CCDS223 – Group Project Report

Employee Attrition Prediction

November 14, 2025

Group 2:

Abdulalah Abdullah Nasser 2340466

Abdullah Atiah Alzahrani 2340088

Mohannad Abdullah Alamri 2340047

Tariq Zaid Alqhtani 2340458

Instructor: Abdullah Alghoson

TABLE OF CONTENTS

1. Introduction

The purpose of this stage is to perform a full data cleaning and preprocessing for the IBM HR Analytics Employee Attrition dataset.

This step ensures that the dataset is complete, consistent, and ready for further analysis in the next milestone (Exploratory Data Analysis) and Milestone 4 (Modeling).

During this milestone, we imported the dataset into SAS Model Studio, handled missing values, detected outliers, removed irrelevant variables, and prepared the final cleaned dataset for modeling.

2. Data Collection and cleaning

We used the IBM HR Analytics Employee Attrition dataset from Kaggle.

The original dataset contains:

• 1470 rows (employees)

• 35 variables (demographic, job roles, performance, income, experience, satisfaction, and attrition)

Because the project requires working with 12–20 variables, we cleaned and reduced the dataset to 20 selected variables that are relevant, non-redundant, and meaningful for attrition prediction. Dataset Link:

ATTRITION.xlsx

Attribute Types:

Attribute	Type
Age	Numeric
Attrition	Binary
Department	Nominal
DistanceFromHome	Numeric
Education	Ordinal
EmployeeNumber	Numeric
EnvironmentSatisfcation	Numeric
Gender	Nominal
JobInvolvement	Ordinal
JobRole	Nominal
JobSatisfaction	Ordinal
MaritalStatus	Nominal
MonthlyIncome	Numeric
NumCompaniesWorked	Numeric
OverTime	Nominal (Binary)
PerformanceRating	Numeric
TotalWorkingYears	Numeric
WorkLifeBalance	Ordinal
YearsAtCompany	Numeric
YearsSinceLastPromotion	Numeric

3. Data Understanding

This dataset aims to predict employee attrition (whether an employee will leave the company). The target variable is Attrition, which has two possible values: "Yes" or "No".

After importing the dataset, we defined each variable's:

• Role (Input, Target, ID, Rejected)

• Level (Interval, Nominal, Binary, Ordinal)

