Improving Employee Retention by Predicting Employee Attrition Using Machine Learning

Created by:
Aziz Prabowo
azizprabowo128@gmail.com
linkedin.com/in/aziz-prabowo

A fresh graduate in Data Science with high interest artificial intelligence, data science, and business analytics. Experienced in data cleaning, exploratory data analysis, visualization, machine learning, and basic deep learning through academic, bootcamps, courses, and personal projects.

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

Project Overview

Background

Human resources (HR) are the company's most valuable asset and must be managed effectively to achieve business goals. In this project, we focus on understanding how to retain employees and reduce turnover, which otherwise leads to high recruitment and training costs. By identifying the key factors that drive attrition, the company can design relevant programs to improve employee satisfaction and retention.

Goal

Understand and analyze employee behavior using historical HR data to identify the key factors influencing attrition.

Objective

Develop a predictive model that identifies employees at risk of resigning by leveraging demographic, performance, and engagement data, allowing the company to implement proactive, data-driven HR interventions that reduce turnover and optimize workforce management.

Link to Jupyter Notebook

Dataset Information

```
<class 'pandas.core.frame.DataFrame'>
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
```

The dataset contains **25 columns** and **287 rows**. Several columns, such as Employee Satisfaction Score, Number of Projects Joined, Recent Monthly Lateness, Absenteeism, Participation in the LOP Program, and Resignation Reason, contain missing values. These missing values must be handled before moving forward with eda.

Data Cleaning for EDA

Username	0
EnterpriseID	0
StatusPernikahan	0
JenisKelamin	0
StatusKepegawaian	0
Pekerjaan	0
JenjangKarir	0
PerformancePegawai	0
AsalDaerah	0
HiringPlatform	0
SkorSurveyEngagement	0
SkorKepuasanPegawai	5
JumlahKeikutsertaanProjek	3
JumlahKeterlambatanSebulanTerakhir	1
JumlahKetidakhadiran	6
NomorHP	0
Email	0
TingkatPendidikan	0
PernahBekerja	0
IkutProgramLOP	258
AlasanResign	66
TanggalLahir	0
TanggalHiring	0
TanggalPenilaianKaryawan	0
TanggalResign	0

dtype: int64

The column *IkutProgramLOP* (Participation in LOP Program) has over 90% missing values and was removed from the dataset. For the AlasanResign column, missing values were imputed with "Unknown." Columns with smaller portions of missing data, such as SkorKepuasanPegawai, JumlahKeikutsertaanProjek, JumlahKeterlambatanSebulanTerakhir, dan JumlahKetidakhadiran, were imputed using statistical measures.

Data Cleaning for EDA

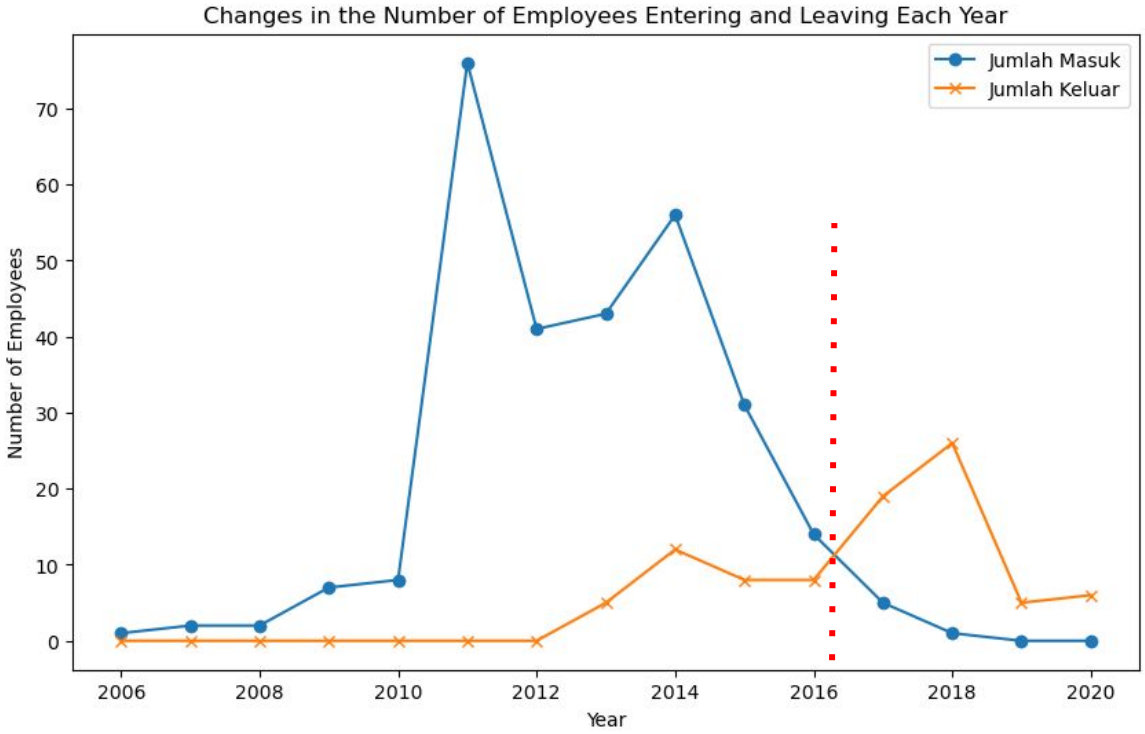
```
Value Counts - PernahBekerja
1      286
yes      1
Name: count, dtype: int64
```

```
Value Counts - StatusPernikahan
Belum_menikah    132
Menikah           57
Lainnya           48
Bercerai          47
-                  3
Name: count, dtype: int64
```

