REPORT

Name: Thenujan N. **Index no: 200647R**

1. Problem Overview:

Marvelous Construction, a major construction firm in Sri Lanka, has been experiencing a high rate of employee resignations. The Human Resources department has recognized the need to analyze employee data to understand the underlying reasons for attrition. As the data scientist, the objective is to perform an in-depth analysis of the provided dataset, which includes employee details, attendance records, leave records, and salary information. By leveraging this data, the aim is to gain insights into the factors contributing to employee attrition and provide recommendations to address the issue.

2. Dataset Description:

The dataset provided for analysis consists of the following files:

- employee: This file contains 997 records with details such as Employee_No, Employee_Code, Name, Title, Year_of_Birth, Gender, Religion_ID, Marital_Status, Designation_ID, Date_Joined, Date_Resigned, Reporting_emp_1, Reporting_emp_2, Employment_Category, Employment_Type, Religion, and Designation. The Date_Joined and Date_Resigned fields will be crucial for training and testing attrition prediction models.
- leaves: This file consists of 1019 records and includes information on Employee_No, leave_date, Type (Half day/Full Day), Applied Date, Remarks, and apply_type (Annual/Casual). The data provides insights into employee leave patterns, including the frequency and type of leaves.
- salary: This file contains 9036 records and provides details on Employee_No, Amount, month, year, and various factors related to monthly additions and deductions. The data offers an overview of employee salaries, including any additional payments or deductions made.
- attendance: This file consists of 224058 records and includes fields such as id, project_code, date, out_date, employee_no, in_time, out_time, Hourly_Time, Shift_Start, and Shift_End. It captures employee attendance information, including their arrival and departure times, hours worked, and any late minutes compared to the scheduled shift start time.

Column	Data Type	Description
Employee_No	nominal	Employee's unique identification number
Employee_Code	nominal	Employee's unique code
Name	nominal	Employee's name
Title	nominal	Employee's job title
Gender	nominal	Employee's gender
Religion_ID	nominal	ID for employee's religion
Marital_Status	nominal	Employee's marital status
Designation_ID	nominal	ID for employee's job designation
Date_Joined	nominal	Date when the employee joined the company
Date_Resigned	nominal	Date when the employee resigned
Status	nominal	Employee's status
Inactive_Date	nominal	Date when the employee became inactive
Reporting_emp_1	nominal	Employee's first reporting manager
Reporting_emp_2	nominal	Employee's second reporting manager
Employment_Category	nominal	Category of employee's employment
Employment_Type	nominal	Type of employment
Religion	nominal	Employee's religion
Designation	nominal	Employee's job designation

Column	Data Type	Description
Year_of_Birth	descrete	Employee's year of birth
Year_Joined	descrete	Year when the employee joined the company
Experience	descrete	Total experience of the employee in years
Active_Months	descrete	Total number of active months for the employee
Leaves_Count	descrete	Number of leaves taken by the employee
Experience_Months	descrete	Total experience of the employee in months
months_resigned	descrete	Number of months the employee worked before resigning
months_joined	descrete	Number of months the employee has been with the company
average_leaves_per_m	continuous	Average number of leaves taken per month by the employee
Net Salary_x	continuous	Net salary of the employee
Total Deduction	continuous	Total deductions from the salary
Hourly_Time	continuous	Hourly time of the employee
Late_Minutes	continuous	Minutes of lateness of the employee

3. Data Pre-processing:

The provided dataset underwent several pre-processing steps to ensure data quality and prepare it for analysis. The following steps were performed:

- Handling Inconsistent Reporting Employee Numbers: In order to address inconsistent reporting employee numbers, the code identified rows where the "Employee_No" matched the "Reporting_emp_1" and replaced the corresponding entries in the "Reporting_emp_1" column with NaN values. This ensured consistency in the reporting hierarchy.

- Managing Resignation Dates: Certain values in the "Date_Resigned" column had special markers such as '\N' and '0000-00-00'. These markers were replaced with the corresponding values from the "Inactive_Date" column to ensure accurate and consistent information.

- Standardizing Titles based on Gender: The code addressed inconsistencies in the "Title" column by standardizing titles based on gender. If the title was 'Mr' or 'Miss' and the gender was 'Female', it was changed to 'Ms'. Similarly, if the title was 'Ms' or 'Miss' and the gender was 'Male', it was changed to 'Mr'. This ensured uniformity in title representation.

