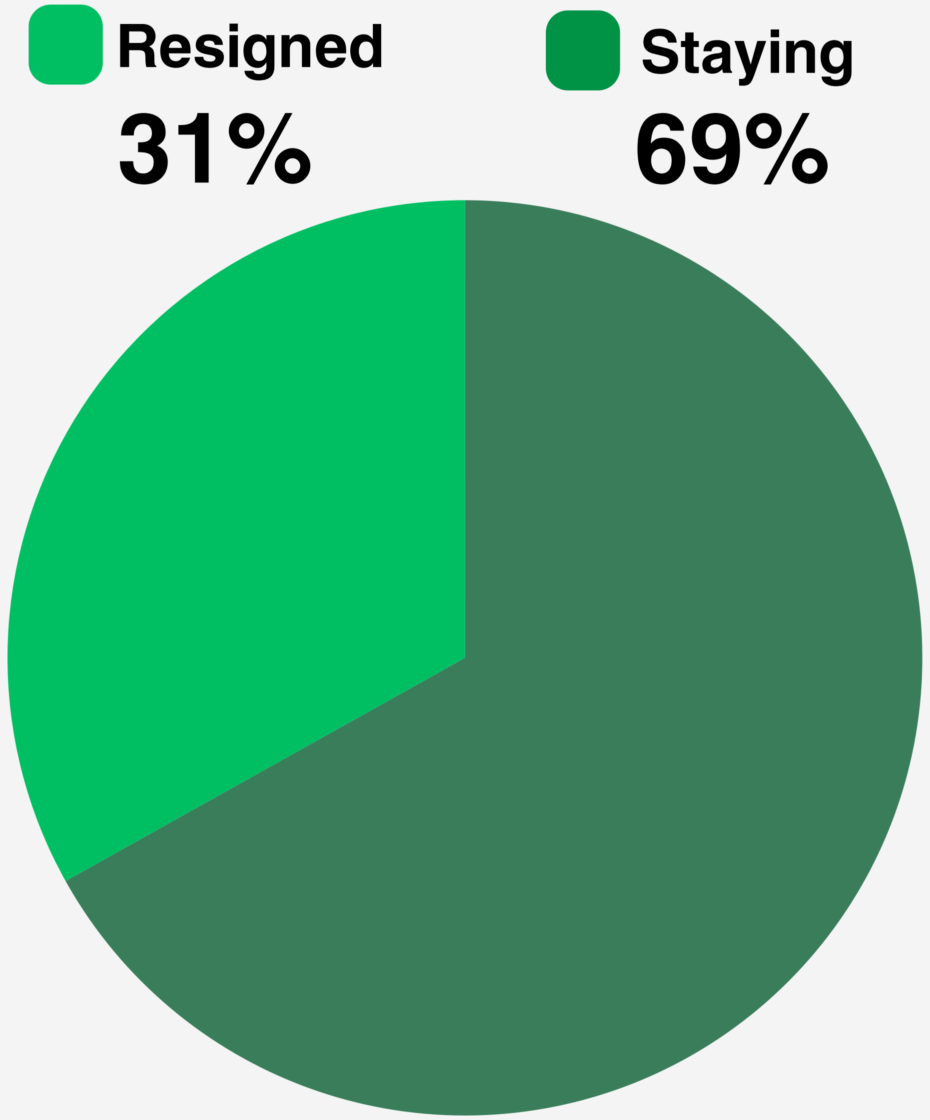
Mini Project Report:

Improving Employee Retention by Predicting Employee Attrition Using Machine Learning

by Bhima Fairul Rifqi

Business Overview



XYZ's 30%+ attrition since inception has hurt its projects, operations, and reputation. To address this issue, they decide to use machine learning to identify at-risk employees.