One category in the PernahBekerja column only contained a single value, so the column was removed. Additionally, the “–” category in StatusPernikahan was merged with Lainnya since both categories carry similar meaning.

Exploratory Data Analysis

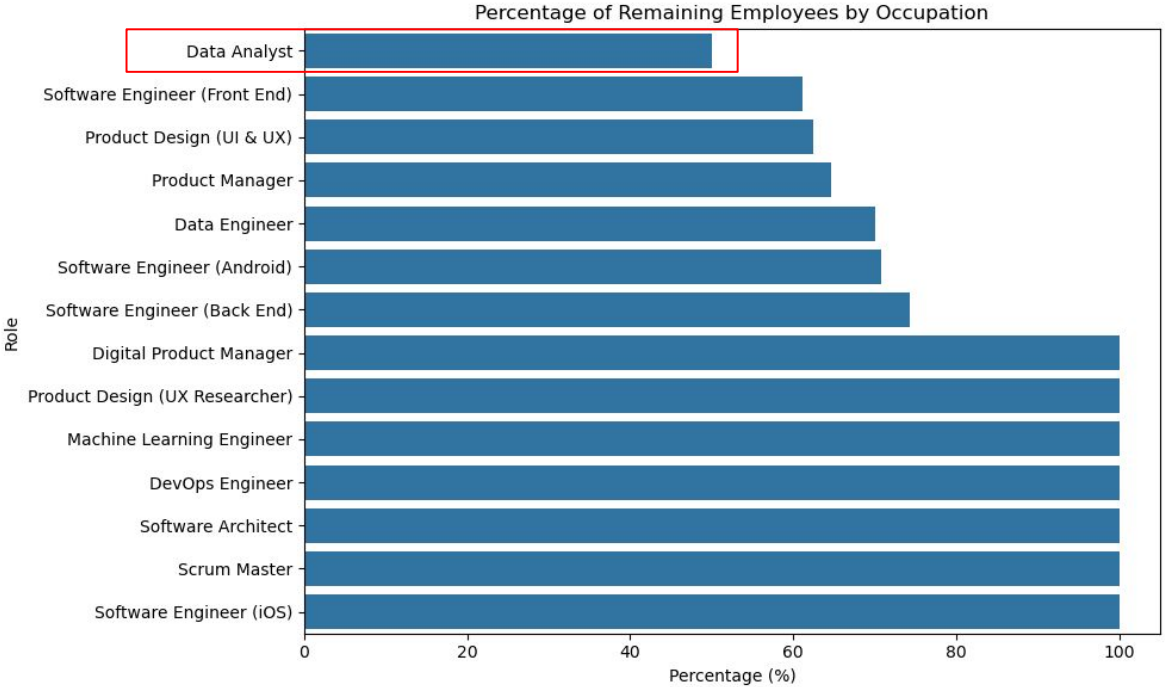
2017 Was a Turning Point for Employee Retention



Between 2006 and 2016, the company consistently experienced positive growth in employee numbers, as new hires always exceeded resignations. However, **2017 marked a turning point**: the number of resignations began to surpass new hires.