- Converting Data Types and Calculating Experience: The "Year_of_Birth" column was converted to numeric type, with any non-numeric values replaced with NaN. The "Date_Joined" column was converted to datetime type, and the corresponding year was extracted as "Year_Joined". The "Experience" column was then calculated by subtracting "Year_Joined" from 2022, providing insight into the number of years of experience.

- Handling Missing Values: Rows with "Year_of_Birth" values equal to '0000' were considered as missing values and replaced with NaN. Missing values in the dataset were imputed using a decision tree classifier model trained on non-missing values. The missing values in the "Year_of_Birth" column were predicted based on the features "Experience", "Designation", and "Marital_Status". Similarly, missing values in the "Marital_Status" column were predicted using the features "Year_of_Birth" and "Gender.

- Handling Data Quality Issues: Duplicate rows were dropped from the dataset to avoid biases and ensure each employee's information was represented only once.

- Changing the N/A values to nan: This process ensures that missing or unknown values in the specified columns are treated as NaN, which is a standard representation for missing data in Pandas. Once the replacements are done, the 'final' DataFrame will have NaN values in the respective columns wherever '\N' was present.

- Aggregating Data: I used other csv files such as ‘salary.csv’, ‘leaves.csv’ and ‘attendance.csv’. I grouped them using employee_no and mapped them into the employee_df dataframe.

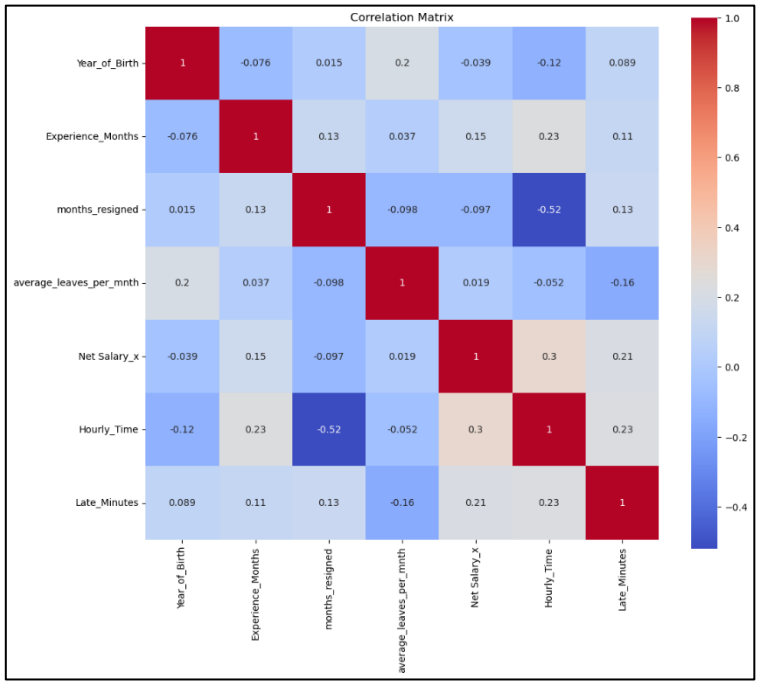
- Handling Outliers: As I have introduced several numerical features from other csv files. I removed the outliers using the z-score method.

- Data Transformation: To validate the data quality I used decision tree classifier. For that I encoded the categorical variables using One Hot Encoder. I encoded Status using the Label Encode.

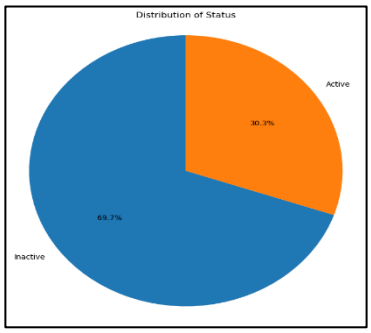
4. Insights from data analysis:

I performed descriptive analysis and exploratory analysis to gain insights about the given data. The following is the summaries of some features.

