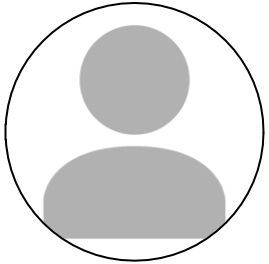
Improving Employee Retention by Predicting Employee Attrition Using Machine Learning

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

Rakamin Academy



Created by:
Bintang Phylosophie
bintangphy@gmail.com
https://www.linkedin.com/in/bintang-phylosophie/

A results-driven Data Scientist and Data Analyst with a strong foundation in machine learning, data analysis, and visualization. Proficient in Python, SQL, and advanced analytics tools such as Power BI, and Looker Studio. Experienced in handling large datasets, optimizing queries, and building data-driven solutions through project-based internships in several company. Holds a Data Science certification from Rakamin Academy with hands-on expertise in statistical modeling, predictive analytics, and interactive dashboard creation. Passionate about data storytelling and leveraging insights to drive business decisions.

Overview

Background

Human resource (HR) is the key asset that needs effective management to help companies achieve their business goals. In this project, we are faced with an issue related to human resources within a company. Our focus is to understand how retain employees to prevent the swelling of recruitment and training costs for new hires. By identifying factors causing employees to leave, we can promptly address these concerns by creating relevant employee programs.

Dataset & Business Understanding

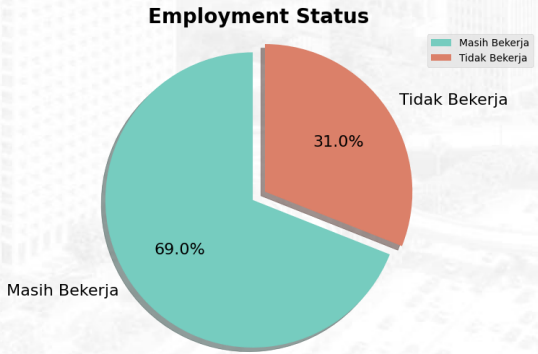
Dataset Information:

The dataset contains employee attrition information from a fictional company from 2006 to 2020.

Attribute Information:

- **Identifier**
  Username and EnterpriseID is unique ID that each of which is an ID for each employee
- **Target**
  StatusKerja has extracted from TanggalResign. If `TanggalResign` contain a datetime value it is indicated that the employee **'Tidak Bekerja'** (already resign). And NaN value indicated is an employee is **'Masih Bekerja'** (still working) in the company.
- **Company Goals:**
  - Enhance Employee Retention: Strengthen strategies to retain employees and reduce turnover.
  - Leverage Data for Decision-Making: Use historical employee data to gain insights into attrition trends.
  - Implement Predictive Analytics: Develop models to anticipate resignations and enable proactive interventions.
- **Problem:**
  Employee attrition threatens organizational stability, productivity, and long-term growth. Data indicates a sharp rise in resignations, particularly in 2018, alongside a declining trend in new hires. This imbalance underscores the need for proactive retention strategies to mitigate turnover.
- **Objectives:**
  - Analyze Employee Data: Examine historical trends, employee attributes, and resignation patterns to uncover key drivers of turnover.
  - Generate Actionable Insights: Identify factors influencing retention and recommend strategies for improvement.
  - Develop Predictive Models: Create data-driven models to forecast resignations and help the company address employee concerns proactively.

