INDIVIDUAL ASSIGNMENT

TECHNOLOGY PARK MALAYSIA

CT045-3-M-ABAV-L-3

APPLIED MACHINE LEARNING

ASSIGNMENT (PART -C)

HAND OUT DATE: 4 OCTOBER 2021

HAND IN DATA: 20 DECEMBER 2021

INSTRUCTIONS TO CANDIDATES:

1 Submit your assignment at the administrative counter.

2 Students are advised to underpin their answers with the use of references (cited using the Harvard Name System of Referencing).

3 Late submission will be awarded zero (0) unless Extenuating Circumstances (EC) are upheld.

4 Cases of plagiarism will be penalized.

5 The assignment should be bound in an appropriate style (comb bound or stapled).

6 Where the assignment should be submitted in both hardcopy and softcopy, the softcopy of the written assignment and source code (where appropriate) should be on a CD in an envelope / CD cover and attached to the hardcopy.

7 You must obtain 50% overall to pass this module.

1.0 Abstract

Many businesses are beginning to benefit from the use of cutting-edge technology, where data management and analysis are proving to be a strategic advantage in terms of effectiveness and market competitiveness. Artificial intelligence is beginning to affect organization decisions regarding their employees, based on the evaluation of objective data rather than subjective considerations, as it becomes more widely adopted in the sales and marketing sectors (Gupta, Fernandes and Jain, 2018). Finding dedicated staff is difficult for any organization, but finding replacements is even more challenging. Not only does this increase HR costs, but it also lowers the company's market worth. In this study, we used SAS Enterprise Miner (EM) to analyze the IBM Employee Attribution dataset of 1470 records to identify the primary reasons why employees quit a company by identifying top attributes that significantly led to attrition.

1.1 Aim

The goal was to compare the outcomes of two distinct model techniques and choose the optimal model that produced the most reliable predictions on the training and validation data. For this project, we will be performing decision tree and HP forest models to predict the attrition rates in IBM.

2.0 Exploratory Data Analysis

2.1 Description of Dataset

For this project, we examined the IBM HR Analytics Employee Attrition dataset, which comprises 1470 observations and 35 features that are publicly available on the Kaggle website, for this research (Pavansubhash, 2017). There are 34 independent features in the dataset that may or may not contribute to attrition, as well as one targeted variable, the 'attrition' variable shown in table 1.

Table 1: Description of Dataset

Features	**Description**
Age	Age of the employees
Attrition	Employees leaving the company (No/Yes)
Business Travel	'Non- Travel', 'Travel Frequently', 'Travel Rarely'
Daily Rate	Daily salary
Department	'Human Resources', 'Research and development', Sales'
Distance from home	Distance between work and home
Education	1- Below college 2- College 3- Bachelor 4- Master 5- Doctorate
Education Field	'Human Resources', 'Life Science' 'Medical', 'Other', 'Marketing', 'Technical Degree'
Employee Count	Number of Employees
Employee Number	Employee ID
Environment satisfaction	1- Low 2- Medium 3- High 4- Very high
Gender	'Female', 'Male'
Hourly Rate	Hourly salary
Job Involvement	1- Low 2- Medium 3- High 4- Very high

Job Level	Numerical value
Job Role	'Lab Technician', 'Healthcare Representative', 'Manufacturing Director', 'Human Resources', 'Manager', 'Research Director', 'Research Scientist', 'Sales Executive', 'Sales Representative'
Job Satisfaction	1- Low 2- Medium 3- High 4- Very high
Marital Status	'Divorced' 'Married', 'Single'
Monthly Income	Income of employee on monthly basis
Monthly Rate	Rate of the employee on monthly basis
Num Companies Worked	Number of companies previously worked at
Over18	Employees with age over 18
Overtime	Yes/No
Percent Salary Hike	Percentage increase in salary
Performance Rating	1- Low 2- Good 3- Excellent 4- Outstanding
Relationship Satisfaction	1- Low 2- Medium 3- High 4- Very high
Standard Hours	Standard working hours
Stock Option Level	Stocks of the company owned by the employee
Total Working Years	Numbers of years worked
Training Times Last Year	Hours spent training
Work-Life Balance	1- Bad 2- Good 3- Better 4- Best
Years at Company	Total number of years at the company
Years in Current Role	Number of years with the same role
Years since Last Promotion	Number of years since last promotion
Years with Current Manager	Years spent with current manager

2.3 Data Exploration

Data exploration is an important approach in the analytical operation of data to understand and identify underlying patterns in the dataset. In this work, we have conducted statistical and visualization methodologies to study the data and emphasize important aspects leading to attrition on both demographic and job-related information. In this study, we used statistical and visualization approaches to analyze the data and focus on key factors that contribute to attrition in both demographic and job-related data.

2.3.2 Graph Analysis

The continuous and categorical variables were examined in this section. The distribution of each continuous variable was examined for the continuous variables. The categorical variables, on the other hand, were examined using a frequency distribution to better explain their distribution. We used the bar chart as a visualization tool for the study.

Data Role=TRAIN Variable Name=BusinessTravel

Target	Target Level	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
Attrition	No	4	3	Travel_Rarely	71.94	Travel_Frequently	16.87
Attrition	Yes	3	0	Travel_Rarely	65.82	Travel_Frequently	29.11
OVERALL		4	3	Travel_Rarely	70.95	Travel_Frequently	18.84

Figure 1: Summary Statistics of Business Travel

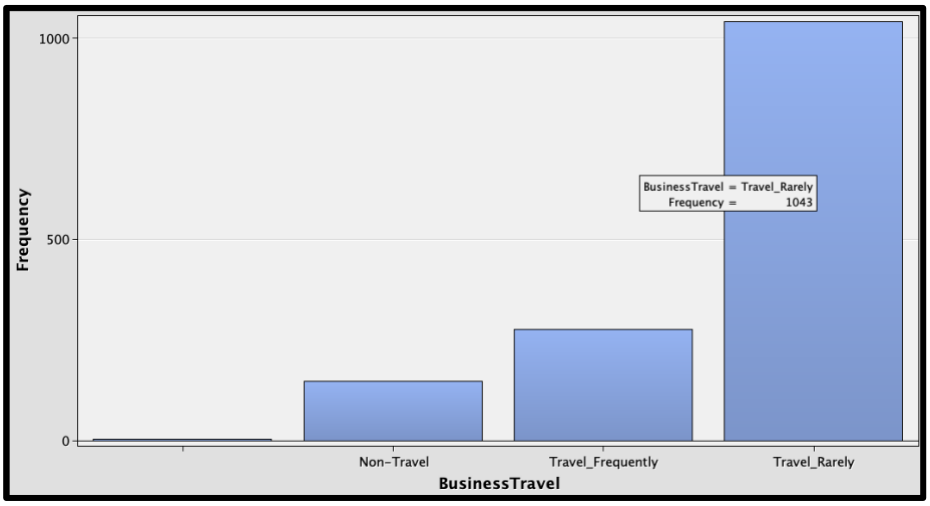

Figure 2: Histogram of Business Travel

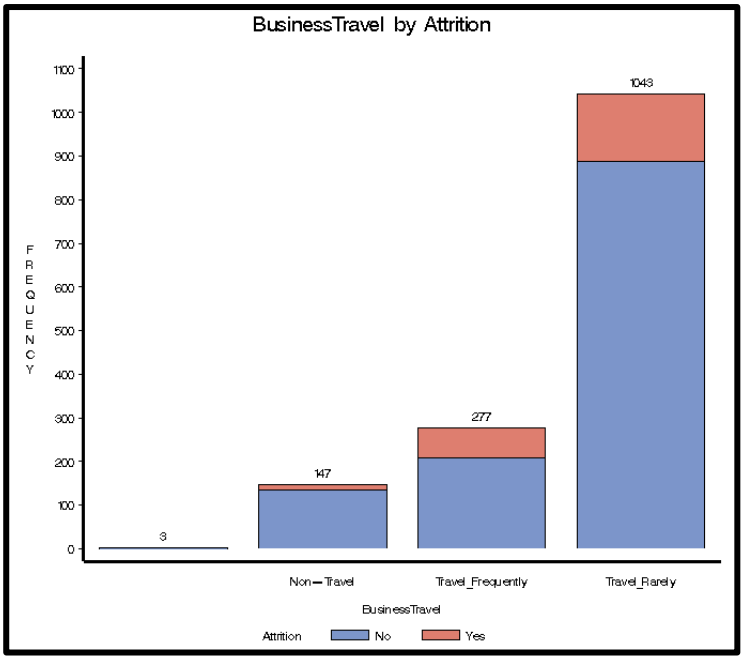

Figure 3: Histogram of Business Travel by Attrition

From figure 1, we can deduce that there are overall 3 missing values from the age variable. Moreover, we can say that 1043 (70.95%) employees in the company rarely travel. Furthermore, in figure 3, we can also observe that 65.82% of the employees that left reported also traveling rarely.

Data Role=TRAIN Variable Name=Department

Target	Target Level	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
Attrition	No	3	0	Research & Development	67.15	Sales	28.71
Attrition	Yes	3	0	Research & Development	56.12	Sales	38.82
OVERALL		3	0	Research & Development	65.37	Sales	30.34

Figure 4: Summary Statistics of Department

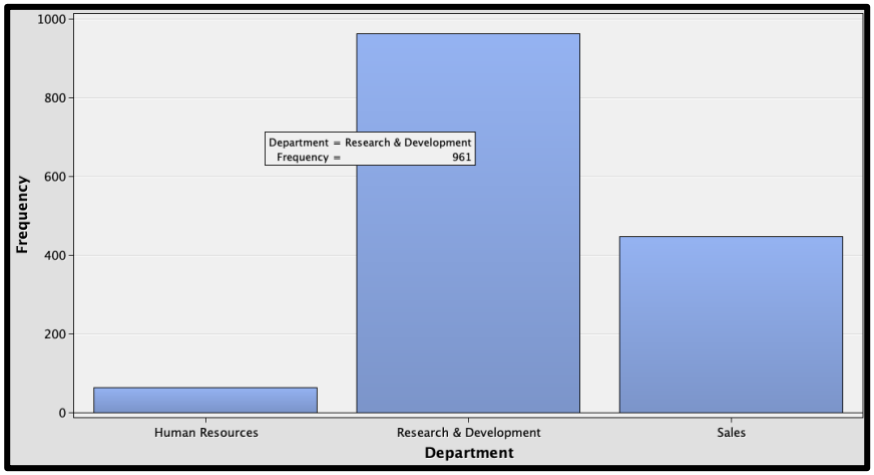

Figure 5: Histogram of Department

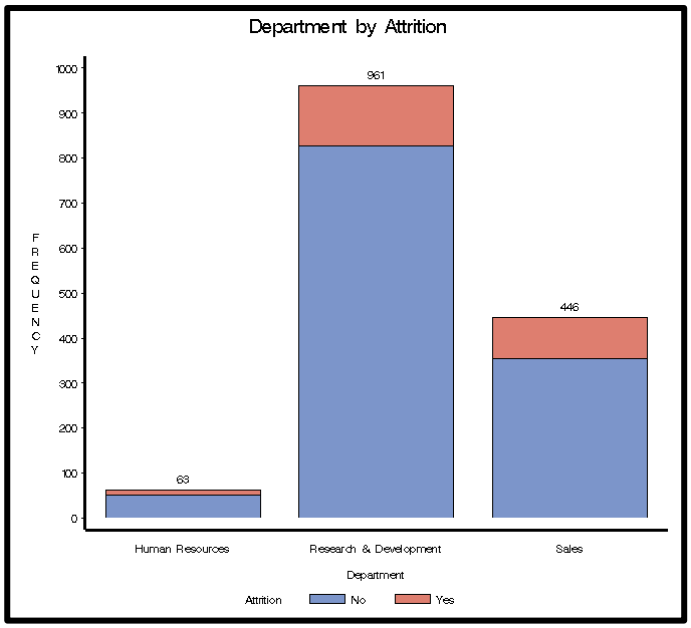

Figure 6: Histogram of Department By Attrition

From figure 4, we can deduce that there are no missing values from the department variable. Moreover, we can say that 961 (65.37%) employees are from the research and development department followed by 446 (30.34%) employees from the sales department. Furthermore, in figure 6, we can also observe that 56.12% of the employees that left were from the research and development department.