Exploratory Data Analysis

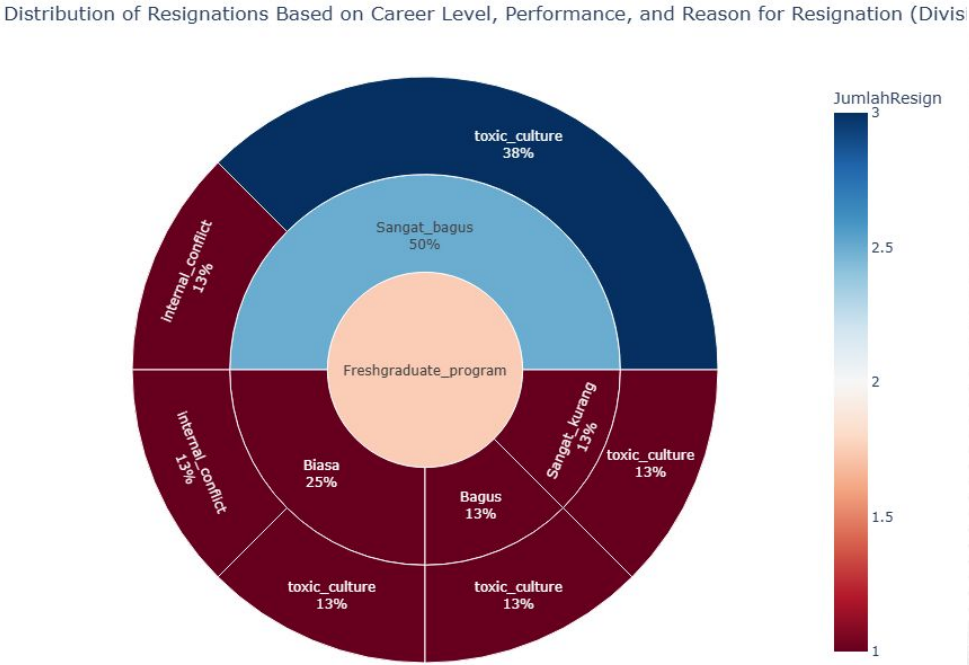
50% of Data Analyst Employees Resign



The ***Data Analyst*** role has the highest resignation rate, with **50% of employees leaving**, followed by other strategic positions such as *Front End Engineer* (40%) and *UI/UX Designer* (37.5%). High attrition in these technical roles signals serious retention challenges, risking project continuity and increasing the workload for remaining staff.

Exploratory Data Analysis

Toxic Culture is the Main Reason Top Talent Resigns



A large proportion of employees who resigned actually came from **high-performance groups**. The leading cause is ***toxic work culture*** **(38%)**, followed by ***internal conflict*** **(13%)**. Toxic culture appears consistently across most categories, making it the primary driver of attrition among top talent.

Data Cleaning for Modeling

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 287 entries, 0 to 286
Data columns (total 18 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   StatusPernikahan                287 non-null    object
 1   JenisKelamin                    287 non-null    object
 2   StatusKepegawaian               287 non-null    object
 3   Pekerjaan                       287 non-null    object
 4   JenjangKarir                    287 non-null    object
 5   PerformancePegawai              287 non-null    object
 6   AsalDaerah                      287 non-null    object
 7   HiringPlatform                  287 non-null    object
 8   SkorSurveyEngagement            287 non-null    int64
 9   SkorKepuasanPegawai             287 non-null    float64
 10  JumlahKeikutsertaanProjek       287 non-null    float64
 11  JumlahKeterlambatanSebulanTerakhir  287 non-null    float64
 12  JumlahKetidakhadiran            287 non-null    float64
 13  TingkatPendidikan               287 non-null    object
 14  TanggalLahir                    287 non-null    object
 15  TanggalHiring                   287 non-null    object
 16  TanggalPenilaianKaryawan        287 non-null    object
 17  TanggalResign                   287 non-null    object
dtypes: float64(4), int64(1), object(13)
memory usage: 40.5+ KB
```

From 25 to 18 features

The IkutProgramLOP feature was removed due to excessive missing values. Non-informative features such as Username, EnterpriseID, NomorHP, Email, AlasanResign were also dropped. For skewed numeric features such as JumlahKeikutsertaanProjek, JumlahKeterlambatanSebulanTerakhir, and JumlahKetidakhadiran missing values were imputed with the median to avoid distortion from outliers. Categorical features with missing values were imputed with the mode. Features like AlasanResign and TanggalResign were excluded to prevent **data leakage**.

Feature Engineering

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 287 entries, 0 to 286
Data columns (total 16 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   StatusPernikahan                 287 non-null    object
 1   JenisKelamin                     287 non-null    object
 2   StatusKepegawaian                287 non-null    object
 3   Pekerjaan                        287 non-null    object
 4   JenjangKarir                     287 non-null    object
 5   PerformancePegawai               287 non-null    object
 6   AsalDaerah                       287 non-null    object
 7   HiringPlatform                   287 non-null    object
 8   JumlahKeterlambatanSebulanTerakhir 287 non-null  float64
 9   TingkatPendidikan                287 non-null    object
 10  Attrition                        287 non-null    int64
 11  Usia                             287 non-null    int64
 12  LamaBekerja                      287 non-null    int64
 13  BulanSejakPenilaian              287 non-null    int64
 14  AktifScore                       287 non-null    float64
 15  SkorGabungan                     287 non-null    float64
dtypes: float64(3), int64(4), object(9)
memory usage: 36.0+ KB
```

