Human Capital Analytics

# Introduction

In the last decades, having the best machines was enough to be competitive or to dominate an industrial sector. Nowadays, the company that has more engaged and productive employees will have a better chance of winning market competition. For this reason, companies cannot lose important employees and when that begins to happen you need to understand why, to prevent this from happening. The Human Capital Analytics dataset (Employee.csv), is used to explain the first steps in the data analysis path. In this project, first part is presented how to get familiarize ourselves with the dataset by performing the descriptive analytics. Techniques such as exploratory data analysis (EDA) allow us to present the data in a more meaningful way, applying general statistical methods and exploratory graphics, that allow a simpler interpretation before engage in machine learning algorithms for data mining.

# The Human Capital Dataset

The Human Capital Analytics dataset (Employee.csv) is a simulated dataset from a medium size engineering company and the focus is to understand why the best and most experienced employees are leaving the company. By the exploration of this dataset its possible to extract good insights of a problems that the Human Capital department deals daily. In many industries retaining their best employees is a question of long term strategy, and can impact the company’s growth or put in financial risk, mainly if the employees leave to work for the competitor.

# Exploratory Data Analysis (EDA)

Exploratory data analysis employs a variety of techniques (mostly statistical graphics) before making inferences from data. It is essential to examine all variables in the dataset to:

* Catch mistakes or adjusting labels
* Summary Statistics interpretation
* Generate hypotheses (already generated)
* See patterns in the data through data visualizations
* Extract important variables’ features
* Detect outliers, anomalies, and missing values
* Gain deep familiarity with the dataset

Refine selection of features that will be used to build the machine learning models.

Special attention to not skip the EDA process, because can generate inaccurate models or accurate models on the wrong data. This dataset contains 14999 objects and 10 attributes described below:

|  |  |  |
| --- | --- | --- |
| Column Number | Variables | Descriptions |
| 1 | satisfaction\_level | Satisfaction Level |
| 2 | last\_evaluation | Last evaluation |
| 3 | number\_project | Number of projects |
| 4 | average\_montly\_hours | Average monthly hours |
| 5 | time\_spend\_company | Time spent at the company |
| 6 | work\_accident | Whether they have had a work accident |
| 7 | left | Whether the employee has left |
| 8 | promotion\_last\_5years | Whether had a promotion in the last 5 years |
| 9 | sales | Departments (column sales) |
| 10 | salary | Salary |

# Preprocessing the dataset

Before starting the process, it’s important to answer the following question:

* Is it clear what kind of problem we are dealing with?

because in many cases isn't so simple to identify it. A good understanding of the problem will help to choose the best machine learning algorithm then, right data mining and to make the right predictions or classifications. Thus, the first step, is preprocessing because, in real word, the raw data can be collect from many sources like sensors, websites, public data and many others.

# Variables transformations

Data transformation is the process of converting, cleansing, and structuring data into a usable format. The goal is to prepare data for analysis, reporting, storage, or data mining so it can be used to support decision making and drive organizational growth. Perform transformation if it is necessary

# Descriptive Analysis

The descriptive Analysis is used to simplify and summarize the mainly characteristics of the dataset. In other words, show what kind of information the dataset has. The R Summary method generates a descriptive statistic that summarize the central tendency, dispersion, and shape of the dataset. By using this method in Human Capital dataset, what important insights is possible to see?

# Hypothesis

Now let's extract some more information and testing some hypothesis

**First Hypothesis**

***The first hypothesis: salary is the reason why the employees left the company. Let's see if is this correct***

**Second Hypothesis**

***The second hypothesis is: employees leave the company because work is not safe.***

**Third Hypothesis**

***The third hypothesis is: this company is a good place to grow professionally.***

# Summary of the Exploratory Data Analysis

Prepare a summary of your analysis about HR data

Human Capital Modeling Requirements

# Project Requirements

Project Name: Human Capital Analytics

Dataset: Employee.csv

Method: Data Analytics Modeling

Models: Classifications and Predictions

Approach:

1. Project introduction
2. Business and analytics goals
3. Data preprocessing (such as attributes definition, data exploration, checking missing value, checking zero, and more)
4. Predictor analysis and relevancy
5. Dimension reduction (if needed)
6. Data transformation (if needed)
7. Data partitioning methods
8. Model selection
9. Model fitting, validation accuracy and test accuracy
10. Report models performance
11. Model evaluation (of the selected models)
12. Observation and conclusion

Deadline: Week 5

Delivery:

* Week 2: part 1, 2, and 3
* Week 3: part 4, 5, 6, 7, and 8
* Week 4: part 9, 10, and 11
* Week 5: complete work submission (documentation and presentation)

# Deliverable Policy

Demonstrate your analytics ability as an entry level data analytics professional. Your work should be delivered on time, late work will not be accepted. Do not share your work with other students. Avoid machine learning blunders and fundamental mistakes.