

CoDros Resign Prediction by **ARINI**

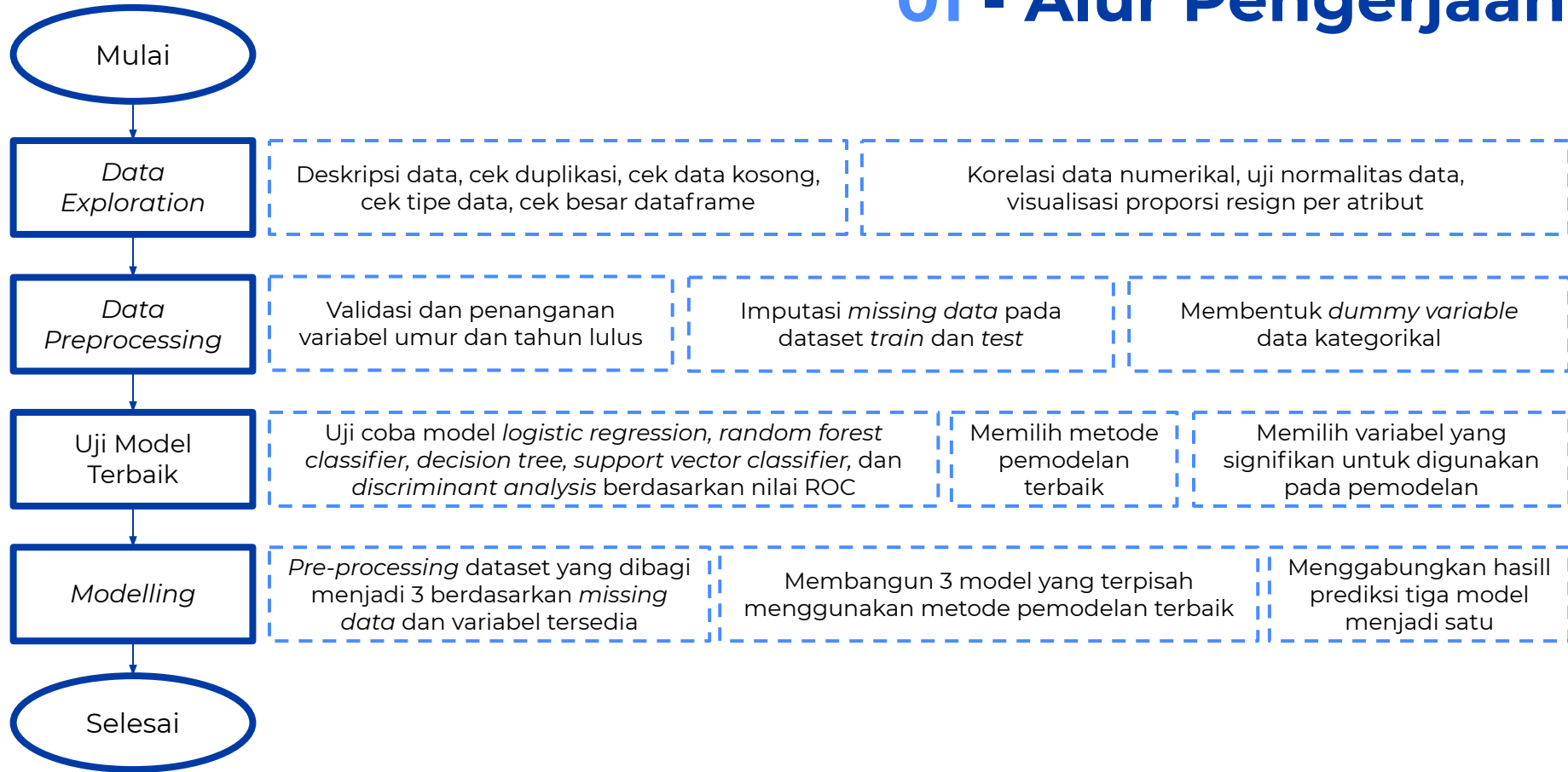
The Team

Amalia Dini Putri Prady
Mochamad **Rizal** Prasetyo
Ghani Murtafa Amal Alaudin

Table of Contents

- 01 Alur Pengerjaan**
How we solve the problem
- 02 Data Exploration**
Understanding the current data situation
- 03 Data Preprocessing**
Preparing the data for further process
- 04 Data Modelling**
Optimizing decision making process