Data Role=TRAIN Variable Name=EducationField

Target	Target Level	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
Attrition	No	7	2	Life Sciences	41.93	Medical	32.52
Attrition	Yes	7	1	Life Sciences	37.55	Medical	26.58
OVERALL		7	3	Life Sciences	41.22	Medical	31.56

Figure 7: Summary Statistics of Education Field

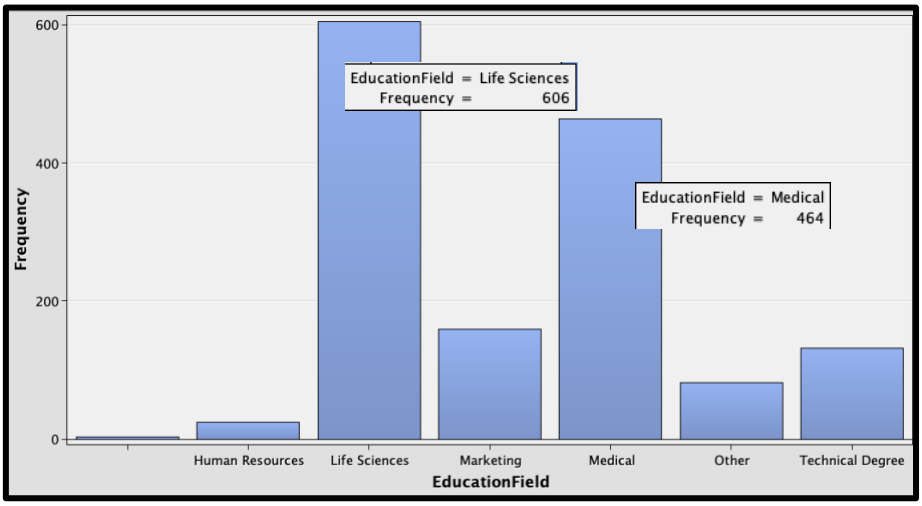

Figure 8: Histogram of Education Field

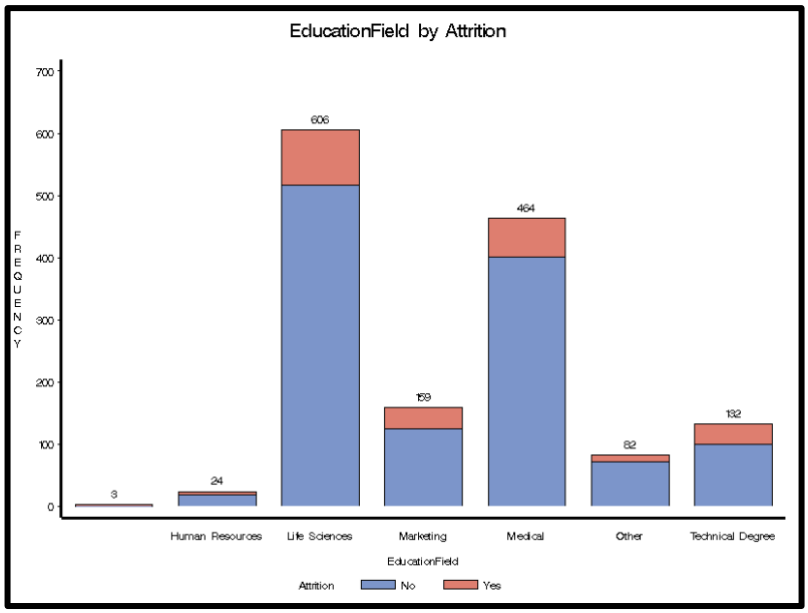

Figure 9: Histogram of Education Field by Attrition

From figure 7, we can deduce that there are 3 missing values from the education field variable. Moreover, we can say that 606 (41.22%) employees are from the life sciences field followed by 464 (31.56%) employees from the medical field. Furthermore, in figure 9, we can also observe that 37.55% of the employees that left were from the life sciences background.

Data Role=TRAIN Variable Name=Gender

Target	Target Level	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
Attrition	No	2	0	Male	59.37	Female	40.63
Attrition	Yes	2	0	Male	63.29	Female	36.71
OVERALL		2	0	Male	60.00	Female	40.00

Figure 10: Summary Statistics of Gender

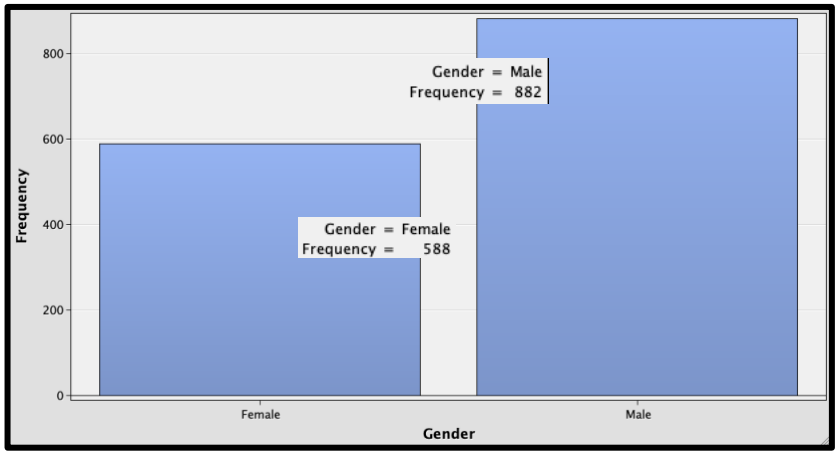

Figure 11: Histogram of Gender

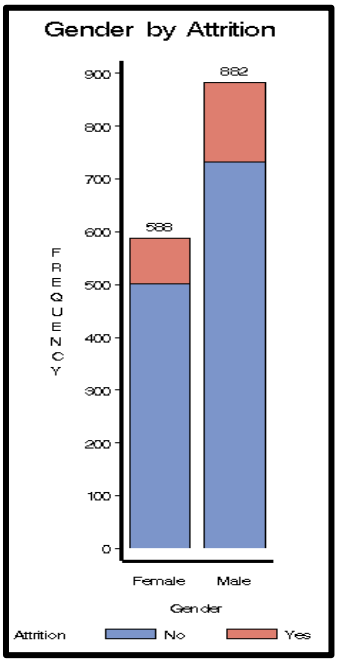

Figure 12: Histogram of Gender by Attrition

From figure 10, we can deduce that there are no missing values from the gender variable. Moreover, we can say that 882 (60%) employees were male followed by 558 (40%) employees who were female. Furthermore, in figure 12, we can also observe that 63.29% of the employees that left were male.

Data Role=TRAIN Variable Name=JobRole

Target	Target Level	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
Attrition	No	9	0	Sales Executive	21.82	Research Scientist	19.87
Attrition	Yes	9	0	Laboratory Technician	26.16	Sales Executive	24.05
OVERALL		9	0	Sales Executive	22.18	Research Scientist	19.86

Figure 13: Summary Statistics of Job Role

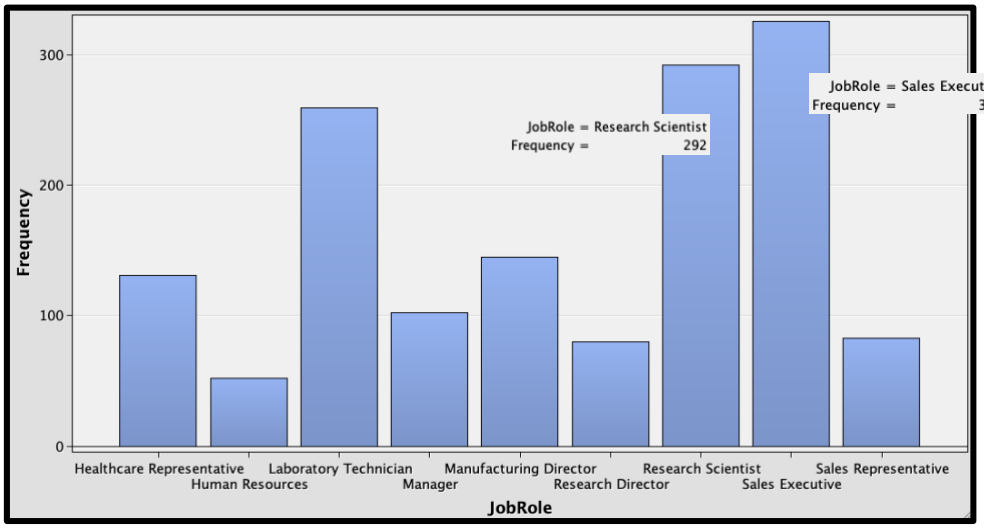

Figure 14: Histogram of Job Role

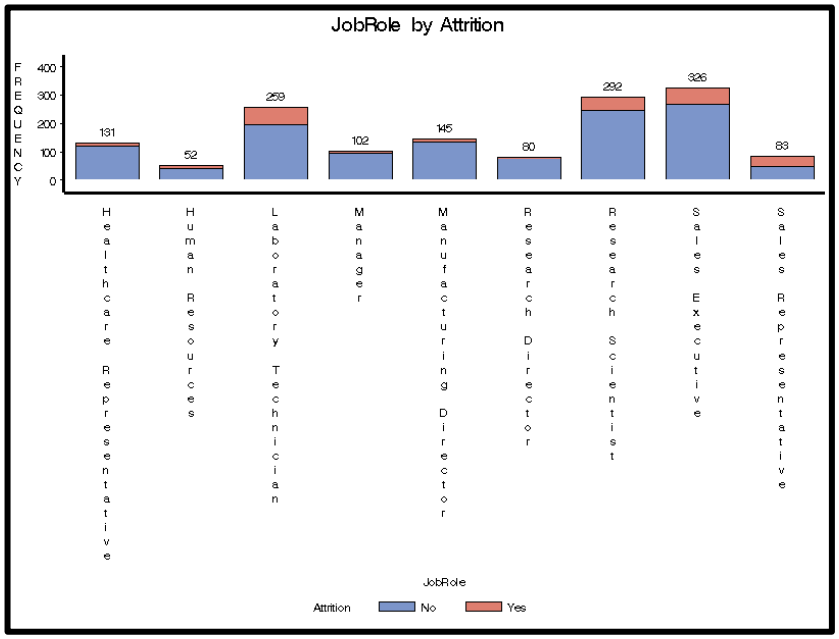

Figure 15: Histogram of Job Role by Attrition

From figure 13, we can deduce that there are no missing values from the job role variable. Moreover, we can say that 326 (22.18%) employees are sales executives followed by 292 (19.86%) employees who were research scientists. Furthermore, in figure 15, we can also observe that 26.16% of the employees that left were laboratory technicians.

Data Role=TRAIN Variable Name=MaritalStatus

Target	Target Level	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
Attrition	No	3	0	Married	47.77	Single	28.39
Attrition	Yes	3	0	Single	50.63	Married	35.44
OVERALL		3	0	Married	45.78	Single	31.97

Figure 16: Summary Statistics of Marital Status

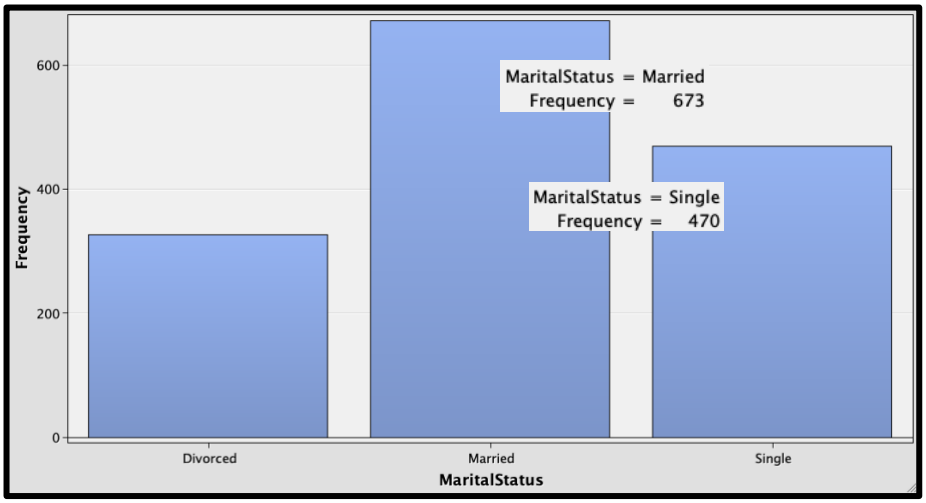

Figure 17: Histogram of Marital Status

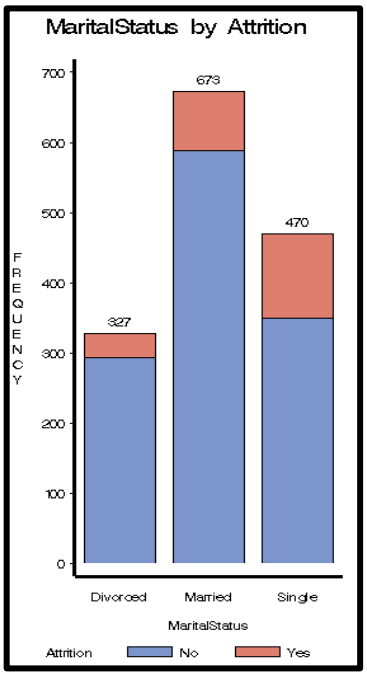

Figure 18: Histogram of Marital Status

From figure 16, we can deduce that there are no missing values from the marital status variable. Moreover, we can say that 673 (45.78%) employees are married followed by 470 (31.97%) employees who were single. Furthermore, in figure 18, we can also observe that 58.63% of the employees that left were single.