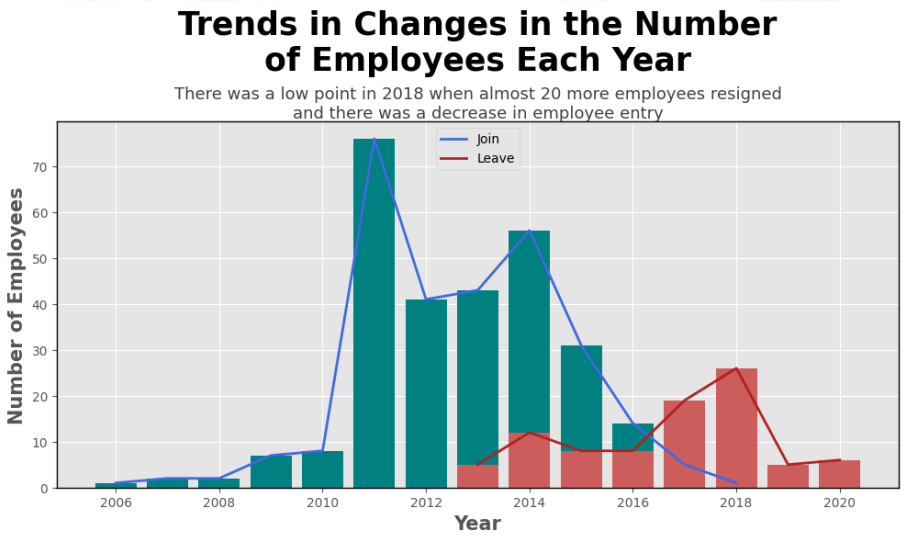
What Happened?



- Since 2006, 31% of employees have parted ways with the company, marking a significant shift in the workforce by 2020

Insight and Business Recommendation

Employee Retention

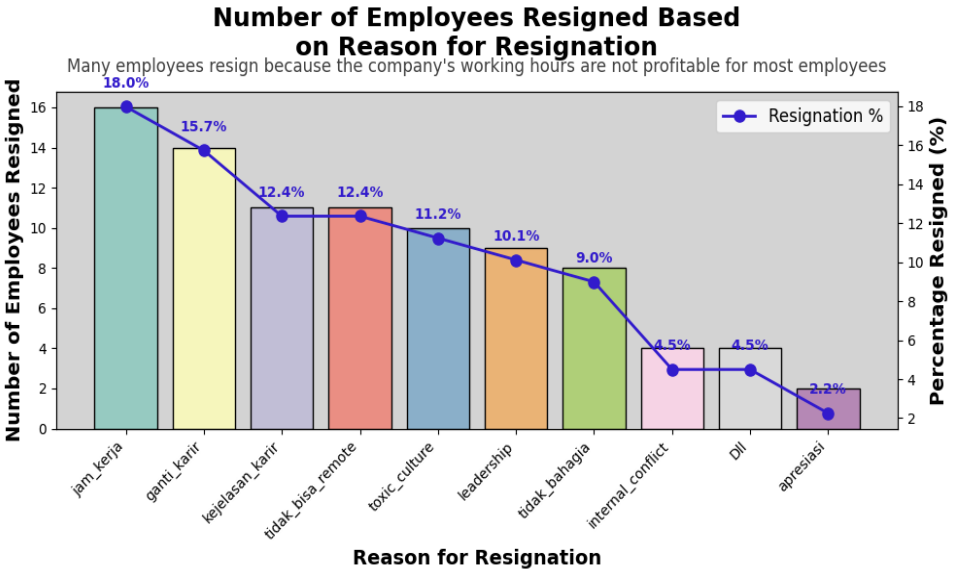


- **2006-2010:** Minimal hiring before 2008, with a small increase in 2009-2010 and almost no resignations during this period.
- **2011-2012:** saw the highest hiring rate (over 70 employees in 2012). The company was in a strong growth phase, possibly due to expansion or new projects. No resignations in these years, indicating employee stability.
- **2014-2016:** Hiring declined significantly after 2012, reaching about half the peak level in 2014. Resignations started increasing, showing early signs of employee turnover issues. By 2016, the hiring trend continued downward while resignations became more frequent.
- **2018:** Resignations peaked (~20 employees left) while new hires dropped significantly. This suggests organizational instability, dissatisfaction, layoffs, or market changes. The company may have faced financial difficulties, restructuring, or cultural shifts that led to higher attrition.
- **2019-2020:** Low hiring levels persisted, with only small numbers of new employees joining. Resignations remained present but were lower than in 2018.The company may have stabilized, but with minimal workforce expansion.

