Predicting Employee Attrition Based on Causal Factors

CSCI 113

Alegarbes, Deekimcheng, Felipe, Lee, Tan

CRISP-DM

Cross-Industry Standard Process for Data Mining

Consists of 6 Phases:

1. Business Understanding - Set Problem Statement and Objectives, Formulate Research Questions
2. Data Understanding - Collecting and Describing the Data, Finding Key Attributes, Checking Data Quality
3. Data Preparation - Cleaning, Transforming, and Selecting key features
4. Modeling - Implementing the appropriate modeling technique, setting key parameter values
5. Evaluation - Evaluate model's performance using testing set
6. Deployment - Summarize results and create a report

BUSINESS UNDERSTANDING

Problem Statement

Business Objective

- Identify and predict which of the current employees in the company are likely to leave the company
- Reduce long-term attrition rate by using the predictive model when hiring new employees, who are less likely to leave the company.

Data Mining Objectives

- Feature Selection- Identifying the factors that contribute to employee attrition.
- Classification - Developing a predictive model that can accurately classify which employees are most likely to leave the company based on known data.

Problem Statement

A company wants to identify the most important factors which lead to employee attrition and predict which of their current employees are more likely to leave the company. Employee attrition is defined as the process of employees leaving an organization over a period of time, either voluntarily or involuntarily. A high attrition rate can signify poor productivity and morale among the employees of the company. Currently, the company has data on current and past employees, which contains information about their age, monthly income, work life balance, performance rating, and other features. The main data mining objective is to create a predictive model for employee attrition.

DATA UNDERSTANDING

Dataset

	A	B	C	D	E	F	G	H	I	J
1	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber
23	36	Yes	Travel_Rarely	1218	Sales	9	4	Accounting	1	27
35	39	Yes	Travel_Rarely	895	Sales	5	3	Technical Degree	1	42
38	50	Yes	Travel_Rarely	869	Sales	3	2	Marketing	1	47
39	35	No	Travel_Rarely	890	Sales	2	3	Marketing	1	49
88	23	No	Travel_Rarely	541	Sales	2	1	Technical Degree	1	113

	K	L	M	N	O	P
	EnvironmentSatisfaction	Gender	HourlyRate	JobInvolvement	JobLevel	JobRole
3	3	Male	82	2	1	Inside Sales Rep
5	4	Male	56	3	2	Inside Sales Rep
8	1	Male	86	2	1	Inside Sales Rep
9	4	Female	97	3	1	Inside Sales Rep
8	3	Male	62	3	1	Inside Sales Rep

	Q	R	S	T	U	V
	JobSatisfaction	MaritalStatus	MonthlyIncome	MonthlyRate	NumCompaniesWorked	Over18
	1	Single	3407	6986	7	Y
	4	Married	2086	3335	3	Y
	3	Married	2683	3810	1	Y
	4	Married	2014	9687	1	Y
	1	Divorced	2322	9518	3	Y

Dataset Summary

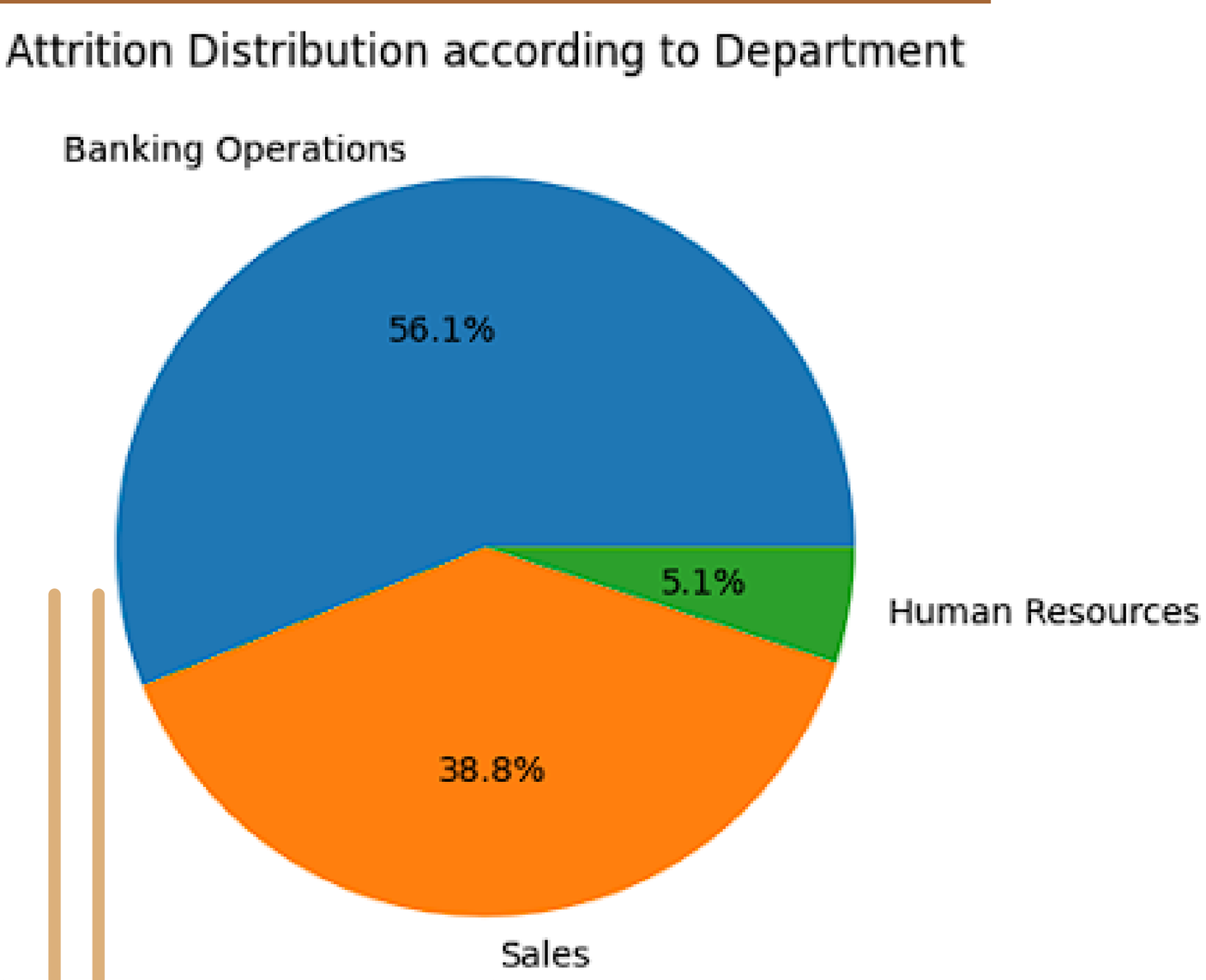
	Age	DailyRate	DistanceFromHome	Education	EmployeeCount	EmployeeNumber	EnvironmentSatisfaction	HourlyRate	JobInvolvement	JobLevel	JobSatisfaction	MonthlyIncome	MonthlyRate
count	1470.000000	1470.000000	1470.000000	1470.000000	1470.0	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000
mean	36.923810	802.485714	9.192517	2.912925	1.0	1024.865306	2.721769	65.891156	2.729932	2.063946	2.728571	6502.931293	14313.103401
std	9.135373	403.509100	8.106864	1.024165	0.0	602.024335	1.093082	20.329428	0.711561	1.106940	1.102846	4707.956783	7117.786044
min	18.000000	102.000000	1.000000	1.000000	1.0	1.000000	1.000000	30.000000	1.000000	1.000000	1.000000	1009.000000	2094.000000
25%	30.000000	465.000000	2.000000	2.000000	1.0	491.250000	2.000000	48.000000	2.000000	1.000000	2.000000	2911.000000	8047.000000
50%	36.000000	802.000000	7.000000	3.000000	1.0	1020.500000	3.000000	66.000000	3.000000	2.000000	3.000000	4919.000000	14235.500000
75%	43.000000	1157.000000	14.000000	4.000000	1.0	1555.750000	4.000000	83.750000	3.000000	3.000000	4.000000	8379.000000	20461.500000
max	60.000000	1499.000000	29.000000	5.000000	1.0	2068.000000	4.000000	100.000000	4.000000	5.000000	4.000000	19999.000000	26999.000000

NumCompaniesWorked	PercentSalaryHike	PerformanceRating	RelationshipSatisfaction	StandardHours	StockOptionLevel	TotalWorkingYears	TrainingTimesLastYear	WorkLifeBalance
1470.000000	1470.000000	1470.000000	1470.000000	1470.0	1470.000000	1470.000000	1470.000000	1470.000000
2.693197	15.209524	3.153741	2.712245	80.0	0.793878	11.279592	2.799320	2.761224
2.498009	3.659938	0.360824	1.081209	0.0	0.852077	7.780782	1.289271	0.706476
0.000000	11.000000	3.000000	1.000000	80.0	0.000000	0.000000	0.000000	1.000000
1.000000	12.000000	3.000000	2.000000	80.0	0.000000	6.000000	2.000000	2.000000
2.000000	14.000000	3.000000	3.000000	80.0	1.000000	10.000000	3.000000	3.000000
4.000000	18.000000	3.000000	4.000000	80.0	1.000000	15.000000	3.000000	3.000000
9.000000	25.000000	4.000000	4.000000	80.0	3.000000	40.000000	6.000000	4.000000

YearsAtCompany	YearsInCurrentRole	YearsSinceLastPromotion	YearsWithCurrManager
1470.000000	1470.000000	1470.000000	1470.000000
7.008163	4.229252	2.187755	4.123129
6.126525	3.623137	3.222430	3.568136
0.000000	0.000000	0.000000	0.000000
3.000000	2.000000	0.000000	2.000000
5.000000	3.000000	1.000000	3.000000
9.000000	7.000000	3.000000	7.000000
40.000000	18.000000	15.000000	17.000000

Preliminary Observations

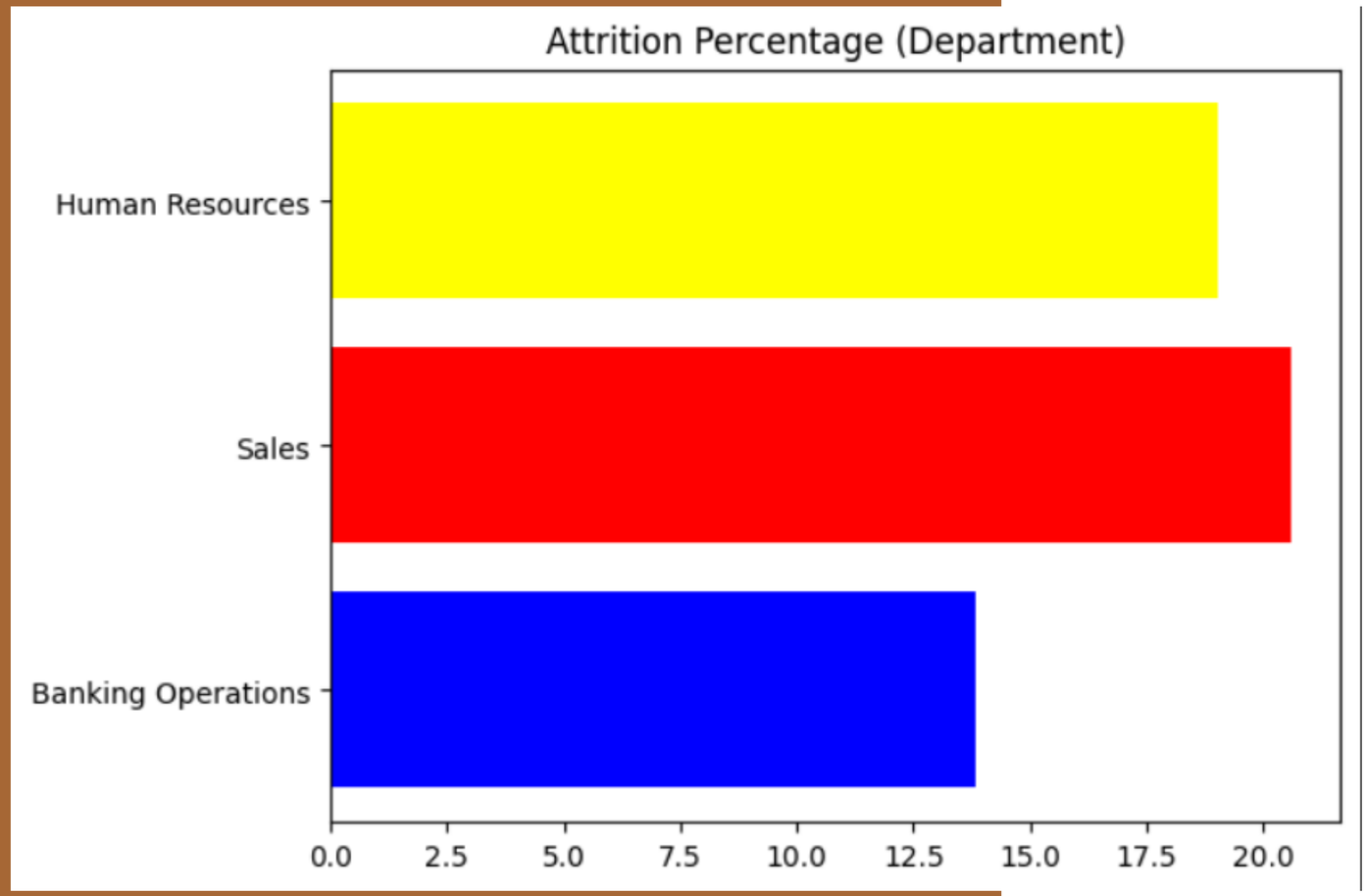
- The values of EmployeeNumber are unique to each employee and can be made the index of the dataframe.
- The attributes: EmployeeCount, StandardHours have the same value for all employees, and thus can be removed from the dataset.
- For the categorical data, Over18 has the same value, "Y" for all employees so it can also be removed.



Descriptive Analysis

56.1%

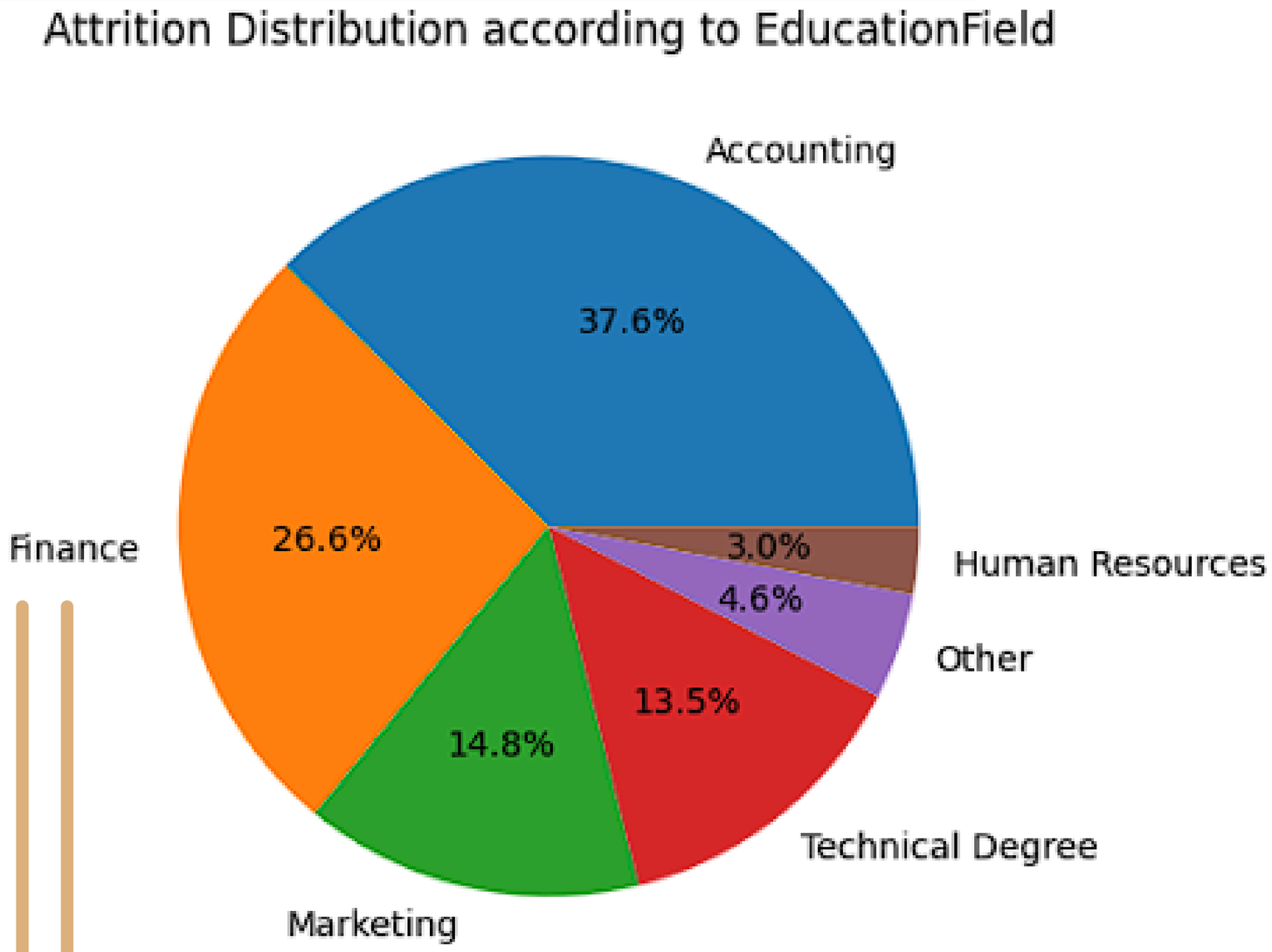
of the employees who left the company came from the Banking Operations Department while only a small percentage came from the Human Resources Department