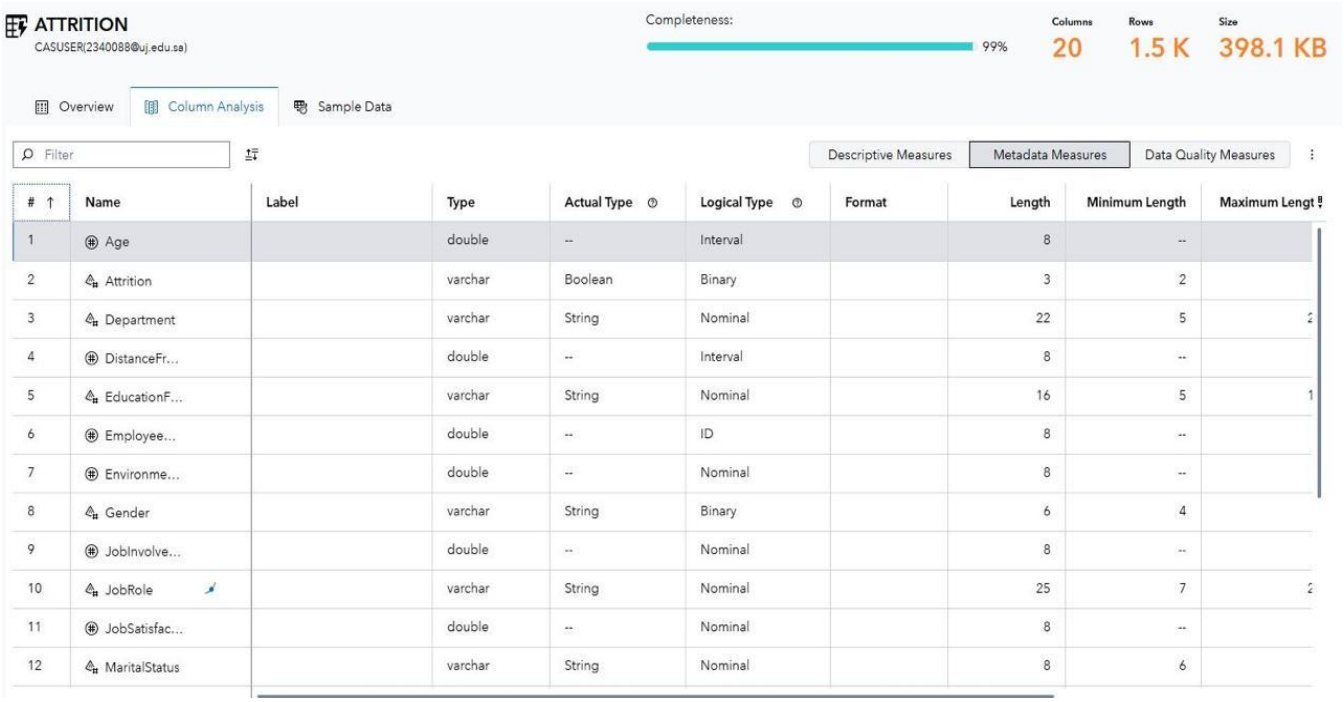
Target Variable

• Attrition → Role: Target, Level: Binary

ID Variable

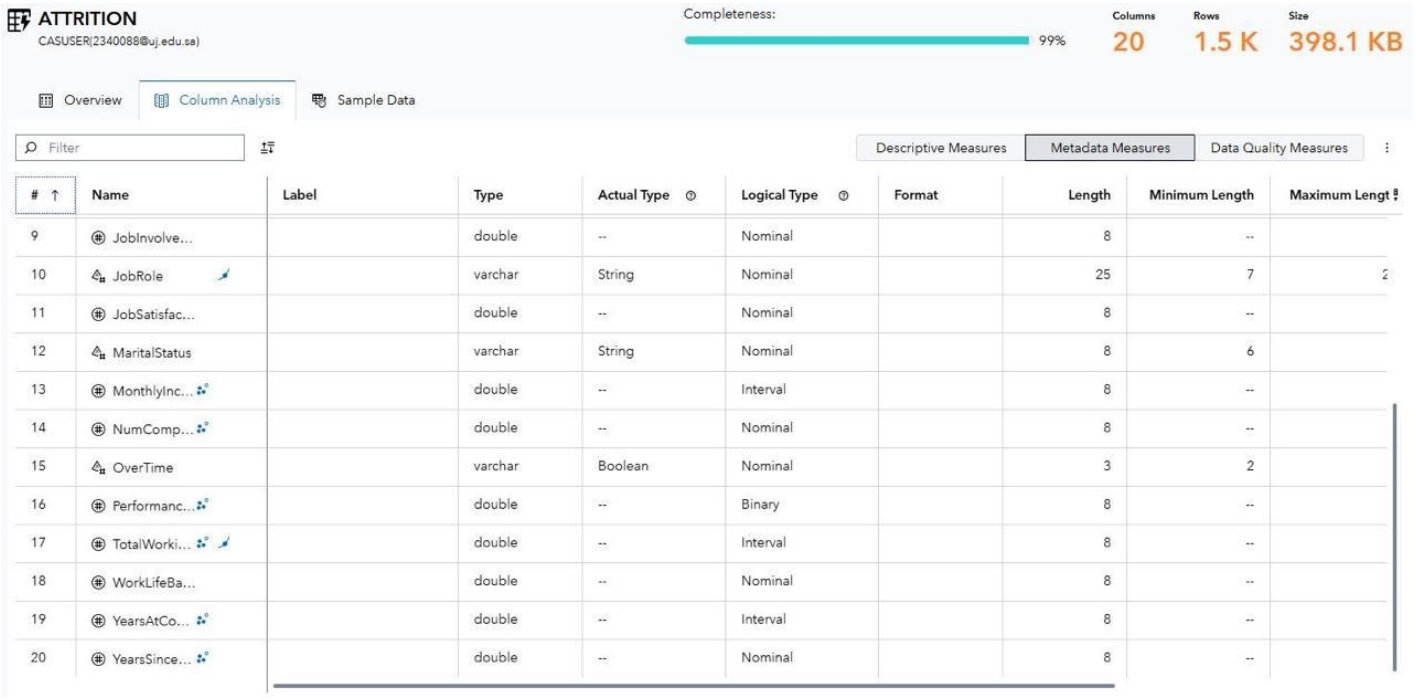
• EmployeeNumber → Role: ID, Level: Interval

As it's shown in the following figures from sas:



#	Name	Label	Type	Actual Type	Logical Type	Format	Length	Minimum Length	Maximum Lengt
1	Age		double	--	Interval		8	--	
2	Attrition		varchar	Boolean	Binary		3	2	
3	Department		varchar	String	Nominal		22	5	2
4	DistanceFr...		double	--	Interval		8	--	
5	EducationF...		varchar	String	Nominal		16	5	1
6	Employee...		double	--	ID		8	--	
7	Environme...		double	--	Nominal		8	--	
8	Gender		varchar	String	Binary		6	4	
9	JobInvolve...		double	--	Nominal		8	--	
10	JobRole		varchar	String	Nominal		25	7	2
11	JobSatisfac...		double	--	Nominal		8	--	
12	MaritalStatus		varchar	String	Nominal		8	6	

Figure 1



#	Name	Label	Type	Actual Type	Logical Type	Format	Length	Minimum Length	Maximum Lengt
9	JobInvolve...		double	--	Nominal		8	--	
10	JobRole		varchar	String	Nominal		25	7	2
11	JobSatisfac...		double	--	Nominal		8	--	
12	MaritalStatus		varchar	String	Nominal		8	6	
13	MonthlyInc...		double	--	Interval		8	--	
14	NumComp...		double	--	Nominal		8	--	
15	OverTime		varchar	Boolean	Nominal		3	2	
16	Performanc...		double	--	Binary		8	--	
17	TotalWorki...		double	--	Interval		8	--	
18	WorkLifeBa...		double	--	Nominal		8	--	
19	YearsAtCo...		double	--	Interval		8	--	
20	YearsSince...		double	--	Nominal		8	--	

Figure 2

4. Removed Variables (Rejected)

We removed 15 irrelevant or redundant variables by setting Role = Rejected in SAS.