Click Here to see my code

Insight and Business Recommendation

Reason for Resignation

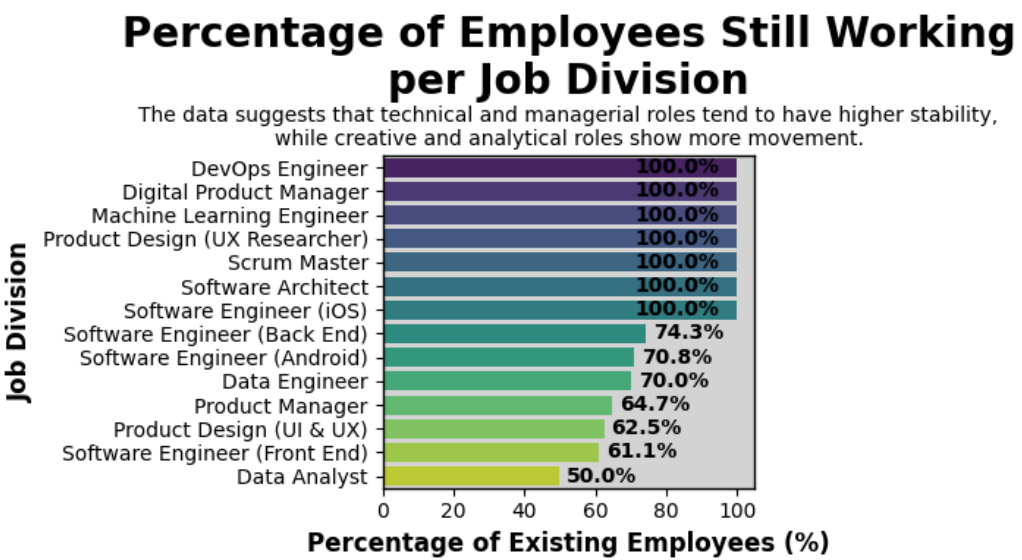


- The highest number of resignations (16 employees) cited Jam Kerja (**working hours)** as the primary reason. This suggests that employees may feel overworked, have poor work-life balance, or find the hours non-beneficial. Consider flexible work schedules, improved overtime compensation, or workload distribution to enhance work-life balance.
- **Ganti Karir (career change) and Kejelasan Karir (career clarity)** rank among the top reasons. Employees might be leaving due to lack of career advancement opportunities, unclear job roles, or stagnant professional development.
- **Toxic Culture:** Indicates dissatisfaction with workplace dynamics, possibly due to poor management, unhealthy work relations, or high stress.
- **Tidak Bisa Remote (remote work limitations):** Suggests employees may seek more flexibility, remote work options, or better work arrangements.
- The top resignation factors (work hours, career issues, and work environment) indicate areas where **companies can improve retention strategies**. Addressing these concerns proactively can help in reducing employee turnover and increasing job satisfaction.