	Year_of_Birth	Experience_Months	months_resigned	average_leaves_per_mnth	Net Salary_x	Hourly_Time	Late_Minutes
count	713.000000	497.000000	497.000000	57.000000	626.000000	713.000000	713.000000
mean	1983.242637	16.291751	12.350101	1.225858	41124.183996	1904.710000	16.602312
std	13.212663	29.354435	5.196362	0.835593	36160.480571	1806.959059	30.602252
min	1946.000000	1.000000	1.000000	0.105263	280.000000	0.000000	0.000000
25%	1974.000000	4.000000	9.000000	0.500000	22234.878409	511.600000	0.045118
50%	1986.000000	8.000000	11.000000	1.000000	31717.260000	1323.510000	3.930292
75%	1994.000000	19.000000	15.000000	1.578947	47734.726355	2727.890000	20.508475
max	2006.000000	407.000000	30.000000	4.200000	319108.330000	7942.010000	262.471429



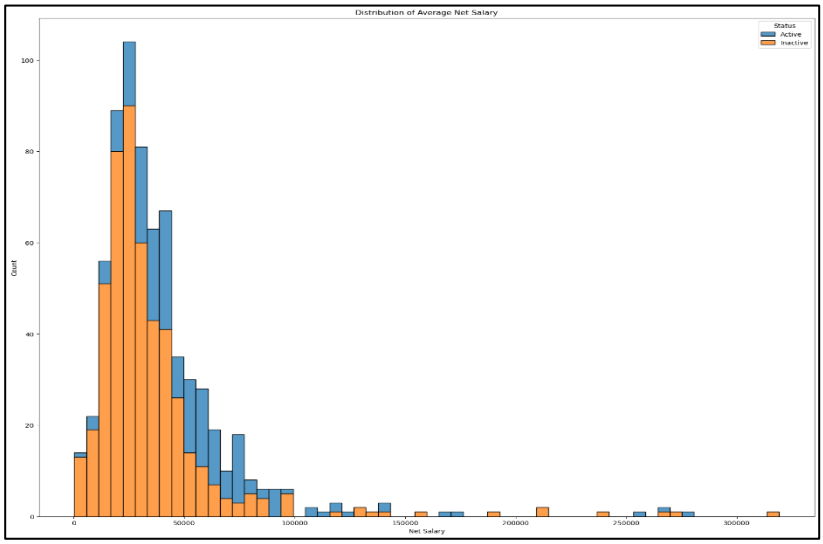
Correlation between Numerical features.



Distribution of Status

Insight 01

- There is a significant relationship between Net salary and the attrition rate. Employees with less salary are more likely to leave the job.



I used the data from the salary.csv file to get the average salary for an employee. I grouped the salary data using employee no and found the mean of the Net salary. I used the salary and visualized using a stacked bar chart. This plot is the output.

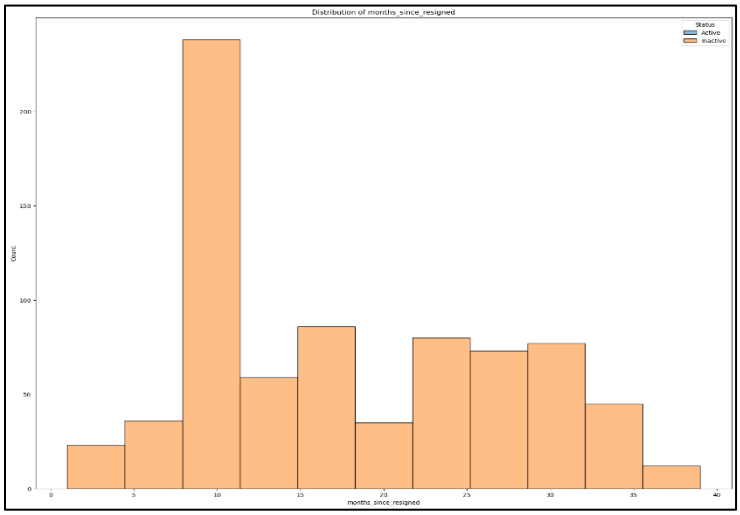
From the bar chart we can see that employees with less salary more likely to leave the job.

```
Two-sample t-test:
t-statistic: 9.580251974060234
p-value: 7.43509953960341e-21
```

Then I tested the relationship between the Net salary and Status using Two-sample t-test. The output of the test is shown below. From the p-value we can say that there is a significant relationship between Net salary and Status.