Data Role=TRAIN Variable Name=OverTime

Target	Target Level	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
Attrition	No	2	0	No	76.56	Yes	23.44
Attrition	Yes	2	0	Yes	53.59	No	46.41
OVERALL		2	0	No	71.70	Yes	28.30

Figure 19: Summary Statistics of Overtime

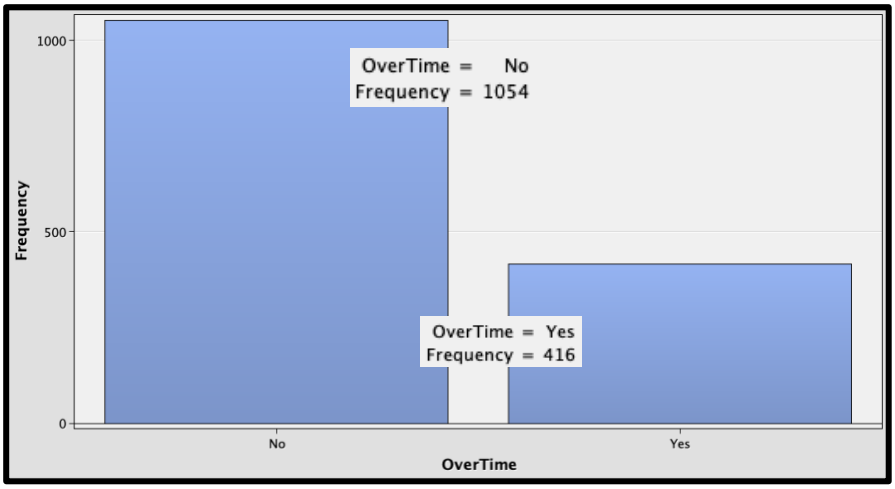

Figure 20: Histogram of Overtime

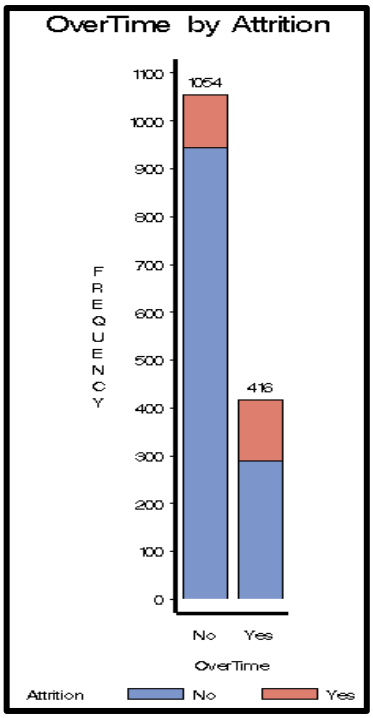

Figure 21: Histogram of Over Time

From figure 19, we can deduce that there are no missing values from the overtime variable. Moreover, we can say that 1054 (71.70%) employees do not work overtime followed by 416 (28.30%) employees who work overtime. Furthermore, in figure 21, we can also observe that 53.59% of the employees that left worked overtime.

Data Role=TRAIN Variable=Age

Target	Target Level	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label
Attrition	No	36	0	1233	18	60	37.56123	8.88836	0.408122	-0.41183	INPUT	Age
Attrition	Yes	31	0	237	18	58	33.60759	9.68935	0.715732	-0.05704	INPUT	Age
OVERALL		36	0	1470	18	60	36.92381	9.135373	0.413286	-0.40415	INPUT	Age

Figure 22: Summary Statistics of Age

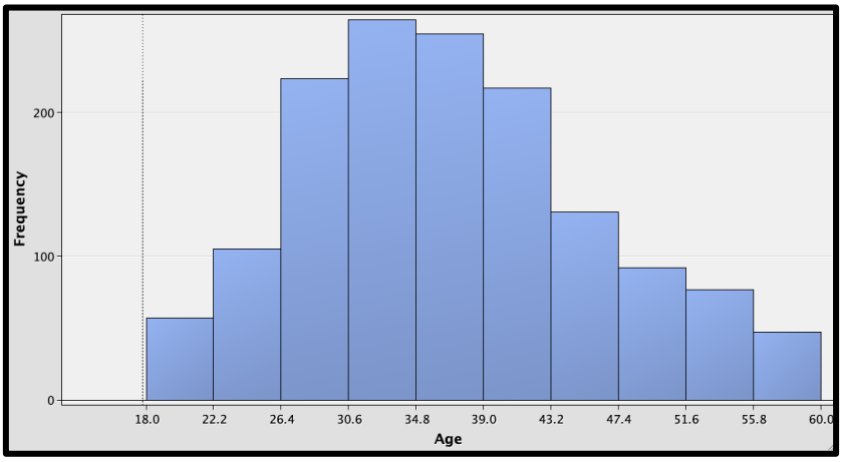

Figure 23: Histogram of Age

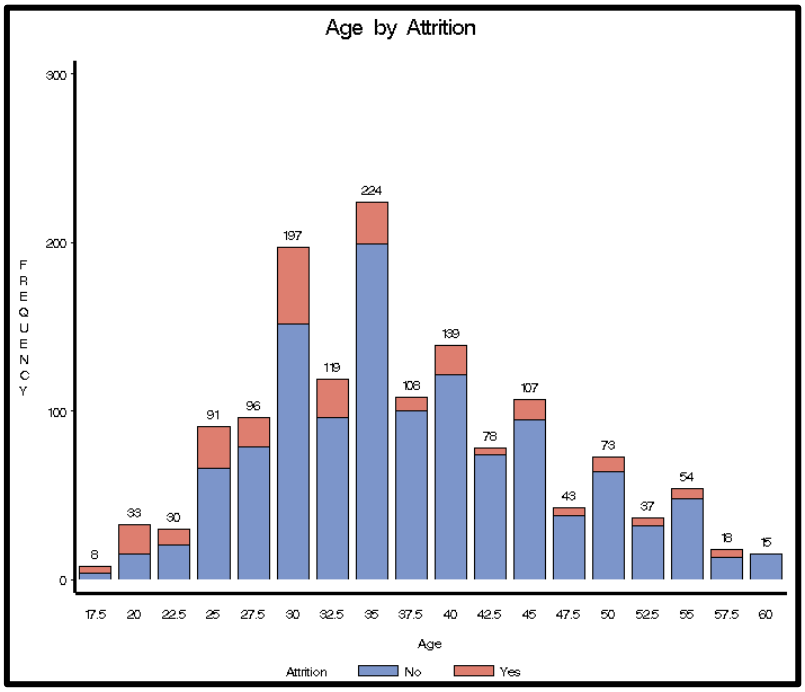

Figure 24: Histogram of Age by Attrition

From figure 22, we can deduce that there are no missing values from the age variable. Moreover, we can say that the youngest employee in the company is 18 years old while the oldest is 60 years old. The mean age of employees in the company is about 37 years of age. Those who left were employees ranging from 18 to 58 years old, where the mean age is about 34 years of age. We can also say that most employees from the company are young employees.

Data Role=TRAIN Variable=Education

Target	Target Level	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label
Attrition	No	3	0	1233	1	5	2.927007	1.027002	-0.28594	-0.56051	INPUT	Education
Attrition	Yes	3	0	237	1	5	2.839662	1.008244	-0.32365	-0.55256	INPUT	Education
OVERALL		3	0	1470	1	5	2.912925	1.024165	-0.28968	-0.55911	INPUT	Education

Figure 25: Summary Statistics of Education

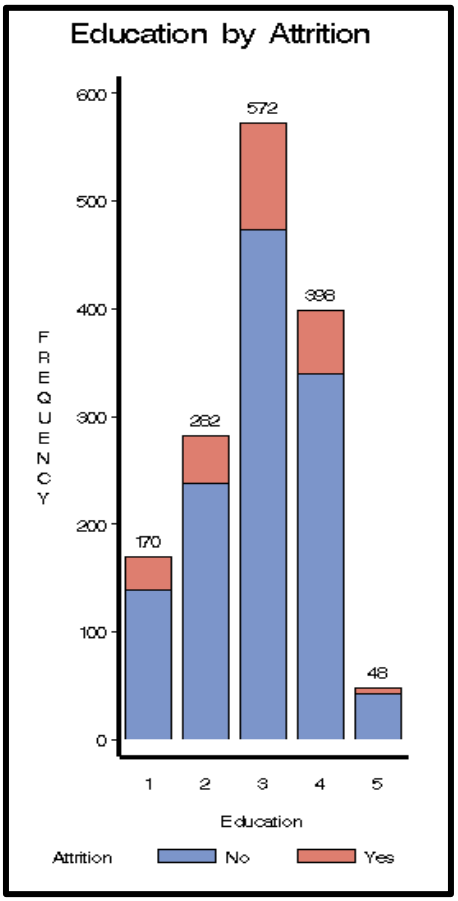

Figure 26: Histogram of Education by Attrition

From figure 25, we can deduce that there are no missing values from the education variable. Moreover, we can say that 572 employees have an education rated as 3 (bachelor's degree) followed by 396 employees with a master's degree rated at 4. Those who left the company was also those with a bachelor's degree.

Data Role=TRAIN Variable=EnvironmentSatisfaction

Target	Target Level	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label
Attrition	No	3	0	1233	1	4	2.77129	1.071132	−0.37669	−1.11492	INPUT	EnvironmentSatisfaction
Attrition	Yes	3	0	237	1	4	2.464135	1.169791	−0.00899	−1.47911	INPUT	EnvironmentSatisfaction
OVERALL		3	0	1470	1	4	2.721769	1.093082	−0.32165	−1.20252	INPUT	EnvironmentSatisfaction

Figure 27: Summary Statistics of Environment Satisfaction

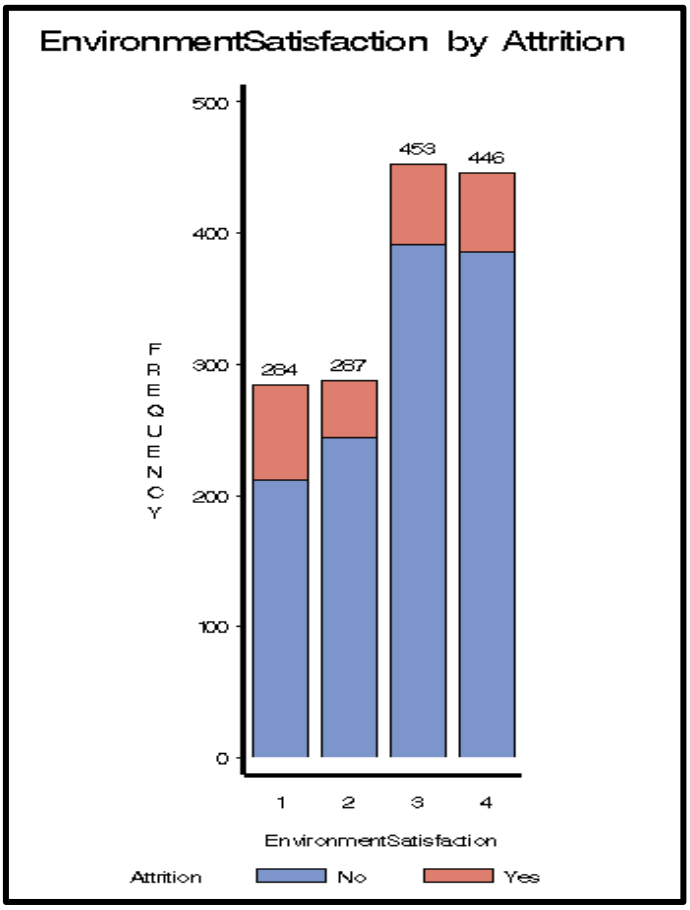

Figure 28: Histogram of Environment Satisfaction by Attrition

From figure 27, we can deduce that there are no missing values from the environment satisfaction variable. Moreover, we can say that 453 employees rated their environment satisfaction as 3 (high satisfaction). In figure 28, those who left were also those who rated their environment satisfaction as 3.