Reasons for removal:

1. Constant variables (no useful info):

• EmployeeCount

• Over18

• StandardHours

2. Salary breakdown features (redundant due to MonthlyIncome):

• DailyRate

• HourlyRate

• MonthlyRate

• PercentSalaryHike

• StockOptionLevel

3. Highly correlated variables (duplicate info):

• JobLevel

• YearsInCurrentRole

• YearsWithCurrManager

4. Low predictive value:

• TrainingTimesLastYear

• BusinessTravel

• Education

• RelationshipSatisfaction

Final remaining variables = 20.

5. Data Exploration in SAS

From the results, we checked:

Initial exploration of the dataset shows the following insights: - The dataset now contains 1470 rows and 20 columns. - The Attrition column is imbalanced, with fewer employees leaving than staying. - Income and age appear to be related to attrition; younger and lower-paid employees tend to leave more often.

Observations:

• No missing values

• Some outliers exist but are valid real-world cases

6. Descriptive Statistics

The Statistics of the data:

ATTRITION
CASUSER(2340088@uj.edu.sa)

Completeness: 100%

Columns	Rows	Size	Status
20	1.5 K	398.1 KB	None

Overview | Column Analysis | Sample Data

Date analyzed:

Filter

Descriptive Measures | Metadata Measures | Data Quality

#	Name	Distinct Values	Mean	Median	Minimum	Maximum	Standard Deviation	Missing	Blanks	Outliers
1	Age	43	36.92381	36	18	60	9.135373	0	--	
2	Attrition	2	--	--	--	--	--	0	0	--
3	Department	3	--	--	--	--	--	0	0	--
4	DistanceFr...	29	9.192517	7	1	29	8.106864	0	--	
5	EducationF...	6	--	--	--	--	--	0	0	--
6	Employee...	1,470	1024.865	1020.5	1	2068	602.0243	0	--	
7	Environme...	4	2.721769	3	1	4	1.093082	0	--	
8	Gender	2	--	--	--	--	--	0	0	--
9	JobInvolve...	4	2.729932	3	1	4	0.711561	0	--	
10	JobRole	9	--	--	--	--	--	0	0	--
11	JobSatisfac...	4	2.728571	3	1	4	1.102846	0	--	
12	MaritalStatus	3	--	--	--	--	--	0	0	--
13	MonthlyInc...	1,348	6498.197	4907.5	1009	19999	4706.089	0	--	

Figure 3

ATTRITION
CASUSER(2340088@uj.edu.sa)

Completeness: 100%

Columns	Rows	Size	Status
20	1.5 K	398.1 KB	None

Overview | Column Analysis | Sample Data

Filter

Descriptive Measures | Metadata Measures | Data Q

#	Name	Distinct Values	Mean	Median	Minimum	Maximum	Standard Deviation	Missing	Blanks	Outliers
8	Gender	2	--	--	--	--	--	0	0	--
9	JobInvolve...	4	2.729932	3	1	4	0.711561	0	--	
10	JobRole	9	--	--	--	--	--	0	0	--
11	JobSatisfac...	4	2.728571	3	1	4	1.102846	0	--	
12	MaritalStatus	3	--	--	--	--	--	0	0	--
13	MonthlyInc...	1,348	6498.197	4907.5	1009	19999	4706.089	0	--	
14	NumComp...	10	2.693197	2	0	9	2.498009	0	--	
15	OverTime	2	--	--	--	--	--	0	0	--
16	Performanc...	2	3.153741	3	3	4	0.360824	0	--	
17	TotalWorki...	40	11.27959	10	0	40	7.780782	0	--	
18	WorkLifeBa...	4	2.761224	3	1	4	0.706476	0	--	
19	YearsAtCo...	37	7.008163	5	0	40	6.126525	0	--	
20	YearsSince...	16	2.187755	1	0	15	3.22243	0	--	

Figure 4

7. Handling Outliers (Transformation / Replacement Node)

We used transformation/scaling techniques such as:

- Standardization
- Min-Max Scaling

8. Final Clean Dataset

After all steps, the cleaned dataset contains:

- 1470 rows
- 20 variables

9. Conclusion

- Data imported into SAS
- Roles and levels defined
- 20 variables selected
- 15 variables rejected
- Missing values handled
- Outliers assessed
- Dataset scaled and prepared for modeling