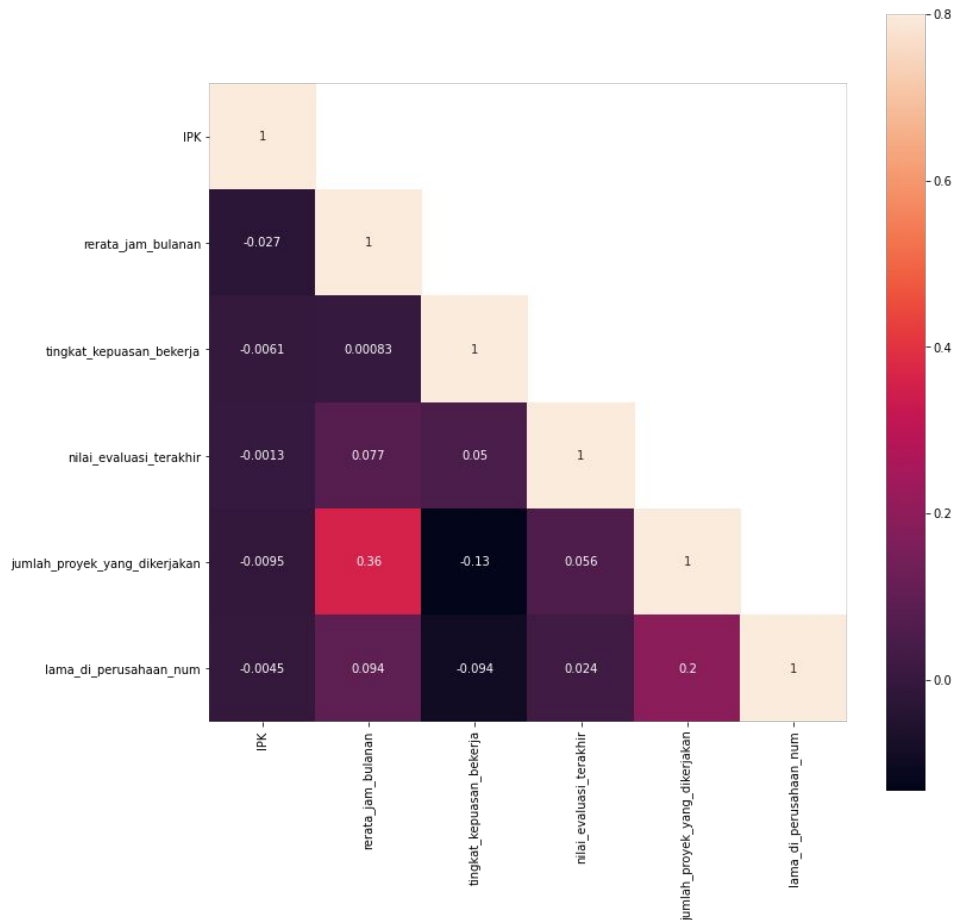
01 - Alur Pengerjaan



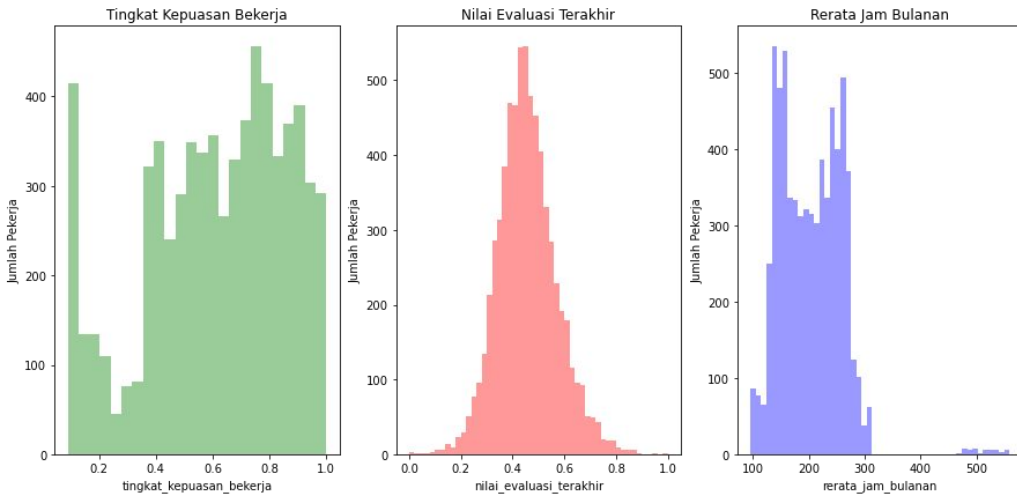
02 - Data Exploration

Uji Korelasi Data Numerik

Tidak terdapat multikolinearitas
pada dataset *train*.
(cut-off > 0.8)



1. IPK
2. rerata_jam_bulanan
3. tingkat_kepuasan_bekerja
4. Nilai_evaluasi_terakhir
5. Jumlah_proyek_yang_dikerjakan
6. lama_di_perusahaan_num