From 18 to 16 features

The target variable Attrition was created based on whether the TanggalResign exists (1 = resigned, 0 = stayed). Date fields were converted to datetime and transformed into new features: Usia, LamaBekerja, and BulanSejakPenilaian.

Additional features included:

- *AktifScore* = JumlahKeikutSertaanProyek – JumlahKetidakhadiran
- *SkorGabungan* = 60% SkorSurveyEngagement + 40% SkorKepuasanPegawai
- Irrelevant or redundant columns were then removed to simplify the dataset.

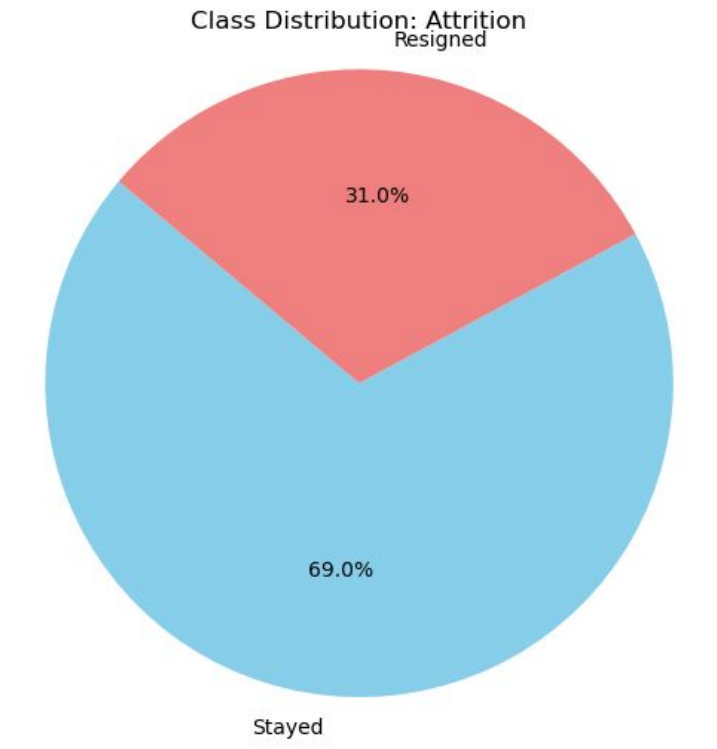
Preprocessing

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 287 entries, 0 to 286
Data columns (total 40 columns):
 #   Column                                          Non-Null Count  Dtype
---  ------                                          --------------  -----
 0   scale__JumlahKeterlambatanSebulanTerakhir       287 non-null    float64
 1   scale__Usia                                     287 non-null    float64
 2   scale__LamaBekerja                              287 non-null    float64
 3   scale__BulanSejakPenilaian                      287 non-null    float64
 4   scale__AktifScore                               287 non-null    float64
 5   scale__SkorGabungan                             287 non-null    float64
 6   ord_enc__JenjangKarir                           287 non-null    float64
 7   ord_enc__PerformancePegawai                     287 non-null    float64
 8   ord_enc__TingkatPendidikan                      287 non-null    float64
 9   nom_enc__JenisKelamin_Wanita                    287 non-null    float64
 10  nom_enc__StatusPernikahan_Bercerai              287 non-null    float64
 11  nom_enc__StatusPernikahan_Lainnya               287 non-null    float64
 12  nom_enc__StatusPernikahan_Menikah               287 non-null    float64
 13  nom_enc__Pekerjaan_Data Engineer                287 non-null    float64
 14  nom_enc__Pekerjaan_DevOps Engineer              287 non-null    float64
 15  nom_enc__Pekerjaan_Digital Product Manager      287 non-null    float64
 16  nom_enc__Pekerjaan_Machine Learning Engineer    287 non-null    float64
 17  nom_enc__Pekerjaan_Product Design (UI & UX)     287 non-null    float64
 18  nom_enc__Pekerjaan_Product Design (UX Researcher) 287 non-null  float64
 19  nom_enc__Pekerjaan_Product Manager              287 non-null    float64