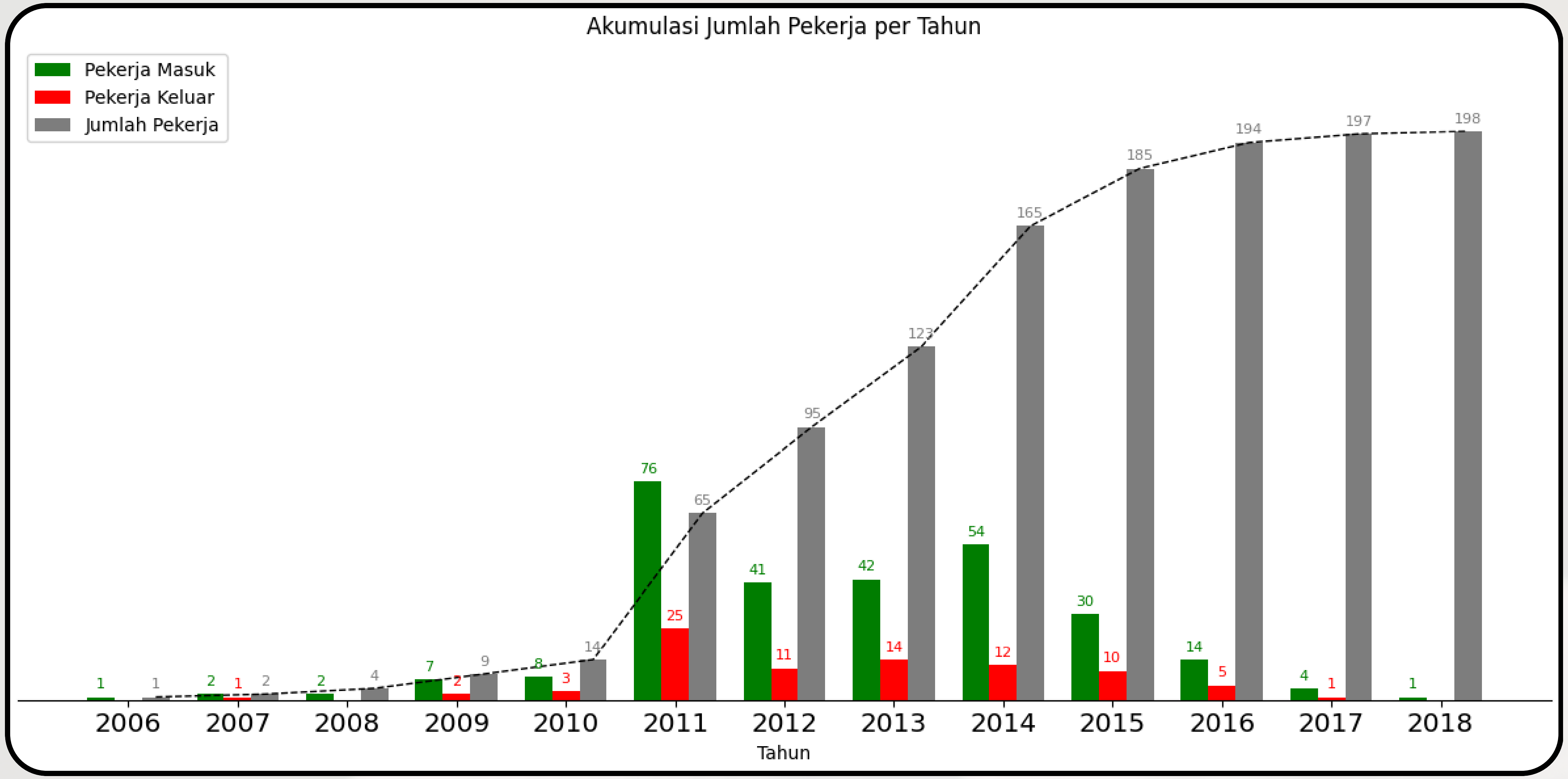
Business Metric

Rate of Employees' Attrition

*the focus target are labeled as 'StatusResign' in the dataset.

Insights

The heatmap shows that all features have weak correlations with resignation status (StatusResign), indicating no strong linear predictors of resignation.