Click Here to see my code

Insight and Business Recommendation

Existing Employees

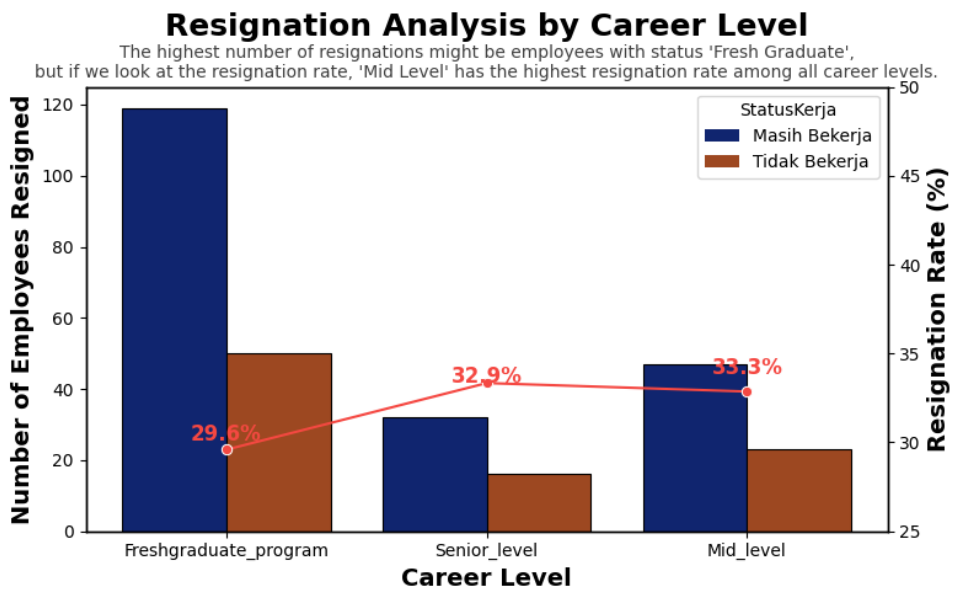


- **100% Retention in Several Roles:**.This suggests no attrition in these roles, possibly due to high job satisfaction, strong demand, or limited hiring in these categories. Company can maintain job satisfaction by offering career progression, competitive salaries, and challenging projects.
- **Moderate Retention (70%-75%):** These roles experience some level of turnover, likely due to competitive job markets, career transitions, or alternative opportunities.
- **Lower Retention (50%-65%):** These roles show higher attrition, especially Data Analysts (50%), which might indicate limited career growth opportunities, competitive external job offers, or role dissatisfaction or burnout.

Click Here to see my code

Insight and Business Recommendation

Resigned Employee Career Level

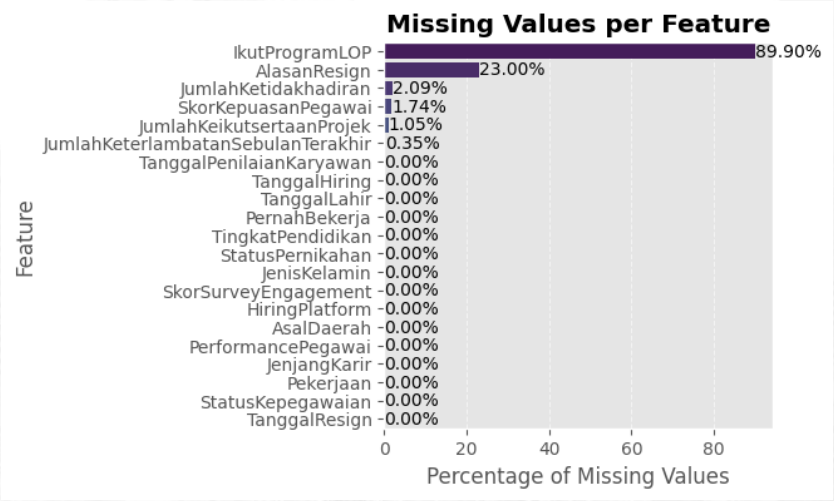


- **The Fresh Graduate Program** has the highest number of resignations compared to other career levels. However, its resignation rate (29.6%) is the lowest, indicating that although many employees resign from this group, the total number of employees in this category is also large
- **The Mid-Level career group** has the highest resignation rate (33.3%), even though the total number of resignations is lower than the Fresh Graduate group. **The Senior-Level career group** has a resignation rate of 32.9%, slightly lower than Mid-Level.
- The company should focus on retaining Mid-Level employees, as they have the highest resignation rate. Mid-level and Senior-Level employees might face career stagnation, job dissatisfaction, or a competitive job market with better external opportunities.