 20  nom_enc__Pekerjaan_Scrum Master                 287 non-null    float64
 21  nom_enc__Pekerjaan_Software Architect           287 non-null    float64
 22  nom_enc__Pekerjaan_Software Engineer (Android)  287 non-null    float64
 23  nom_enc__Pekerjaan_Software Engineer (Back End) 287 non-null    float64
 24  nom_enc__Pekerjaan_Software Engineer (Front End) 287 non-null   float64
 25  nom_enc__Pekerjaan_Software Engineer (iOS)      287 non-null    float64
 26  nom_enc__StatusKepegawaian_Internship           287 non-null    float64
 27  nom_enc__StatusKepegawaian_Outsource            287 non-null    float64
 28  nom_enc__AsalDaerah_Jakarta Pusat               287 non-null    float64
 29  nom_enc__AsalDaerah_Jakarta Selatan             287 non-null    float64
 30  nom_enc__AsalDaerah_Jakarta Timur               287 non-null    float64
 31  nom_enc__AsalDaerah_Jakarta Utara               287 non-null    float64
 32  nom_enc__HiringPlatform_Diversity_Job_Fair      287 non-null    float64
 33  nom_enc__HiringPlatform_Employee_Referral       287 non-null    float64
 34  nom_enc__HiringPlatform_Google_Search           287 non-null    float64
 35  nom_enc__HiringPlatform_Indeed                  287 non-null    float64
 36  nom_enc__HiringPlatform_LinkedIn                287 non-null    float64
 37  nom_enc__HiringPlatform_On-line_Web_application 287 non-null    float64
 38  nom_enc__HiringPlatform_Other                   287 non-null    float64
 39  nom_enc__HiringPlatform_Website                 287 non-null    float64
dtypes: float64(40)
memory usage: 89.8 KB
```

- Numerical features (e.g., JumlahKeterlambatanSebulanTerakhir, Usia, LamaBekerja) → scaled with **RobustScaler** to reduce outlier impact.
- Ordinal features (e.g., JenjangKarir, PerformancePegawai, dan TingkatPendidikan) → encoded with **OrdinalEncoder**.
- Nominal features (e.g., JenisKelamin, StatusPernikahan, Pekerjaan) → encoded with **OneHotEncoder**, dropping the first category to avoid dummy trap.

Data Sampling

Imbalance Class



The dataset is imbalanced, with 31% attrition vs. 69% non-attrition. To address this, oversampling and undersampling methods were tested. The evaluation focused on the **F2-score**, prioritizing recall (detecting employees at risk of resigning) as an early warning system.

Data Sampling

Rakamin Academy

Data Sampling Result

	Sampling Method	precision	recall	f1	f2	roc_auc