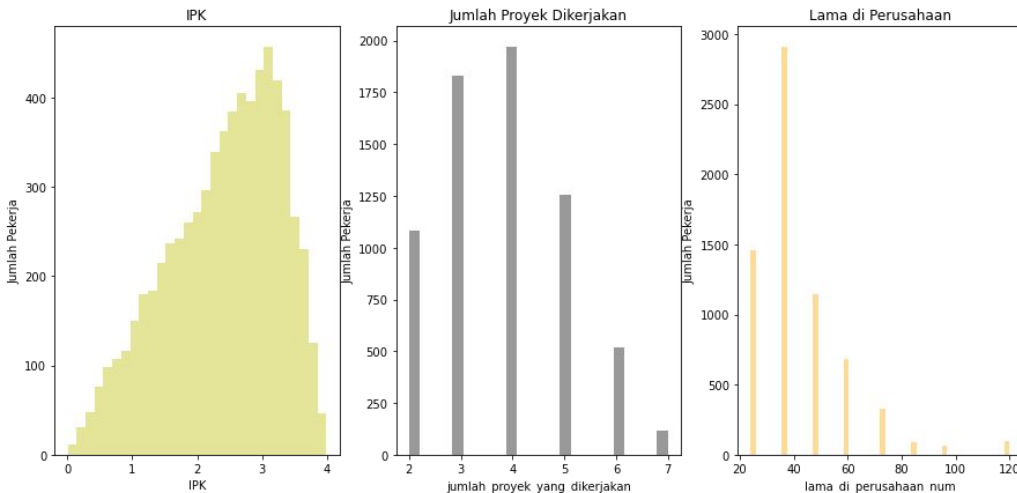


Uji Normalitas

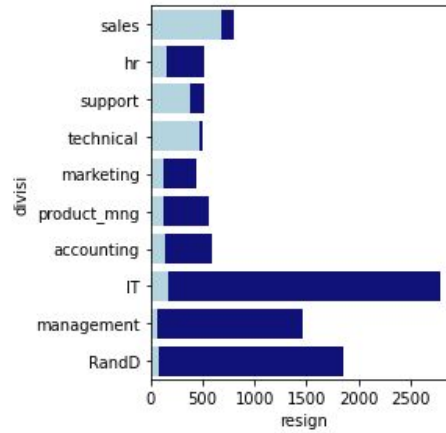
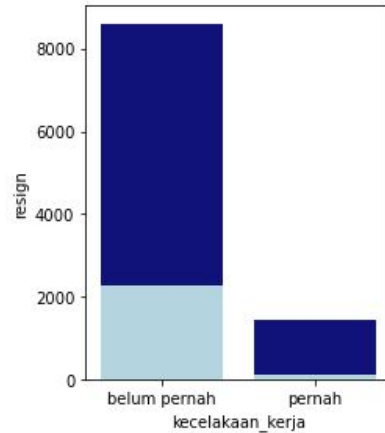
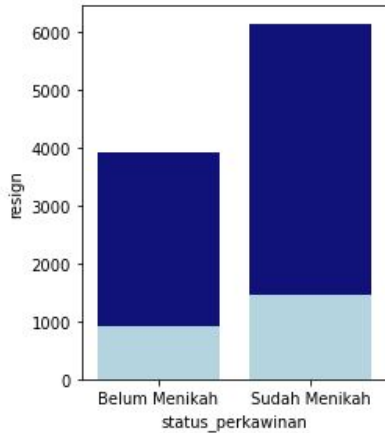
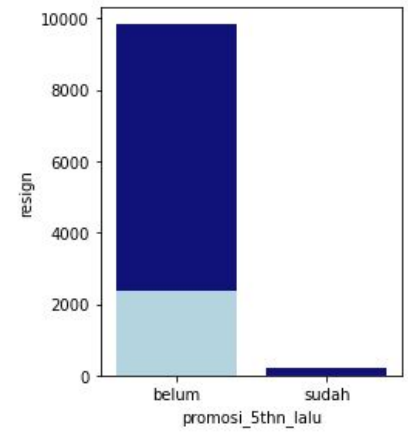
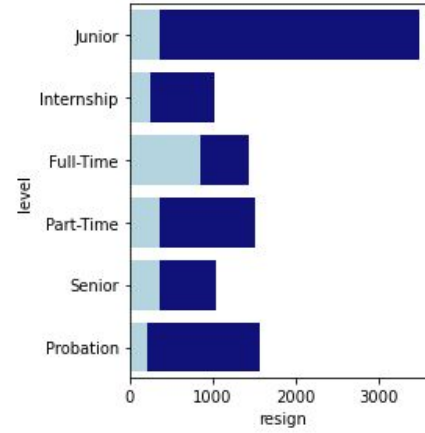
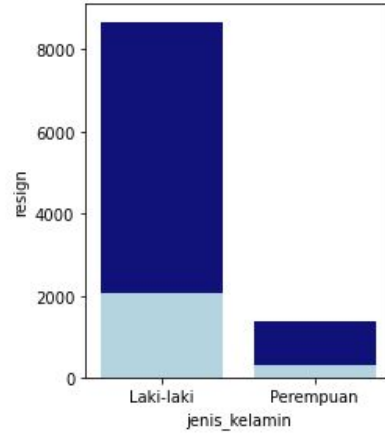
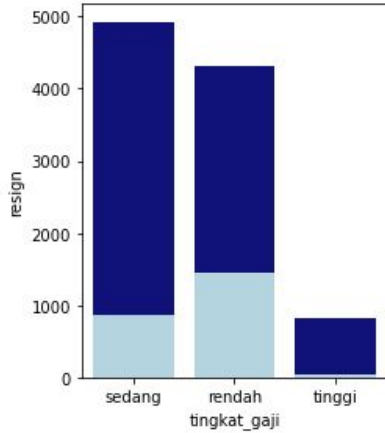
Data *train* untuk kategori numerik berdistribusi normal.

(untuk mengecek perubahan karakteristik setelah imputasi, dasar asumsi model statistik)

$pvalue < 0.05 \rightarrow$ Berdistribusi Normal*
 *(data setelah menghilangkan baris mengandung null)



Data Numerik	<i>pvalue</i>
IPK	4.804e-97
rerata_jam_bulanan	0
nilai_evaluasi_terakhir	2.209e-122
tingkat_kepuasan_bekerja	9.06e-38
jumlah_proyek_yang_dikerjakan	2.402e-53
lama_di_perusahaan_num	0



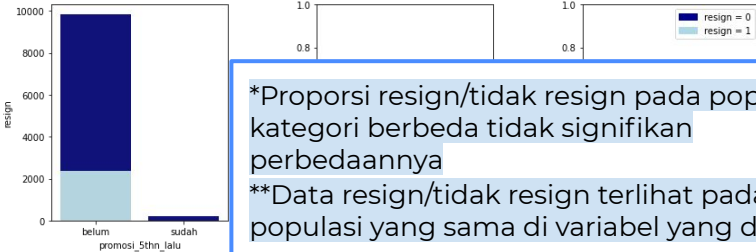
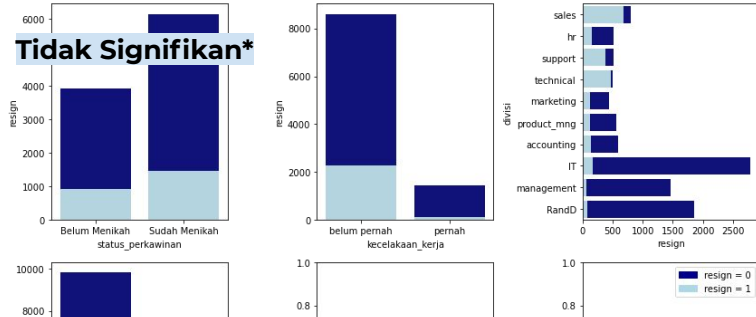
Karyawan resign didominasi