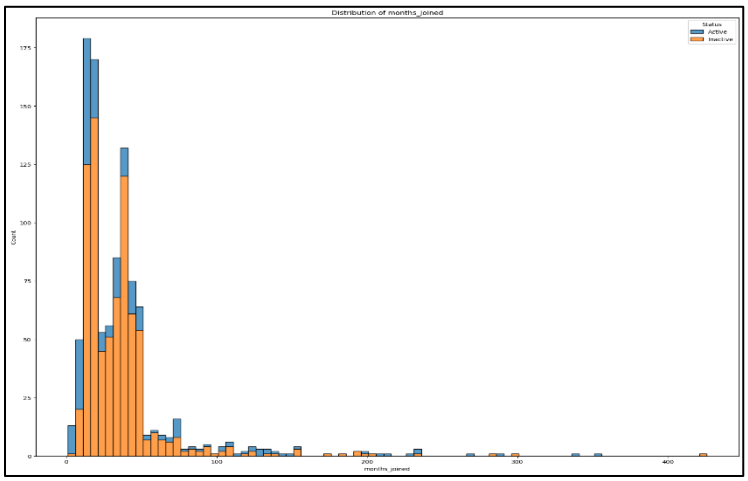
Insight 02

- Most employees are hired in the last 4 years. Employees started leaving the company in the last 3 years and most employees left the job as soon as they joined the job.

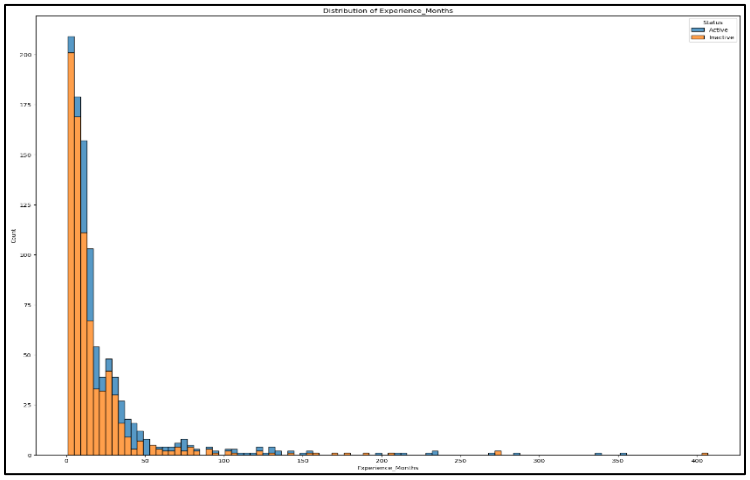


Distribution of no of months since employee resigned. (Only Inactive employees)

I made a plot showing how many months since the employees resigned. This plot explains it well. I selected all the Inactive employees and calculated the number of months since they have resigned. From the following plot we can see that employees were started leaving the company in the last 3 years. There are no resignations before that as per the data.



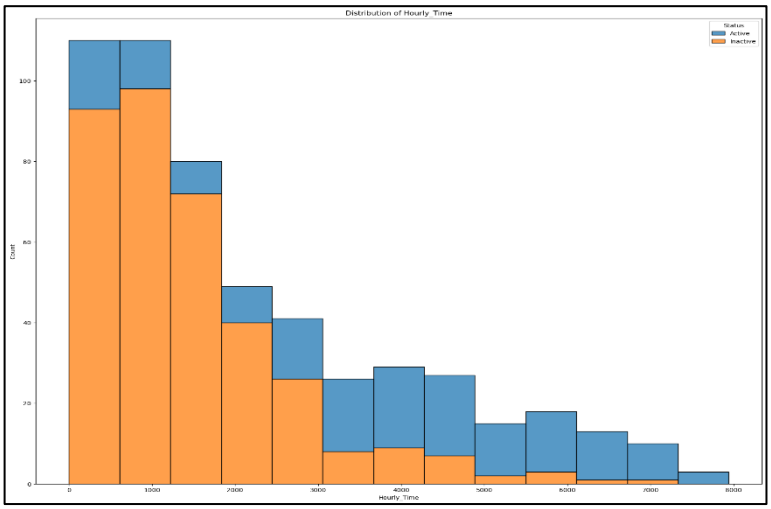
This plot shows the distribution of number of months since employee joined. In this plot, we can see that most of the employees are hired in the last 4 years. And they are more likely to leave the company.



The plot shows the number of months an employee worked. As you can see most of the employees who are Inactive, left the company within a short period of time. Most employees left the job within a year.