Click Here to see my code

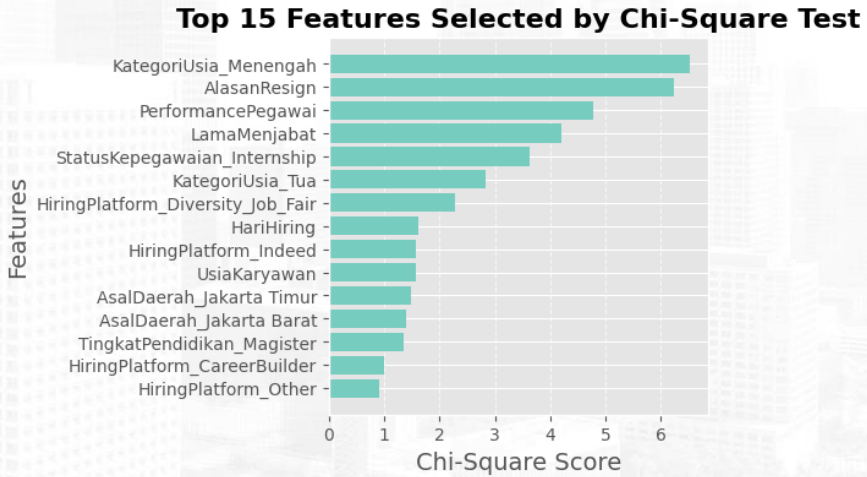
Data Preprocessing

Data Preprocessing



- When building machine learning models, we must address data skewness by **imputing missing values using the median** for 'AlasanResign', 'JumlahKetidakhadiran', 'ScoreKepuasanPegawai', and 'JumlahKeikutsertaanProjek'.
- However, 'IkutProgramLOP' **will be dropped** due to a high percentage of missing values.
- We create several columns from **extracting existed features**
- Data distribution seems normal, we can conclude **that the dataset does not have outliers**..
- Several inconsistent data values have been **tidied up**.

Feature Selection



- Since we are working with categorical features and a classification problem, **SelectKBest** with the chi² (chi-square) test is a great choice because it is fast, helps remove irrelevant features, and works well with encoded categorical data. I will **select the top 15 features** for this machine learning model.
- **Label encoding** is applied to columns with higher cardinality, while **one-hot encoding** is used for columns with lower cardinality.
- Additionally, I perform scaling using **MinMaxScaler.**

Click Here to see my code

Modelling

Model Development

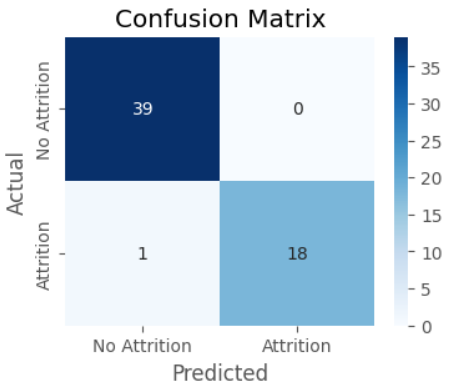
I use Support Vector Machine (SVM), Gradient Boosting, Decision Tree, Random Forest, and Logistic Regression

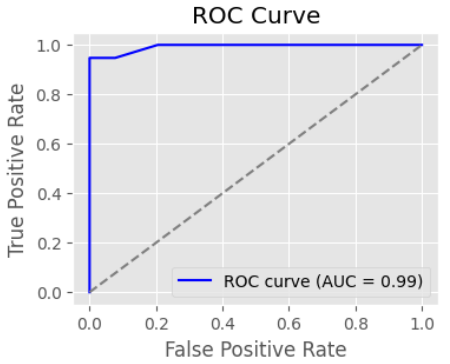
Model Evaluation

I choose **ROC AUC** in employee attrition prediction ensures that the model is:

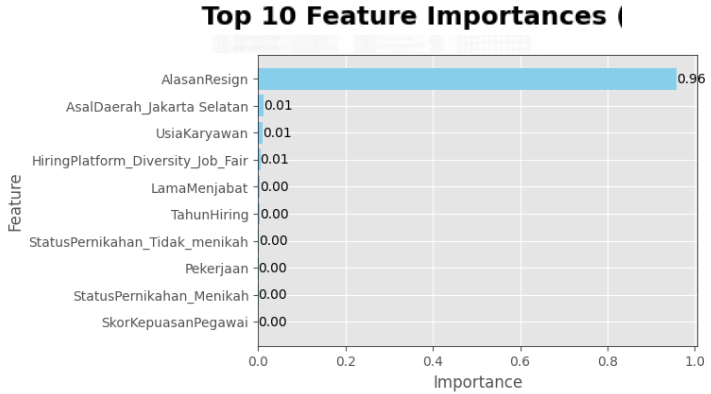
- Fair & robust across imbalanced classes.
- Optimized for real-world HR decisions with a proper trade-off between recall & precision.