Descriptive Analysis

Sales

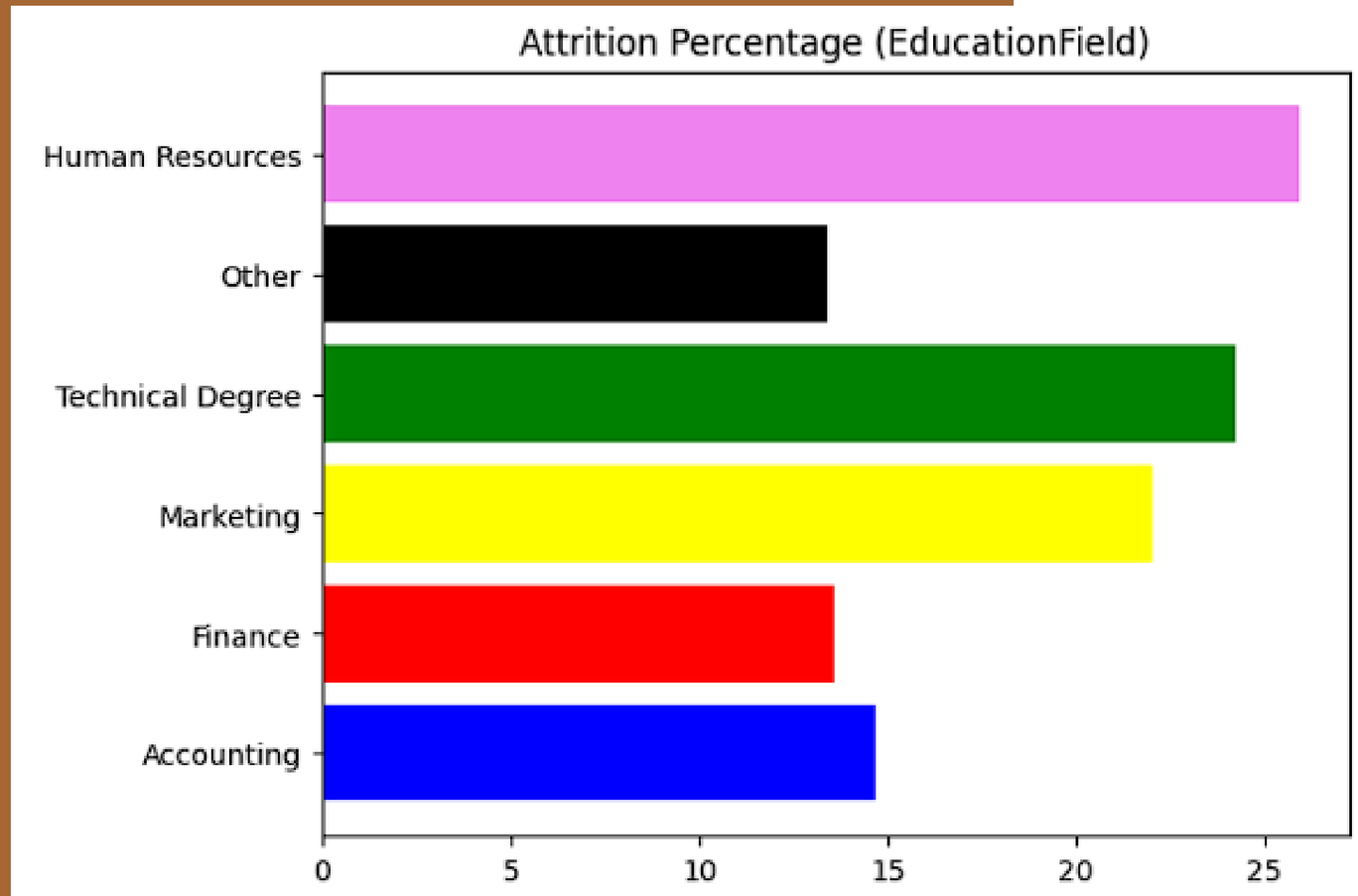
is the department with the highest percentage of their employees leaving the company.



Descriptive Analysis

37.6%

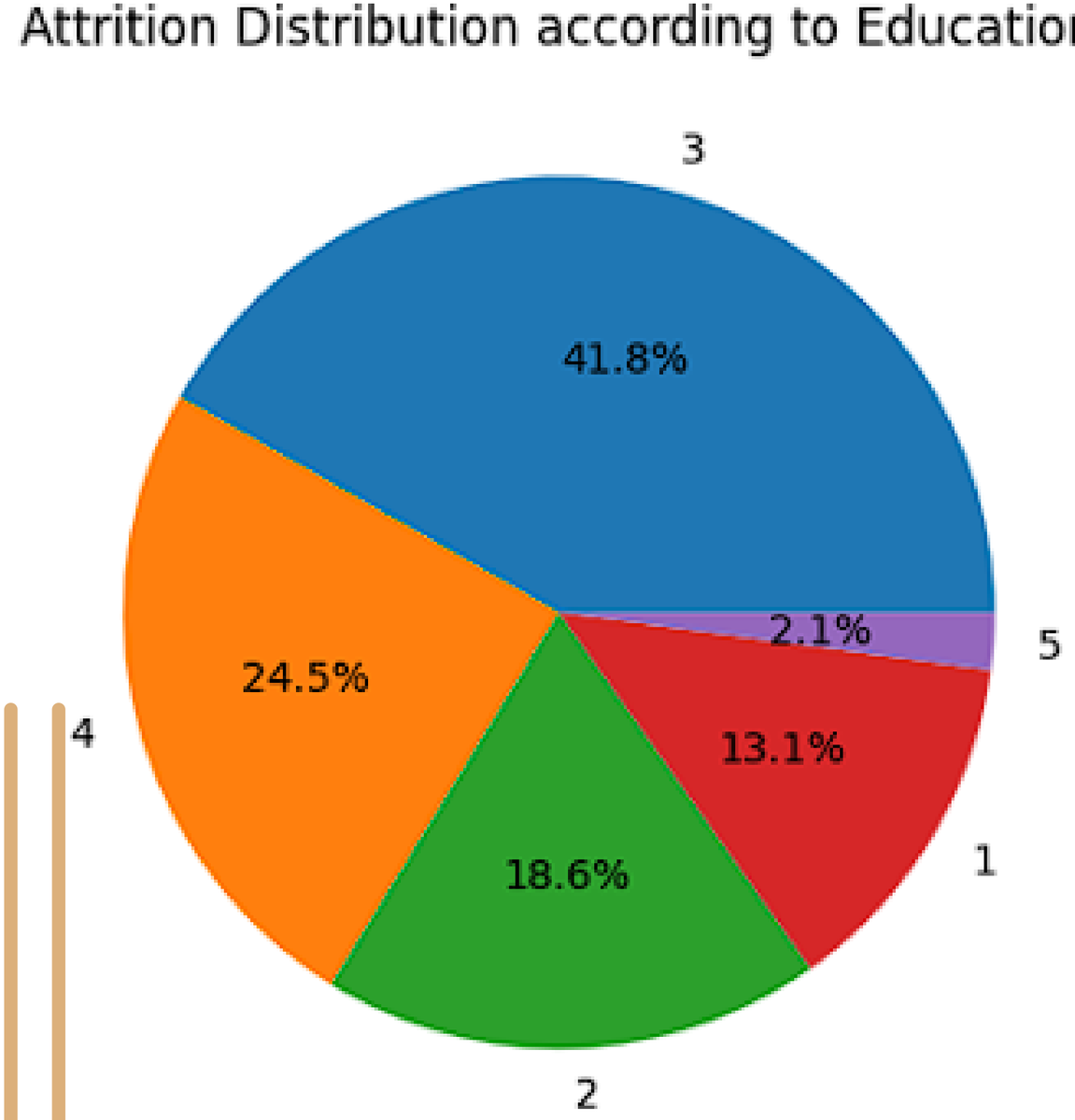
of the employees who left the company came from a course in the Accounting Field, followed by the Finance Field, while only a small percentage studied a course in the Human Resources Field.



Descriptive Analysis

HR

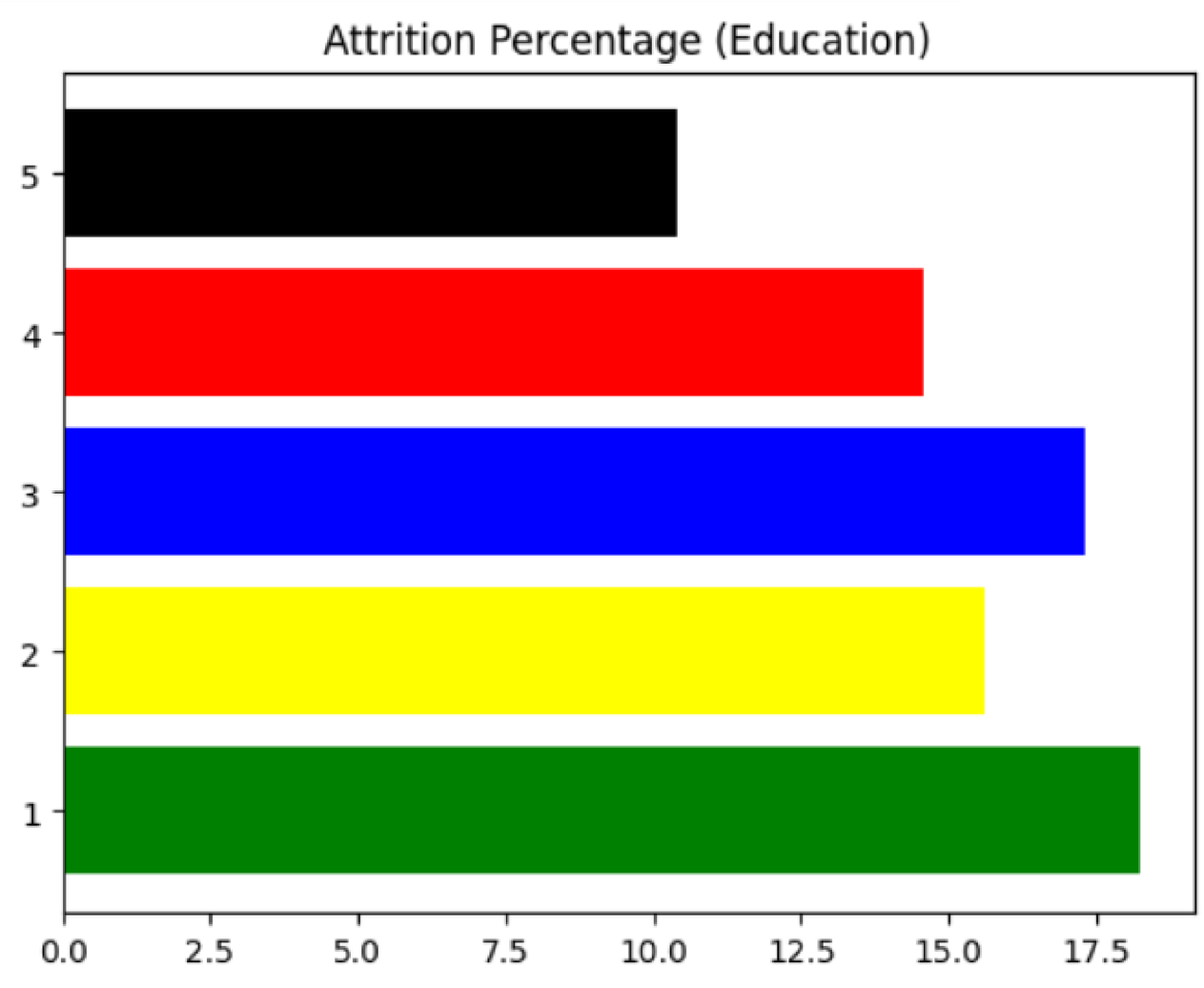
is the department with the highest percentage of their employees leaving the company; despite having the lowest number of employees who left.



Descriptive Analysis

41.8%

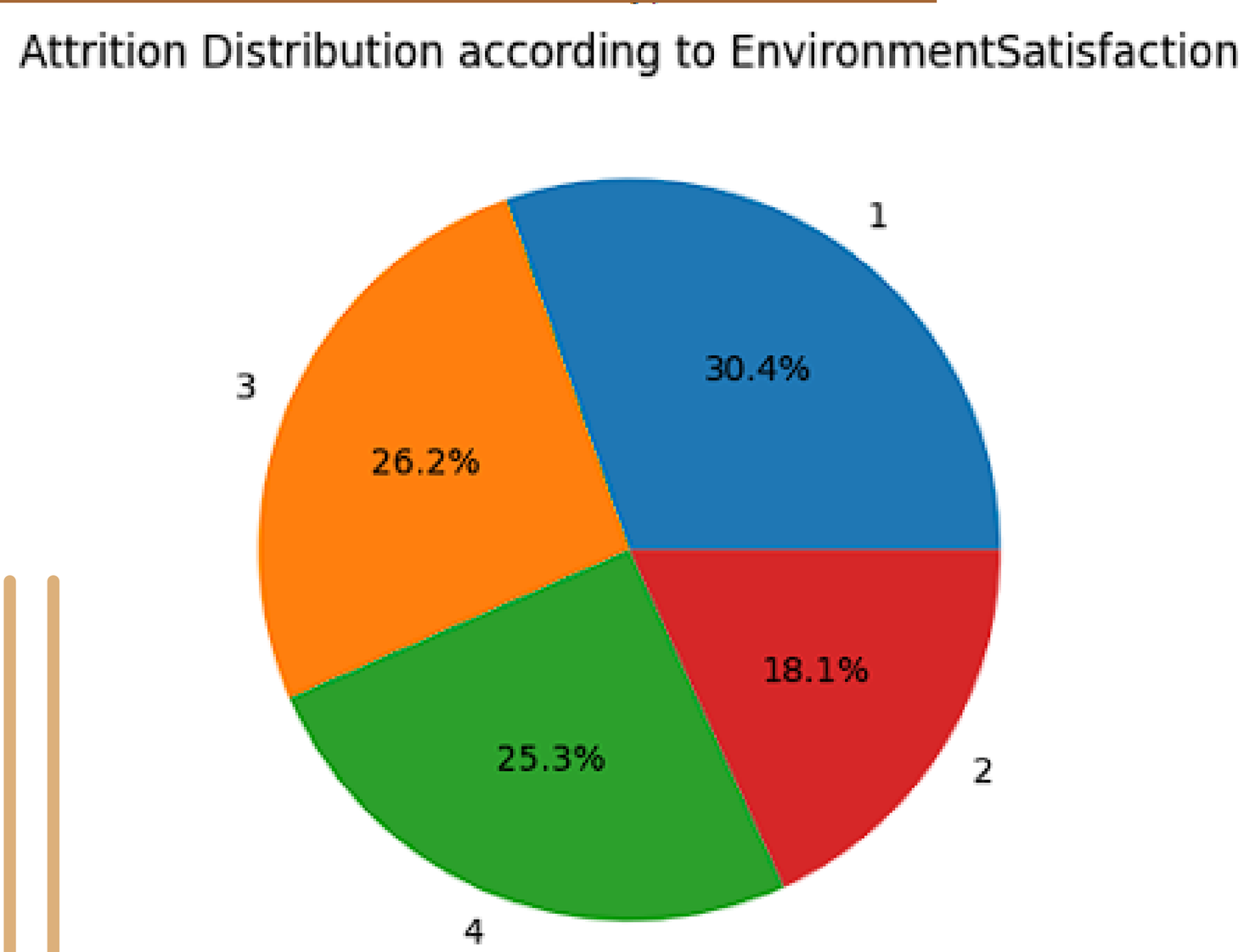
of the employees who left the company have a Level 3 education, followed by those with a Level 4 education, while only a small percentage have a Level 5 education.



Descriptive Analysis

Level 1

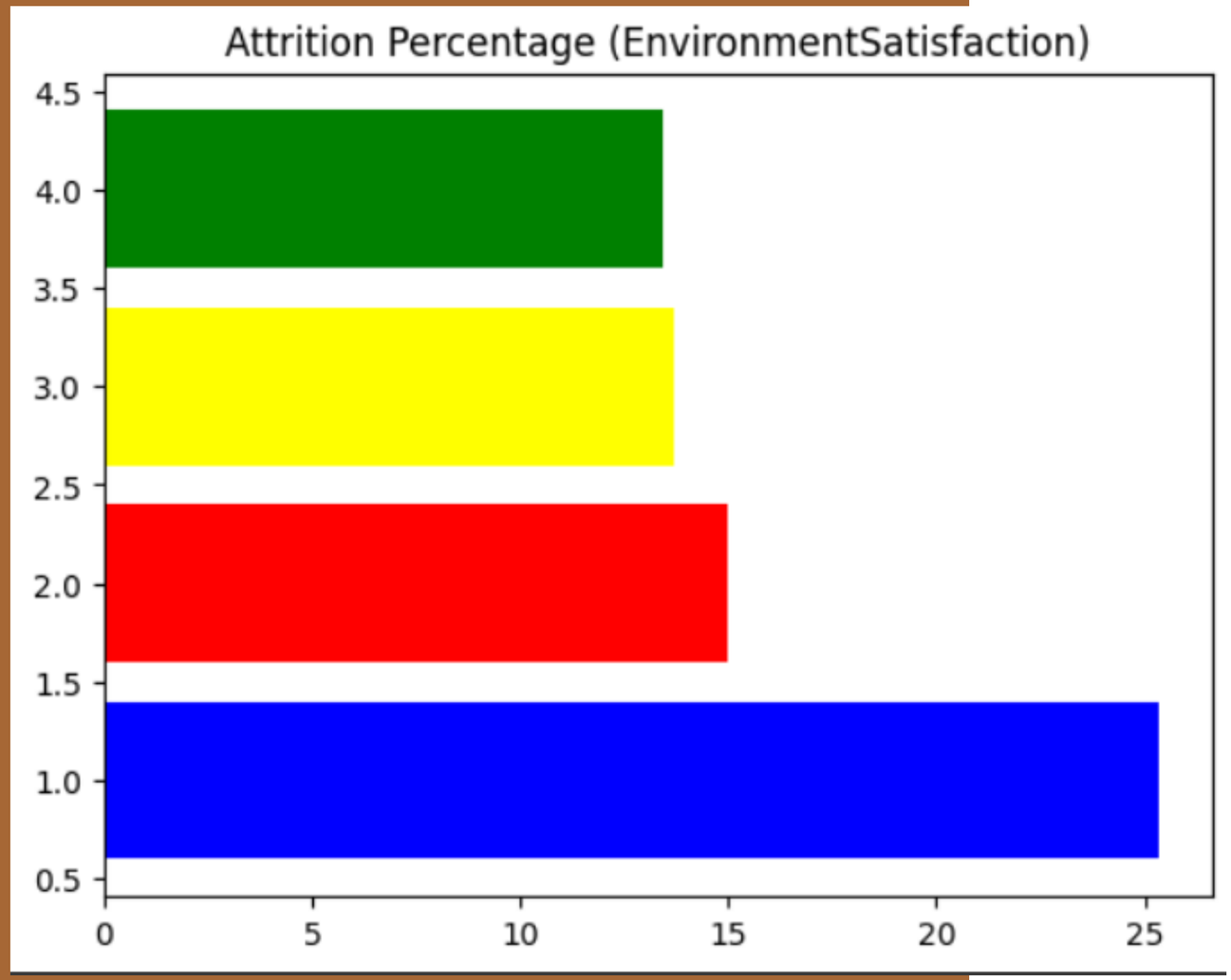
is the educational level with the highest percentage of their group leaving the company.