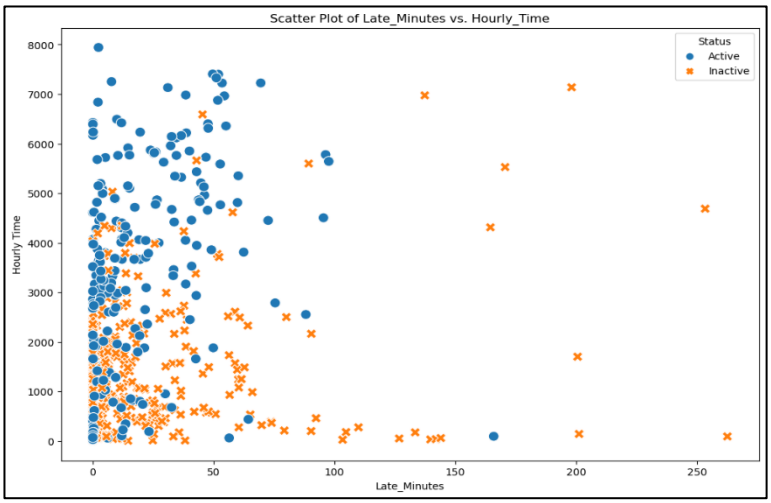
Insight 03

- Employees who have large number of working hour and a smaller number of late minutes are less likely to leave the job.



I calculated the late minutes and got Hourly Time by grouping by employee no from attendance.csv file. Then I merged those with the employee data frame.

This plot also shows that employees who worked for many hours are less likely to leave the job. And employees who worked for few hours are more likely to leave the job.



This plot shows that employees who worked for large number of hours and have few minutes as late minutes are less likely to leave the job. And employees who have less late minutes are more likely to leave the job.

Insight 04

- Contract basis employees are more likely to leave the job than Permanent employees.

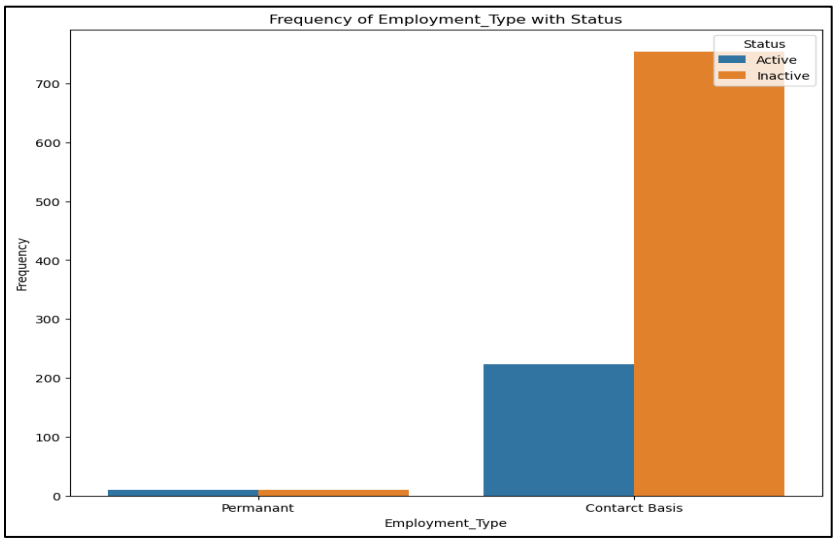
Here is a contingency table and a column wise normalized contingency table. We can see that most of the employees are in contract basis and attrition percentage of Contract Basis is greater than Permanent.

Employment_Type	Contarct Basis	Permanant
Status		
Active	223	10
Inactive	754	10

Normalized Contingency table for Status vs Employment_Type:

Employment_Type	Contarct Basis	Permanant
Status		
Active	22.824974	50.0
Inactive	77.175026	50.0

Here is a Bar chart showing the distribution of Employment Type.