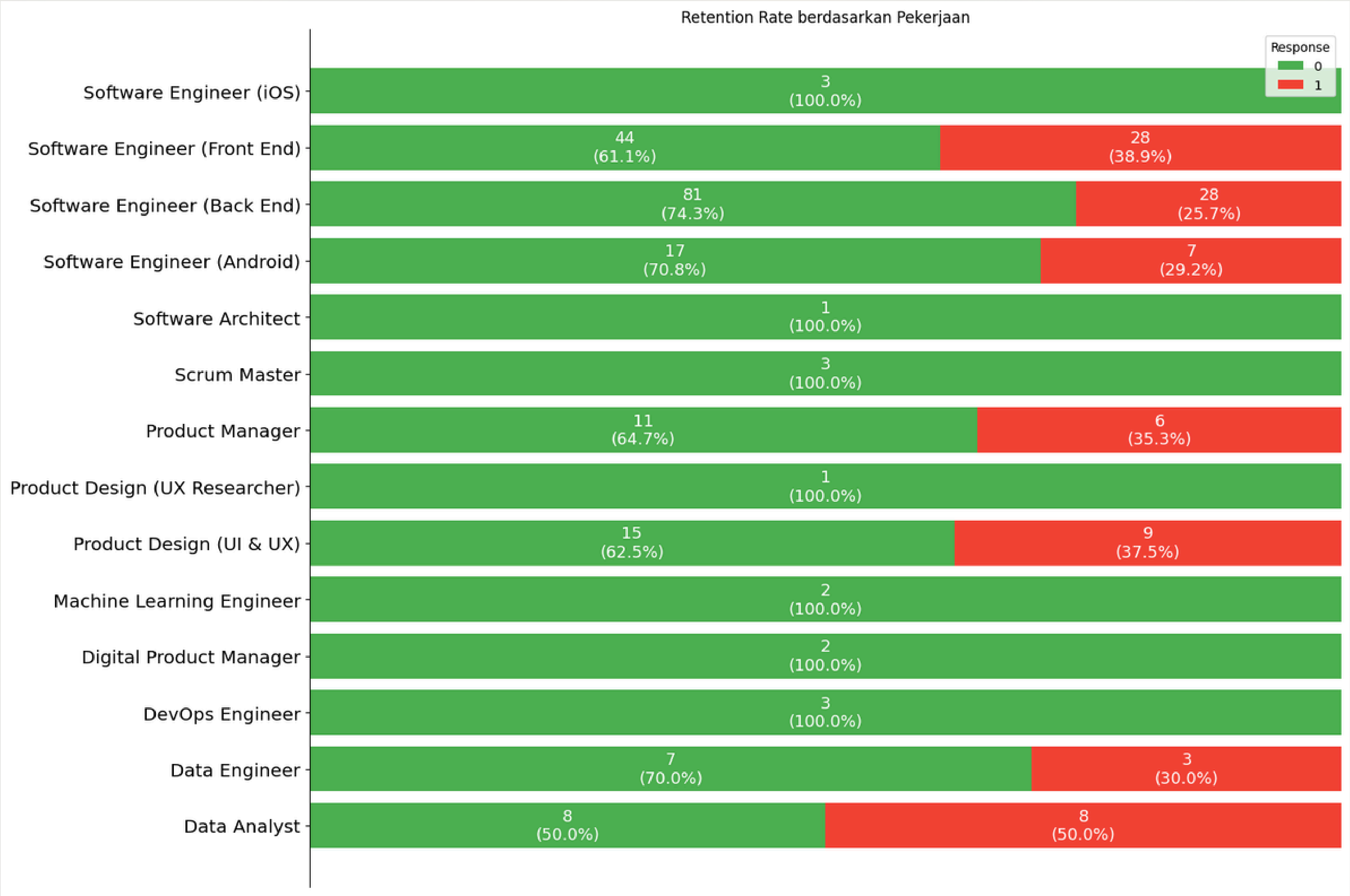


Insights

Despite aggressive hiring between 2011 and 2014, high employee turnover limited long-term growth, with workforce numbers stabilizing after 2015, signaling a need to focus on retention over recruitment.

Recommendation

Shift focus from aggressive hiring to strengthening employee retention through targeted initiatives like career development, engagement programs, and predictive analytics to identify and address attrition risks early.