1. Tingkat gaji rendah
2. Laki - laki
3. Level full-time
4. Sudah menikah
5. Belum pernah mengalami kecelakaan kerja
6. Divisi sales
7. Belum mendapat promosi dalam 5 tahun terakhir

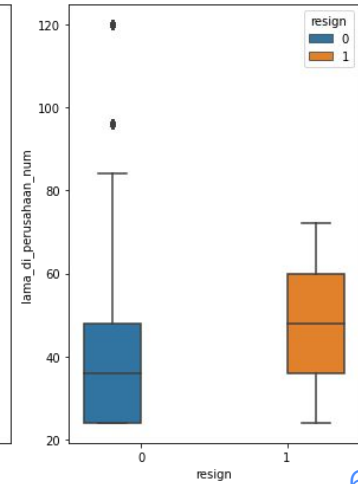
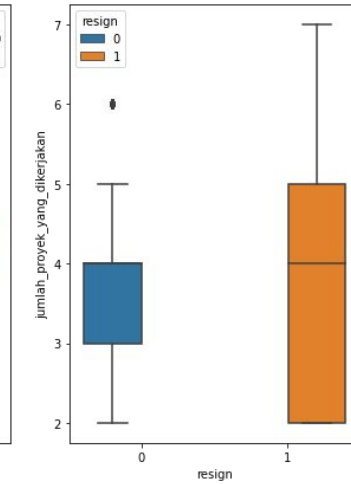
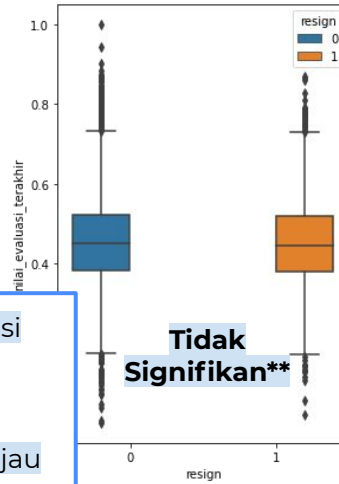
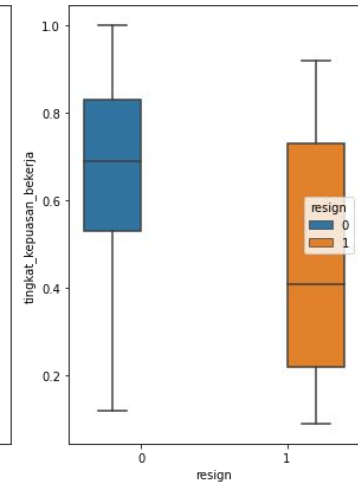
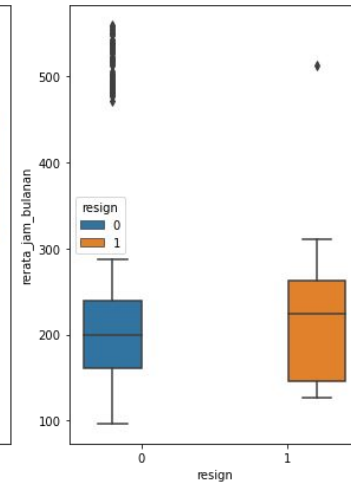
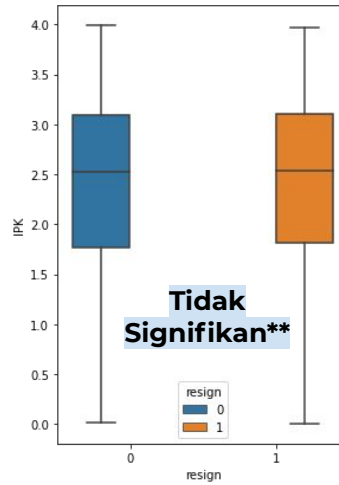


Evaluasi Signifikansi Variabel

Menilai variabel berdasarkan kemampuan membedakan keputusan resign 0 dan 1



*Proporsi resign/tidak resign pada populasi kategori berbeda tidak signifikan perbedaannya
 **Data resign/tidak resign terlihat pada populasi yang sama di variabel yang ditinjau



Validasi variabel umur dan tahun lulus → terlihat banyak data yang tidak masuk akal

03 - Data Preprocessing

```
print(df2.loc[(df2['lama_di_perusahaan_num'] == 120) & (df2['tahun_lulus'] > 2010)])
```

	tahun_lulus	lama_di_perusahaan_num
1000	2011.0	120.0
1185	2019.0	120.0
1322	2018.0	120.0
2158	2012.0	120.0
2304	2020.0	120.0
3242	2014.0	120.0
3950	2019.0	120.0
4643	2014.0	120.0
5008	2020.0	120.0
5500	2016.0	120.0
6116	2016.0	120.0
6427	2016.0	120.0
6808	2018.0	120.0
7102	2017.0	120.0
7111	2015.0	120.0
8456	2012.0	120.0
8912	2016.0	120.0
9139	2018.0	120.0
9227	2016.0	120.0
9509	2019.0	120.0
10015	2015.0	120.0

Lulusan 2011 namun lama kerja di perusahaan 120 bulan (10 tahun)