Data Role=TRAIN Variable=JobInvolvement

Target	Target Level	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label
Attrition	No	3	0	1233	1	4	2.770479	0.69205	-0.47049	0.367203	INPUT	JobInvolvement
Attrition	Yes	3	0	237	1	4	2.518987	0.773405	-0.47958	-0.32406	INPUT	JobInvolvement
OVERALL		3	0	1470	1	4	2.729932	0.711561	-0.49842	0.270999	INPUT	JobInvolvement

Figure 29: Summary Statistics of Job Involvement

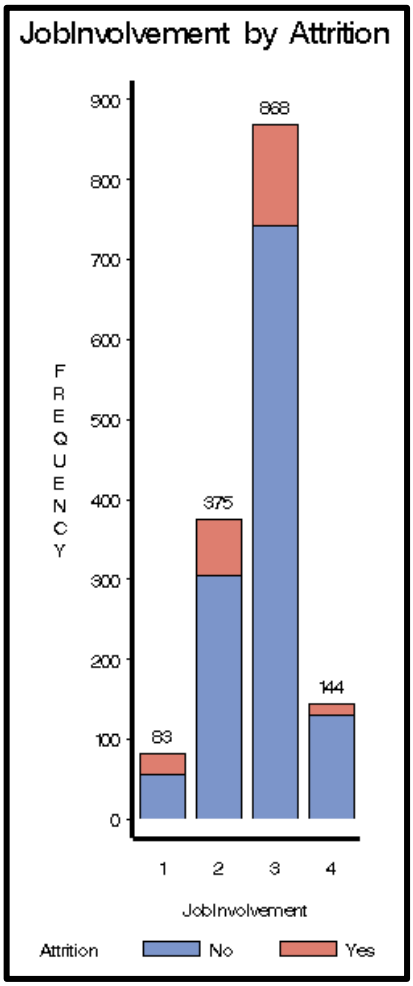

Figure 30: Histogram of Job Involvement by Attrition

From figure 29, we can deduce that there are no missing values from the job involvement variable. Moreover, we can say that 868 employees rated their job involvement as 3 (high involvement). In figure 30, those who left were also those who rated their job involvement as 3.

Data Role=TRAIN Variable=JobLevel

Target	Target Level	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label
Attrition	No	2	0	1233	1	5	2.145985	1.117933	0.956543	0.245313	INPUT	JobLevel
Attrition	Yes	1	0	237	1	5	1.637131	0.940594	1.554019	2.126678	INPUT	JobLevel
OVERALL		2	0	1470	1	5	2.063946	1.10694	1.025401	0.399152	INPUT	JobLevel

Figure 31: Summary Statistics of Job Level

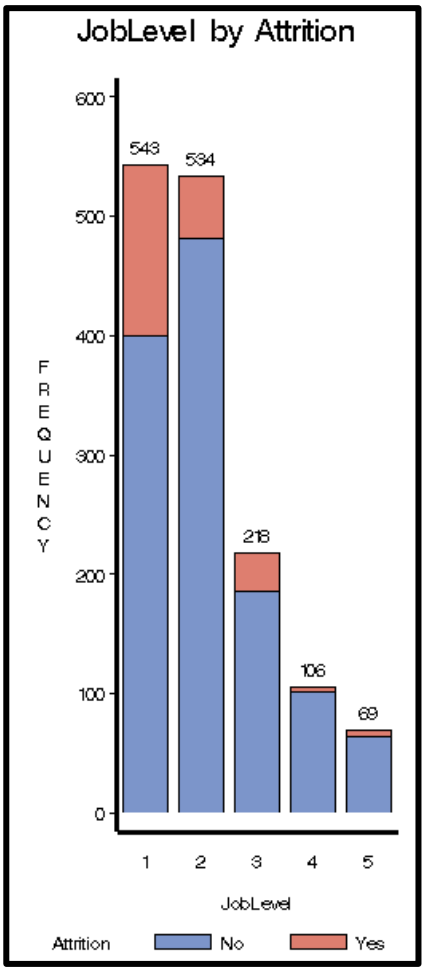

Figure 32: Histogram of Job Level by Attrition

From figure 31, we can deduce that there are no missing values from the job level variable. Moreover, we can say that 543 employees were coming from the entry-level. In figure 32, those who left were also those who were entry level (job level 1-2)

Data Role=TRAIN Variable=MonthlyIncome

Target	Target Level	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label
Attrition	No	5204	0	1233	1051	19999	6832.74	4818.208	1.286231	0.671641	INPUT	MonthlyIncome
Attrition	Yes	3172	0	237	1009	19859	4787.093	3640.21	1.921147	4.181845	INPUT	MonthlyIncome
OVERALL		4908	0	1470	1009	19999	6502.931	4707.957	1.369817	1.005233	INPUT	MonthlyIncome

Figure 33: Summary Statistics of Monthly Income

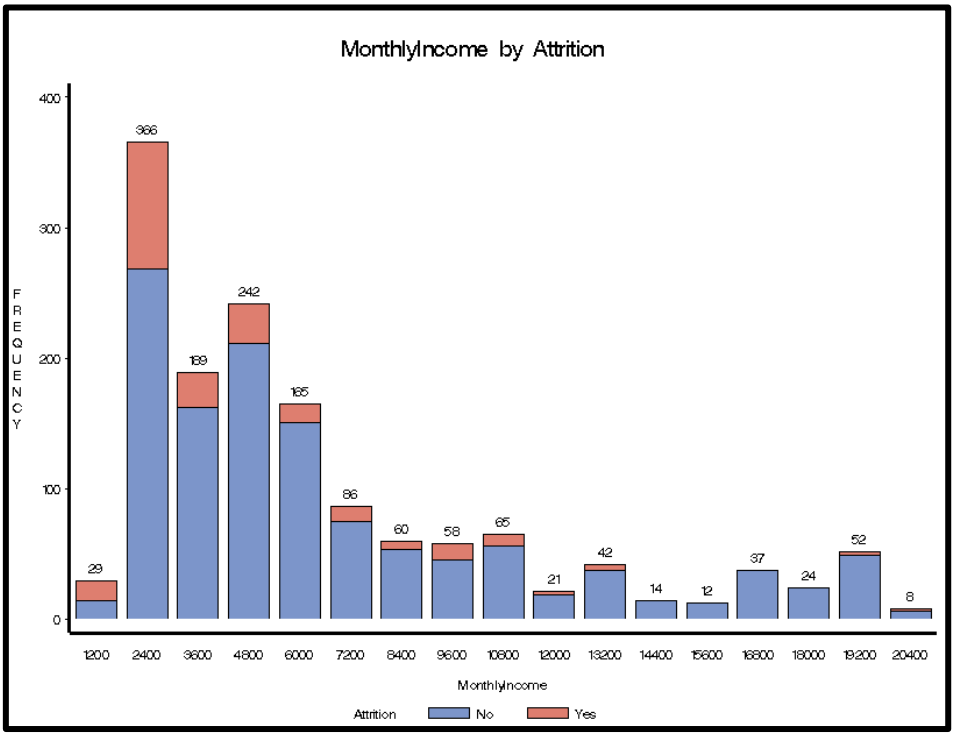

Figure 34: Histogram of Monthly Income by Attrition

From figure 33, we can deduce that there are no missing values from the monthly income variable. The monthly income of employees ranges from 1009 dollars to 19,999 dollars monthly. Moreover, we can say that 366 employees are receiving pay of 2400 dollars monthly. The average salary of employees in the company was 6502.93 dollars and the average salary of employees that left was 4787.09 dollars. In figure 34, those who left were those who were paid below the average salary of 2400 dollars monthly.

Data Role=TRAIN Variable=NumCompaniesWorked

Target	Target Level	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label
Attrition	No	2	0	1233	0	9	2.64558	2.46009	1.058795	0.141433	INPUT	NumCompaniesWorked
Attrition	Yes	1	0	237	0	9	2.940928	2.678519	0.86417	-0.54306	INPUT	NumCompaniesWorked
OVERALL		2	0	1470	0	9	2.693197	2.498009	1.026471	0.010214	INPUT	NumCompaniesWorked

Figure 35: Summary Statistics of Number of Companies Worked At

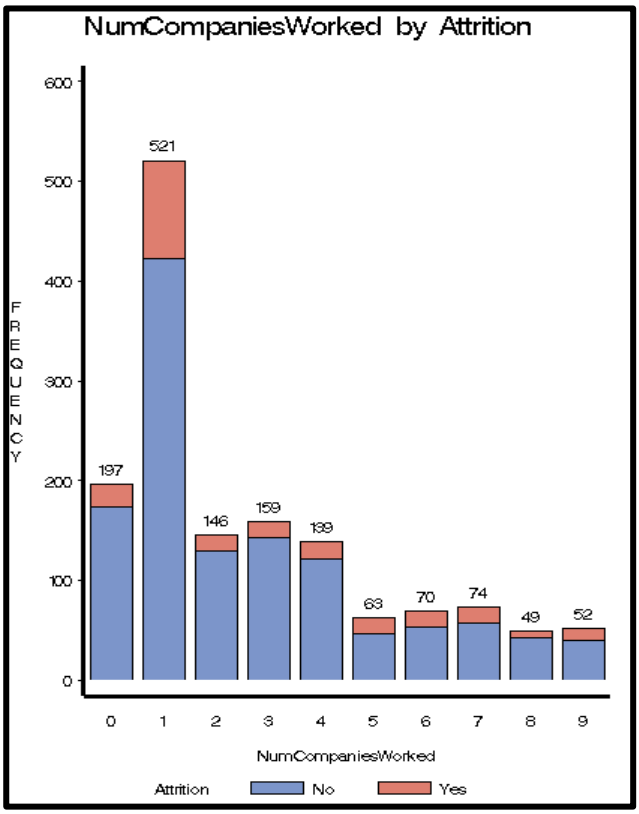

Figure 36: Histogram of Number of Companies Worked At by Attrition

From figure 35, we can deduce that there are no missing values from the number of companies worked variable. Overall, most employees worked at only 1 company and 3 companies on average with the maximum being 9 companies. In figure 36, those who left were those who have worked at only 1 company.

Data Role=TRAIN Variable=StockOptionLevel

Target	Target Level	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label
Attrition	No	1	0	1233	0	3	0.845093	0.841985	0.870974	0.259176	INPUT	StockOptionLevel
Attrition	Yes	0	0	237	0	3	0.527426	0.856361	1.690106	2.067347	INPUT	StockOptionLevel
OVERALL		1	0	1470	0	3	0.793878	0.852077	0.96898	0.364634	INPUT	StockOptionLevel

Figure 37: Summary Statistics of Stock Option Level

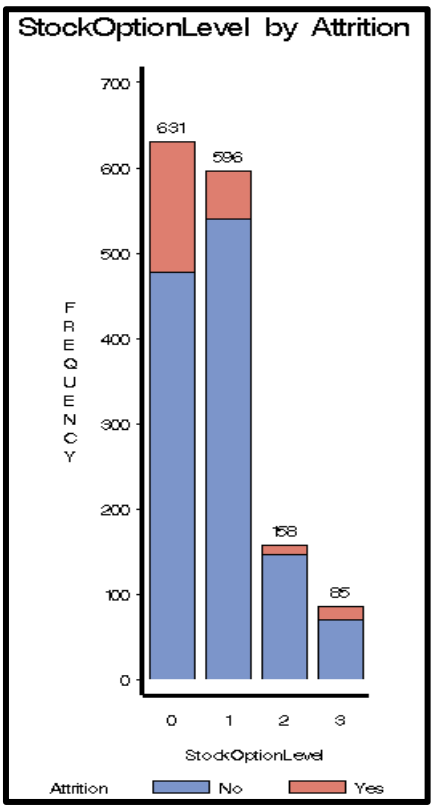

Figure 38: Histogram of Stock Options Level by Attrition

From figure 37, we can deduce that there are no missing values from the stock options level variable. Overall, most employees have zero to only 1 stock options level. In figure 38, those who left were also those who have received 0 to only 1 stock options levels.