Descriptive Analysis

30.4%

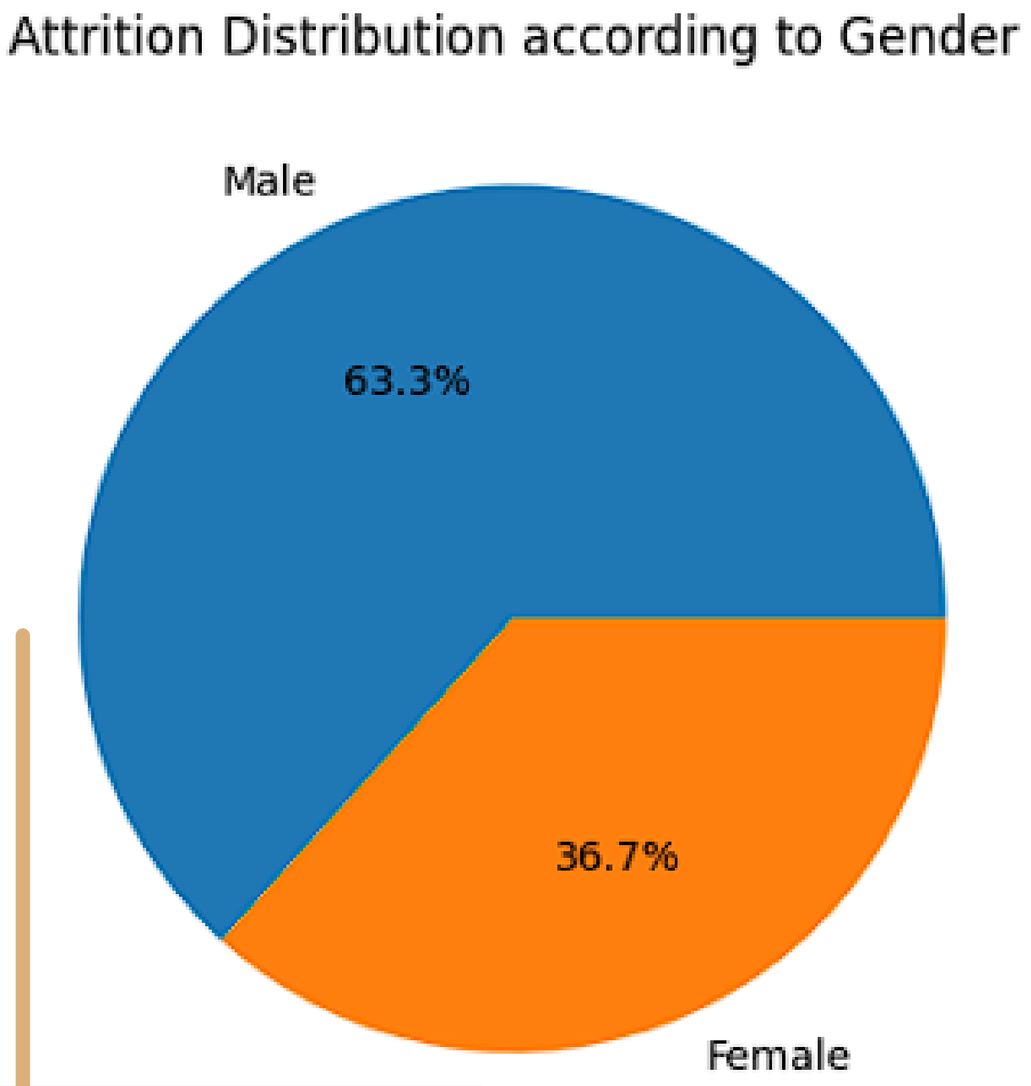
of the employees who left the company have the lowest level of environment satisfaction while the percentage difference is not that huge than those who have higher environment satisfaction



Descriptive Analysis

Level 1

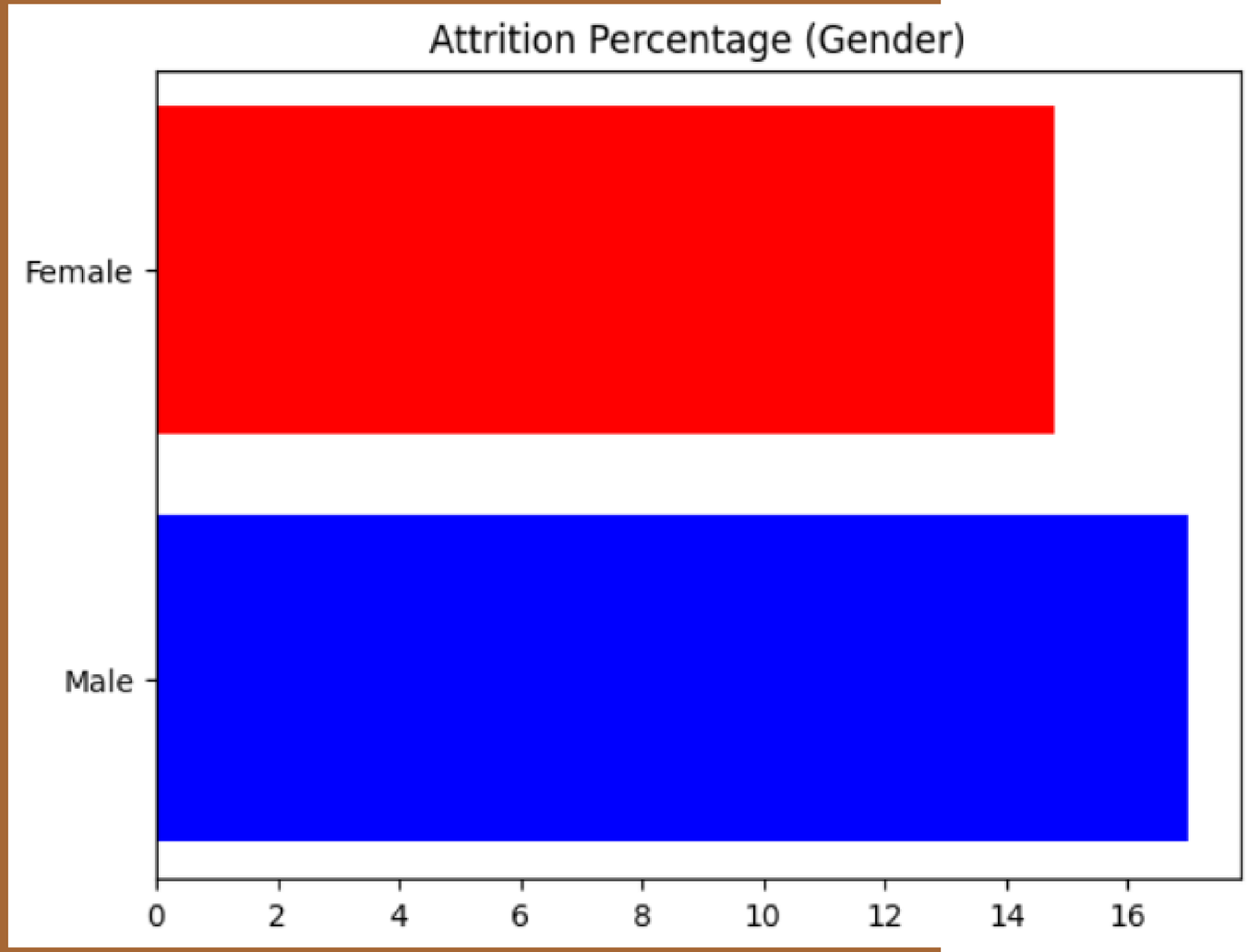
is the environment satisfaction level with the highest percentage of their group leaving the company.



Descriptive Analysis

63.3%

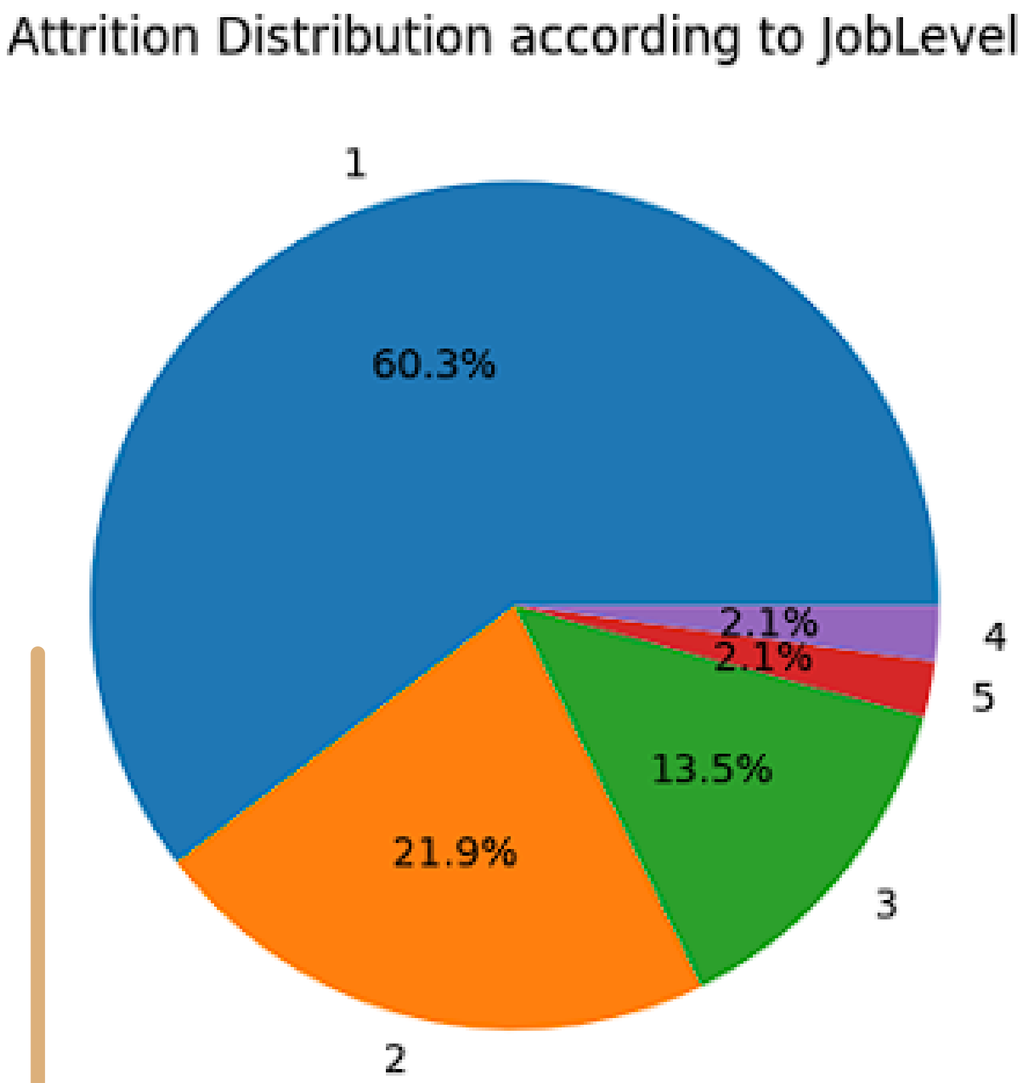
of the employees who left the company are male which could mean that those who are males are likely to leave the company.



Descriptive Analysis

Male

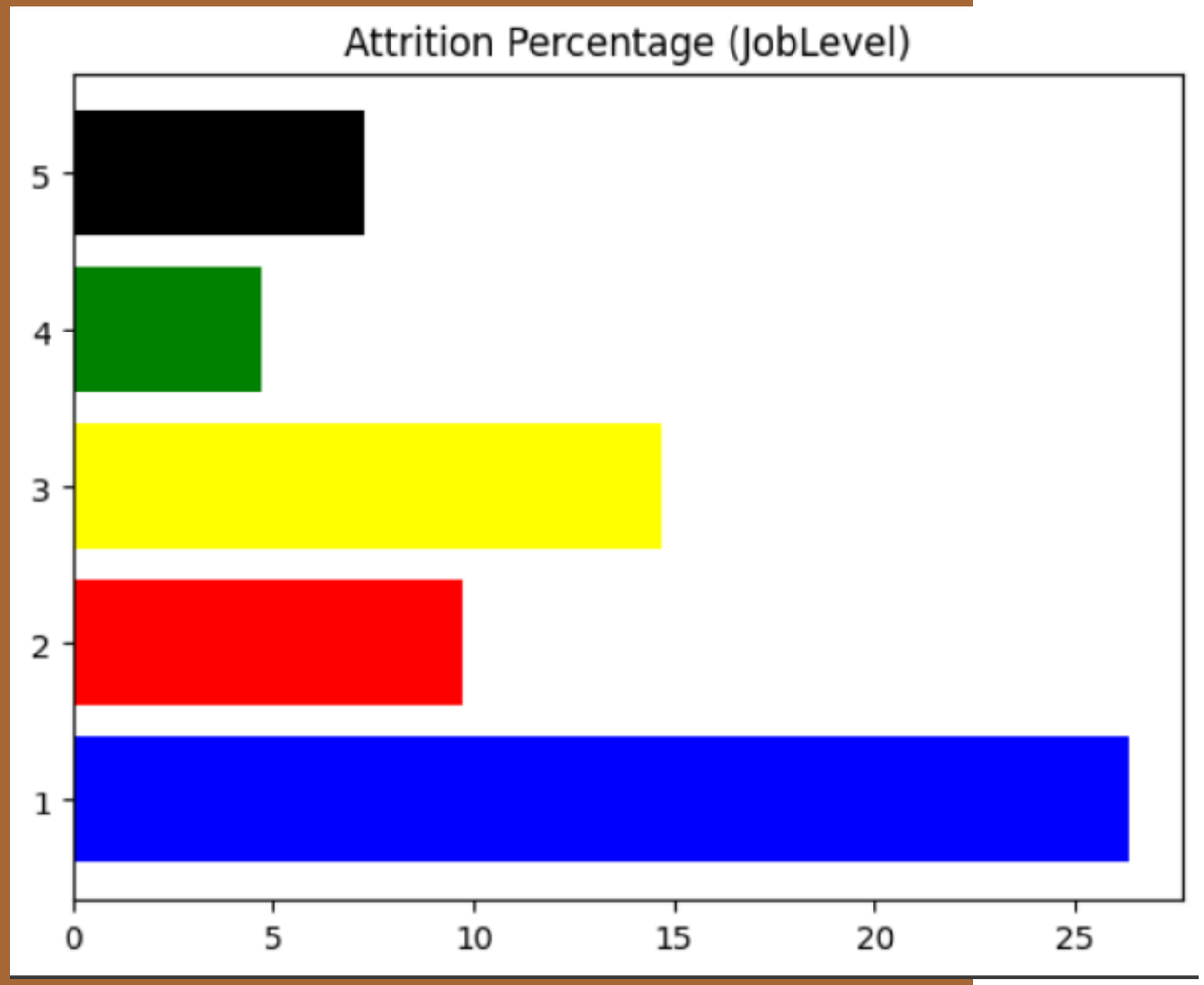
A larger percentage of male employees leave the company as compared to females.