Tidak memungkinkan

```
print(df1.loc[(df1['lama_di_perusahaan_num'] == 120) & (df1['umur'] < 30)])
```

	umur	lama_di_perusahaan_num
1185	23.0	120.0
1322	22.0	120.0
2304	21.0	120.0
3242	28.0	120.0
3950	24.0	120.0
4643	29.0	120.0
5008	20.0	120.0
5500	26.0	120.0
6116	25.0	120.0
6427	26.0	120.0
6808	24.0	120.0
7102	26.0	120.0
7111	26.0	120.0
8912	25.0	120.0
9139	24.0	120.0
9227	25.0	120.0
9509	22.0	120.0
10015	26.0	120.0

Lulusan umur kurang dari 30 namun sudah bekerja 120 bulan (10 tahun)

Tidak memungkinkan
(kerja sejak di bawah umur)

	umur	tahun_lulus	lama_di_perusahaan_num
--	------	-------------	------------------------

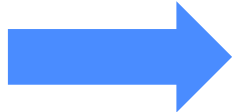
umur	1.000000	-0.993589	-0.013734
tahun_lulus	-0.993589	1.000000	0.015372
lama_di_perusahaan_num	-0.013734	0.015372	1.000000

umur dan tahun lulus berkorelasi tinggi (0,99), mengindikasikan informasi yang tersedia sama

Keputusan:
kolom tahun lulus dan umur dihapus

```
train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10049 entries, 0 to 10048
Data columns (total 17 columns):
#   Column                Non-Null Count  Dtype
---  -
0   employee_id            10049 non-null  int64
1   umur                   2338 non-null   float64
2   jenis_kelamin          10049 non-null  object
3   IPK                    10049 non-null  float64
4   level                  10049 non-null  object
5   tahun_lulus            2338 non-null   float64
6   status_perkawinan      10049 non-null  object
7   divisi                 10049 non-null  object
8   rerata_jam_bulanan     8700 non-null   float64
9   tingkat_kepuasan_bekerja 9253 non-null   float64
10  nilai_evaluasi_terakhir 8499 non-null   float64
11  jumlah_proyek_yang_dikerjakan 10049 non-null  int64
12  lama_di_perusahaan     10049 non-null  object
13  kecelakaan_kerja       10049 non-null  object
14  promosi_5thn_lalu      10049 non-null  object
15  tingkat_gaji           10049 non-null  object
16  resign                 10049 non-null  int64
dtypes: float64(6), int64(3), object(8)
memory usage: 1.3+ MB
```



Tindakan yang dilakukan

1. Menghapus kolom umur
2. Menghapus kolom tahun_lulus
3. Mengubah tipe lama_di_perusahaan menjadi float64 (kolom lama_di_perusahaan_num)

```
test.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4950 entries, 0 to 4949
Data columns (total 16 columns):
#   Column                Non-Null Count  Dtype
---  -
0   employee_id            4950 non-null  int64
1   umur                   1070 non-null   float64
2   jenis_kelamin          4950 non-null  object
3   IPK                    4950 non-null  float64
4   level                  4950 non-null  object
5   tahun_lulus            1151 non-null   float64
6   status_perkawinan      4950 non-null  object
7   divisi                 4950 non-null  object
8   rerata_jam_bulanan     4362 non-null   float64
9   tingkat_kepuasan_bekerja 4484 non-null   float64
10  nilai_evaluasi_terakhir 4094 non-null   float64
11  jumlah_proyek_yang_dikerjakan 4950 non-null  int64
12  lama_di_perusahaan     4950 non-null  object
13  kecelakaan_kerja       4950 non-null  object
14  promosi_5thn_lalu      4950 non-null  object
15  tingkat_gaji           4950 non-null  object
dtypes: float64(6), int64(2), object(8)
memory usage: 618.9+ KB
```

```
trainrev.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10049 entries, 0 to 10048
Data columns (total 15 columns):
#   Column                Non-Null Count  Dtype
---  -
0   employee_id            10049 non-null  int64
1   jenis_kelamin          10049 non-null  object
2   IPK                    10049 non-null  float64
3   level                  10049 non-null  object
4   status_perkawinan      10049 non-null  object
5   divisi                 10049 non-null  object
6   rerata_jam_bulanan     10049 non-null  float64
7   tingkat_kepuasan_bekerja 10049 non-null  float64
8   nilai_evaluasi_terakhir 10049 non-null  float64
9   jumlah_proyek_yang_dikerjakan 10049 non-null  int64
10  kecelakaan_kerja       10049 non-null  object
11  promosi_5thn_lalu      10049 non-null  object
12  tingkat_gaji           10049 non-null  object
13  resign                 10049 non-null  int64
14  lama_di_perusahaan_num 10049 non-null  float64
dtypes: float64(5), int64(3), object(7)
memory usage: 1.2+ MB
```

```
testrev.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4950 entries, 0 to 4949
Data columns (total 14 columns):
#   Column                Non-Null Count  Dtype
---  -
0   employee_id            4950 non-null  int64
1   jenis_kelamin          4950 non-null  object
2   IPK                    4950 non-null  float64
3   level                  4950 non-null  object
4   status_perkawinan      4950 non-null  object
5   divisi                 4950 non-null  object
6   rerata_jam_bulanan     4362 non-null   float64
7   tingkat_kepuasan_bekerja 4484 non-null   float64
8   nilai_evaluasi_terakhir 4094 non-null   float64
9   jumlah_proyek_yang_dikerjakan 4950 non-null  int64
10  kecelakaan_kerja       4950 non-null  object
11  promosi_5thn_lalu      4950 non-null  object
12  tingkat_gaji           4950 non-null  object
13  lama_di_perusahaan_num 4950 non-null   float64
dtypes: float64(5), int64(2), object(7)
memory usage: 541.5+ KB
```



```
testrev.isna().sum()
```