Data Role=TRAIN Variable=TotalWorkingYears

Target	Target Level	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label
Attrition	No	10	0	1233	0	38	11.86294	7.760719	1.066923	0.678079	INPUT	TotalWorkingYears
Attrition	Yes	7	0	237	0	40	8.244726	7.169204	1.688158	3.784098	INPUT	TotalWorkingYears
OVERALL		10	0	1470	0	40	11.27959	7.780782	1.117172	0.91827	INPUT	TotalWorkingYears

Figure 39: Summary Statistics of Total Working Years

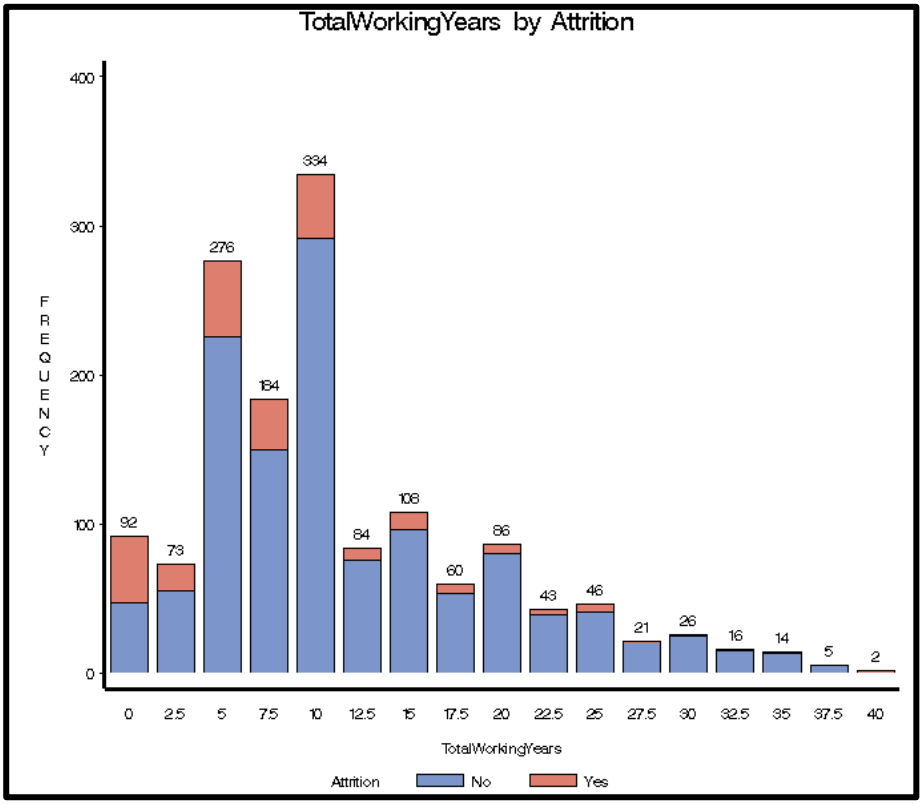

Figure 40: Histogram of Total Working Years by Attrition

From figure 39, we can deduce that there are no missing values from the total working years variable. Overall, 334 employees (majority) have experience working over 10 years. The number of working years ranges from 0 to 40 years. In figure 40, those who left were those who have an average of about 7-8 years of working experience.

Data Role=TRAIN Variable=TrainingTimesLastYear

Target	Target Level	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label
Attrition	No	3	0	1233	0	6	2.832928	1.293585	0.590064	0.447276	INPUT	TrainingTimesLastYear
Attrition	Yes	2	0	237	0	6	2.624473	1.254784	0.337787	0.658225	INPUT	TrainingTimesLastYear
OVERALL		3	0	1470	0	6	2.79932	1.289271	0.553124	0.494993	INPUT	TrainingTimesLastYear

Figure 41: Summary Statistics of Training Times Last Year

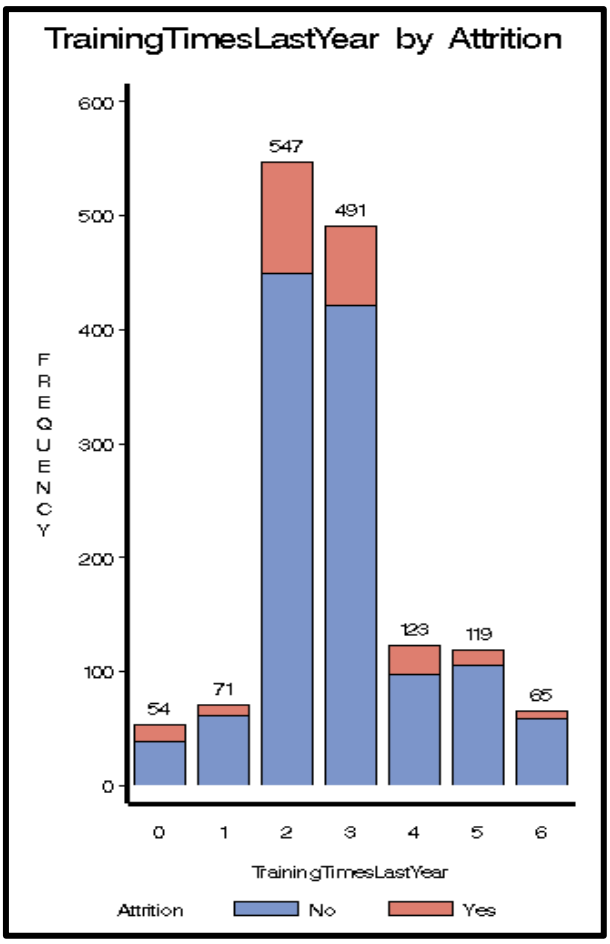

Figure 42: Histogram of Training Times Last Year by Attrition

From figure 41, we can deduce that there are no missing values from the training times last year variable. The number of training times ranges from 0 to 6 times. Overall, most employees have only 2 training times, which is considered as quite low. In figure 42, those who left were also those who only had 2 training times last year.

Data Role=TRAIN Variable=YearsAtCompany

Target	Target Level	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label
Attrition	No	6	0	1233	0	37	7.369019	6.096298	1.657958	3.353473	INPUT	YearsAtCompany
Attrition	Yes	3	0	237	0	40	5.130802	5.949984	2.682244	9.608029	INPUT	YearsAtCompany
OVERALL		5	0	1470	0	40	7.008163	6.126525	1.764529	3.935509	INPUT	YearsAtCompany

Figure 43: Summary Statistics of Years At Company

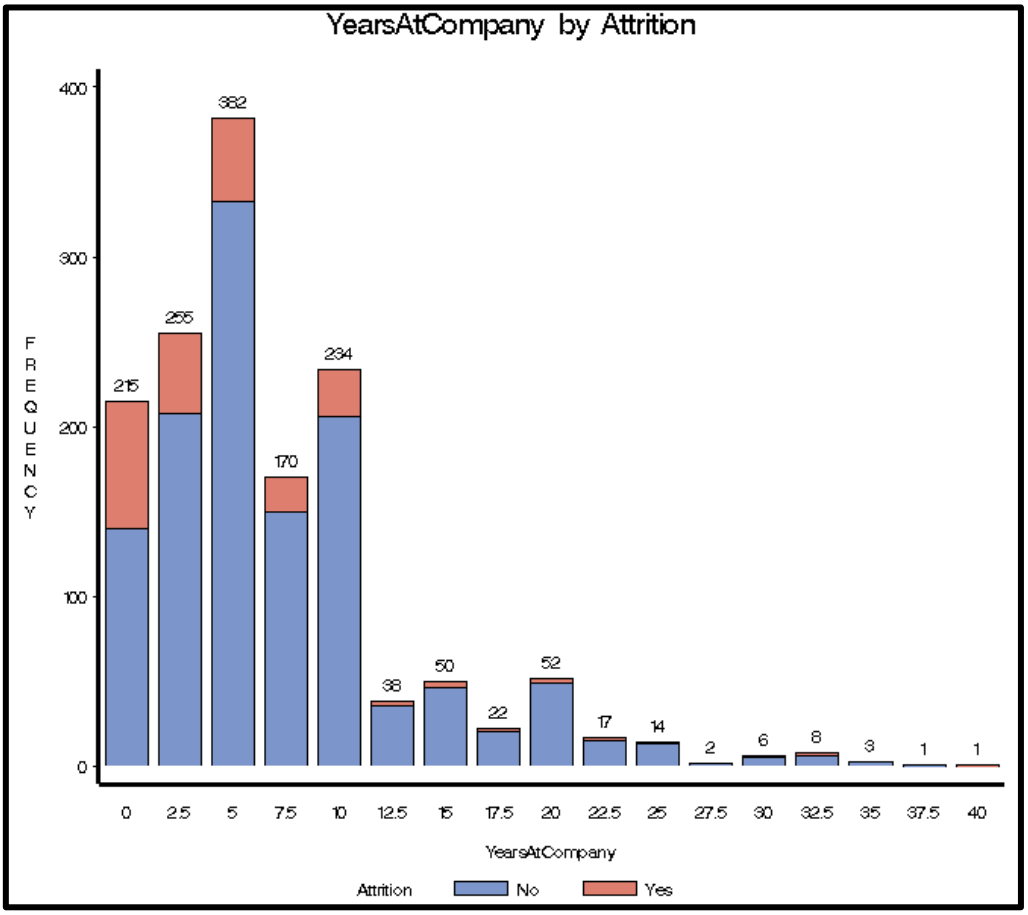

Figure 44: Histogram of Years at Company by Attrition

From figure 43, we can deduce that there are no missing values from the years at the company variable. The number of years at the company ranges from 0 to 40 years, with the average being 7 years. The majority of the employees in the company have 5 years of experience in the company. In figure 44, those who left were also those who have worked at the company for an average of 5 years, and the majority had 3 years of experience at the company.

Data Role=TRAIN Variable=YearsSinceLastPromotion

Target	Target Level	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label
Attrition	No	1	0	1233	0	15	2.234388	3.234762	1.94671	3.430452	INPUT	YearsSinceLastPromotion
Attrition	Yes	1	0	237	0	15	1.945148	3.153077	2.217563	4.861144	INPUT	YearsSinceLastPromotion
OVERALL		1	0	1470	0	15	2.187755	3.22243	1.98429	3.612673	INPUT	YearsSinceLastPromotion

Figure 45: Summary Statistics of Years Since Last Promotion

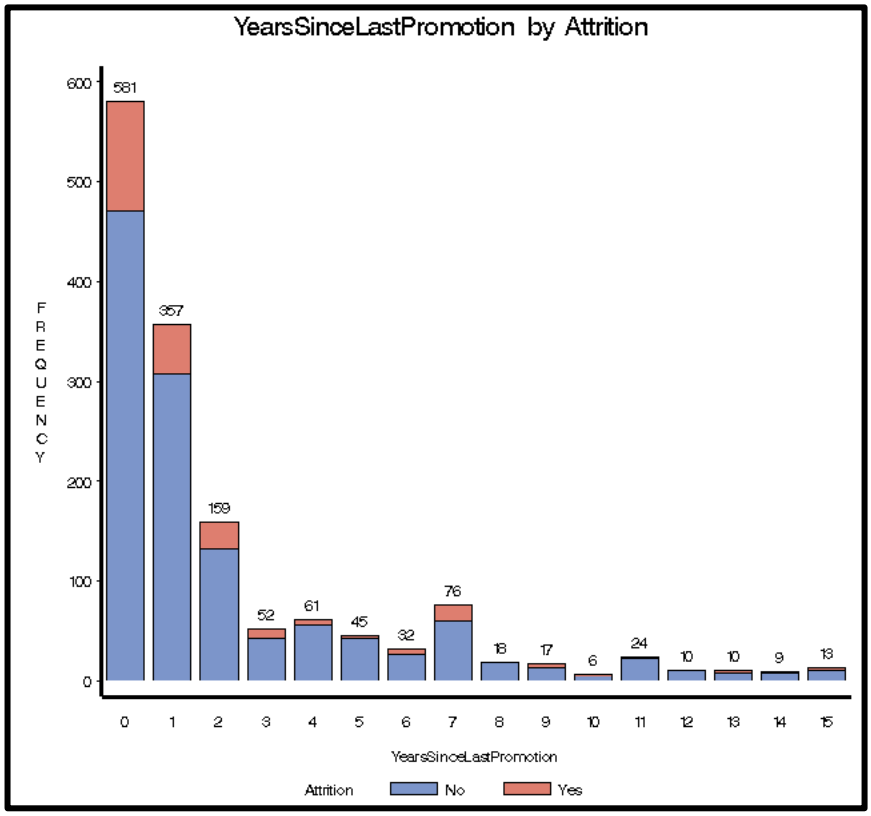

Figure 46: Histogram of Years Since Last Promotion

From figure 45, we can deduce that there are no missing values from the years since the last promotion variable. The number of years since the last promotion ranges from 0 to 15 years, with the average being 2 years. Overall, the majority of the employees in the company have 0 years since last promotion followed by 1 year since their last promotion. In figure 46, the majority of those who left were also those who have 0 to 1 year since their last promotion.

