The dataset was created by IBM employees and was downloaded from Kaggle. The dataset is fictional and that data does not actually represent any actual IBM employees.

Attrition: It is basically the turnover rate of employees inside an organization.

This can happen for many reasons:

Employees looking for better opportunities. A negative working environment. Bad management Sickness of an employee (or even death) Excessive working hours

The objective is to see what influences the attrition

Initially, I studied my dataset and cleaned it.

Then I plotted a box plot where It can be observed that the value ranges of columns (MonthlyIncome, MonthlyRate, EmployeeNumber, DailyRate) are significantly higher than the remaining numeric columns. This can be corrected using normalization.

So, I normalized my dataset. After which I observed that The previous percentages show that almost 84% of the employees included in the dataset did not suffer from attrition. Also, it can be observed that the data is imbalanced between the two class labels (83.8% for 'No' and 16.1% for 'Yes') of the 'Attrition' target column. Thus, there is a need to balance the sampling ratio during the training process of a classifier algorithm.

I then plotted the correlation of various variables and saw that The correlation analysi shows interesting findings First, there is a high positive correlation between the “TotalWorkingYears” column and the “JobLevel” and “MonthlyIncome”, which reflects a sort of fairness in promoting and paying people in the company based on their experience level. Second, there was a high positive correlation between “PerformanceRating” and “PercentSalaryHike” columns, which again confirms that the increase in salary is based on the increase in the performance level. Third, the

“JobSatisfaction” column does not have any correlation with the reminder of the numeric columns, which is somehow unexpected as it would be reasonable to have it increased with the increase in “MonthlyIncome” or “JobLevel” columns. Similarly I plotted various visualizations to see what all affects attrition the most.

List of resources: Kaggle, Stackoverflow, GeeksForGeeks