employee_id	0
jenis_kelamin	0
IPK	0
level	0
status_perkawinan	0
divisi	0
rerata_jam_bulanan	588
tingkat_kepuasan_bekerja	466
nilai_evaluasi_terakhir	856
jumlah_proyek_yang_dikerjakan	0
kecelakaan_kerja	0
promosi_5thn_lalu	0
tingkat_gaji	0
lama_di_perusahaan_num	0
dtype: int64	

```
trainrev.isna().sum()
```

employee_id	0
jenis_kelamin	0
IPK	0
level	0
status_perkawinan	0
divisi	0
rerata_jam_bulanan	1349
tingkat_kepuasan_bekerja	796
nilai_evaluasi_terakhir	1550
jumlah_proyek_yang_dikerjakan	0
kecelakaan_kerja	0
promosi_5thn_lalu	0
tingkat_gaji	0
resign	0
lama_di_perusahaan_num	0
dtype: int64	

Metode Imputasi dan Scaling

Metode imputasi menggunakan *mean* pada dataset train dan test. Ditujukan untuk melakukan uji model terbaik dan juga melakukan scaling untuk data numerik supaya memudahkan pemodelan (diantara 0 dan 1).

Dummy Variable

Mengubah data kategorikal menjadi variabel 0 dan 1

```
#Membuat variabel dummy untuk kategorikal - train
catt = ['kecelakaan_kerja', 'tingkat_gaji', 'resign']
numm = ['rerata_jam_bulanan', 'tingkat_kepuasan_bekerja', 'jumlah_proyek_yang_dikerjakan', 'lama_di_perusahaan_num']
categorical_df = pd.get_dummies(modeltrainrev[catt], drop_first=True)
numerical_df = modeltrainrev[numm]

modeldf = pd.concat([categorical_df, numerical_df], axis=1)
modeldf.head()

#Membuat variabel dummy untuk kategorikal - test
catttest = ['kecelakaan_kerja', 'tingkat_gaji']
nummtest = ['rerata_jam_bulanan', 'tingkat_kepuasan_bekerja', 'jumlah_proyek_yang_dikerjakan', 'lama_di_perusahaan_num']
categorical_dftest = pd.get_dummies(modeltestrev[catttest], drop_first=True)
numerical_dftest = modeltestrev[nummtest]

modeldftest = pd.concat([categorical_dftest, numerical_dftest], axis=1)
modeldftest.head()
```

Pemilihan metode *modelling* terbaik

Dataset dipersiapkan untuk menguji 4 metode pemodelan

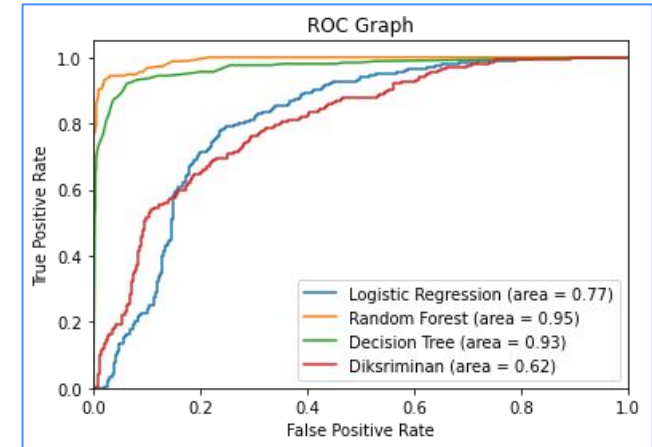
```
target_name = 'resign'  
X = modeldf0.drop('resign', axis = 1)  
y = modeldf0[target_name]  
X_train0, X_val0, y_train0, y_val0 = train_test_split(X,y,test_size=0.2, random_state=42)
```

Pemodelan

---Logistic Model---					---Random Forest Model---				
Logistic AUC = 0.77					Random Forest AUC = 0.95				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.92	0.74	0.82	755	0	0.97	0.99	0.98	755
1	0.51	0.79	0.62	250	1	0.96	0.90	0.93	250
accuracy					accuracy				
macro avg					macro avg				
weighted avg					weighted avg				
	0.71	0.77	0.76	1005		0.97	0.95	0.97	1005
	0.81	0.76	0.77	1005		0.97	0.97	0.97	1005
---Decision Tree Model---					---Diskriminan---				
Decision Tree AUC = 0.93					Diskriminan AUC = 0.62				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.97	0.93	0.95	755	0	0.80	0.93	0.86	755
1	0.81	0.92	0.87	250	1	0.59	0.30	0.40	250
accuracy					accuracy				
macro avg					macro avg				
weighted avg					weighted avg				
	0.89	0.93	0.91	1005		0.70	0.62	0.63	1005
	0.93	0.93	0.93	1005		0.75	0.77	0.75	1005