3.0 Data Pre- Processing

3.1 The StatExplore Node

The StatExplore node is a multifunctional tool that is used to investigate the training dataset's variable distributions and statistics.

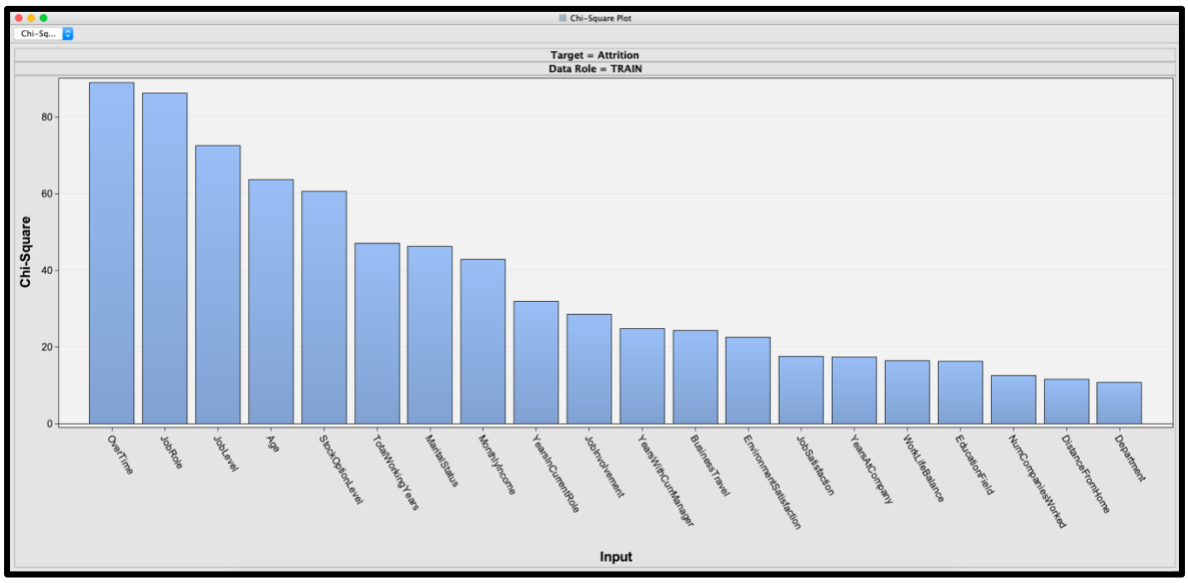

Figure 47: Chi Square Plot

Chi-Square Statistics
(maximum 500 observations printed)

Data Role=TRAIN Target=Attrition

Input	Chi-Square	Df	Prob
OverTime	89.0439	1	<.0001
JobRole	86.1903	8	<.0001
JobLevel	72.5290	4	<.0001
Age	63.6208	4	<.0001
StockOptionLevel	60.5983	3	<.0001
TotalWorkingYears	47.0783	4	<.0001
MaritalStatus	46.1637	2	<.0001
MonthlyIncome	42.8923	4	<.0001
YearsInCurrentRole	31.8168	4	<.0001
JobInvolvement	28.4920	3	<.0001
YearsWithCurrManager	24.7154	4	<.0001
BusinessTravel	24.3273	3	<.0001
EnvironmentSatisfaction	22.5039	3	<.0001
JobSatisfaction	17.5051	3	0.0006
YearsAtCompany	17.3471	4	0.0017
WorkLifeBalance	16.3251	3	0.0010
EducationField	16.1616	6	0.0129
NumCompaniesWorked	12.5714	4	0.0136
DistanceFromHome	11.5089	4	0.0214
Department	10.7960	2	0.0045
TrainingTimesLastYear	9.7741	4	0.0444
YearsSinceLastPromotion	7.8535	4	0.0971
DailyRate	6.7678	4	0.1487
HourlyRate	5.4465	4	0.2445
RelationshipSatisfaction	5.2411	3	0.1550
Education	3.0740	4	0.5455
EmployeeNumber	2.9979	4	0.5582
PercentSalaryHike	2.5053	4	0.6437
MonthlyRate	2.1662	4	0.7052
Gender	1.2752	1	0.2588
PerformanceRating	0.0123	1	0.9118

Figure 48: Chi Square Statistics

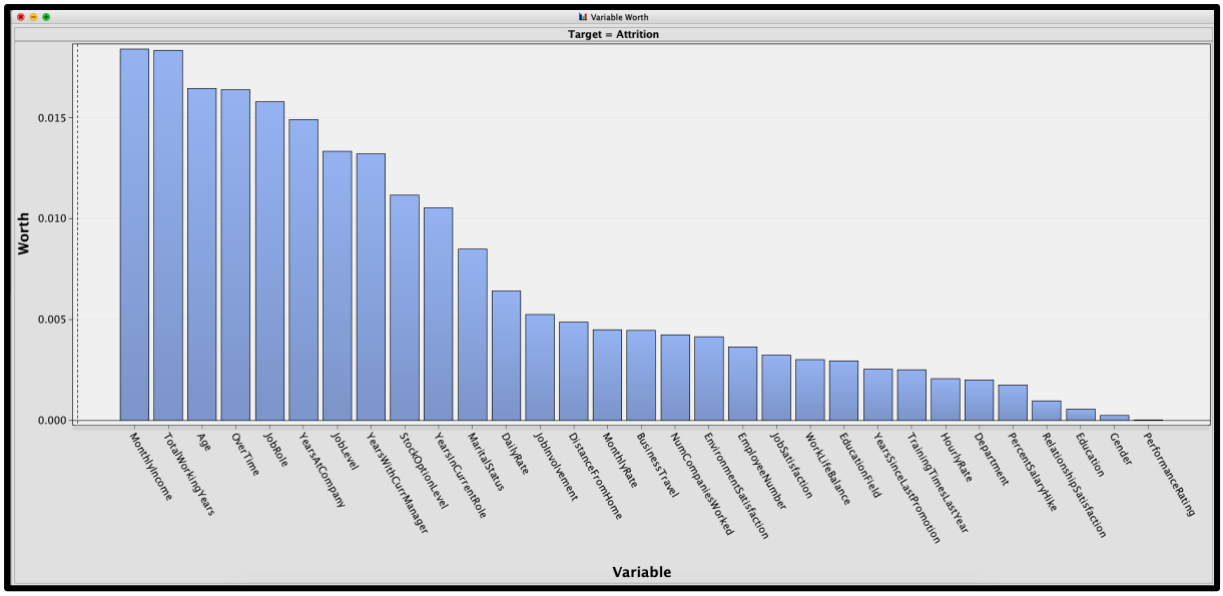

Figure 49: Variable Worth

From the StatExplore node observation, the chi-square plot, chi-square statistics, and variable worth tell us the most significant and contributing independent variables to predict the dependent variable, which is attrition. From the chi-square plot, we can observe that the overtime variable is the most significant with the largest chi-square value and prob of less than 0.05. When we observe the chi-square statistics, many variables were rejected as they had a prob value of more than 0.05. However, we will not be dropping most of the suggested variables as this will result in a poor modelling output as it is biased. Thus, we will also be dropping the insignificant variables manually. From the variable worth section, the last four variables, which are relationship satisfaction, education, gender, and performance rating will be dropped as their worth value is near zero.

3.2 Dropping Variables Manually

Name	Drop	Role	Level
Age	Default	Input	Interval
Attrition	Default	Target	Nominal
BusinessTrav	Default	Input	Nominal
DailyRate	Yes	Input	Interval
Department	Default	Input	Nominal
DistanceFror	Default	Input	Interval
Education	Yes	Input	Interval
EducationFie	Default	Input	Nominal
EmployeeNu	Yes	Input	Interval
Environment	Default	Input	Interval
Gender	Yes	Input	Nominal
HourlyRate	Yes	Input	Interval
JobInvolvem	Default	Input	Interval
JobLevel	Default	Input	Interval
JobRole	Default	Input	Nominal
JobSatisfacti	Default	Input	Interval
MaritalStatus	Default	Input	Nominal
MonthlyInco	Default	Input	Interval
MonthlyRate	Yes	Input	Interval
NumCompar	Default	Input	Interval
OverTime	Default	Input	Nominal
PercentSalar	Default	Input	Interval
Performance	Yes	Input	Interval
Relationship	Yes	Input	Interval
StockOption	Default	Input	Interval
TotalWorkin	Default	Input	Interval
TrainingTim	Default	Input	Interval
WorkLifeBala	Default	Input	Interval
YearsAtCom	Default	Input	Interval
YearsInCurre	Default	Input	Interval
YearsSinceLa	Default	Input	Interval
YearsWithCu	Default	Input	Interval

Figure 50: Variables that have been dropped

To provide an accurate analytical output, removing irrelevant variables is a crucial element of the data cleansing process. According to a study by (Yang and Islam, 2020) features including 'daily rate,' 'hourly rate,' and 'monthly rate' were deleted because they had a limited relationship with other variables. The 'employee number' variable had been eliminated by (Yadav, Jain and Singh, 2018) since it was unsuitable for the study. In addition, as demonstrated in the variable worth figure, we have deleted the relationship satisfaction, education, gender, and performance rating factors.

3.3 Missing Values Treatment

Data recording faults, inadequate client responses, actual system, or measurement issues can all lead to missing results. Dismissing all missing data may lead to the omission of relevant or important data that is still present in the non-missing variables. Furthermore, eliminating all missing data could bias the sample because missing values data may share other characteristics. The model could exclude records with missing values, however, reducing the size of the training data set and lowering the models' predictive potential.

Using the impute node, we can replace the missing values from the business travel and education field variables, where each had 3 missing values. The values of missing variables are imputed using a count setting to replace missing class variable values with the most frequently occurring class variable value. In the case of 'business travels and education field, it is travel rarely and life sciences respectively.

