UNIVERSITY OF MALAYA

WIE3007 Data Mining & Warehousing

GROUP ASSIGNMENT

Semester I 2022/2023

Absenteeism at Work Analysis

Lecturer: PROF. DR. TEH YING WAH

NAME	MATRIC NO
Muhammad Rafi Azhar Salman	S2003599
Zahid Fajri Ramadhan	17216872/1
Muhammad Farhan Bin Mohd Latif	17203850
Muhammad Danial Bin Azni Jaya	U2005301
Muhammad Amirul Amin Arman	17202969

Table of Contents

1. Introduction

Absenteeism at work is a complicated issue that is affected by many things. Figuring out the patterns and reasons behind absences is important for managing the productivity of the workforce and making the workplace a healthy and supportive place to be. This report adopts the SEMMA (Sample, Explore, Modify, Model, Assess) methodology, a robust framework for data mining, to dissect absenteeism patterns in a diverse workforce. By utilising the "Absenteeism at work" dataset sourced from the UCI Machine Learning Repository, which gives a detailed picture of when employees don't show up for work. The dataset groups absences by a lot of different factors, such as personal demographics, work-related factors, and health conditions, as defined by the International Classification of Diseases (ICD).

1.1 Objectives

The core aim of this report is to dissect absenteeism in the workplace through a dual analytical lens:

- Classification Analysis: To identify key factors that predict absenteeism, thereby enabling the development of targeted strategies to mitigate unwarranted absences.
- Sequence Analysis: To determine common sequences of absenteeism events among employees. This includes identifying if certain types of absences tend to occur in a particular order, which could suggest underlying causal relationships or dependencies.
- Association Analysis: To uncover associations between various attributes within the data, such as the reasons for absence, demographic characteristics, and other recorded metrics, with the severity of absenteeism.

These analyses will facilitate a deeper understanding of absenteeism trends, contributing to more effective human resource management and policy formulation.

1.2 Methodology

The analysis will be done in the following steps, following the SEMMA method:

- Sample: Use stratified sampling to choose a subset of the data that is representative of the whole set of data. This will make sure that the different types of absences are fairly represented.
- Explore: Look through the data to understand how it is distributed, find outliers, and find the first patterns.
- Modify: Modify the data and get it ready for modelling by dealing with missing values, making derived variables, and normalising the data as needed.
- Model: Create and use classification models to make predictions, and use clustering and association rule methods to find patterns.
- Assess: Assess how well the models worked and how important the patterns found were, making sure that the results are reliable and correct.

2. Data Overview

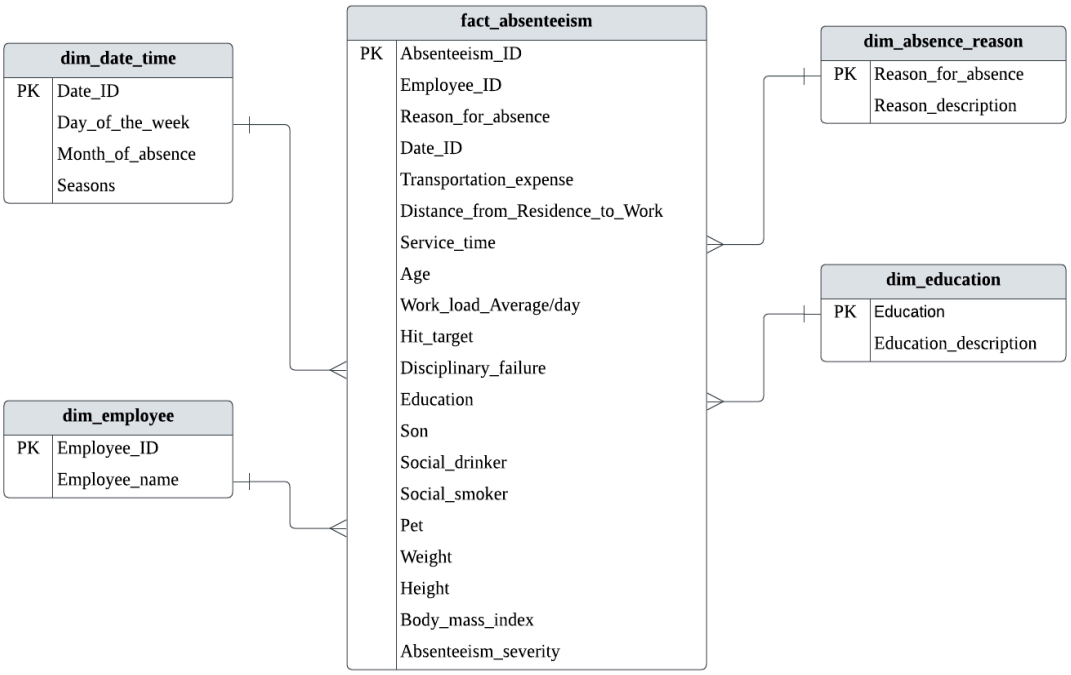
The Absenteeism at Work dataset comprises data entries related to employee absenteeism in a workplace setting. This dataset, obtained from the UC Irvine Machine Learning Repository, is a valuable resource for analysing patterns and factors that influence absenteeism in the workplace. The dataset consists of a range of attributes, encompassing demographic information, work-related variables, and personal lifestyle preferences of employees.