12	SMOTEENN	0.310 ± 0.027	0.744 ± 0.139	0.436 ± 0.047	0.579 ± 0.082	0.511 ± 0.114
9	AllKNN	0.278 ± 0.108	0.675 ± 0.258	0.393 ± 0.151	0.524 ± 0.200	0.489 ± 0.105
7	EditedNearestNeighbours	0.322 ± 0.059	0.574 ± 0.112	0.411 ± 0.074	0.495 ± 0.091	0.510 ± 0.083
11	InstanceHardnessThreshold	0.334 ± 0.063	0.554 ± 0.163	0.412 ± 0.083	0.485 ± 0.117	0.511 ± 0.098
8	RepeatedEditedNearestNeighbours	0.299 ± 0.051	0.565 ± 0.229	0.376 ± 0.098	0.465 ± 0.153	0.502 ± 0.070
5	RandomUnderSampler	0.310 ± 0.096	0.475 ± 0.161	0.373 ± 0.116	0.427 ± 0.137	0.493 ± 0.119
10	NeighbourhoodCleaningRule	0.306 ± 0.056	0.463 ± 0.130	0.365 ± 0.077	0.417 ± 0.103	0.487 ± 0.102
2	SMOTE	0.310 ± 0.103	0.440 ± 0.151	0.362 ± 0.120	0.405 ± 0.135	0.492 ± 0.126
3	BorderlineSMOTE	0.315 ± 0.092	0.418 ± 0.149	0.357 ± 0.113	0.391 ± 0.132	0.505 ± 0.106
13	SMOTETomek	0.292 ± 0.119	0.407 ± 0.183	0.338 ± 0.142	0.375 ± 0.164	0.491 ± 0.127
1	RandomOverSampler	0.289 ± 0.106	0.393 ± 0.158	0.330 ± 0.122	0.364 ± 0.140	0.512 ± 0.143
4	SVMSMOTE	0.313 ± 0.124	0.351 ± 0.164	0.328 ± 0.137	0.341 ± 0.151	0.499 ± 0.113
6	TomekLinks	0.227 ± 0.248	0.126 ± 0.141	0.153 ± 0.163	0.135 ± 0.146	0.493 ± 0.098
0	No Sampling	0.194 ± 0.267	0.081 ± 0.122	0.103 ± 0.143	0.087 ± 0.126	0.497 ± 0.133

The experimental results show that the **SMOTEENN** data sampling method delivers the best performance compared to other methods, with an **F2 score of 0.679**, a **recall of 0.744**, and a **ROC AUC of 0.511**.

Considering that the main focus is to maximize the detection of employees who are likely to resign (recall) while maintaining a balanced penalty through the F2 score, **SMOTEENN is selected as the sampling method to be used in the next stage of model training.**

Model Selection

Rakamin Academy

Model Selection Result

	F1-score (mean)	F2-score (mean)	Precision (mean)	Recall (mean)	ROC-AUC (mean)
KNN	0.4547	0.6389	0.3074	0.8773	0.4763
SVM	0.4242	0.5903	0.2895	0.8028	0.4363
LogReg	0.4273	0.5728	0.3011	0.7444	0.4897
CatBoost	0.4293	0.5689	0.3062	0.7306	0.5081
Ridge	0.4189	0.5532	0.2996	0.7083	0.4869
XGBoost	0.4315	0.5380	0.3267	0.6481	0.5341
RandomForest	0.4154	0.5366	0.3032	0.6699	0.5026
MLP	0.4197	0.5350	0.3105	0.6593	0.4934
ExtraTrees	0.4182	0.5329	0.3100	0.6569	0.5017
AdaBoost	0.4155	0.5275	0.3093	0.6477	0.5129
PA	0.3985	0.5113	0.2958	0.6398	0.4813
GradientBoost	0.4048	0.5091	0.3034	0.6181	0.4947