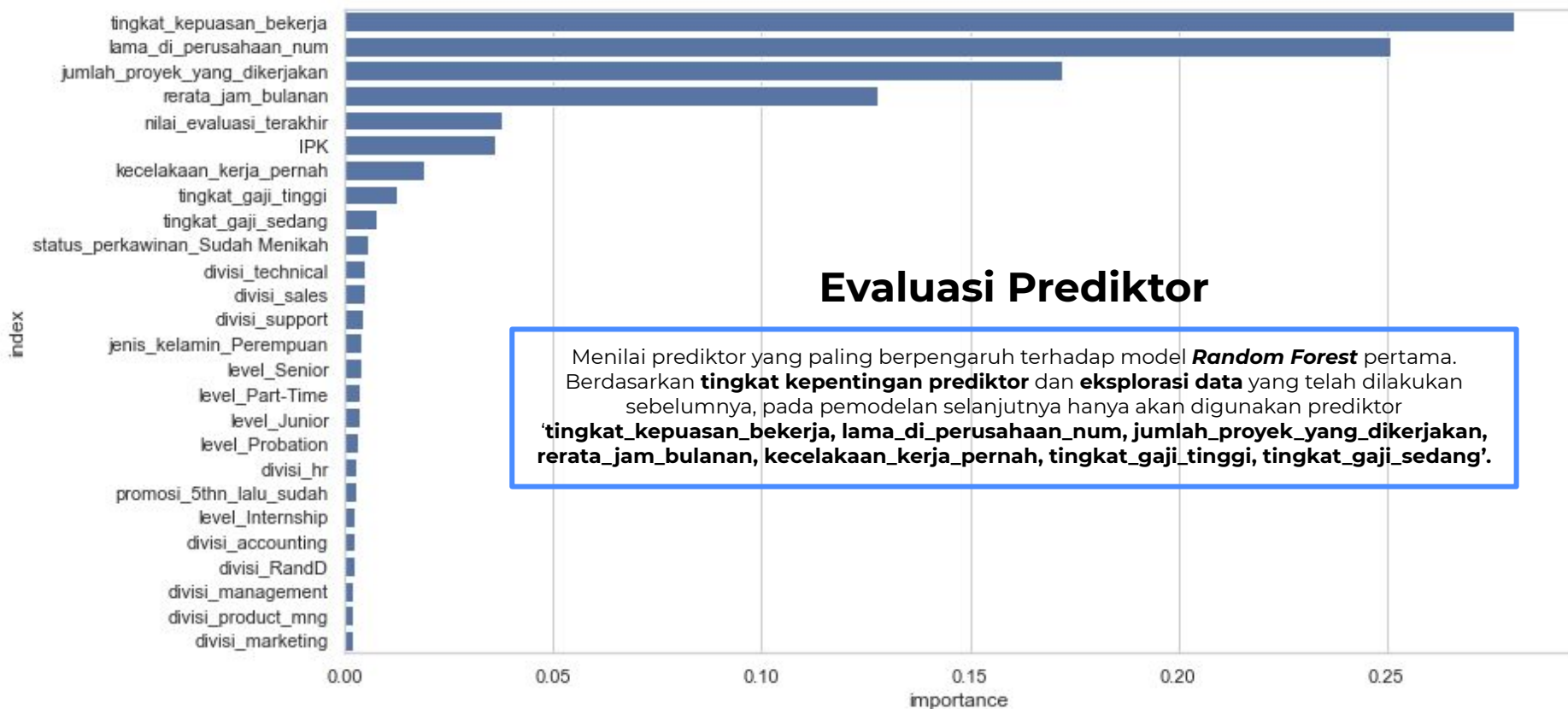
04 - Data Modelling

Evaluasi Model



Keputusan:

Random Forest dipilih sebagai metode pemodelan



Membangun Model Baru

Membangun model baru untuk melakukan pemodelan *resign prediction* dengan membagi dataset *test* menjadi 3 yaitu:

- *No missing value*, data dengan variabel lengkap, row dengan missing value dihapus
- *Missing value RJB* tidak menggunakan variabel RJB, row dengan *missing value* dihapus
- *Missing value TKB* tidak menggunakan variabel TKB, row dengan *missing value* diisi dengan imputasi (mean) untuk RJB

Model dipisahkan berdasarkan *missing value* dengan tidak menggunakan variabel dengan *missing value* agar tidak “merusak” model

*TKB = tingkat_kepuasan_bekerja

RJB = rerata_jam_bulanan

No missing value - tanpa imputasi

#	Column
0	kecelakaan_kerja_pernah
1	tingkat_gaji_sedang
2	tingkat_gaji_tinggi
3	rerata_jam_bulanan
4	tingkat_kepuasan_bekerja
5	jumlah_proyek_yang_dikerjakan
6	lama_di_perusahaan_num

Missing value RJB - tanpa imputasi

#	Column
0	kecelakaan_kerja_pernah
1	tingkat_gaji_sedang
2	tingkat_gaji_tinggi
3	tingkat_kepuasan_bekerja
4	jumlah_proyek_yang_dikerjakan
5	lama_di_perusahaan_num

Missing value TKB - dengan imputasi

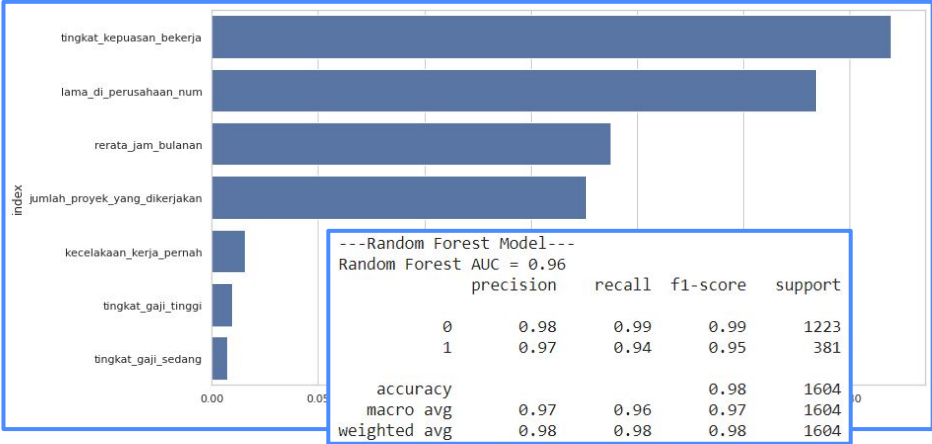
#	Column
0	kecelakaan_kerja_pernah
1	tingkat_gaji_sedang
2	tingkat_gaji_tinggi
3	rerata_jam_bulanan
4	jumlah_proyek_yang_dikerjakan
5	lama_di_perusahaan_num