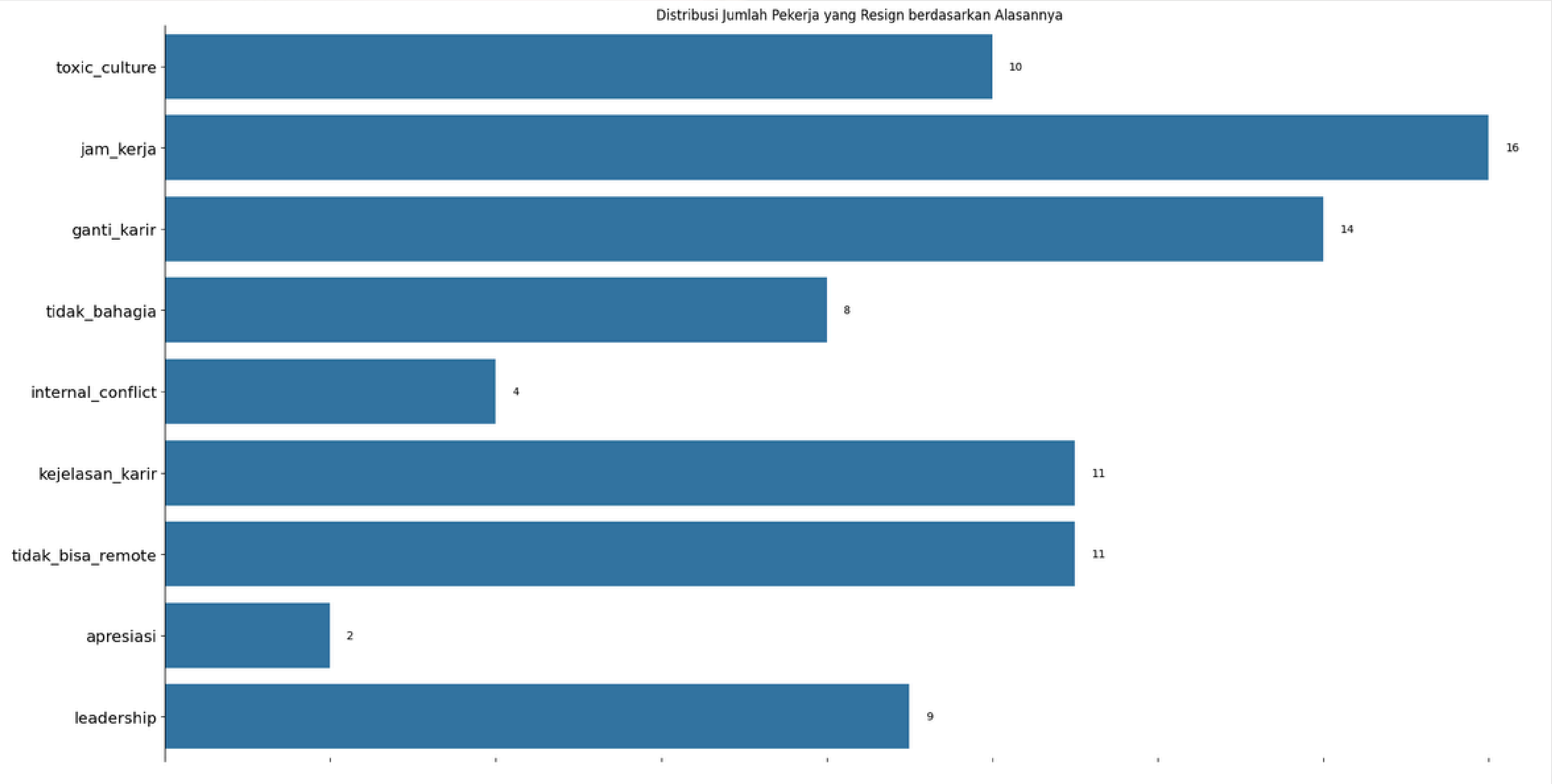


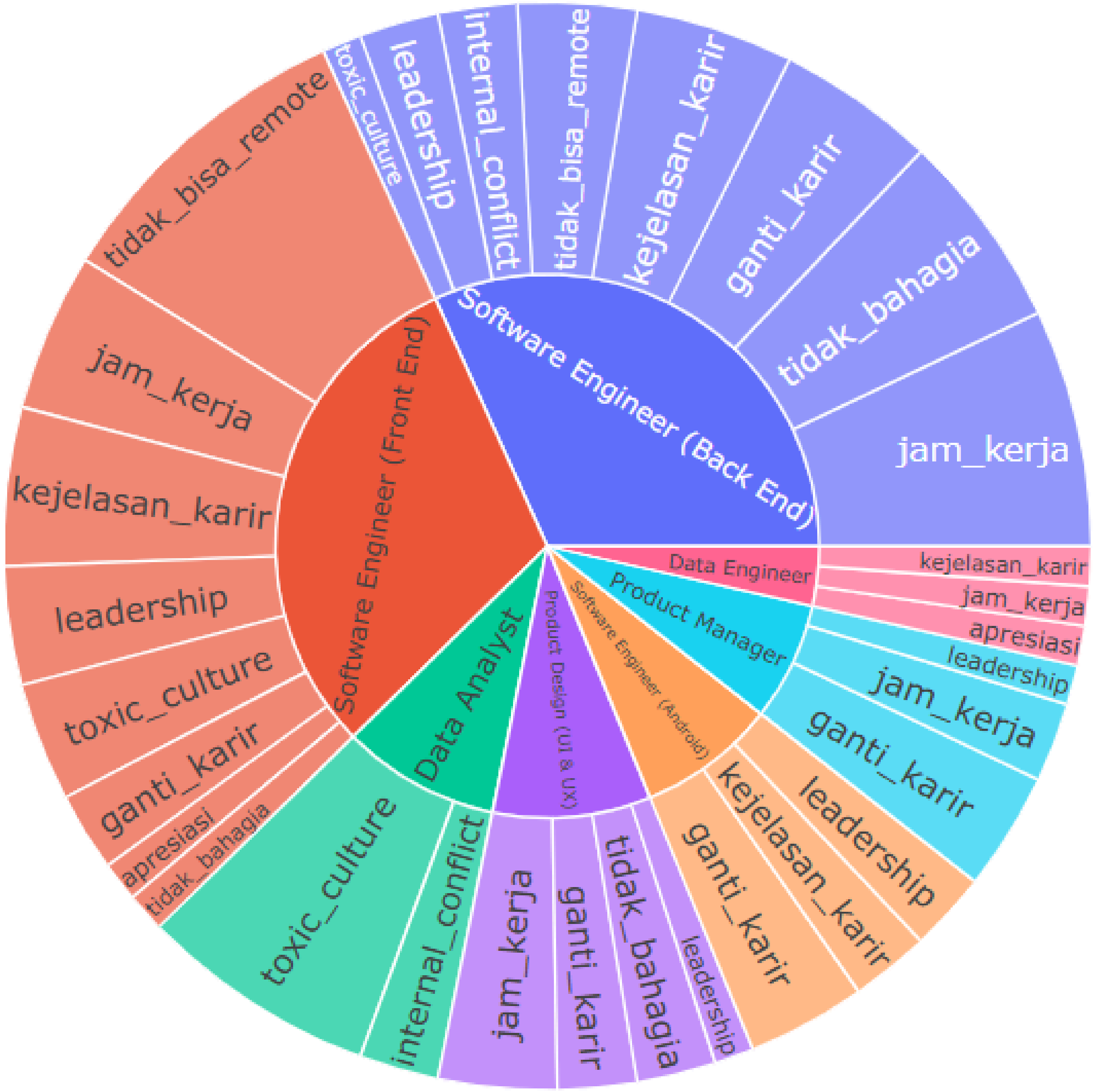
Insights

Retention rates vary widely by role, with Front-End Engineers and Data Analysts showing the highest attrition, indicating these positions may require targeted retention strategies.

Insights

The top reasons employees resign are long working hours, career change, and toxic culture, indicating key areas where the company should improve work-life balance, internal mobility, and organizational culture.





Insights:

Front-End and Back-End Engineers, who account for the most resignations, along with Data Analysts, who have the highest attrition rate, all cite toxic culture as a key reason for leaving. Their satisfaction scores toward the company are moderate, with Data Analysts at 3.75, Front-End Engineers at 3.86, and Back-End Engineers at 3.93 (on a 1 to 5 scale).

Recommendation:

To reduce attrition, the company should address the top three resignation drivers: toxic culture, long working hours, and career change. Improving workplace culture, promoting work-life balance, and offering clearer career development paths can significantly enhance employee retention.