LightGBM	0.4072	0.5062	0.3105	0.6102	0.5003
DecisionTree	0.3973	0.4923	0.3025	0.5889	0.4936

Among tested models, **KNN** performed best with the highest F2-score (0.6389), followed by **SVM** (0.5903). The F2 metric was chosen because it emphasizes recall, minimizing the risk of missing at-risk employees.

Hyperparameter Tuning

Hyperparameter Tuning Result

	Best Parameters	Best Score	F1	F2	ROC-AUC	Precision	Recall
XGBoost	{'classifier__learning_rate': 0.05, 'classifie...	0.656886	0.459459	0.697674	0.407639	0.315789	1.0
SVM	{'classifier__C': 0.1, 'classifier__gamma': 's...	0.691888	0.459459	0.692308	0.386806	0.310345	1.0
LogReg	{'classifier__C': 0.1, 'classifier__l1_ratio':...	0.693247	0.459459	0.692308	0.308333	0.310345	1.0
CatBoost	{'classifier__depth': 4, 'classifier__iteratio...	0.692262	0.459459	0.674603	0.517361	0.314815	0.944444
KNN	{'classifier__metric': 'euclidean', 'classifie...	0.674915	0.459459	0.664062	0.386806	0.303571	0.944444

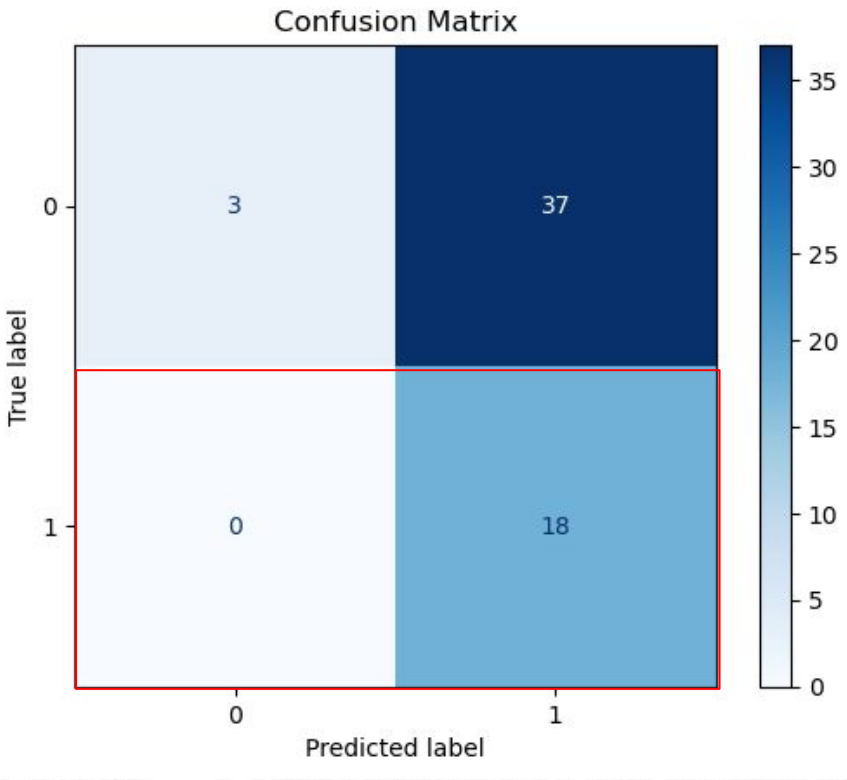
After tuning five top models (KNN, SVM, Logistic Regression, XGBoost, CatBoost), **XGBoost** achieved the best performance with:

- Training F2 = 0.6568
- Testing F2 = 0.6977
- F1 = 0.4600, Precision = 0.3158, Recall = 1.0000

Since recall was the priority, **XGBoost was selected as the final model**.

Model Evaluation

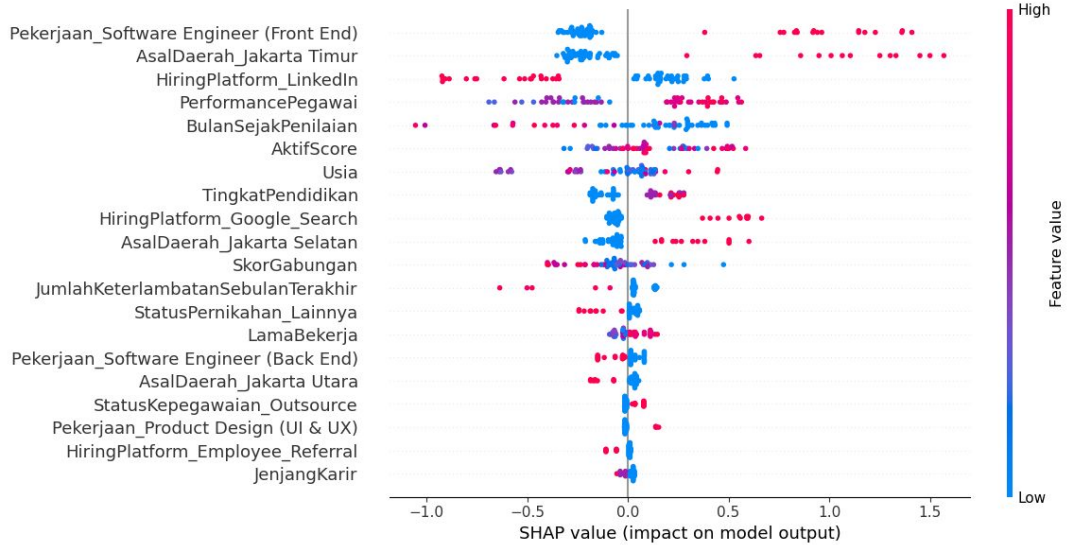
XGBoost Confusion Matrix



The confusion matrix shows **the model achieved perfect recall (1.00)**, correctly identifying all 18 employees who resigned. However, it also produced many false positives (**low precision = 0.33**), misclassifying non-resigners as at-risk. While the model is highly sensitive, **making it useful for early intervention**, it lacks specificity in distinguishing actual stayers.

Model Interpretability

SHAP Value Summary Plot



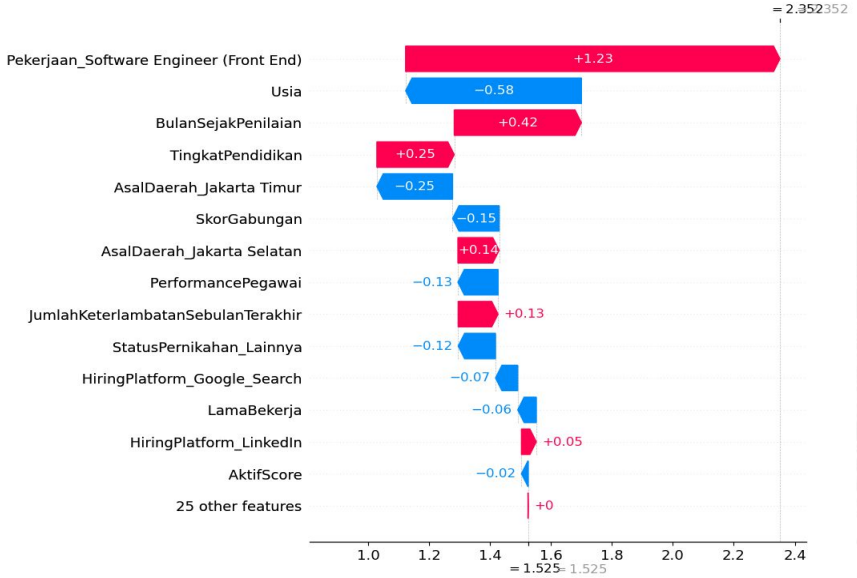
SHAP analysis highlighted key attrition drivers:

- Being a **Front End Engineer** *significantly increases resignation risk*.
- Employees from **East Jakarta** and those recruited via **LinkedIn** showed *higher attrition likelihood.*
- **High-performing employees** paradoxically face *higher risk of leaving*.
- **Longer gaps since the last evaluation** correlate with *higher attrition*.
- **Lower Activity Score** strongly signals *risk*.

Model Interpretability

An Employee's Prediction Simulation