Name	Use	Method	Use Tree	Role	Level
Age	Default	Default	Default	Input	Interval
Attrition	Default	Default	Default	Target	Nominal
BusinessTrav	Yes	Count	Default	Input	Nominal
Department	Default	Default	Default	Input	Nominal
DistanceFror	Default	Default	Default	Input	Interval
EducationFie	Yes	Count	Default	Input	Nominal
Environment	Default	Default	Default	Input	Interval
JobInvolvem	Default	Default	Default	Input	Interval
JobLevel	Default	Default	Default	Input	Interval
JobRole	Default	Default	Default	Input	Nominal
JobSatisfacti	Default	Default	Default	Input	Interval
MaritalStatus	Default	Default	Default	Input	Nominal
MonthlyInco	Default	Default	Default	Input	Interval
NumCompar	Default	Default	Default	Input	Interval
OverTime	Default	Default	Default	Input	Nominal
PercentSalary	Default	Default	Default	Input	Interval
StockOption	Default	Default	Default	Input	Interval
TotalWorking	Default	Default	Default	Input	Interval
TrainingTime	Default	Default	Default	Input	Interval
WorkLifeBala	Default	Default	Default	Input	Interval
YearsAtCom	Default	Default	Default	Input	Interval
YearsInCurre	Default	Default	Default	Input	Interval
YearsSinceLa	Default	Default	Default	Input	Interval
YearsWithCu	Default	Default	Default	Input	Interval

Figure 51: Variables Chosen for Imputation

Imputation Summary
Number Of Observations

Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
BusinessTravel	COUNT	IMP_BusinessTravel	Travel_Rarely	INPUT	NOMINAL	BusinessTravel	3
EducationField	COUNT	IMP_EducationField	Life Sciences	INPUT	NOMINAL	EducationField	3

Figure 52: Imputation Summary

3.4 Outlier treatment

Outliers are readings that appear to be out of the norm when compared to other existing points. The variables are assumed to be regularly distributed in many statistical techniques. However, the issue is that a few outliers can sometimes be sufficient to affect results by increasing data variance. In this section, we present one way of outlier treatments which is the log10 transformation method (Maisuradze, 2017). Outliers can be interpreted through the value of skewness from the output section in the StatExplore node.

	Variable	Role	Mean	Deviation	Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
73											
74	Age	INPUT	36.92381	9.135373	1470	0	18	36	60	0.413286	-0.40415
75	DailyRate	INPUT	802.4857	403.5091	1470	0	102	802	1499	-0.00352	-1.20382
76	DistanceFromHome	INPUT	9.192517	8.106864	1470	0	1	7	29	0.958118	-0.22483
77	Education	INPUT	2.912925	1.024165	1470	0	1	3	5	-0.28968	-0.55911
78	EmployeeNumber	INPUT	1024.865	602.0243	1470	0	1	1019	2068	0.016574	-1.22318
79	EnvironmentSatisfaction	INPUT	2.721769	1.093082	1470	0	1	3	4	-0.32165	-1.20252
80	HourlyRate	INPUT	65.89116	20.32943	1470	0	30	66	100	-0.03231	-1.1964
81	JobInvolvement	INPUT	2.729932	0.711561	1470	0	1	3	4	-0.49842	0.270999
82	JobLevel	INPUT	2.063946	1.10694	1470	0	1	2	5	1.025401	0.399152
83	JobSatisfaction	INPUT	2.728571	1.102846	1470	0	1	3	4	-0.32967	-1.22219
84	MonthlyIncome	INPUT	6502.931	4707.957	1470	0	1009	4908	19999	1.369817	1.005233

Figure 53: Summary Statistics

If skewness is less than -1 or larger than 1, the distribution is strongly skewed, according to a general rule. The distribution is significantly skewed if the skewness is between -1 and -0.5 or between 0.5 and 1. The distribution is nearly symmetric if the skewness is between -0.5 and 0.5.

Data transformations are mathematical alterations made to the values of a variable in principle (Osborne, 2003). It minimizes skewness and produces uniform distributions among the data, which is very useful in datasets with outlier values. As a result, we've decided to use log transformation to modify the monthly income variable. The log of each observation is taken, and we can use either base-10 logs or base-e logs, commonly known as natural logs, to do so.

Source	Method	Variable Name	Formula	Number of Levels	Non Missing	Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Label
Input	Original	MonthlyIncome		.	1470	0	1009	19999	6502.931	4707.957	1.369817	1.005233	MonthlyIncome
Output	Computed	LG10_Monthl...	log10(Monthl...	.	1470	0	3.004321	4.30103	3.714413	0.288508	0.286448	-0.69756	Transformed...

Figure 54: Transformation statistics

We can observe after log transformation that the skewness of the monthly income variable has reduced significantly from 1.369817 to 0.286448, indicating an approximately symmetric distribution.

3.4 Data Partition

Partition Summary

Type	Data Set	Number of Observations
DATA	EMWS2.Trans_TRAIN	1470
TRAIN	EMWS2.Part_TRAIN	1028
VALIDATE	EMWS2.Part_VALIDATE	442

Figure 55: Data Partition Summary

The data is divided into training and validation data sets using the data partition node. Partitioning is necessary for maintaining the model's reliability when fitting. The training data is used to fit the model first, whereas the validation data is used to evaluate the model empirically without overfitting the data. In the properties panel, under the data set allocations section, the training set was set to 70% while the validation set was set to 30%. In this project, the test set was ignored and set to 0%.

4.0 Modelling

4.1 Decision Tree

Since they are straightforward to use and evaluate, decision trees have become famous. Tree-based models partition the data many rounds based on specified feature cutoff values. Various subsets of the dataset are formed by splitting. Leaf nodes are the final subsets, whereas split nodes are the intermediate subsets.

4.1.1 Subtree Assessment Plot

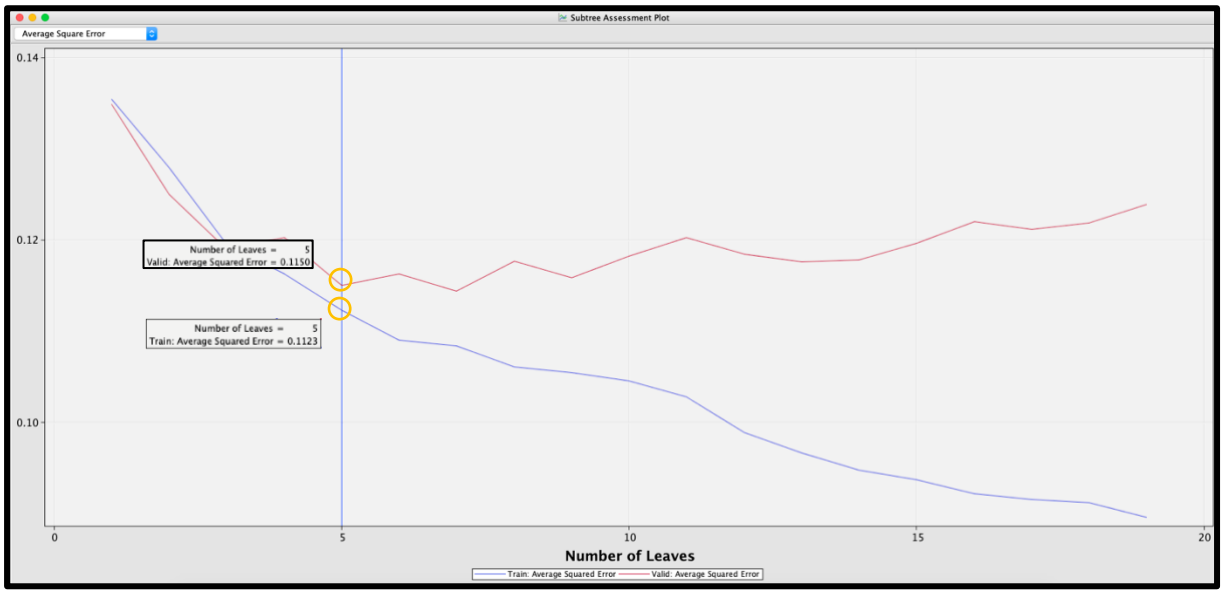

Figure 56: Subtree Assessment Plot

Based on figure 56, we can observe that the default number of leaves selected initially was 5, where the average squared error for train and valid are 0.1150 and 0.1123 respectively.

4.1.2 Tree Plot

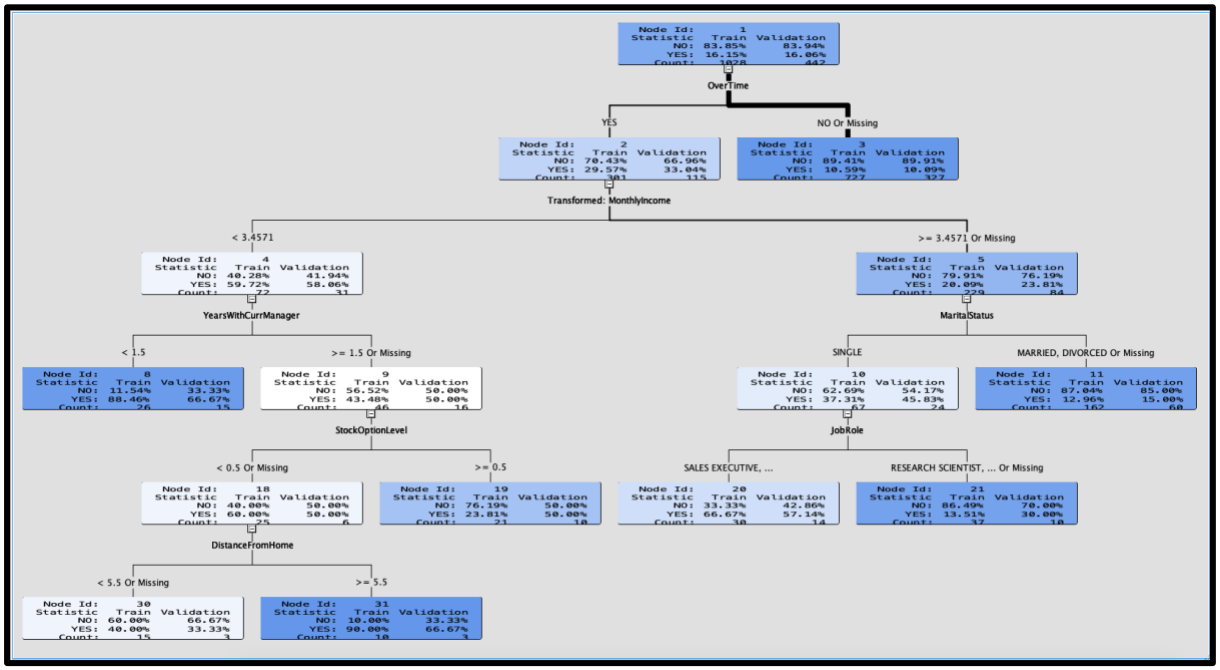

Figure 57: Tree plot

Overtime is the predictor variable used for the primary split in node 1. From the statistics of node 1, we can observe that the accuracy predicted for the train and validation sets are not too far off with 16.15% for those who voted 'yes' in the training set and 16.06% for those who voted 'yes' in the validation set. Node 2 represents those who voted 'yes' for overtime and node 3 represents those who voted 'no'. Node 2 is further split using monthly income.

Monthly income is then split into node 4 and node 5. If the monthly income is less than 3.45712 and years with the current manager is less than 1.5 years, node 8 is produced, where predicted attrition as 'yes' is 88.46% and 66.67% for train and validation sets respectively. As training prediction is significantly higher than validation, this may indicate an overfit.

We can also observe that if the monthly income is more than or missing and the marital status is either married or divorced, then the predicted attrition as 'yes' is 12.96% and 15% for train and validation sets respectively.

Furthermore, when years with the current manager is more than 1.5 or missing and the stock options level is above 0.5 with monthly income less than 3.45712 in node 19, then the predicted attrition as 'yes' is 23.81% and 50% for train and validation sets respectively. This may indicate an underfit as the validation prediction is significantly more than the train set.

Moreover, if monthly income is more than 3.45712 or missing, marital status is single and job role is either sales, executive, or laboratory technician, then the tree node identifies is 20 with predicted attrition as 'yes' is 66.67% and 57.14% for train and validation sets respectively. However, in node 21, where the job role is either research scientist, manufacturing director, healthcare representative, or missing, the predicted attrition as 'yes' is 13.51% and 30% for train and validation sets respectively. This may indicate an underfit as the validation prediction is significantly more than the train set.

4.1.2 Output

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
LG10_MonthlyIncome	Transformed: MonthlyIncome	1	1.0000	0.7849	0.7849
OverTime	OverTime	1	0.9438	1.0000	1.0595
JobRole	JobRole	1	0.7375	0.2176	0.2950
YearsWithCurrManager	YearsWithCurrManager	1	0.6250	0.0000	0.0000
MaritalStatus	MaritalStatus	1	0.5715	0.5974	1.0453
DistanceFromHome	DistanceFromHome	1	0.4175	0.1074	0.2572
StockOptionLevel	StockOptionLevel	1	0.4168	0.0000	0.0000

Figure 58: Variable Importance

From the variable importance section, we can observe that the most significant variable in predicting the dependent variable in the decision tree model is the 'Over time' variable with an importance of 0.9428 and validation importance of 1.0000 in the validation importance. The ratio of validation to training importance is also the highest with a value of 1.0595. Next is the monthly income variable with importance of 1.0000, validation importance of 0.7849, and ratio of validation to training importance of 0.7849.

