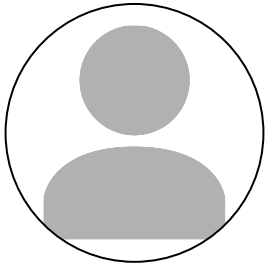
Improving Employee Retention by Predicting Employee Attrition Using Machine Learning

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



Created by:
Tsabitah Karimah
tsabitahkarimah@gmail.com
LinkedIn

"I am a dietetics graduate with a newfound interest for data science, seeking to leverage analytical skills and healthcare knowledge in a data scientist role. Completed an intensive 5-month data science course, gaining proficiency in key programming languages and machine learning techniques. Eager to apply newly acquired skills in data manipulation, visualization, and predictive modeling to extract meaningful insights from complex datasets. Combines technical acumen with strong problem-solving abilities to drive data-informed decision-making in a data scientist role."

Overview

"Human resources (HR) are the primary assets that a company must manage effectively in order to achieve its business objectives efficiently and effectively. In this instance, we will be addressing a specific HR problem within the company. Our focus will be on determining how to retain employees within the company to prevent the ballooning costs associated with employee recruitment and training for new hires. By identifying the primary factors causing employees to leave, the company can promptly address these issues by implementing programs that are relevant to the employees' concerns."

Data Preprocessing

- The data has a total of 25 columns and 287 rows.
- There are a total of 6 columns with missing values;
  - 'SkorKepuasanPegawai' has minimal missing values (5) and the overall data distribution is near normal, hence, the missing values will be filled with median as it is more robust.
  - 'JumlahKeikutsertaanProjek', 'JumlahKeterlambatanSebulanTerakhir' and 'JumlahKetidakhadiran' also have minimal missing values, (3), (1) and (6), respectively. Median will be used to fill in missing values.
  - In 'AlasanResign' columns, there are 66 missing values, which may implies to the employees that are still working in the company. The missing value (66) added up with the 'masih_bekerja' value (132), equal to 198 which is in line with the column 'TanggalResign' that has '-' values. Hence, the missing value will be filled with 'masih_bekerja'
  - 'IkutProgramLOP' has the most missing values, almost 90% of its data. Hence, the missing value will be replaced with unknown value.

- Other values
  - In column 'StatusPernikahan', there is 3 '-' values which will be replaced with the column's mode.
  - In TanggalResign' column, there is 198 '-' values which represents the employees that are still working with the company, hence, the value wont be change as of now

- Column ‘PernahBekerja’ has 286 ‘1’ value and 1 ‘yes’ value which represents the same value, hence, this column will be dropped considering that it has constant/ one unique value.

For further details, please refer to this Jupyter Notebook