Descriptive Analysis

60.3%

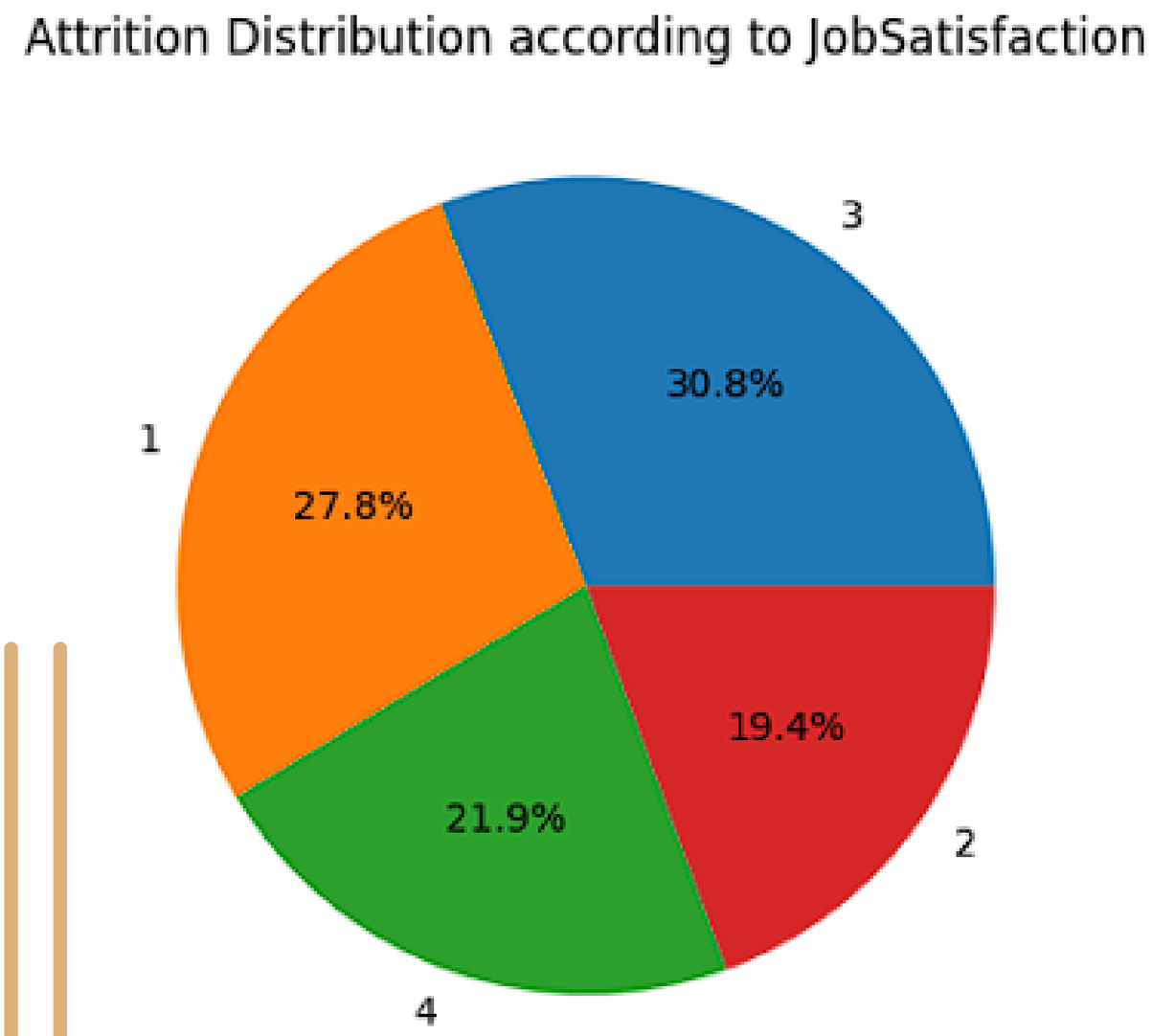
of the employees who left the company have a Level 1 Job while only about 4.2% of them have job levels 4 and 5.



Descriptive Analysis

Job Level 1

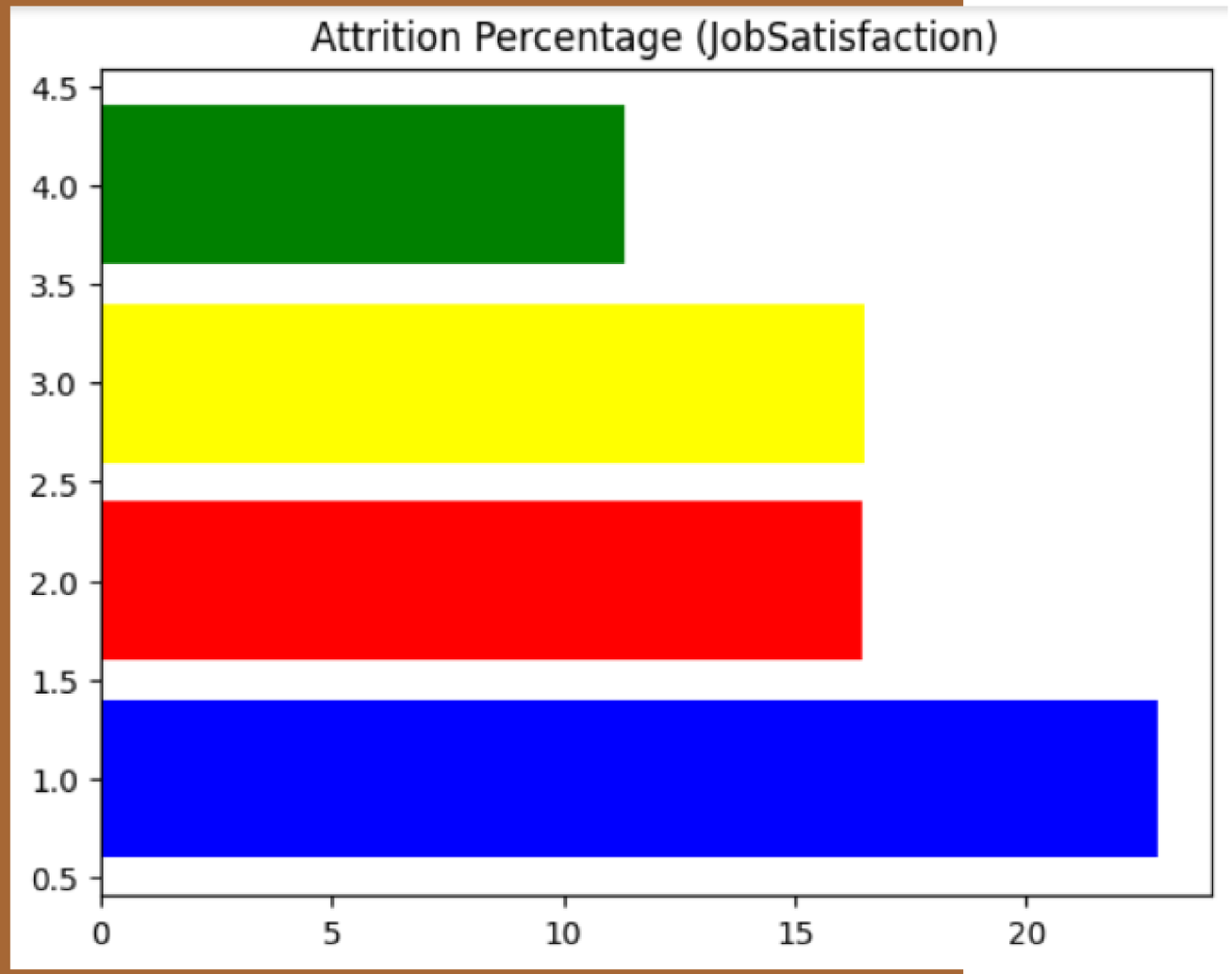
The entry-level jobs showed the largest percentage of employee turnover. This supports the claim that entry level jobs are more replaceable compared to managerial positions.



Descriptive Analysis

30.8%

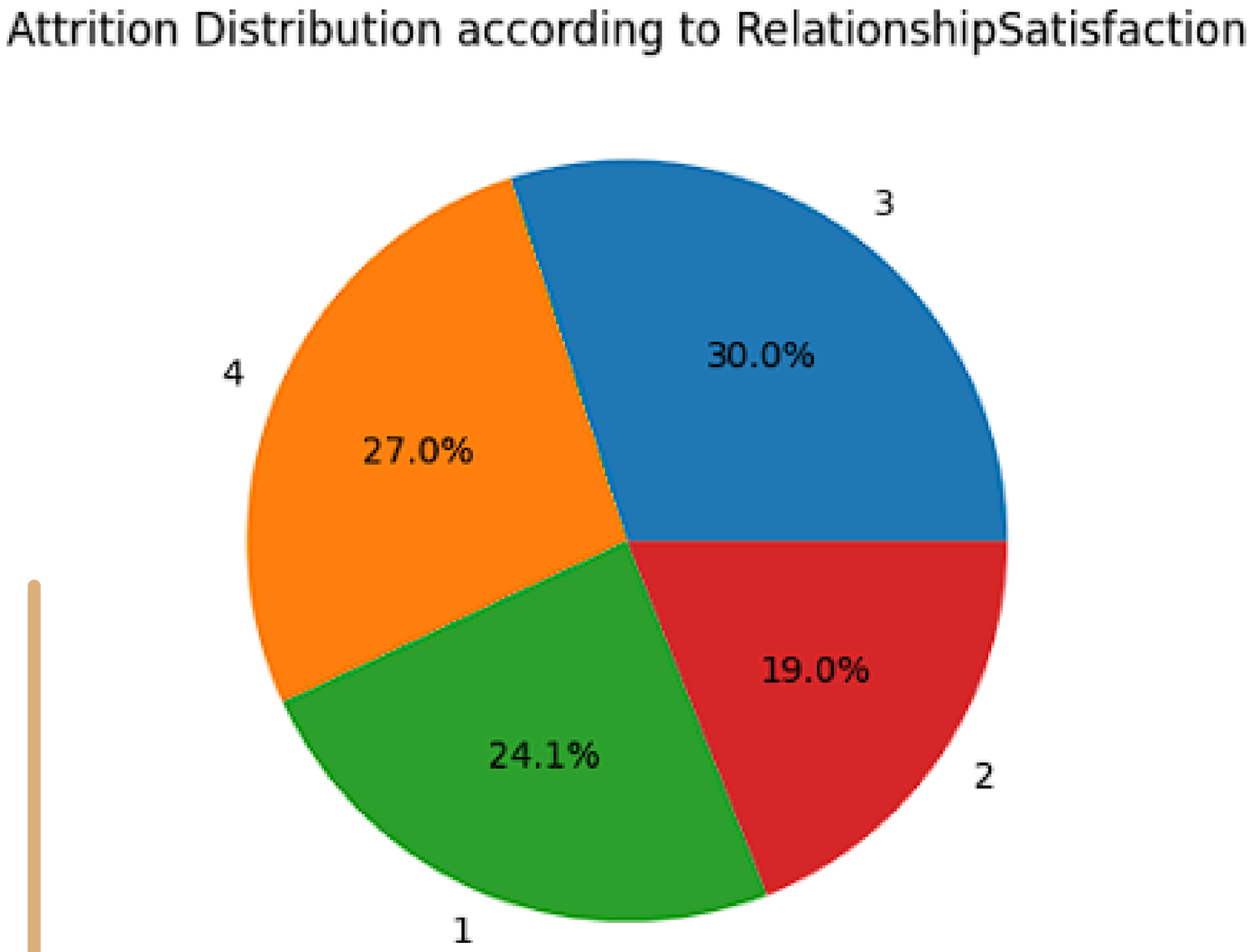
of the employees who left the company have a Job Satisfaction rating of 3 followed by those who have a rating of 1, 4, and lastly, 2.



Descriptive Analysis

Level 1

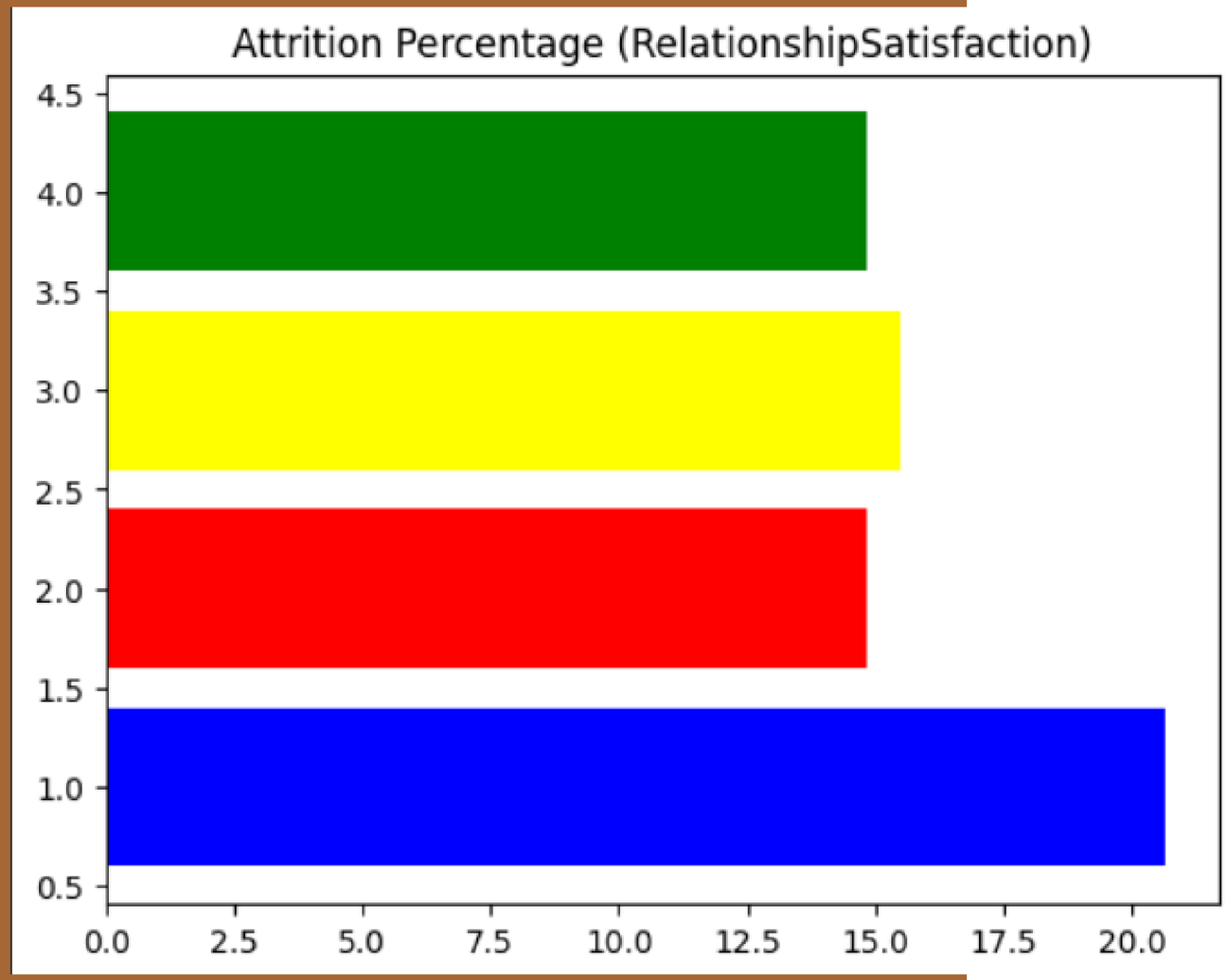
is the Job Satisfaction level with the highest percentage of their group leaving the company



Descriptive Analysis

30.0%

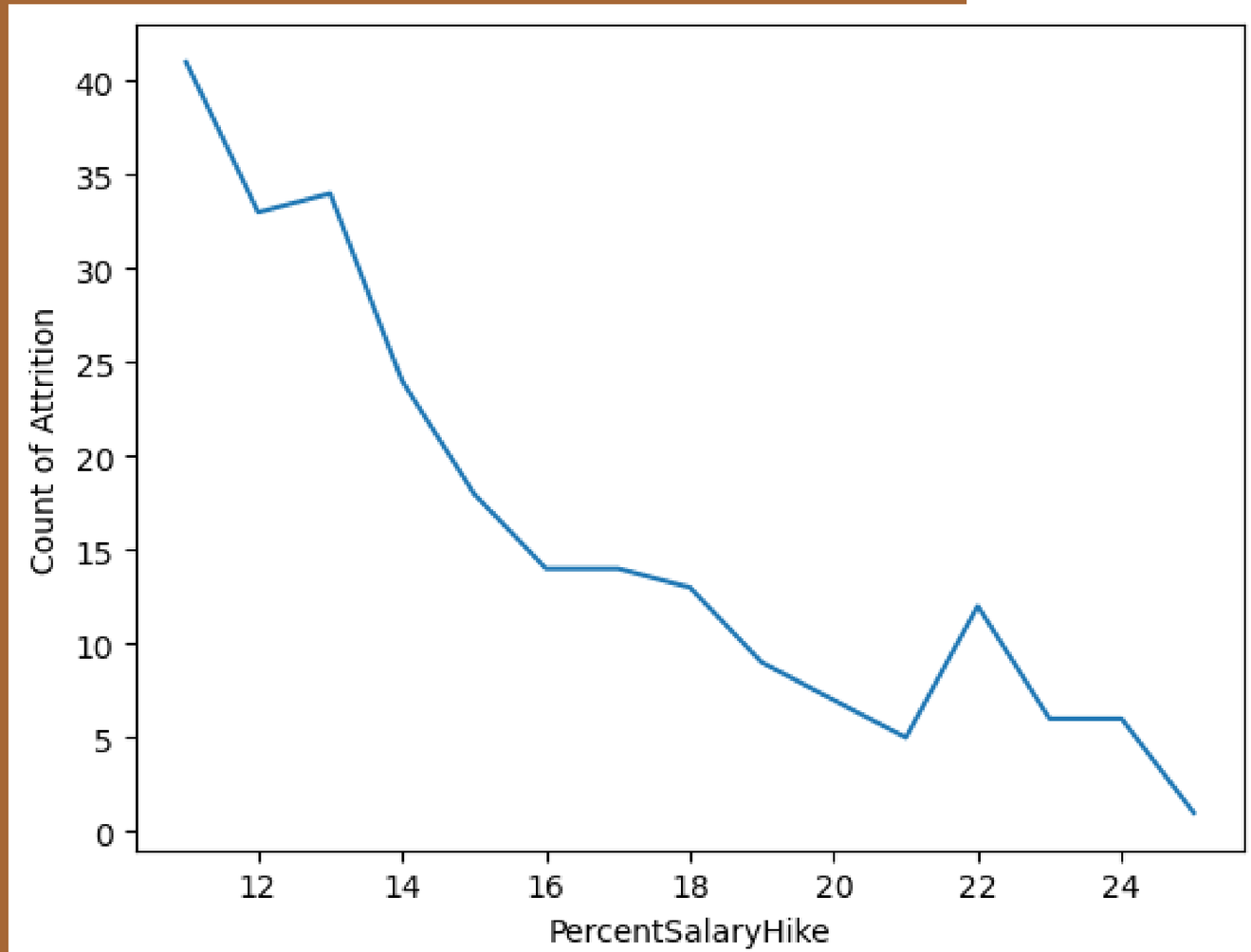
of the employees who left the company have a Relationship Satisfaction rating of 3 followed by those who have a rating of 4, 1, and lastly, 2.