```
Informasi Karyawan:
-----------
StatusPernikahan                          Lainnya
JenisKelamin                               Wanita
StatusKepegawaian                        FullTime
Pekerjaan                 Software Engineer (Front End)
JenjangKarir                            Mid_level
PerformancePegawai                          Biasa
AsalDaerah                          Jakarta Selatan
HiringPlatform                              Indeed
JumlahKeterlambatanSebulanTerakhir             0.0
TingkatPendidikan                         Doktor
Usia                                          46
LamaBekerja                                   12
BulanSejakPenilaian                           69
AktifScore                                  -6.0
SkorGabungan                                4.0
```



In an example case, **an employee showed a high attrition score** (**2.35** vs. **baseline 1.52**). Main risk drivers: *role as Front End Engineer*, *long gap since last evaluation*, *higher education level*, *frequent lateness*, and *certain regional background*. Protective factors included *age*, *good performance*, *engagement score*, and *longer tenure*—but **these were insufficient to offset the risks**.

Recommendations

1. **Focus on Front End Engineers & Young Talent**
   Provide clear career paths, technical mentoring, and balanced workloads.
2. **Accelerate Evaluations & Career Development**
   Conduct quarterly reviews with constructive feedback and offer promotion/project opportunities for high performers and highly educated staff.
3. **Location & Performance-Based Interventions**
   Address region-specific issues (e.g., cost of living, commuting in South Jakarta) with flexible policies. Monitor performance and lateness metrics for early warnings, applying mentoring for weaker staff and incentives for top performers.

Thank You