About The Data: Complete Dataset

Feature Name	Feature Group	Type	To Do
Username	Identifier	Categorical	Drop
EnterpriseID	Identifier	Categorical	Drop
NomorHP	Identifier	Categorical	Drop
Email	Identifier	Categorical	Drop
DomainEmail	Identifier	Categorical	Drop
AlasanResign	Categorical	Categorical	Feature Engineering & Drop
TanggalLahir	Date	Date	Feature Engineering & Drop
TanggalHiring	Date	Date	Feature Engineering & Drop
TanggalResign	Date	Date	Feature Engineering & Drop
TanggalPenilaianKaryawan	Date	Date	Drop
PernahBekerja	Constant	Categorical	Drop
JenisKelamin	Binary	Categorical	Binary Encode
StatusResign (Target)	Binary	Categorical	Binary Encode
StatusPernikahan	Categorical	Categorical	Drop
LamaBekerja	Numerical	Numerical	Drop

Number of Observations: 287

Feature Name	Feature Group	Type	To Do
StatusKepegawaian	Categorical	Categorical	Ordinal Encode
Pekerjaan	Categorical	Categorical	OH Encode
JenjangKarir	Categorical	Categorical	Ordinal Encode
PerformancePegawai	Categorical	Categorical	Ordinal Encode
AsalDaerah	Categorical	Categorical	OH Encode
HiringPlatform	Categorical	Categorical	Drop
TingkatPendidikan	Categorical	Categorical	Ordinal Encode
SkorSurveyEngagement	Numerical	Numerical	Rescale
SkorKepuasanPegawai	Numerical	Numerical	Rescale
JumlahKeikutsertaanProjek	Numerical	Numerical	Rescale
JumlahKeterlambatanSebulanTerakhir	Numerical	Numerical	Rescale
JumlahKetidakhadiran	Numerical	Numerical	Rescale
Usia	Numerical	Numerical	Rescale
IkutProgramLOP	Binary	Categorical	Drop

Data Preprocessing

The data are splitted into train and test set on 80:20 ratio respectively.

Data Imputation

Mode imputation for categoric features
Median imputation for numeric features

Data Encode

One Hot Encoding for nominal categoric features
Ordinal Encoding for ordinal categoric features

Rescale

Rescaling the numeric features using robust scaler

Data Balancing

Applied SMOTE on the target of the data train

Applied Models and Its Evaluation

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC	Best Threshold
cat	0,563636	0,382353	0,8125	0,52	0,592949	0,12
knn	0,527273	0,368421	0,875	0,518519	0,612179	0,33
rf	0,509091	0,358974	0,875	0,509091	0,548878	0,27
xgb	0,709091	0,5	0,5	0,5	0,600962	0,4
logreg	0,418182	0,326087	0,9375	0,483871	0,491977	0,28
ada	0,363636	0,313725	1	0,477612	0,487981	0,26
nb	0,490909	0,323529	0,6875	0,44	0,520833	0,29
gb	0,618182	0,368421	0,4375	0,4	0,5625	0,31
dt	0,618182	0,333333	0,3125	0,322581	0,47516	0,67

Selected Model:

XGB, since it’s balanced between Precision and Recall.

Model Impact Analysis

Let’s assume that HR Intervention Success Rate is 69%.

	Predicted Resign	Predicted Stay
Actually Resigned	11 (TP)	5 (FN)
Actually Stayed	33 (FP)	6 (TN)
Total Employees	58	

Metric	Before Model	After Model + HR Intervention
Resigned employees	16	8
Resignations prevented	–	8
Attrition rate	27,6%	13,8%

Prevented resignations = $11 \times 69\% = 7.59 \approx 8$ employees

50%

Model Explainability Analysis

Feature	Weight
Pekerjaan	0,1767
Usia	0,1498
AsalDaerah	0,1462
JumlahKetidakhadiran	0,1034
SkorKepuasanPegawai	0,0734
PerformancePegawai	0,0727
TingkatPendidikan	0,0685
SkorSurveyEngagement	0,0482
JenjangKarir	0,0476
JumlahKeikutsertaanProjek	0,0366
JenisKelamin	0,0319
StatusKepegawaian	0,0307
JumlahKeterlambatanSebulanTerakhir	0,0143

Insight

Job role and age are the top factors influencing attrition. These insights guide the company to prioritize support and engagement initiatives for high-risk job roles and age groups.

Conclusions

Things to Figure Out

Working Hours, Culture, &
Career Growth.

Impact of Model Implementation

Developed model successfully reduced employees’ attrition rate.

Main Factors of Leaving Tendency

Pekerjaan dan Usia.

THANK YOU

Project Resource:

GitHub / Repo