Descriptive Analysis

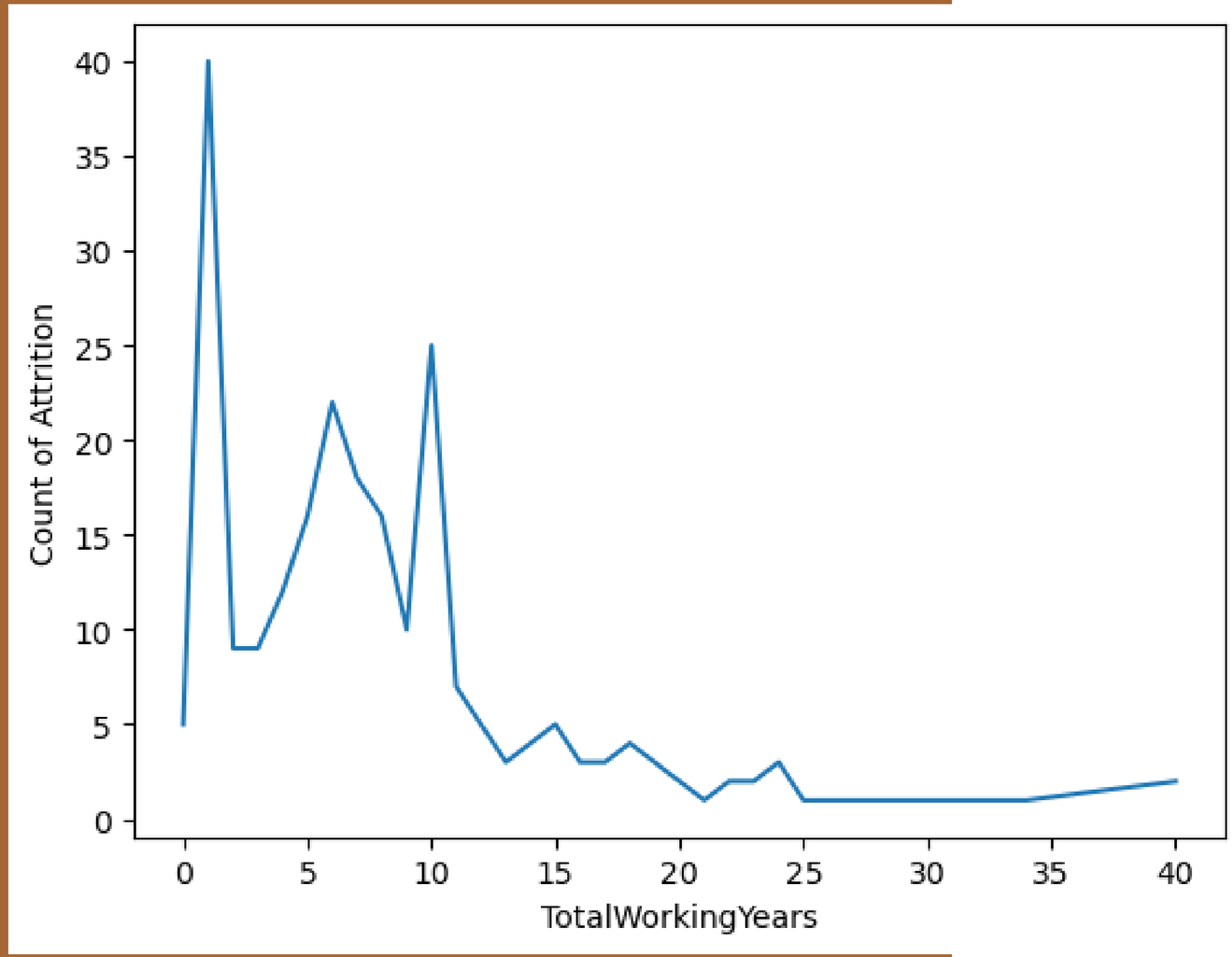
Level 1

is the relationship satisfaction level with the highest percentage of their group leaving the company.



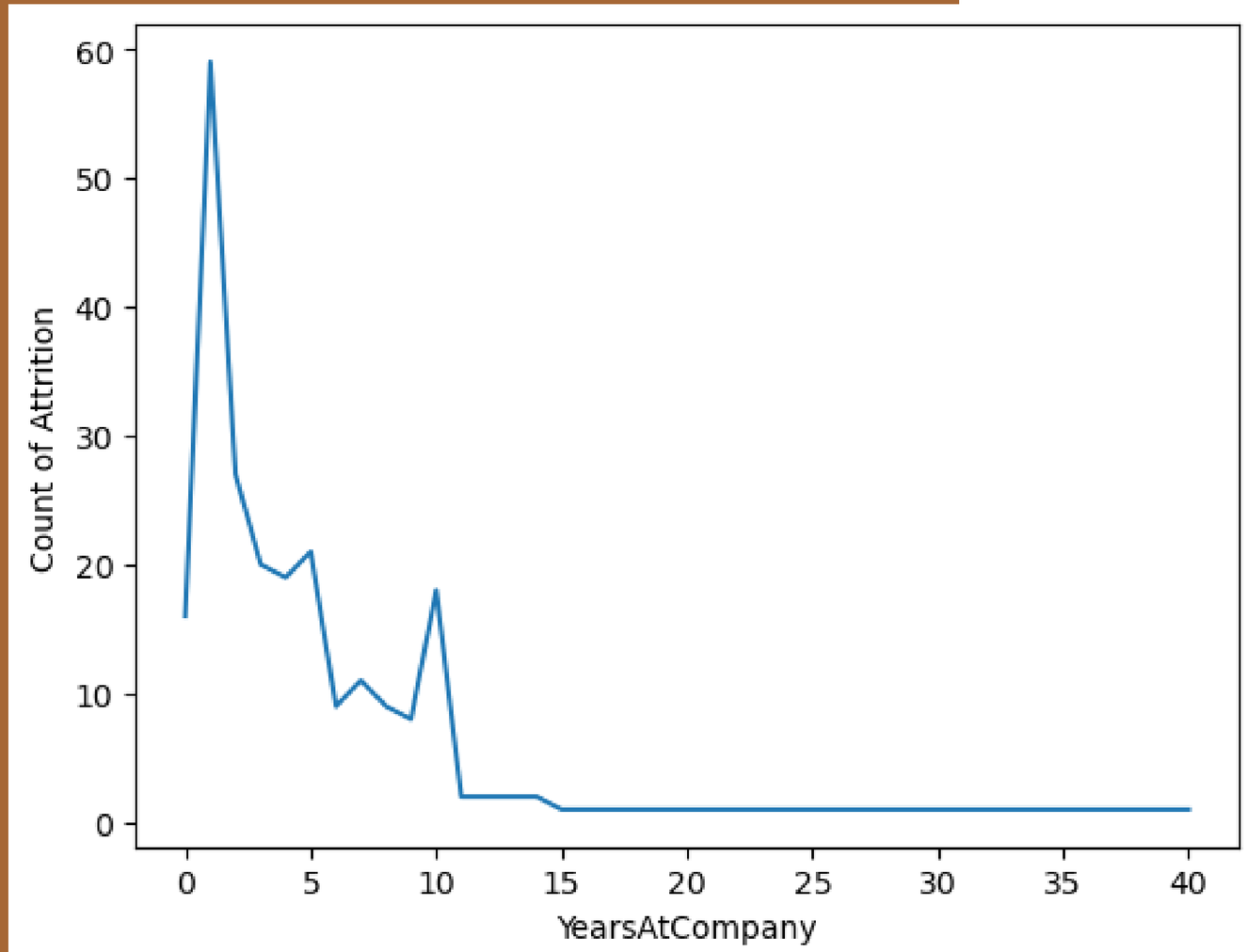
Descriptive Analysis

The line chart illustrates that the higher the percent salary hike of the employee, the less likely they are to leave the company.



Descriptive Analysis

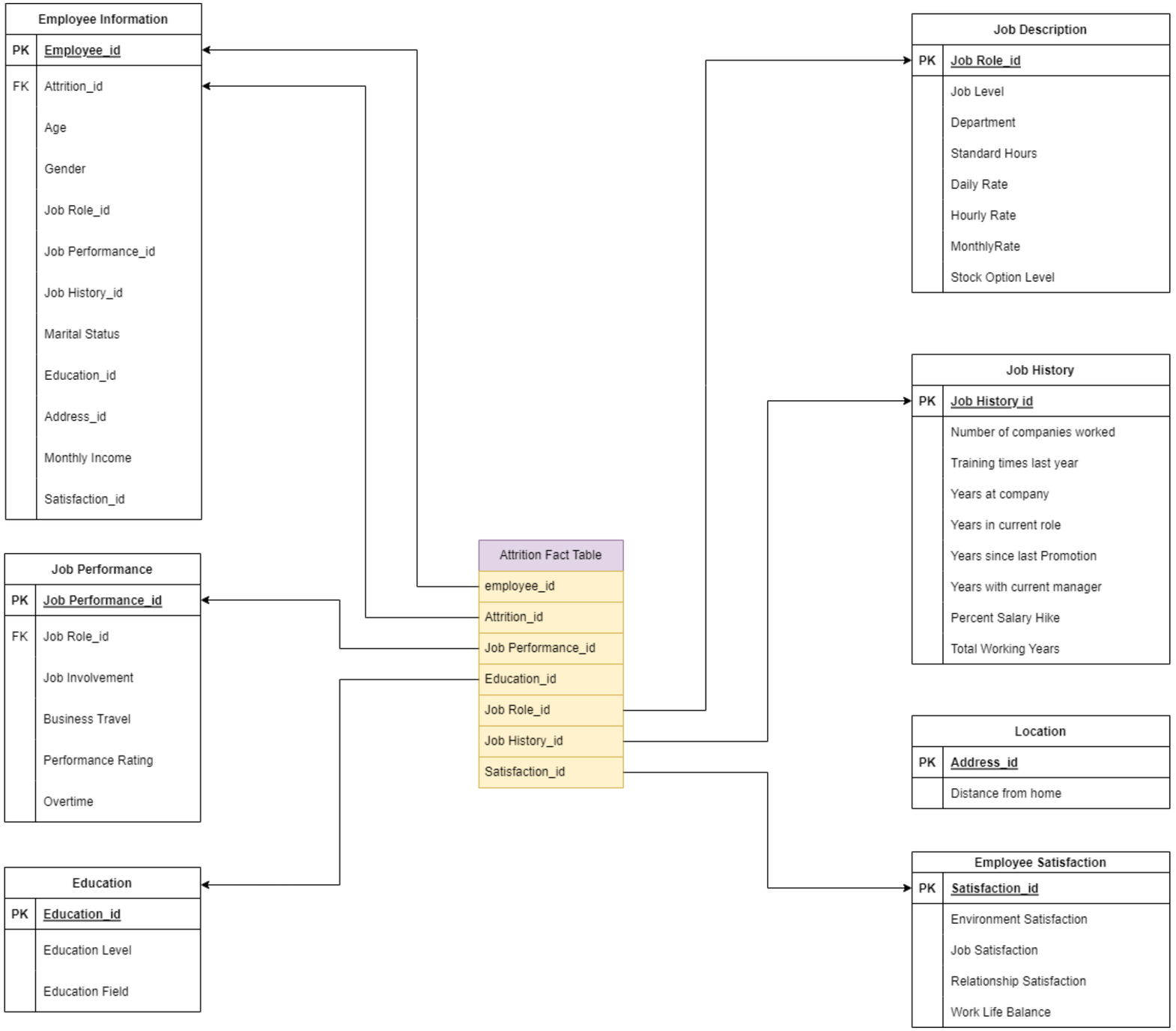
In this line chart, it can be observed that the higher the total working years of the employee, the less likely they are to leave the company.



Descriptive Analysis

The chart demonstrates that the higher the number of years at the company, the less likely they are to leave the company.

DATA PREPARATION



OLAP DESIGN

Based on the business objective, the following snowflake schema is developed. Features were grouped into different dimension tables to better understand the factors that might contribute to employee attrition

Pre-Processing

One-Hot Encoding

- To effectively use the categorical variables, data points were pre-processed by one-hot encoding all the categorical variables.

Normalization (min-max scaling)

- Moreover, data normalization is essential in modeling. Normalization helps to eliminate data redundancy and outliers.
- Normalization reduces the risk of data inconsistencies and errors, which can have severe consequences for the accuracy of your model. The researchers used min-max scaling was used to normalize the data.

Pre-Processed Data

	Attrition	Age	DailyRate	DistanceFromHome	Education	EnvironmentSatisfaction	HourlyRate	JobInvolvement	JobLevel	JobSatisfaction	MonthlyIncome	MonthlyRate	NumCompaniesWorked	PercentSalaryHike	PerformanceRating
1	1	0.547619	0.715820	0.000000	0.25	0.333333	0.914286	0.666667	0.25	1.000000	0.262454	0.698053	0.888889	0.000000	0.0
2	0	0.738095	0.126700	0.250000	0.00	0.666667	0.442857	0.333333	0.25	0.333333	0.217009	0.916001	0.111111	0.857143	1.0
4	1	0.452381	0.909807	0.035714	0.25	1.000000	0.885714	0.333333	0.00	0.666667	0.056925	0.012126	0.666667	0.285714	0.0
5	0	0.357143	0.923407	0.071429	0.75	1.000000	0.371429	0.666667	0.00	0.666667	0.100053	0.845814	0.111111	0.000000	0.0
7	0	0.214286	0.350036	0.035714	0.00	0.000000	0.142857	0.666667	0.00	0.333333	0.129489	0.583738	1.000000	0.071429	0.0

RelationshipSatisfaction	StockOptionLevel	TotalWorkingYears	TrainingTimesLastYear	WorkLifeBalance	YearsAtCompany	YearsInCurrentRole	YearsSinceLastPromotion	YearsWithCurrManager	BusinessTravel_Non-Travel	BusinessTravel_Travel_Frequently
0.000000	0.000000	0.200	0.000000	0.000000	0.150	0.222222	0.000000	0.294118	0	0
1.000000	0.333333	0.250	0.500000	0.666667	0.250	0.388889	0.066667	0.411765	0	1
0.333333	0.000000	0.175	0.500000	0.666667	0.000	0.000000	0.000000	0.000000	0	0
0.666667	0.000000	0.200	0.500000	0.666667	0.200	0.388889	0.200000	0.000000	0	1
1.000000	0.333333	0.150	0.500000	0.666667	0.050	0.111111	0.133333	0.117647	0	0

BusinessTravel_Travel_Rarely	Department_Banking Operations	Department_Human Resources	Department_Sales	EducationField_Accounting	EducationField_Finance	EducationField_Human Resources	EducationField_Marketing	EducationField_Other	EducationField_Technical Degree	Gender_Female	Gender_Male	JobRole_Bank Manager
1	0	0	1	1	0	0	0	0	0	1	0	0
0	1	0	0	1	0	0	0	0	0	0	1	0
1	1	0	0	0	0	0	0	1	0	0	1	0
0	1	0	0	1	0	0	0	0	0	1	0	0
1	1	0	0	0	1	0	0	0	0	0	1	0

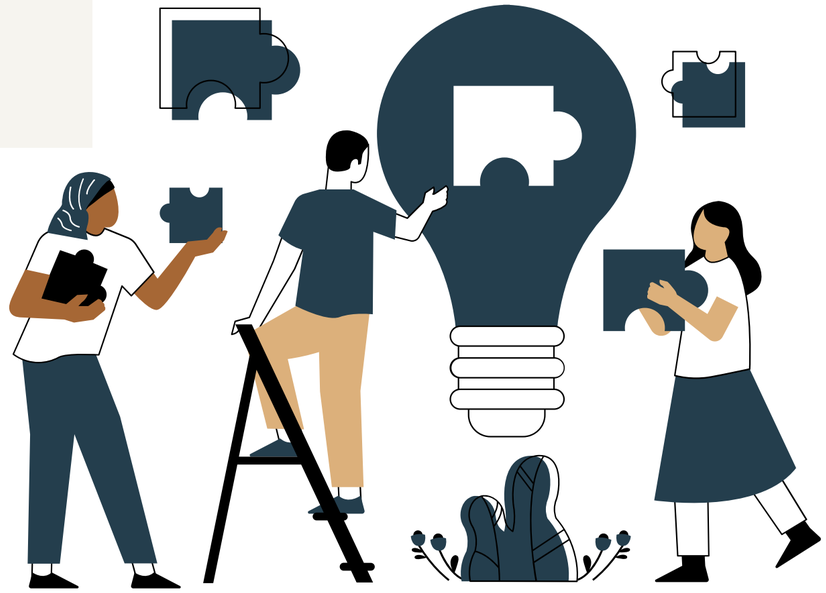
JobRole_Bank Specialist	JobRole_Bank Strategist	JobRole_Bank Teller	JobRole_Client Executive	JobRole_HR Manager	JobRole_Human Resources	JobRole_Inside Sales Rep	JobRole_Loans Officer	JobRole_Loans Supervisor	JobRole_Wealth Manager	MaritalStatus_Divorced	MaritalStatus_Married	MaritalStatus_Single	OverTime_No	OverTime_Yes
0	0	0	0	0	0	0	0	0	1	0	0	1	0	1
0	0	0	0	0	0	0	1	0	0	0	1	0	1	0
0	0	1	0	0	0	0	0	0	0	0	0	1	0	1
0	0	0	0	0	0	0	1	0	0	0	1	0	0	1
0	0	1	0	0	0	0	0	0	0	0	1	0	1	0