Fit Statistics

Target=Attrition Target Label=Attrition

Fit Statistics	Statistics Label	Train	Validation
NOBS	Sum of Frequencies	1028.00	442.000
MISC	Misclassification Rate	0.12	0.143
MAX	Maximum Absolute Error	0.90	0.900
SSE	Sum of Squared Errors	218.15	104.044
ASE	Average Squared Error	0.11	0.118
RASE	Root Average Squared Error	0.33	0.343
DIV	Divisor for ASE	2056.00	884.000
DFT	Total Degrees of Freedom	1028.00	.

Figure 59: Fit statistics

In summary, the misclassification rate of the model is 12.0% and 14.3% respectively for the train and validation sets. Thus, the accuracy is 88% and 85.7% respectively for the train and validation sets, indicating a good fit for the model.

4.2 Random Forest

A HP Forest node generates a forest, a predictive model made up of numerous decision trees that vary in two ways. To begin, the training data for a tree is a random selection of all accessible data. Second, the input variables taken into account when splitting a node are chosen at random from all available inputs.

Some of the accessible data is excluded from the training data for a single tree. The out-of-bag sample is the data that is omitted from training. To make a prediction, an individual tree only needs the out-of-bag sample.

4.2.1 Iteration Plot

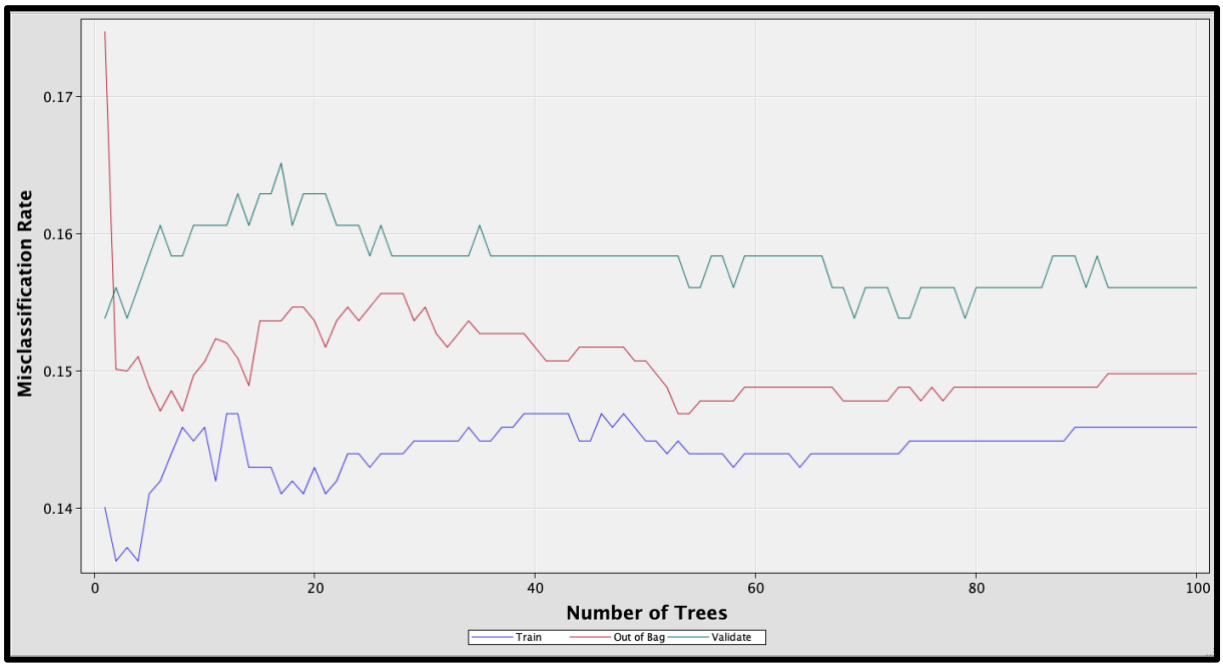

Figure 60: Iteration Plot

As in figure 60, the iteration plot shows the number of trees versus the misclassification rate. By default, the number of trees set in the properties panel is 100. The plot seems to be constant at the 100th tree onwards, thus we will remain with the default setting.

4.2.2 Fit Statistics

Fit Statistics

Target=Attrition Target Label=Attrition

Fit Statistics	Statistics Label	Train	Validation
ASE	Average Squared Error	0.10	0.111
DIV	Divisor for ASE	2056.00	884.000
MAX	Maximum Absolute Error	0.92	0.934
NOBS	Sum of Frequencies	1028.00	442.000
RASE	Root Average Squared Error	0.32	0.332
SSE	Sum of Squared Errors	207.99	97.714
DISF	Frequency of Classified Cases	1028.00	442.000
MISC	Misclassification Rate	0.15	0.156
WRONG	Number of Wrong Classifications	150.00	69.000

Figure 61: Fit Statistics

From figure 61, we can deduce that the misclassification rate of the train and validation sets are 15% and 15.6% respectively. Thus, the accuracy rate is 85% and 84.4% for the train and validation sets respectively. This shows as a good model fit.

4.2.3 Lost Reduction Variable Importance

Loss Reduction Variable Importance

Variable	Number of Rules	Gini	OOB Gini	Valid Gini	Margin	OOB Margin	Valid Margin
OverTime	102	0.009558	0.00724	0.01125	0.019116	0.01726	0.02437
JobLevel	55	0.003421	0.00157	0.00277	0.006843	0.00494	0.00493
StockOptionLevel	67	0.003990	0.00149	0.00185	0.007981	0.00599	0.00785
MaritalStatus	41	0.002652	0.00108	0.00187	0.005304	0.00354	0.00503
IMP_BusinessTravel	66	0.003620	0.00101	-0.00361	0.007240	0.00460	0.00072
YearsAtCompany	39	0.003467	0.00094	0.00257	0.006934	0.00443	0.00565
LG10_MonthlyIncome	40	0.003275	0.00090	0.00152	0.006550	0.00419	0.00444
YearsWithCurrManager	40	0.002991	0.00070	-0.00030	0.005983	0.00363	0.00201
TotalWorkingYears	34	0.003206	0.00058	0.00116	0.006412	0.00400	0.00391
JobRole	25	0.001808	0.00026	0.00078	0.003616	0.00250	0.00264
YearsInCurrentRole	22	0.001195	0.00024	0.00050	0.002391	0.00148	0.00149
Age	40	0.003961	0.00003	0.00042	0.007921	0.00391	0.00443
Department	24	0.000883	-0.00019	-0.00021	0.001766	0.00070	0.00090
TrainingTimesLastYear	8	0.000303	-0.00022	-0.00027	0.000607	-0.00001	-0.00004
PercentSalaryHike	15	0.000548	-0.00034	-0.00058	0.001095	0.00031	-0.00005
IMP_EducationField	19	0.000759	-0.00041	-0.00059	0.001519	0.00051	0.00045
JobSatisfaction	22	0.000748	-0.00051	-0.00014	0.001496	0.00012	0.00067
YearsSinceLastPromotion	30	0.001305	-0.00073	-0.00093	0.002610	0.00057	0.00016
NumCompaniesWorked	17	0.000845	-0.00073	-0.00030	0.001690	0.00007	0.00052
JobInvolvement	31	0.001529	-0.00075	-0.00068	0.003058	0.00081	0.00097
EnvironmentSatisfaction	33	0.001152	-0.00088	-0.00019	0.002305	0.00036	0.00024
WorkLifeBalance	27	0.001065	-0.00091	-0.00038	0.002130	0.00011	0.00093
DistanceFromHome	19	0.000913	-0.00098	-0.00048	0.001826	0.00010	0.00071

Figure 62: Loss Reduction Variable Importance

4.2.4. Variable Importance

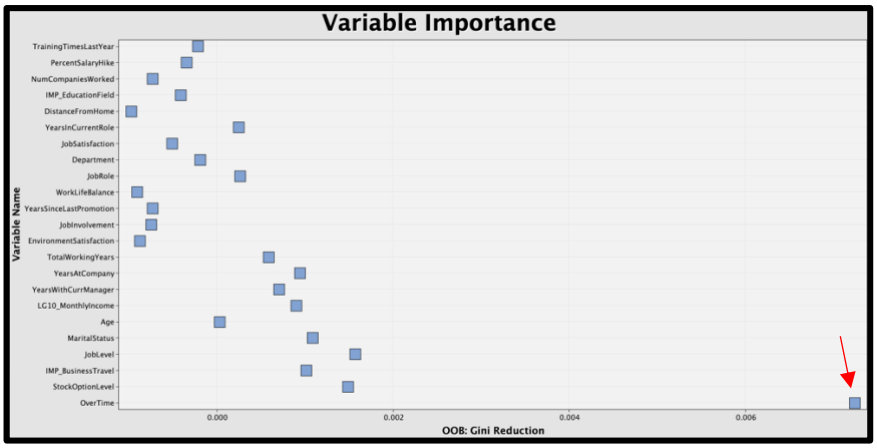

Figure 63: Variable Importance Plot

From figures 62 and 63, we can report that the most important variable for this model is the overtime variable, like the decision tree model as it has the highest Gini value for the train, OOB, and validation sets of 0.009558, 0.00724, and 0.01125 respectively.

4.3 Model Comparison

Model	Misclassification Rate (Train)	Misclassification Rate (Validation)
Decision Tree	0.120%	0.143%
HP Forest	0.150%	0.156%

Thus, the model that outperforms the other is the Decision tree with a misclassification rate of 14.3% and accuracy of 85.7%. Although an ensemble model such as the HP Forest is usually known to be better than their 'single' counterparts, they are only better if and only if the single model suffer from instability, since our dataset is not that huge, to begin with, it is a comfortable training sample size situation in which even a decision tree may get reasonably stable.

5.0 Conclusion

To summarize, we have successfully predicted the attrition rate using the decision tree model, which resulted in an accuracy of 85.7%. The model also indicated that the two most important variables contributing to the prediction of the dependent variable are the 'overtime' and 'monthly income' variable. The majority of the employees that left may be unsatisfied with the below-average salary of 2400 dollars monthly. We recommend that businesses assure that their employees' wages do not drop below the market average. According to Educba, corporations that offer their employees less than 10% compared to their competitors are more likely to lose them over time (Russel, 2018).

Furthermore, statistics revealed that staff was working above 15 hours per week in 2016 when IBM was still improving its organizational structure, which finally contributed to an increase in resignations (McLaren, 2019). It is possible that those employees were suffering from work stress. Reviewing the employees' payroll and checking their timecard data is one way for resolving this problem. Look for times of the year, such as public holidays, when there is likely to be a lot of overtime.

For future work, we suggest performing parameter tunings for the two models to possibly increase their accuracies and lower their misclassification rates. Moreover, more models should be implemented to select the best model.

Reference

Gupta, P., Fernandes, S. F. and Jain, M. (2018) 'Automation in recruitment: a new frontier', *Journal of Information Technology Teaching Cases*, 8(2), pp. 118–125. doi: 10.1057/s41266-018-0042-x.

McLaren, S. (2019) 'Here's How IBM Predicts 95% of Its Turnover Using Data | LinkedIn Talent Blog', LinkedIn. Available at: https://business.linkedin.com/talent-solutions/blog/artificial-intelligence/2019/IBM-predicts-95-percent-of-turnover-using-AI-and-data.

Osborne, J. W. (2003) 'Notes on the use of data transformations', *Practical Assessment, Research and Evaluation*, 8(6).

Russel, Ma. (2018) 'Why Employees Quit: 20 Stats Employers Need to Know', Medium. Available at: https://medium.com/@checkli/why-employees-quit-20-stats-employers-need-to-know-b921c253f767.

Yadav, S., Jain, A. and Singh, D. (2018) 'Early Prediction of Employee Attrition using Data Mining Techniques', *Proceedings of the 8th International Advance Computing Conference, IACC 2018*, 8(2882), pp. 349–354. doi: 10.1109/IADCC.2018.8692137.