Evaluasi Model Baru

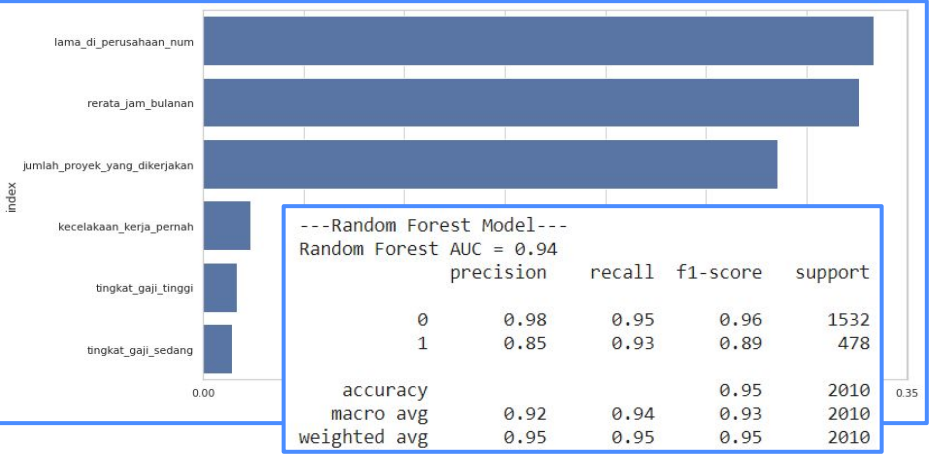
Dengan menggunakan pemodel *random forest*, dengan tiga kondisi model baru yang dibangun didapat masing - masing *classification report*.

*TKB = tingkat_kepuasan_bekerja
RJB = rerata_jam_bulanan

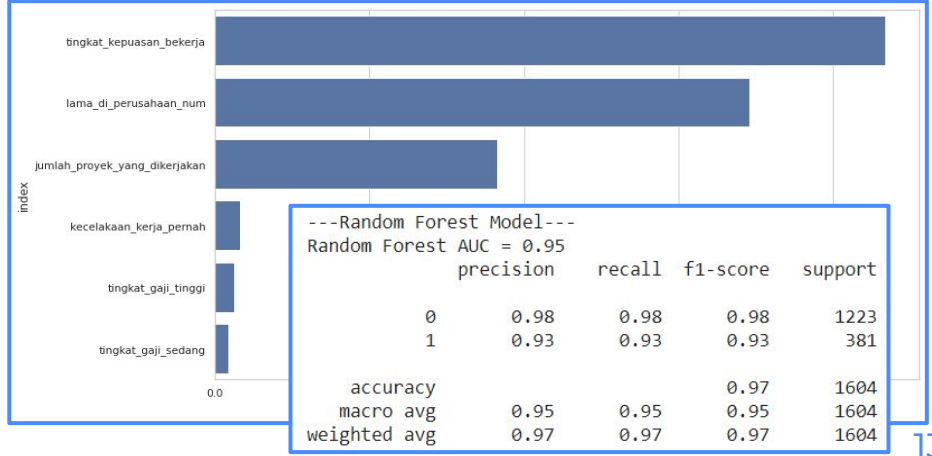
No missing value - tanpa imputasi



Missing value TKB - dengan imputasi



Missing value RJB - tanpa imputasi



Performansi Model

Tiga model yang telah dibangun memiliki **accuracy score** dan **ROC AUC score** sebagai berikut

```
pred = rf.predict(X_val)
print("Accuracy score")
print(accuracy_score(y_val,pred))
print("ROC AUC Score")
print(roc_auc_score(y_val,pred))
```

Accuracy score
0.945771144278607
ROC AUC Score
0.9406769940024251

```
pred2 = rf2.predict(X_val2)
print("Accuracy score")
print(accuracy_score(y_val2,pred2))
print("ROC AUC Score")
print(roc_auc_score(y_val2,pred2))
```

Accuracy score
0.9669576059850374
ROC AUC Score
0.9539373297879874

```
pred3 = rf3.predict(X_val3)
print("Accuracy score")
print(accuracy_score(y_val3,pred3))
print("ROC AUC Score")
print(roc_auc_score(y_val3,pred3))
```

Accuracy score
0.9775561097256
ROC AUC Score
0.9635979680790

Tiga model digabungkan menjadi satu model bertama outputall


```
outputall.groupby("resign").size()
```

resign	
0	3798
1	1152
dtype: int64	1152/4950
	0.23272727272727273

Kesimpulan

Berdasarkan model yang telah dibangun, resign dapat diprediksi menggunakan prediktor (dari prediktor paling penting)

1. tingkat_kepuasan_bekerja (semakin rendah semakin tinggi kemungkinan resign)
2. lama_di_perusahaan_num (semakin rendah semakin tinggi kemungkinan resign)
3. jumlah_proyek_yang_dikerjakan (terlalu banyak/sedikit jumlah proyek semakin tinggi kemungkinan resign)
4. rerata_jam_bulanan (terlalu tinggi/rendah rerata jam bulanan tinggi kemungkinan resign)
5. Kecelakaan_kerja (pernah kecelakaan cenderung tidak resign)
6. Tingkat_gaji (gaji tinggi cenderung tidak resign, gaji rendah cenderung resign)



“Pemodelan data adalah proses yang iteratif yang tiap iterasinya menghasilkan pembelajaran baru”