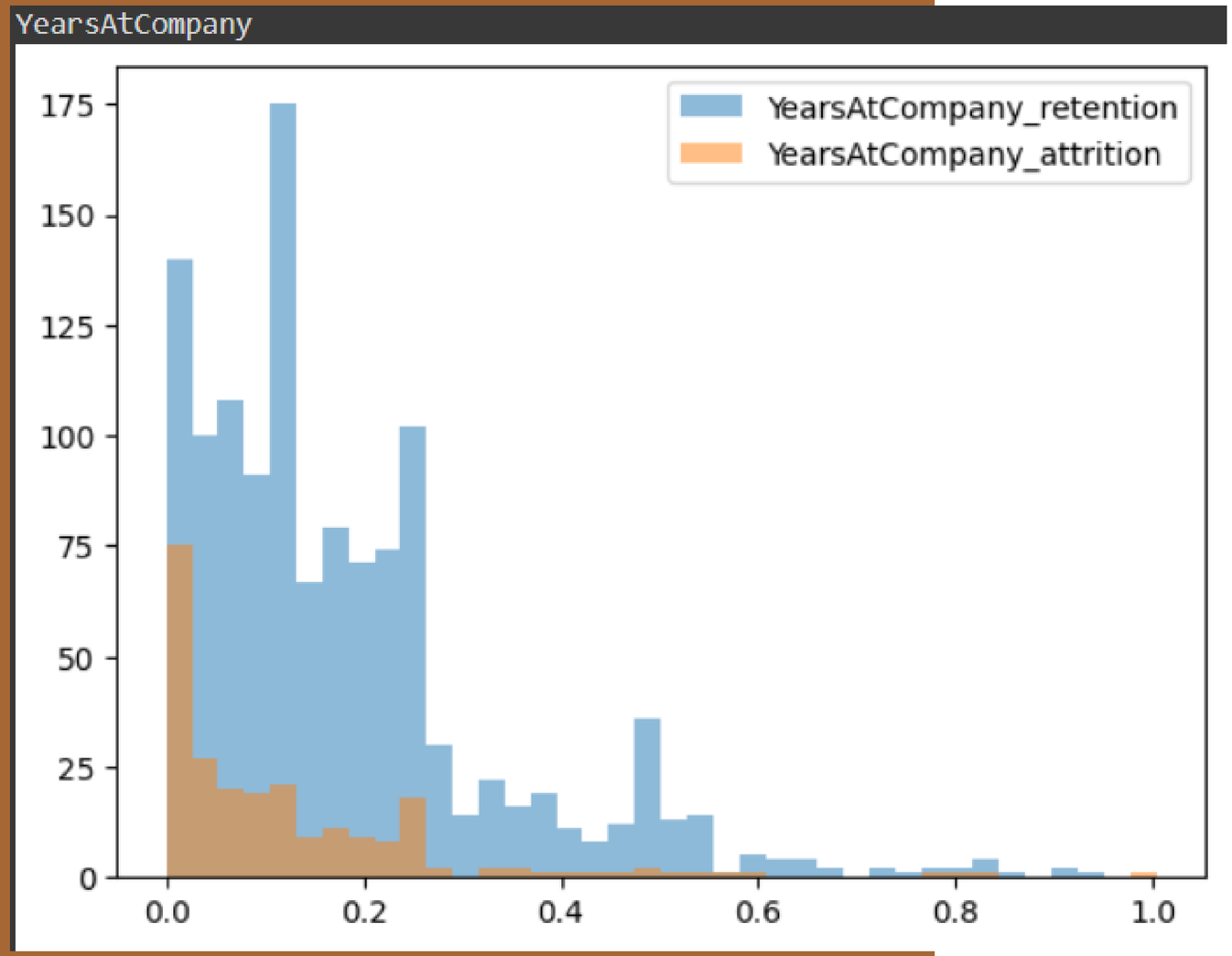
Feature Selection

A **Overlapping Histograms -** The feature with the smallest overlapping area is a better feature because it implies that the feature has a different distribution for its classes and may be a better indicator in prediction.

B **Scatterplot Matrix -** Based on statistics solutions, there is a high degree of correlation between two variables if the correlation coefficient is between ±0.50 and ±1.

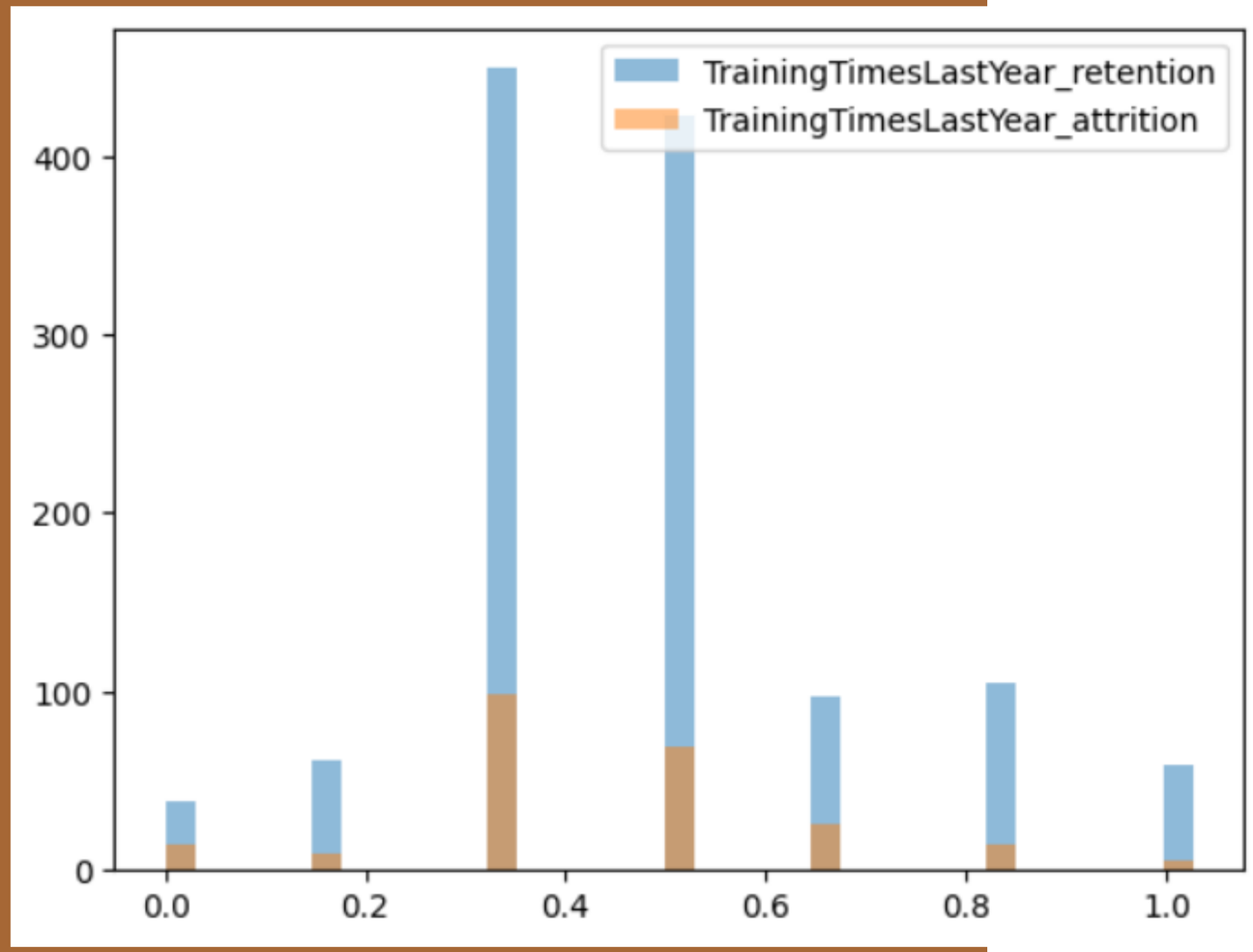
C **Parallel Lines -** Features that intersect the same point on the parallel lines are highly correlated and should be removed from the dataset.





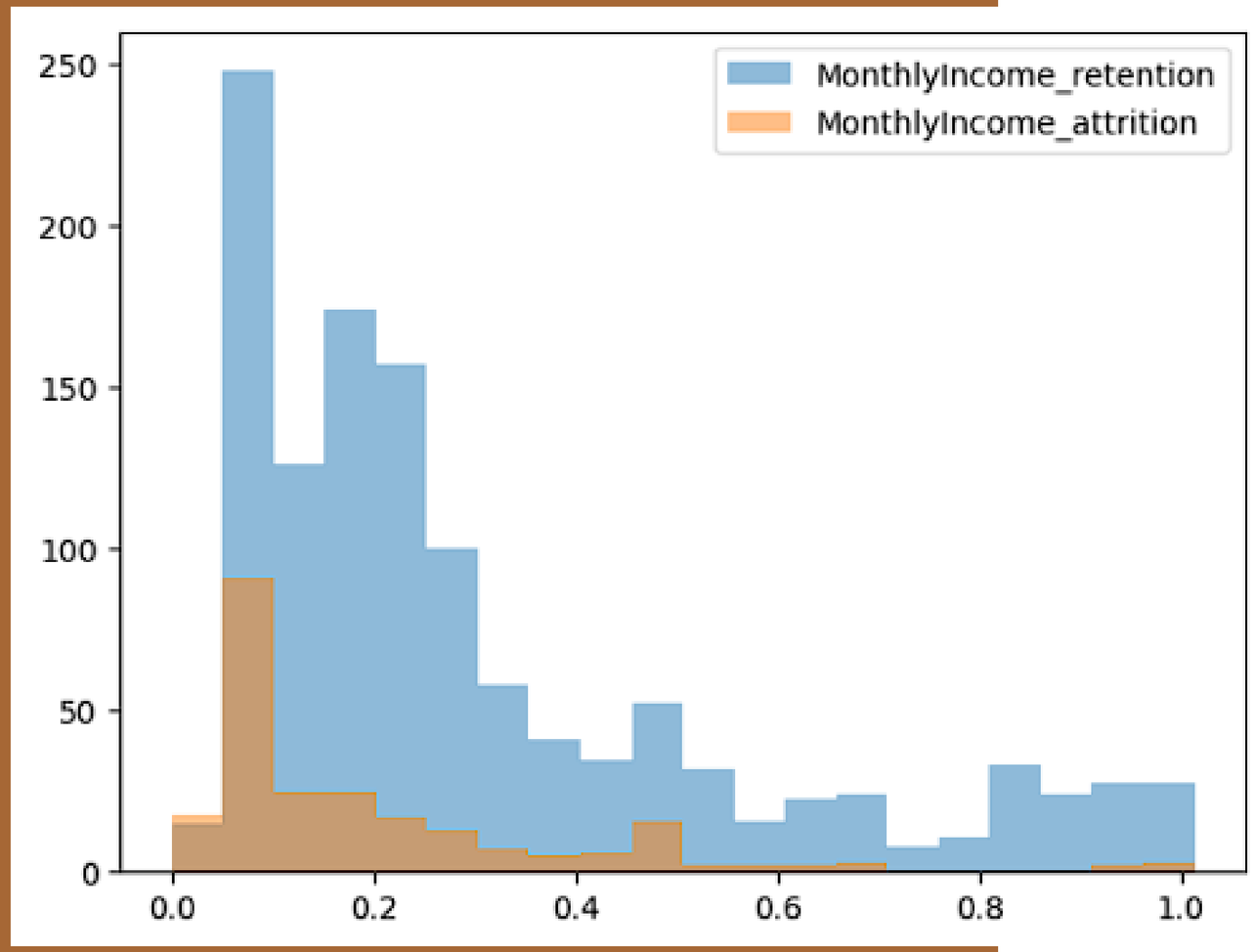
Overlapping Histograms

The following histograms have the lowest overlapping area:



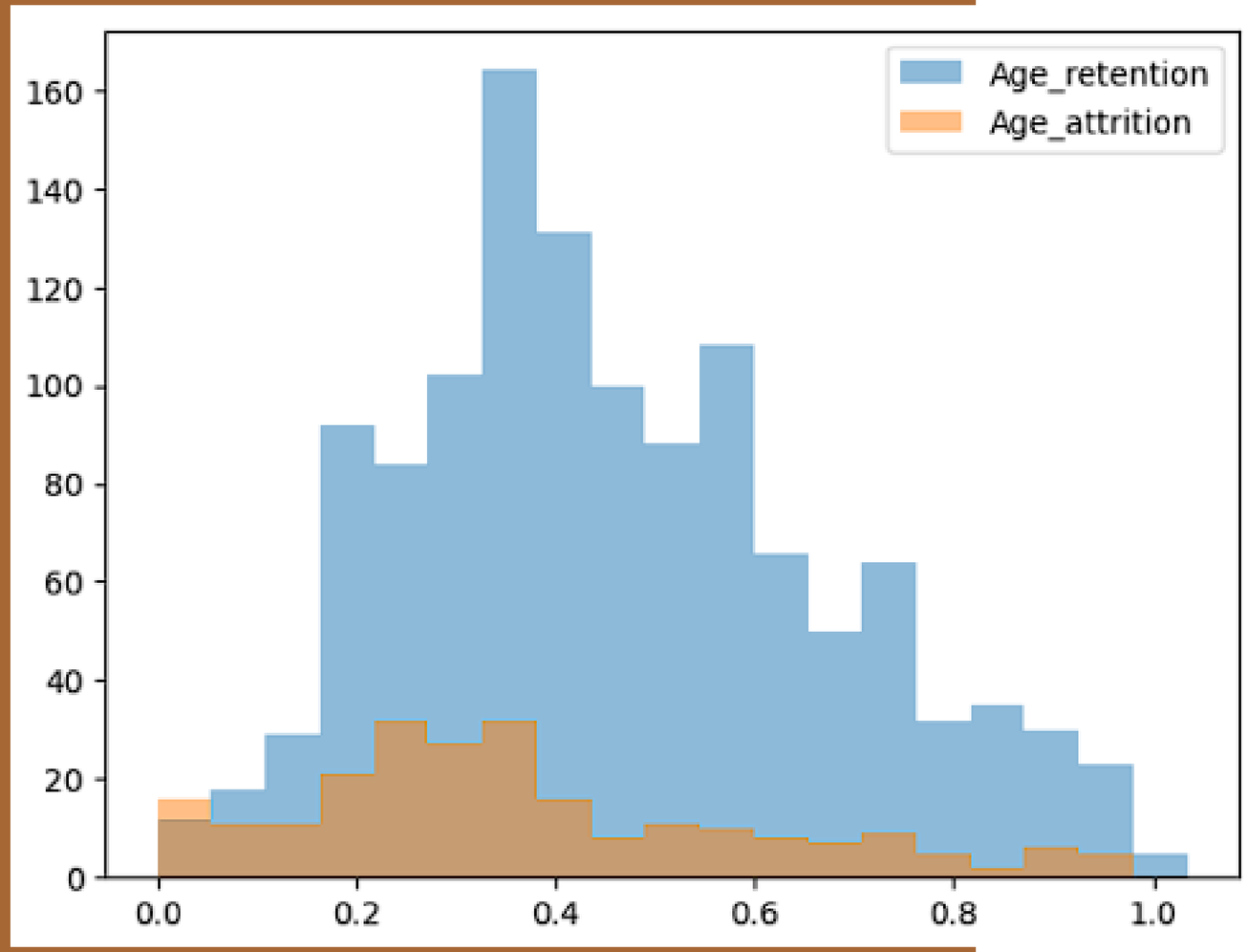
Overlapping Histograms

The following histograms have the lowest overlapping area:



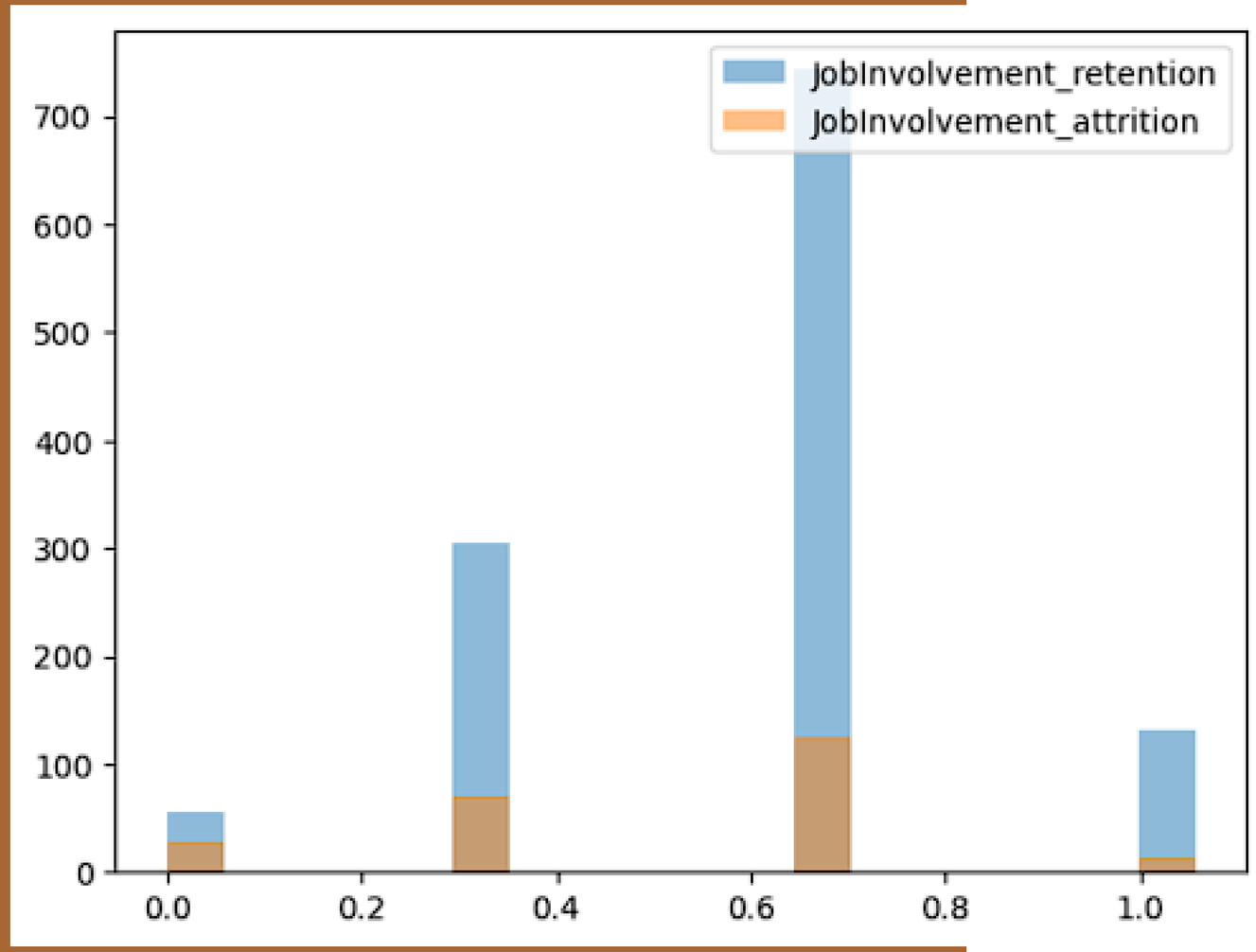
Overlapping Histograms

The following histograms have the lowest overlapping area:



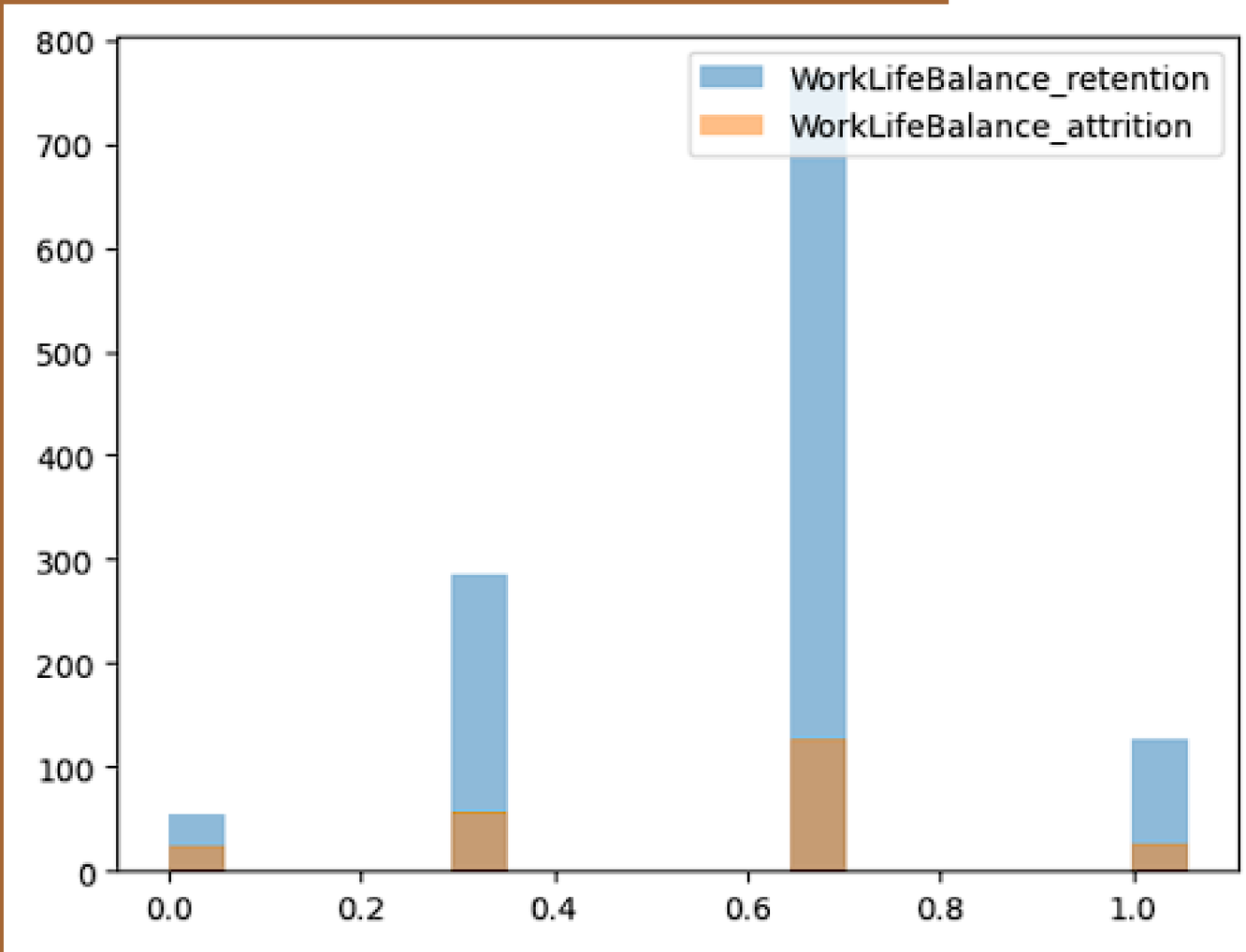
Overlapping Histograms

The following histograms have the lowest overlapping area:



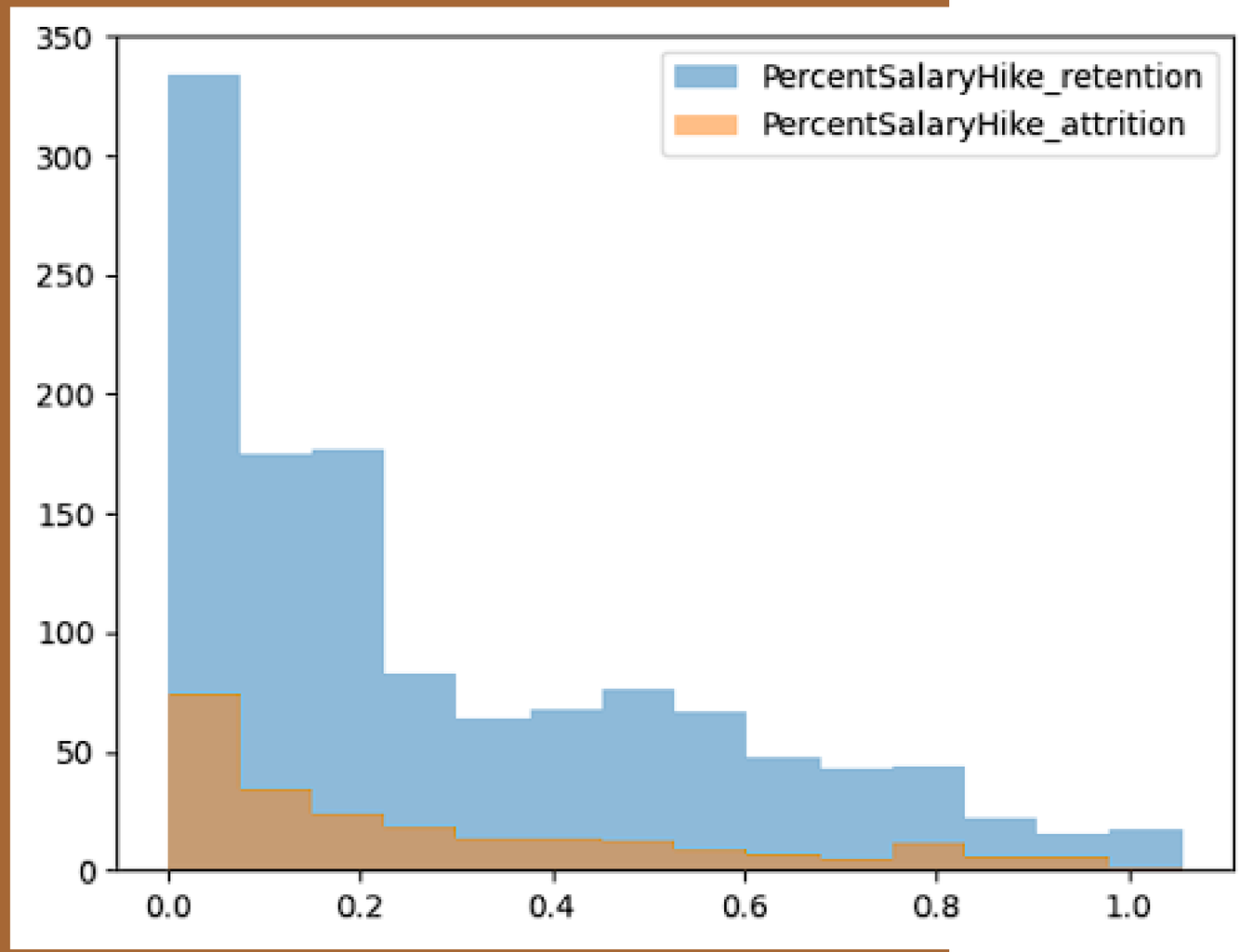
Overlapping Histograms

The following histograms have the lowest overlapping area:



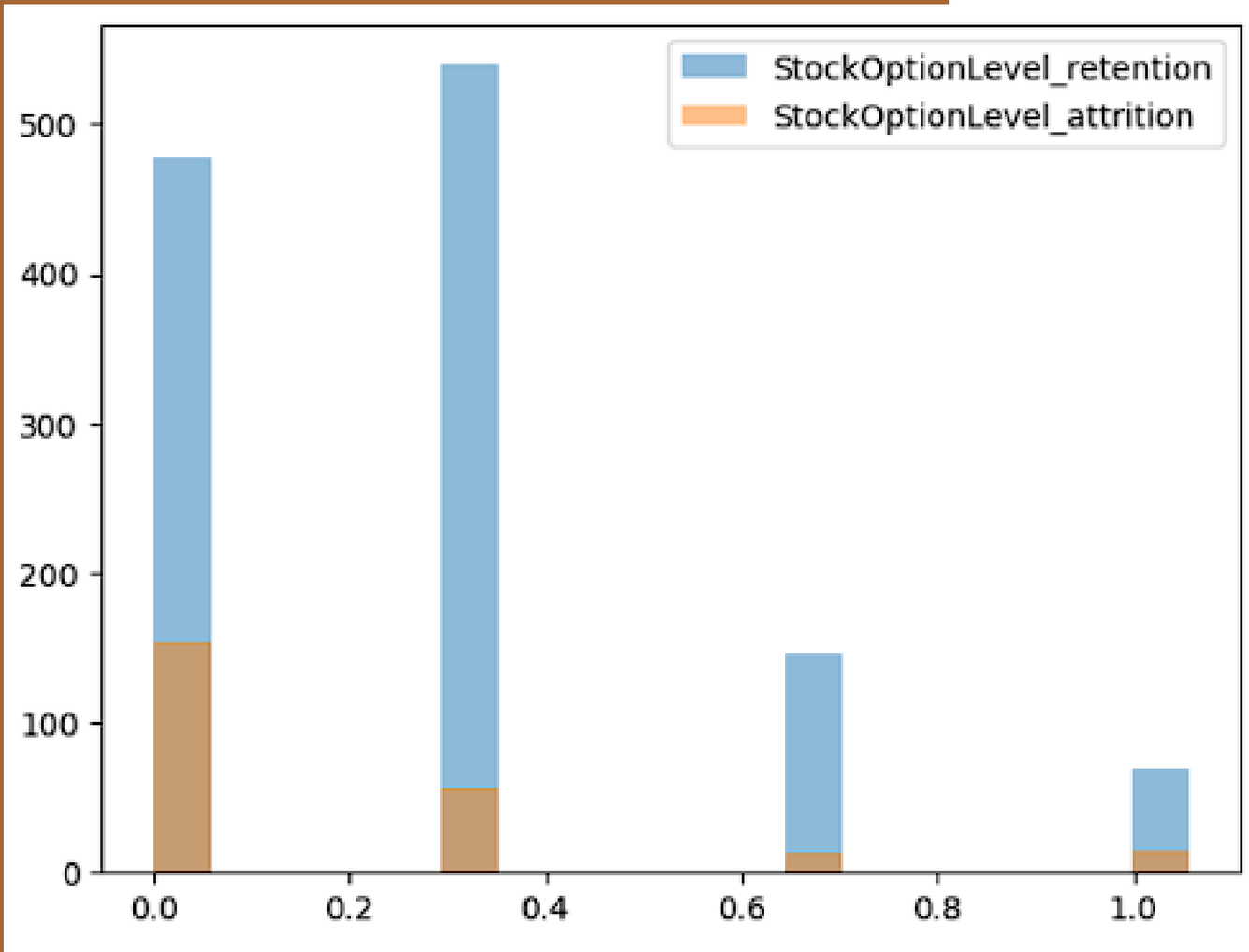
Overlapping Histograms

The following histograms have the lowest overlapping area:



Overlapping Histograms

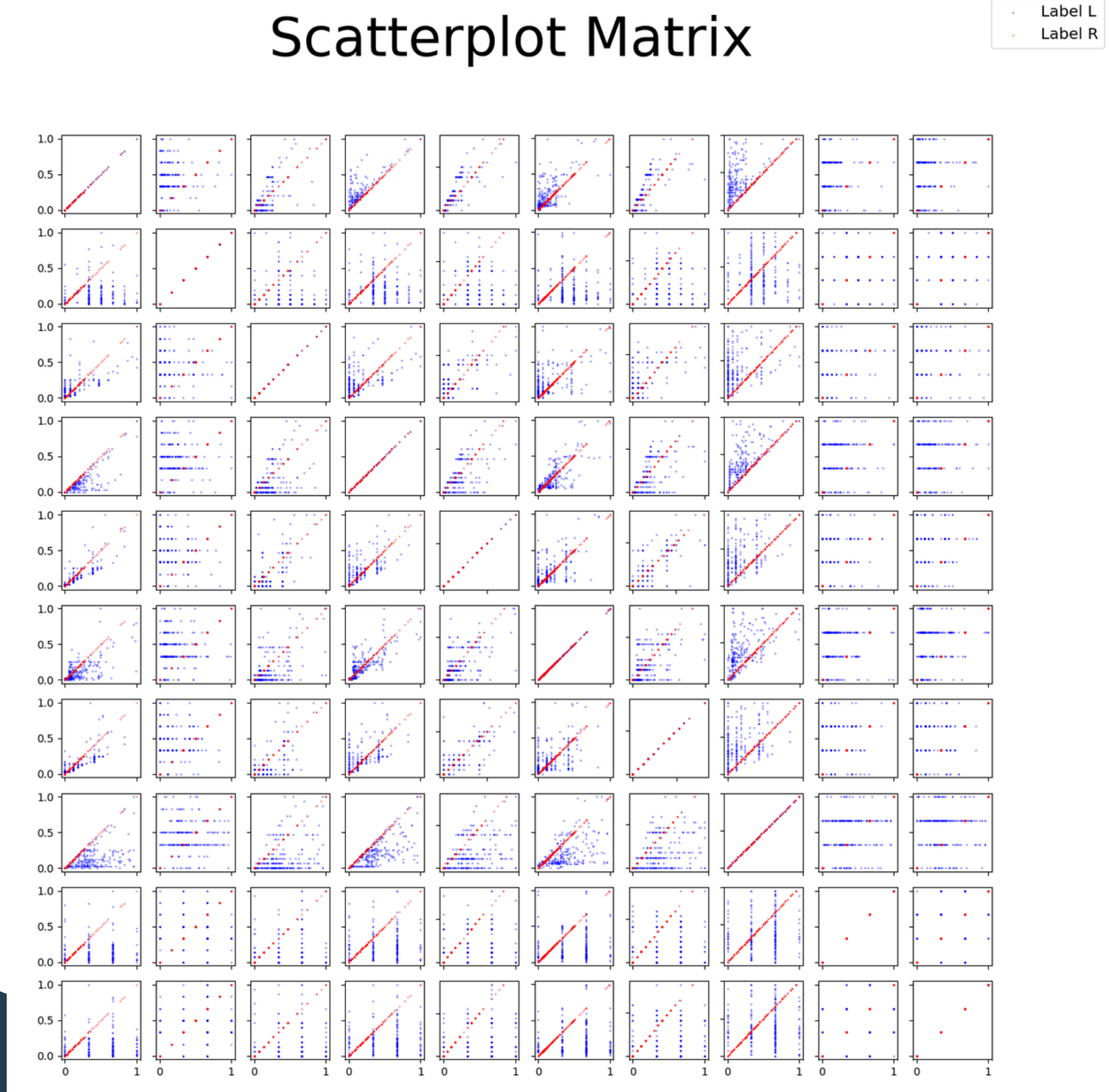
The following histograms have the lowest overlapping area:



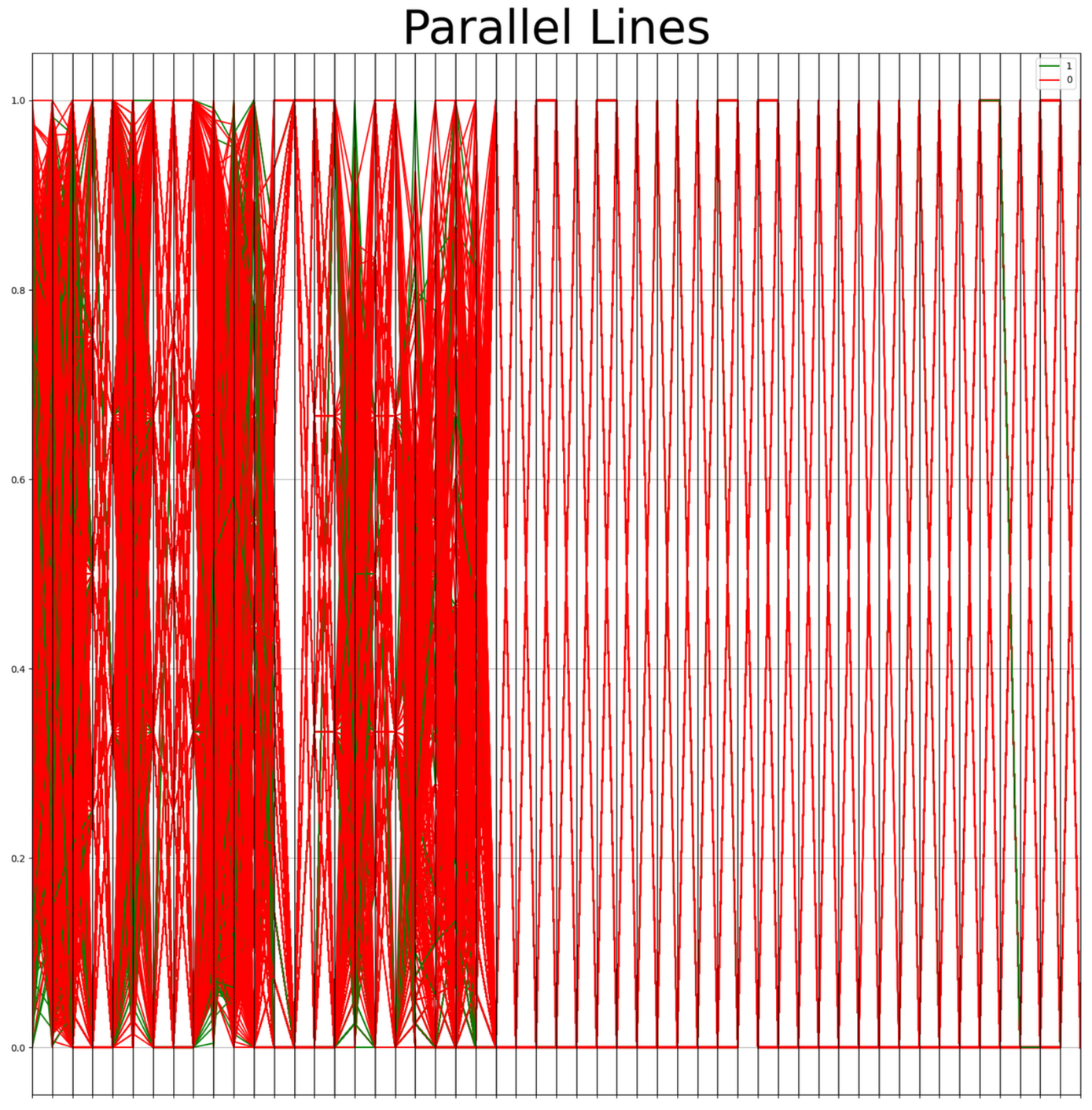
Overlapping Histograms

The following histograms have the lowest overlapping area:

SCATTERPLOT MATRIX



PARALLEL LINES



Final Features Selected

- Years at Company
- Training Times Last Year
- Monthly Income
- Age
- Job Involvement
- Work Life Balance
- Percent Salary Hike
- Stock Option Level

MODELING

Classification Models

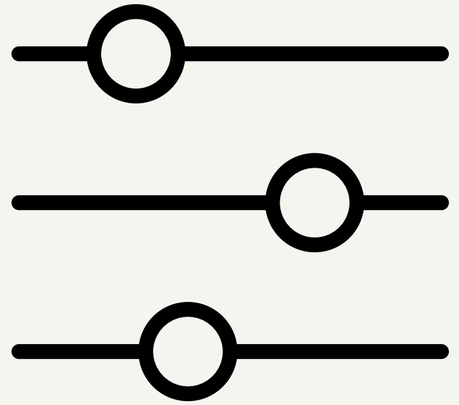
Gaussian Naive Bayes with Feature Selection

Multilayer Perceptron with Feature Selection

Multilayer Perceptron without Feature Selection

Data Partition

- 75% training data and 25% testing data
- Stratified by 'Attrition'



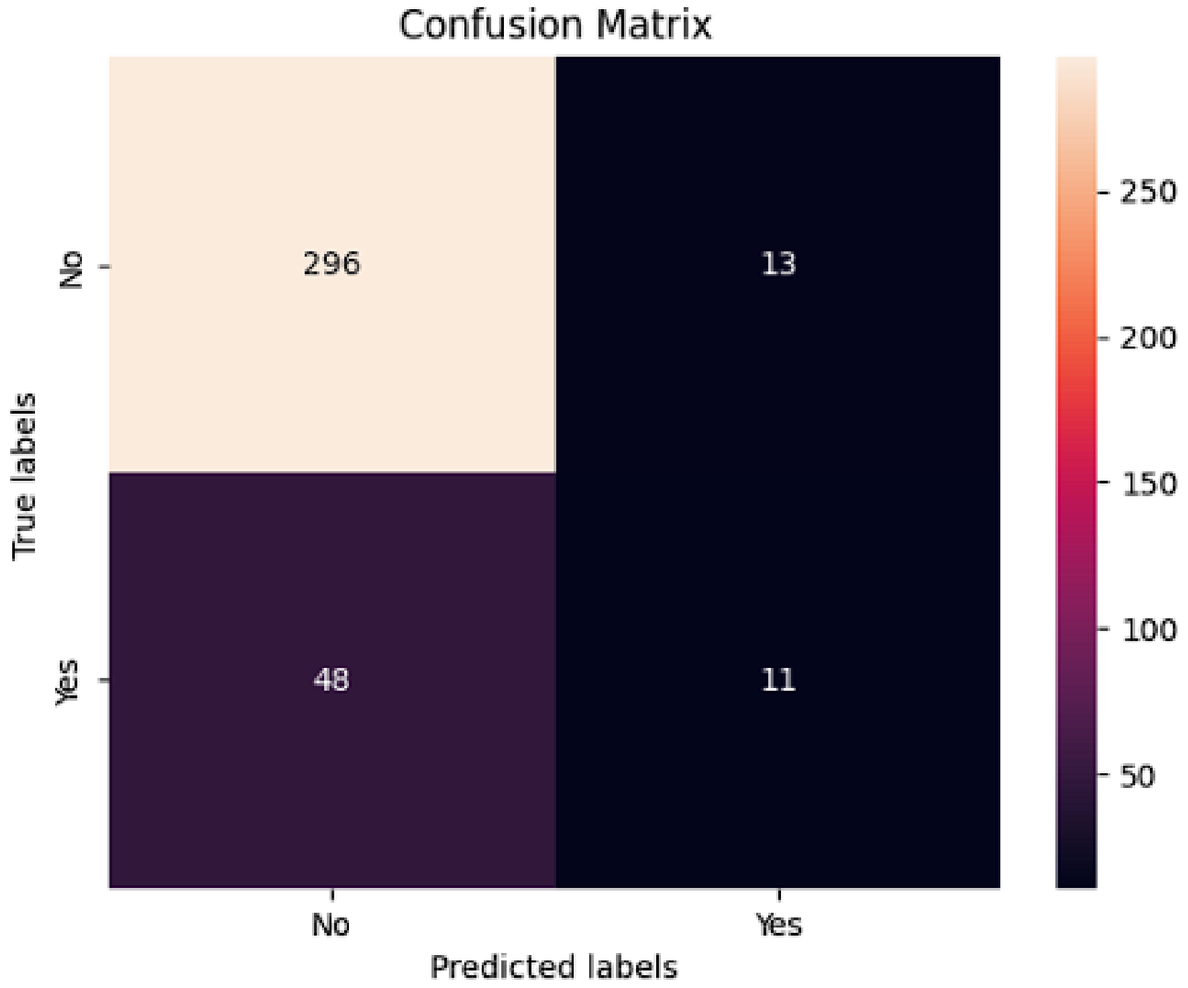
MLP Classifier Parameters

- Hidden Layer Sizes: [200,200] - 2 hidden layers each with 200 neurons
- Activation Function - Relu
- Solver - lbfgs (Limited-memory Broyden-Fletcher-Goldfarb-Shanno)

EVALUATION

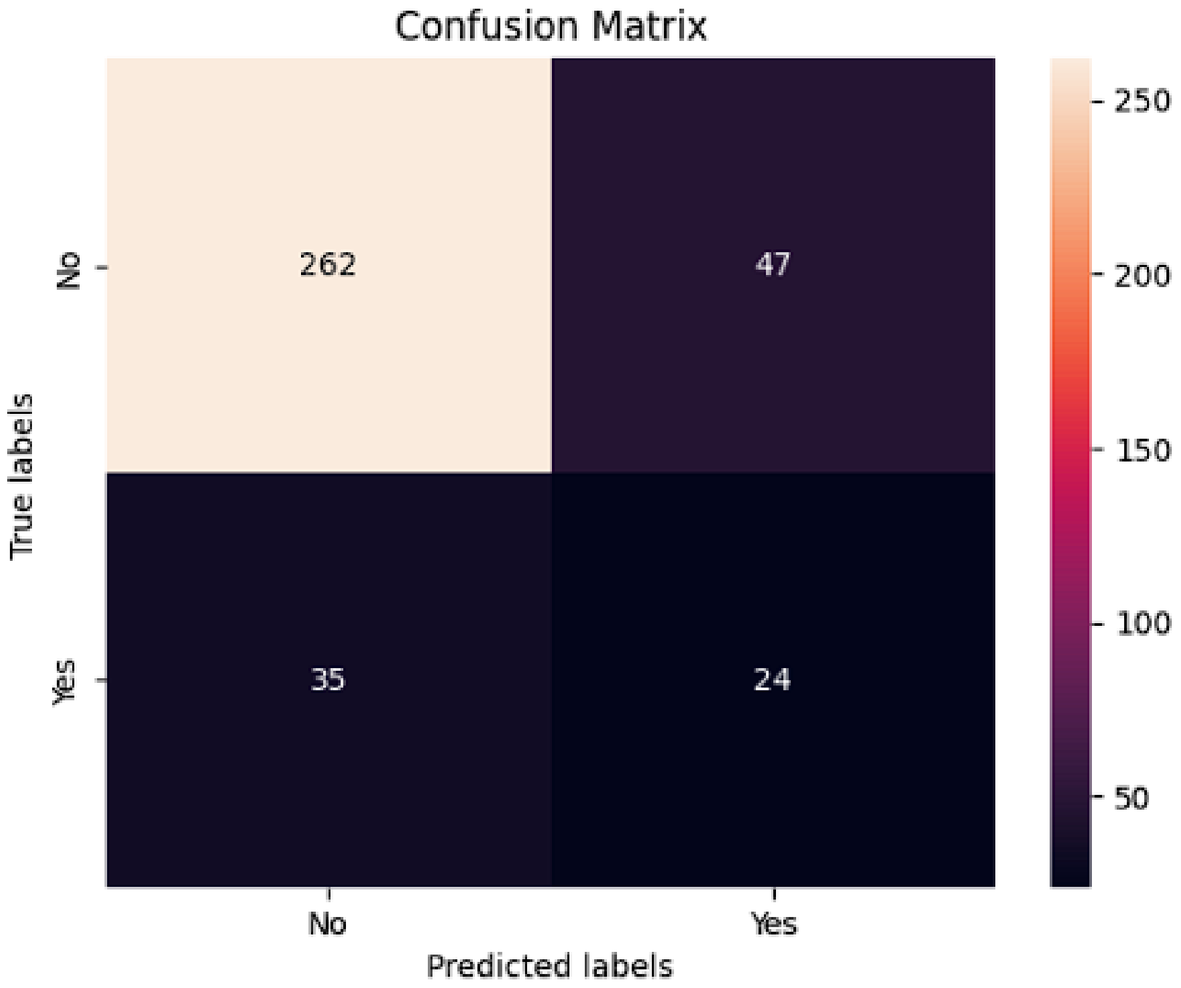
Confusion Matrix

Gaussian Naive Bayes with Feature Selection (FS)



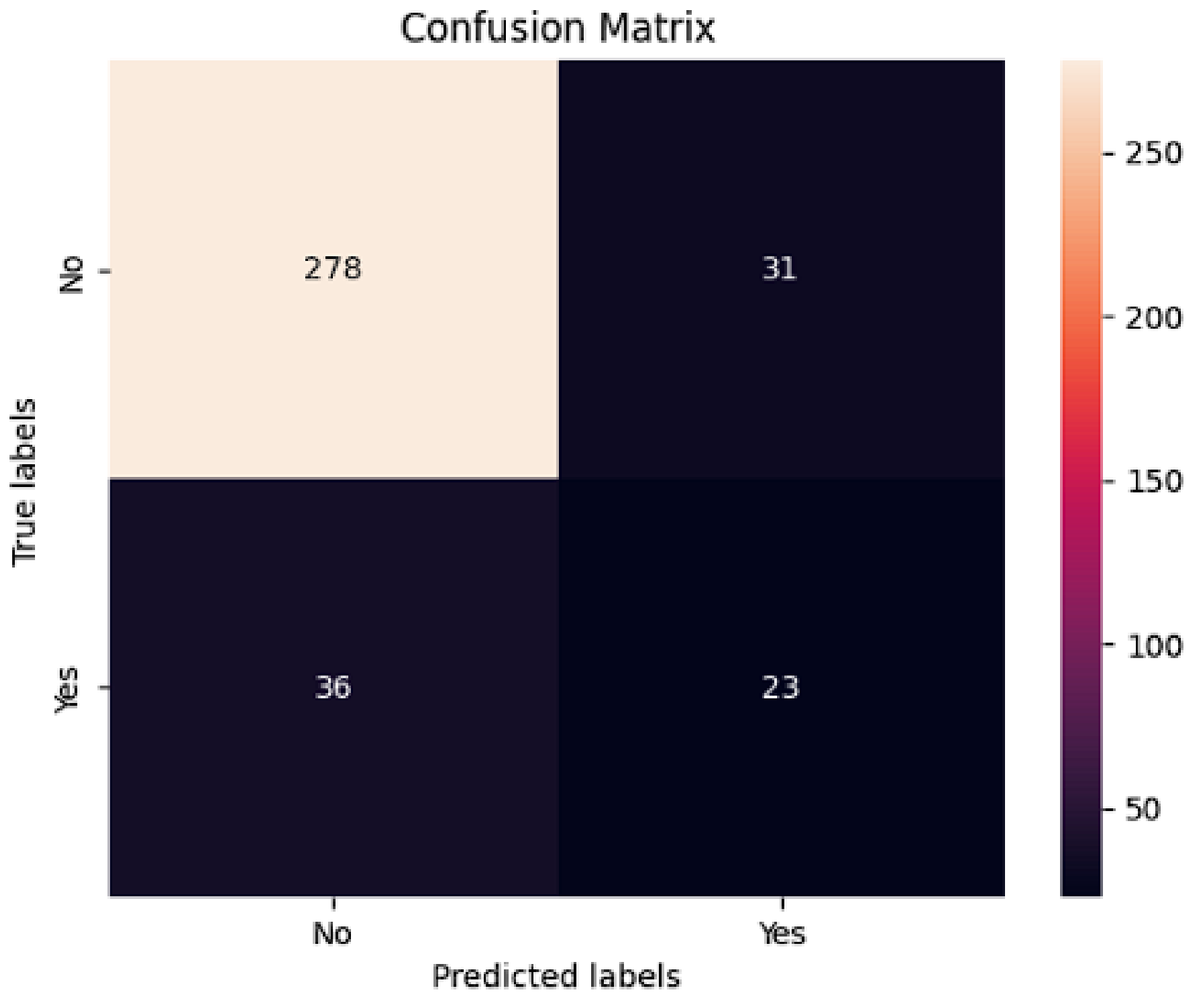
Confusion Matrix

Multilayer Perceptron with Feature Selection (FS)



Confusion Matrix

Multilayer Perceptron without Feature Selection



Metrics

Accuracy

- ratio of correct predictions to the total number of predictions

Recall

- how many of the true positive cases the model is able to correctly identify

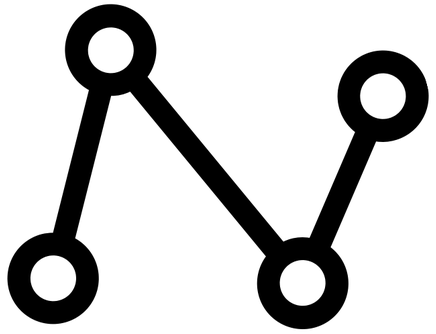
Precision

- how many of the positive predictions made by the model are actually correct

F1 Score

- metric that balances changes in Precision and Recall in one ratio

Metrics



Metrics	Gaussian Naive Bayes with Feature Selection	MultiLayer Perceptron with Feature Selection	Multilayer Perceptron without Feature Selection
Accuracy	0.834239	0.777174	0.817935
Precision	0.458333	0.338028	0.425926
Recall	0.186441	0.406780	0.389831
F1 Score	0.265060	0.369231	0.407080

Interpretation

- The Gaussian Naive Bayes model with feature selection had the highest accuracy and precision.
- The Multilayer perceptron with feature selection had the highest recall.
- The Multilayer Perceptron without feature selection had the highest F1 score.

Key Insights

Selected Features are significant indicators

- Tradeoff for efficiency still provides relevant predictions for attrition.
- The Gaussian Naive Bayes and MLP with FS models have better accuracy, precision, and recall than MLP without FS.

Features lead to better attrition prediction

- While MLP without FS provided better accuracy, MLP with FS showed better prediction of attrition.
- Good precision in Gaussian Naive Bayes shows capabilities of features to predict non-attrition.

CONCLUSION & RECOMMENDATION

Recommendations

LARGER DATASET

Even with stratified sampling, the proportions of the dataset seems to still have a significant bias added to the model weights.

CONSIDER REVALIDATION

To further verify the chosen features, an additional step of removal and observing a significant change in the metrics can be taken.

DIFFERENT NORMALIZATION

Minimax normalization was used to normalize the dataset, but this may not be the best method for one-hot encoded categorical data.

Thank you for listening!