**Employee Attrition Classifier**

## **1. Introduction**

Employee Attrition occurs when the size of your workforce diminishes over time due to unavoidable factors such as employee resignation for personal or professional reasons. The main objective of this data mining project is to predict employee attrition. Given the set of attributes within the data set we will classify an employee who will leave the company (Attrition = Yes.)? Several attribute selection and classification algorithms will be used in combination. The combination that produces the best evaluation metrics such as Accuracy or area under the ROC curve will be chosen as the future classification module.

## **2. Dataset**

|  |  |
| --- | --- |
| Dataset Name | HR-Employee-Attrition |
| URL | https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset/data |
| Data File used | WA\_Fn-UseC\_-HR-Employee-Attrition.csv |

The data set initially contains 35 attributes and 1470 instances. A brief description of each attribute is as follows:

*Age*: Numeric value, age of the employee.

*Attrition*: Nominal Yes or No values, indicates the employee has left the company if Yes. This attribute is our class label and what we are trying to predict.

*BusinessTravel*: Nominal ‘Travel\_Rarely, Travel\_Frequently, Non-Travel’ values. Describes how often the employee travels.

*DailyRate*:Numeric: value providing daily income rate for the employee

*Department*: Nominal value ‘Sales, Research & Development, Human Resources’ providing which department the employee works in

*DistanceFromHome*: Numeric value describing the distance the employee must travel to/from work. The larger the value the further away the employee lives.

*Education*: Nominal values include ‘1:Below College, 2:Some College, 3: Bachelor, 4:Master, 5:Doctor’

*EducationField*: Nominal values include: ‘1: Life Sciences, 2: Other, 3: Medical, 4: Marketing, 5: Technical Degree, 6: Human Resources’

*EmployeeCount*: A value of 1 counts the employee tuple (this attribute is removed during preprocessing).

*EmployeeNumber*: An identifier associated with a unique employee (this attribute is removed during preprocessing).

*EnvironmentSatisfaction*: Nominal with values (1: Low, 2: Medium, 3: High, 4: Very High) describes how satisfied the employee is with their work environment.

Gender: Nominal values are either Male or Female.

*HourlyRate*: Numeric hourly payrate for employee

*JobInvolvement*: Nominal with values (1: Low, 2: Medium, 3: High, 4: Very High)

*JobLevel*: Nominal values Nominal with values 1 through 5. The higher the value the more experience/responsibility the employee has.

*JobRole*: Nominal, the job roles are ‘Sales Executive, Research Scientist, Laboratory Technician, Manufacturing Director, Healthcare Representative, Manager, Sales Representative, Research Director, Human Resources.’

*JobSatisfaction*: Nominal with values (1: Low, 2: Medium, 3: High, 4: Very High) describing how the satisfied the employee is with their job.

*MaritalStatus*: Nominal with values (‘Single, Married, Divorced’)

*MonthlyIncome*: Numeric, monthly income of the employee.

*MonthlyRate*: Numeric monthly pay rate of the employee.

*NumCompaniesWorked*: Numeric details how many companies the company has worked for in their career.

*Over18*: Nominal, binary Yes or No signifying the employee is over 18 years of age. This attribute is removed during preprocessing.

*OverTime*: Nominal, binary Yes or No, states whether the employee has worked overtime or not.

*PercentSalaryHike*: Numeric value indicating the employees last raise as a percentage of their salary.

*PerformanceRating*: Numeric Performance rating given in last review.

*RelationshipSatisfaction*: Nominal with values (1: Low, 2: Medium, 3: High, 4: Very High) describing how satisfied the employee is with their relationships at work.

*StandardHours*: Numeric number of hours the employee works. This attribute is removed in preprocessing.

*StockOptionLevel*: Numeric describing how much equity an employee was given in compensation.

*TotalWorkingYears*: Number of years the employee has been in the work force.

*TrainingTimesLastYear*: Numeric how much training the employee received last year

*WorkLifeBalance*: Nominal with values (1: Low, 2: Medium, 3: High, 4: Very High) describes the work life balance the role offers the employee

*YearsAtCompany*: Numeric, number of years the employee has been at the compnay

*YearsInCurrentRole*: Numeric, number of years the employee has been in their current role.

*YearsSinceLastPromotion*: Numeric, number of years since the employee’s last promotion

*YearsWithCurrManager*: Numeric, number of years the employee has worked with their current manager.

## **3. Problem Statement**

Employees are leaving the workforce faster than they are hired, and it is often outside the employer’s control. There are several reasons for attrition including the lack of professional growth, a hostile work environment, or declining confidence in the company’s market value. Weak leadership is another factor that often drives attrition among employees.

It will be interesting to see how much a skewed class ratio can affect how the classifiers function. There will be lot of practical problems where the positive class we want to identify is relatively rare so we would need to deal with the class imbalance in a similar manner to this project.

## **4. Data Mining Approach**

Our data mining approach will utilize the Classification Algorithms: --

J48:A classifier that builds a decision tree splitting attributes based on information gain.

Naïve Bayes: Uses Bayes Theorem to calculate probability of a tuple belonging to each class and selects the class with the highest probability. Attributes are assumed to be independent from one another.

Logistic Regression: Fits a model to predict a binary or multiclass dependent variable. Classification depends on the probability of the dependent variable given the known independent variables.

Multilayer Perceptron: Trains a neural network with back propagation to then predict unknown class attributes.

## **5. Data Transformation**

The data mining task will start by converting the raw CSV data into an .ARFF file by importing the file into the WEKA datamining tool. Once the data has been imported the next step is to correctly change attributes that have been imported as numeric values but should counted as nominal ones. Attributes such as *EducationField* despite having the values 1 through 5 should be nominal since each number represents a category such as the value 5 representing employees with a doctoral degree.