The dataset contains a variety of attributes, including personal demographic information and specific work-related details.

- **Attributes:** The dataset comprises 22 distinct attributes, including employee ID, reasons for absence (classified according to the International Classification of Diseases), transportation expenses, distance from work, workload, and absenteeism time measured in hours.
- **Applications:** The dataset is well-suited for predictive modelling and can be used to investigate the factors that influence absenteeism. This can help in developing strategies to reduce excessive absenteeism and enhance productivity in the workplace.
- **Structure:** The data is organised in a tabular format, making it easier to analyse and model using different statistical and machine learning frameworks.

The Absenteeism at Work dataset is a reliable resource for comprehending the dynamics of absenteeism, as well as gaining insights into employee welfare and organisational effectiveness. This resource is highly commendable for researchers, HR professionals, and organisational analysts who are interested in exploring the fields of workforce management and behavioural studies.

2.2. Star Schema



2.2.1 Fact Table: fact_absenteeism

Record_ID (Primary Key): A unique identifier for each absenteeism record.

Employee_ID: Reference to the employee's ID.

Reason_for_absence: Numeric code for the reason for absence.

Date_ID: A unique identifier for each datetime value.

Transportation_expense: Costs associated with transportation.

Distance_from_Residence_to_Work: Distance in kilometres.

Service_time: Total service time in years.

Age: Represents the age of the employee.

Work_load_Average/day: Average amount of work load per day.

Hit_target: Numeric value indicating the hit target.

Disciplinary_failure: Indicates whether there was a disciplinary failure (yes=1; no=0).

Education: Categorical representation of education level (high school (1), graduate (2), postgraduate (3), master and doctor (4))

Son: Number of children.

Social_drinker: Indicates if the employee is a social drinker (yes=1; no=0).

Social_smoker: Indicates if the employee is a social smoker (yes=1; no=0).

Pet: Number of pets.

Weight: The weight of the employee, in kilograms.

Height: The height of the employee, in centimeters.

Body_mass_index: Numerice value indicating the employee BMI.

Absenteeism_time_in_hours: The number of hours absent.

Absenteeism_Severity (Target): Categorical representation of absenteeism severity (1=Severe; 0=Not severe)

2.2.2 Dimension Table: dim_date_time

Date_ID (Primary Key): A unique identifier for each datetime value.

Month of absence: The month during which the absence occurred.

Day of the week: The day of the week on which the absence occurred, coded numerically.

Seasons: The season during which the absence occurred, coded numerically.

2.2.3 Dimension Table: dim_absence_reason

Reason_for_absence: Numeric code for the reason for absence

Reason_description: Description of reason for absence numeric code (

1. Certain infectious and parasitic diseases
2. Neoplasms

3. Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
4. Endocrine, nutritional and metabolic diseases
5. Mental and behavioural disorders
6. Diseases of the nervous system
7. Diseases of the eye and adnexa
8. Diseases of the ear and mastoid process
9. Diseases of the circulatory system
10. Diseases of the respiratory system
11. Diseases of the digestive system
12. Diseases of the skin and subcutaneous tissue
13. Diseases of the musculoskeletal system and connective tissue
14. Diseases of the genitourinary system
15. Pregnancy, childbirth and the puerperium
16. Certain conditions originating in the perinatal period
17. Congenital malformations, deformations and chromosomal abnormalities
18. Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
19. Injury, poisoning and certain other consequences of external causes
20. External causes of morbidity and mortality
21. Factors influencing health status and contact with health services.
22. Patient follow-up
23. Medical consultation

24. Blood donation

25. laboratory examination

26. Unjustified absence

27. Physiotherapy

28. Dental consultation.)

2.2.4 Dimension Table: dim_education

Education: Numeric code for education level.

Education_desciption: Description of education numeric code (

1. High school
2. Graduate
3. Postgraduate
4. Master & Doctor)

4. Sample

Our objective in analysing the Absenteeism at Work dataset is to identify patterns and determinants that impact employee absenteeism. To ensure the representativeness and robustness of our sampling strategy, we must consider a wide range of attributes, including demographic, lifestyle, and work-related factors. This will help us address potential variations within different subgroups.

After careful analysis, we choose stratified sampling as our sampling method. This method involves dividing the entire dataset into distinct strata or subgroups based on key characteristics and then randomly sampling from each of these strata. This method was selected to guarantee that our sample effectively reflects the various subgroups present in our dataset, with a particular emphasis on factors that strongly correlate with the target variable "Absenteeism_severity".

