

**School of InfoComm Technology**

**Machine Learning**

Diploma in Data Science (DS)

Diploma in Information Technology (IT)

October 2022 Semester

**INDIVIDUAL ASSIGNMENT 1**

(30% of Machine Learning Module)

**Deadline for Submission:**

**17th Dec 2022 (Saturday), 2359 Hours**

|  |  |  |
| --- | --- | --- |
| Student Name | : | Tran Huy Minh |
| Student Number | : | S10223485H |
| Video Presentation Link | : |  |

**Penalty for late submission:**

10% of the marks will be deducted every day after the deadline.

**NO** submission will be accepted after 24th Dec 2022, 23:59.

Contents

[*1.* *Summary/Overview* 3](#_Toc122197933)

[*2.* *HR Analytics* 4](#_Toc122197934)

[*2.1 Problem Understanding* 4](#_Toc122197935)

[*2.2 Data Exploration* 4](#_Toc122197936)

[*2.3 Data Cleansing and Transformation* 4](#_Toc122197937)

[*2.4 Correlation Analysis* 4](#_Toc122197938)

[*3.* *Airbnb* 5](#_Toc122197939)

[*3.1 Problem Understanding* 5](#_Toc122197940)

[*3.2 Data Exploration* 5](#_Toc122197941)

[*3.3 Data Cleansing and Transformation* 5](#_Toc122197942)

[*3.4 Correlation Analysis* 5](#_Toc122197943)

[*4.* *Summary and Further Improvements* 6](#_Toc122197944)

## Summary/Overview

This assignment involves exploring and analyzing two datasets, HR analytics and Airbnb price listings, one for classification and one for regression, in order to prepare them for machine learning modeling in the next assignment. The process will involve using visualization and statistical techniques to familiarize oneself with the datasets, cleansing and transforming the data by handling missing values and outliers, transforming categorical data into numerical data, and potentially scaling the data. It also involves conducting a correlation analysis to identify relationships between features and determining which features will be helpful for making predictions, possibly creating or dropping features as needed. The final step is to export the cleaned and transformed data into new CSV files to be used in the next assignment.

## HR Analytics

## 2.1 Problem Understanding

The problem of identifying employees who are likely to be promoted is an important one for human resource departments, as it allows for more efficient use of resources and better overall results. In the past, this has been a manual process with a number of limitations. With the availability of the hr\_data.csv dataset, which includes a range of personal, educational, and performance-related information about employees, there is an opportunity to use machine learning modeling to improve the accuracy and efficiency of this process.

The value of this problem lies in its potential to improve the effectiveness of human resource departments. By accurately predicting which employees are most likely to be promoted, HR can focus their efforts on supporting and developing these individuals, ultimately leading to better retention and productivity within the organization. Additionally, this prediction can inform HR decisions about hiring and training, allowing them to make more informed choices about how to allocate resources. Overall, a successful solution to this classification prediction problem has the potential to greatly benefit the organization.

