memiliki lift yang jauh lebih tinggi dibandingkan dengan kurva oranye dan baseline. Namun, seiring bertambahnya persentase sampel, kedua kurva mendekati satu sama lain dan baseline, menunjukkan penurunan efektivitas model.
Semua ini memberikan wawasan tentang bagaimana model memprediksi peluang karyawan uyang berhenti (churn).

INSIGHT

Dari DEEP DIVE EDA diketahui bahwa :

1. Jumlah keseluruhan karyawan adalah 14068 dengan komposisi karyawan yang loyal 10701 orang dan yang keluar 3367 orang.

2. Rata-rata karyawan yang churn masa kerjanya (tenure) diatas 3 tahun.

3. Rata-rata tingkat kepuasan karyawan yang churn 44% dan jam kerjanya 207 jam/bulan

4. karyawan yang churn terbanyak dari divisi sales 3646 orang, selanjutnya divisi engineering 2615 orang dan divisi IT 1158 orang

5. Karyawan yang churn terbanyak dari yang salarynya rendah (salary low)

MACHINE LEARNING

Model terbaik dari hasil permodelan yang dianalisa adalah Random Forest dengan accuracy 0.95, ROC 97.47 , F-1 Score 0.95, Recall 0.98 dan Precision 0.91.

Lift Curve

cumulative gains, model menunjukkan kinerja yang baik mengidentifikasi karyawan sebesar 98% yang akan churn (berhenti) pada 20% dari populasi.

REKOMENDASI

1. Sebaiknya jam kerja karyawan tidak lebih dari 200 jam/bulan
2. Untuk karyawan dengan masa kerja diatas 3 tahun diberikan insentif diluar gaji.
3. Untuk sales, diberikan award setiap pencapaian setiap bulan
4. Untuk divisi IT dan engineering diberikan pengembangan karir (pendidikan lanjutan/beasiswa)
5. Pengembangan karir untuk semua divisi mulai tahun ke-2
6. Meningkatkan keterlibatan karyawan dalam program-program yang diadakan perusahaan.
7. Memberikan kepemilikan saham kepada karyawan yang sudah bekerja diatas 2 tahun, sesuai dengan perhitungan managemen perusahaan.

Thank you!