Model	Accuracy	ROC AUC
Support Vector Machine	0.67	0.50
Gradient Boosting	**0.94**	**0.94**
Decision Tree	0.93	0.93
Random Forest	0.89	0.85
Logistic Regression	0.70	0.57





Conclusion



- Best model: **Gradient Boosting** with grid search hyperparameter tuning
- The test was carried out using the AUC metric with gradient boosting model resulting accuracy **0.94 and ROC AUC 0.99 after tuning**
- The high importance of **the "AlasanResign" feature suggests that the specific reasons** employees provide for their resignation play a crucial role in predicting attrition

Click Here to see my code

Business Recomendation

Company Name Profile: **XYZ Corp**
A tech company, XYZ Corp, has been struggling with high employee turnover, affecting productivity, morale, and recruitment costs. To improve retention, the company deploys a machine learning model to identify key factors leading to resignations. with AI-driven insights, the company designs targeted retention strategies:

- The model strongly relies on "AlasanResign" (Resignation Reason), meaning that resignation trends must be analyzed deeply to take proactive action. Since **working hour** is the dominant factor, some action recommend:
  - Analyze Overtime Trends: Check which departments consistently work overtime and why.
  - Survey Employees: Get feedback on workload, stress levels, and preferred work hours.
  - Compare Productivity Metrics: Measure performance vs. hours worked to find the optimal balance.
  - Implement Flexible Working Hours: Allow employees to choose a start time (e.g., 7 AM - 3 PM, 9 AM - 5 PM, 11 AM - 7 PM)
  - Implement Hybrid & Remote Work Options: Allow employees to work 2-3 days from home per week.
  - Reduce Unnecessary Meetings & Improve Time Efficiency: Limit Meetings to 30-45 Minutes, Set "No-Meeting Days",Block 1-2 days per week for deep work without interruptions.

- The model detects that employees from **Jakarta Selatan** and **mid-employees** have a higher risk of resigning.
  - Personalized Exit Interviews: Focus on mid-employee and Jakarta Selatan employees to understand their concerns.
  - "Stay Interviews": Monthly check-ins with at-risk employees
  - Internal Mentorship Programs: Pair mid-level employees with senior mentors.
- Machine learning provided valuable insights, but human intervention was key in designing effective retention strategies. XYZ Corp launches AI-powered retention programs for 6 months and tracks the impact.

Business Simulation

To prove employee retention improvement after ML implementation, we will refine the financial impact of AI-driven retention strategies at XYZ Corp.

Pre-AI (Before ML Implementation)

Turnover Rate (2018-Present):

- 30% annually (higher due to resignation spike).
- Average Hiring Cost per Employee: Rp 11,000,000 (increased due to market competition).
- Productivity Loss Per Resigned Employee: Rp 7,000,000 (longer onboarding & knowledge gaps).
- Annual Hiring Costs (Based on 2018-2020 trend):

  = (60 employees× 11,000,000) + (30 resignations×7,000,000)
  = 660,000,000 + 210,000,000
  = 870,000,000 💰

Post-AI (After ML Implementation)

- AI-driven Retention Reduces Turnover by 30% (Retains 9 more employees per year).
- Hiring Cost Savings:

  = 9 × 11,000,000
  = 99,000,000
- Productivity Retention Savings:

  = 9 × 7,000,000
  = 63,000,000
- Total Savings from AI = Rp 162,000,000 per year
- Investment in AI System

  Initial AI Cost : Rp 150,000,000
  Annual AI Maintenance Cost : Rp 30,000,000

Net Savings in First Year:

= 162,000,000 – (150,000,000+30,000,000)
=-18,000,000(Initial investment loss)

Net Savings in Second Year & Beyond:

= 162,000,000 – 30,000,000
= **Rp 132,000,000** 💰

Investing in AI for retention is a long-term strategy. The first year has a break-even period, but from year two onwards, the company sees major financial benefits while reducing workforce instability