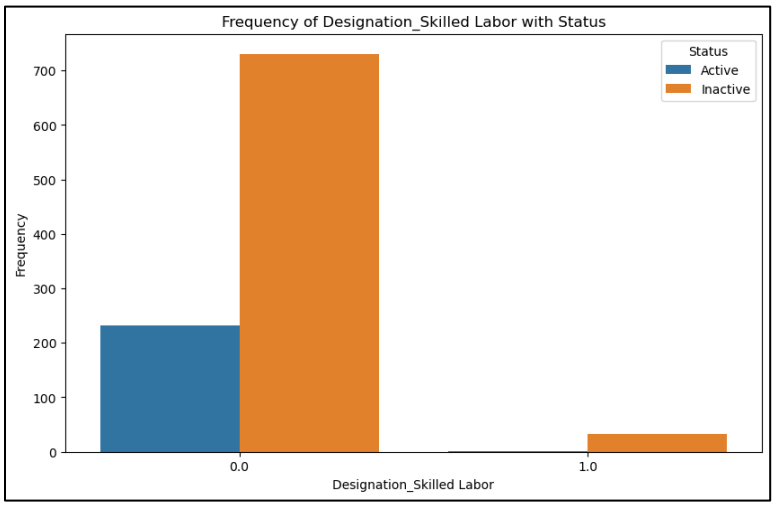
Then I tested the relationship between Employment type and Status using chi-squared test. The output of the test is shown below. From the p-value we can say that there is a significant relationship between Employment Type and Status.

```
Cross-tabulation between Status and Employment_Type:

Employment_Type  Contarct Basis  Permanant
Status
Active                      223         10
Inactive                    754         10

Chi-square Test for Independence (Status vs Employment_Type):
Chi-square statistic: 6.635632103042584
p-value: 0.009995872040024376
```

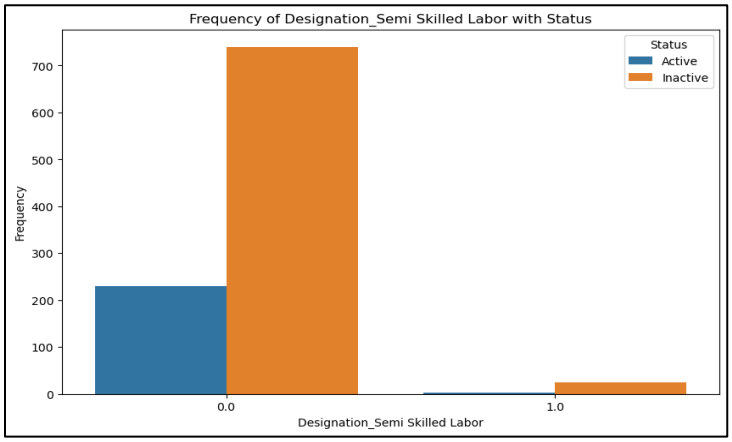
Insight 05

- Attrition rate of the skilled labors is greater than unskilled labors.

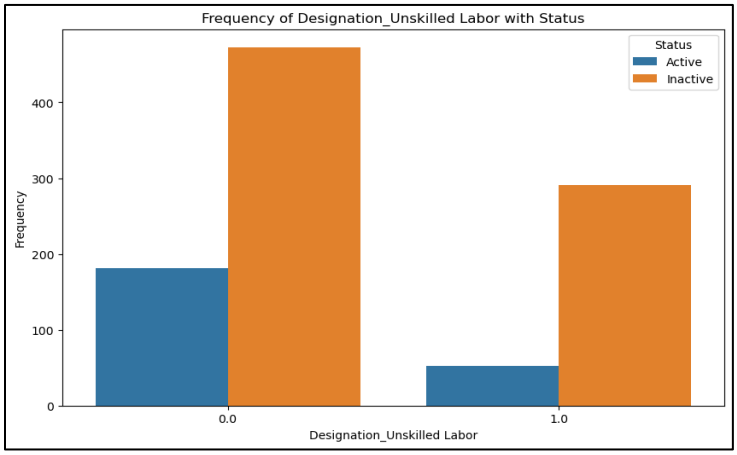
There are many Designation types. Most of the employees are Labors. There are three types of labors which are Skilled, Semi-skilled and Unskilled. I used OneHotEncoder to encode the Designation. The following charts show the attrition difference between the types of labor Employees vs other types of employees.



Skilled labor employee vs other employees. (1 represents the Skilled employee, 0 represents other employees)



Semi-Skilled labor employee vs other employees. (1 represents the Semi-Skilled employee, 0 represents other employees)



Unskilled labor employee vs other employees. (1 represents the Unskilled employee, 0 represents other employees)

05. Predictive analytics to validate the data quality.

I used decision tree classifier to validate the quality of the pre-processed data set. The following is the result.

(0 – Active, 1 – Inactive)

```
Accuracy: 0.95
Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.95      0.88        37
           1       0.99      0.95      0.97       163

    accuracy                           0.95       200
   macro avg       0.90      0.95      0.92       200
weighted avg       0.96      0.95      0.95       200
```