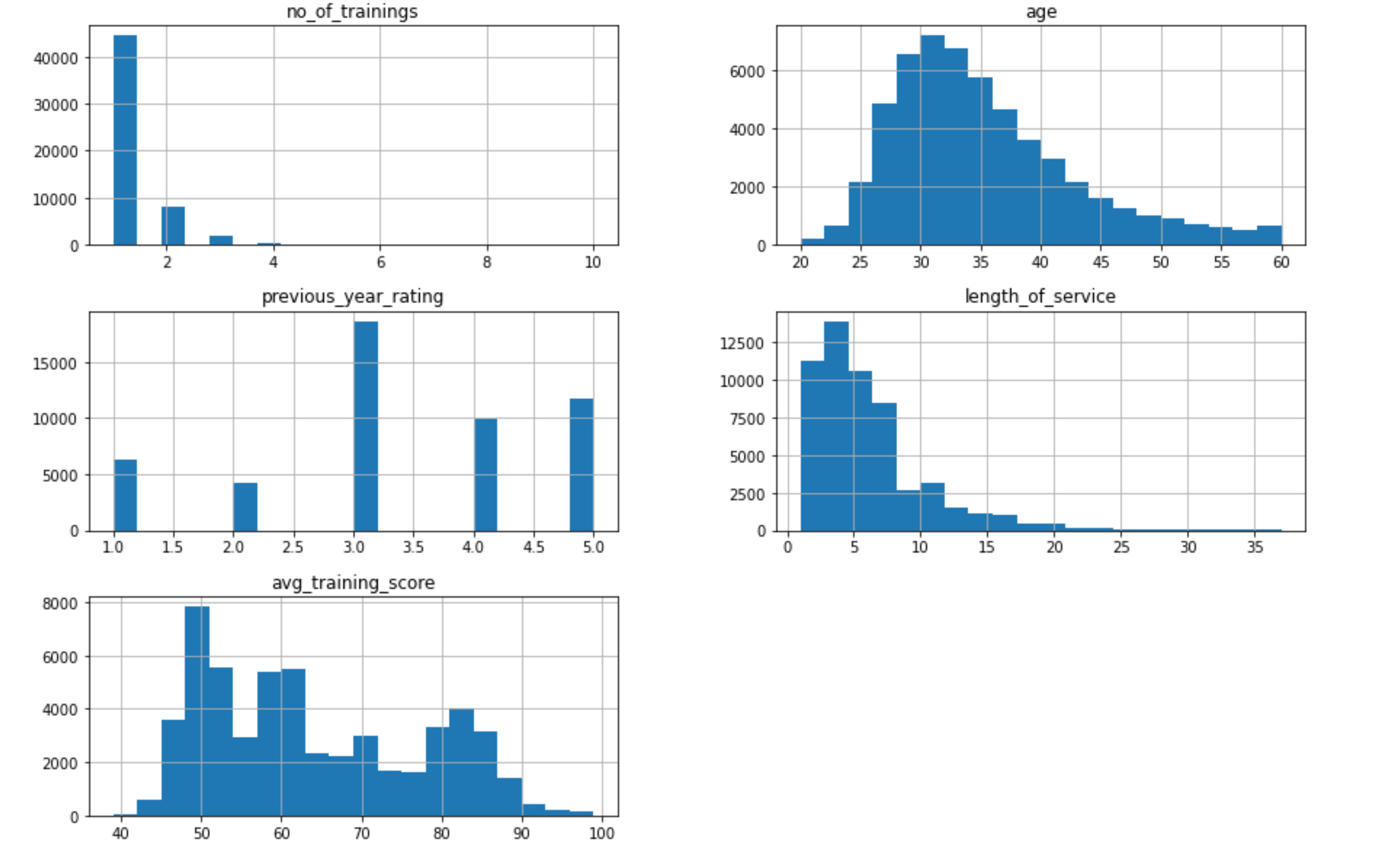
## 2.2 Data Exploration

Before performing any data cleansing and transformation techniques, we must first explore the data to get a sense of its structure, distribution, and any potential issues or inconsistencies.

The HR analytics dataset has 14 features along with 54807 rows of data, each representing an employee based on their employee\_id. The is\_promoted column, defined as the recommendation for the employee’s promotion, is a binary feature and will act as our target variable.

After exploring the count of each unique value inside all the categorical features, I noticed that the education feature is split into Bachelor’s, Master’s & above, and Below Secondary. This is worth taking note of as the employee’s education can be encoded in an ordered ordinal manner later on.

Alongside this, the target variable, is\_promoted does not have equal\_sized samples with 50140 rows representing no promotion recommendation and 4668 representing a promotion recommendation. In general, having equal-sized samples is beneficial for training a machine learning model as it can help to reduce bias and improve the model’s ability to generalize new data. When samples are not equal in size, the model may be more influenced by the larger class, which can lead to poor performance. This can be combatted with by sampling the data to obtain a more equal distribution which will be done in assignment 2.



## 2.3 Data Cleansing and Transformation

## 2.4 Correlation Analysis

## Airbnb

## Problem Understanding

Predicting the rental price of listed properties on Airbnb is a problem that has significant value for both hosts and guests. For hosts, accurately pricing their properties can help to optimize revenue and attract potential renters. On the other hand, guests benefit from knowing the price of a potential rental in advance, allowing them to plan their travels and budget accordingly.

The listings.csv dataset provides a wealth of information about the listed properties, including host and location details, property conditions, and reviews. By utilizing this data, it is possible to create a regression prediction model that can accurately predict the daily rental price of a given property. This solution has the potential to greatly benefit both hosts and guests, allowing them to make more informed decisions about their Airbnb experiences.

## 3.2 Data Exploration

## 3.3 Data Cleansing and Transformation

## 3.4 Correlation Analysis

## Summary and Further Improvements