.. Property	Value
General	
Node ID	Smpl
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Output Type	Data
Sample Method	Stratify
Random Seed	12345
Size	
Type	Percentage
Observations	
Percentage	20.0
Alpha	0.01
PValue	0.01
Cluster Method	Random
Stratified	
Criterion	Proportional
Ignore Small Strata	No
Minimum Strata Size	10
Level Based Options	
Level Selection	Event
Level Proportion	100.0
Sample Proportion	50.0
Oversampling	
Adjust Frequency	No
Based on Count	No
Exclude Missing Levels	No
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	12/26/23 2:22 PM
Run ID	f44c20ff-2763-3e47-a482-a4a000120ade
Last Error	
Last Status	Complete
Last Run Time	12/26/23 2:53 PM
Run Duration	0 Hr. 0 Min. 1.80 Sec.
Grid Host	
User-Added Node	No

Figure 1: Screenshot of "Sample" node parameters in SAS Enterprise Miner

The sampling strategy was implemented in SAS Enterprise Miner using the following settings in the Sample node:

1. **Sample Method = "Stratify"**
    - This choice is crucial for stratified sampling, ensuring that the dataset is divided into meaningful subgroups (strata) and samples are taken from each group.
2. **Percentage of Dataset Sampled = 20%**
    - We set the percentage to 20% of the total dataset. This proportion balances obtaining a sufficiently large sample to represent the dataset and managing computational efficiency. The percentage is adjustable based on the dataset size and specific analysis requirements.
3. **Sampling Criterion = "Proportional"**
    - This criterion maintains the original distribution of the dataset, ensuring each stratum is represented in the sample in proportion to its size in the full dataset. This approach is particularly beneficial for preserving the underlying patterns and relationships within the data.
4. **Treatment of Small Strata = "No"**
    - This decision ensures the inclusion of even the smaller strata in our sample, recognizing their potential significance in the analysis, especially given the diverse range of factors influencing absenteeism.
5. **Minimum Strata Size = 10**
    - This threshold is slightly above the default, aiming to ensure that each stratum in the sample is adequately large to be representative. This parameter was selected considering the distribution and size of the strata in our dataset, and it is adjustable based on strata distribution in future analyses.

5. Explore

5.1 Variable Selection

Property	Value
General	
Node ID	Varsel
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Max Class Level	100
Max Missing Percentage	50
Target Model	Default
Manual Selector	
Rejects Unused Input	Yes
Bypass Options	
Variable	None
Role	Input
Chi-Square Options	
Number of Bins	50
Maximum Pass Number	6
Minimum Chi-Square	3.84
R-Square Options	
Maximum Variable Number	3000
Minimum R-Square	0.005
Stop R-Square	5.0E-4
Use AOV16 Variables	No
Use Group Variables	Yes
Use Interactions	No
Use SPD Engine Library	Yes
Print Option	Default
Score	
Hides Rejected Variables	Yes
Hides Unused Variables	Yes

Figure 2: Screenshot of "Variable Selection" node parameters in SAS Enterprise Miner

1. **Max Missing Percentage = 50%**
    - A crucial parameter, especially since no data imputation has been performed. This higher threshold includes variables with substantial missing data, which may still offer valuable insights.
2. **Target Model = Default**
    - To maintain methodological agility, the target model parameter remains in the default setting, empowering SAS Enterprise Miner to select the most appropriate model based on the data's inherent structure.
3. **Rejects Unused Input = Yes**

- This parameter remains activated to streamline the model by excluding variables that do not provide meaningful contributions, emphasizing the importance of parsimony and interpretability.

4. **Chi-Square Options (Number of Bins = 50, Maximum Pass Number = 6, Minimum Chi-Square = 3.84)**
    - The chi-square options are retained at default values to ensure a statistically robust and computationally efficient evaluation of categorical variables within the selection framework.
5. **R-Square Options (Maximum Variable Number = 3000, Minimum R-Square = 0.005, Stop R-Square = 5.0E-4)**
    - These settings are meticulously chosen to underscore the significance of each variable's contribution to the model's R-square, aligning variable selection with the enhancement of explanatory power.

5.2 Association Rules

An essential part of our data mining study that examines workplace absenteeism trends is association analysis. Finding significant correlations and linkages between the different features in the dataset is the goal of this analytical phase. Through an examination of the relationships among the reasons for absence, demographic traits, and other recorded metrics, we can acquire important knowledge about the elements influencing the degree of absenteeism.

There are 2 variables to explore which are abseentism_severe and Reason_for_absence because these 2 variables are related to the objectives. The input is 'Day_of_the_week', 'Education', 'Social_drinker' and 'Social_smoker'.

1. **Target: Reason_for_absenteeism**

Name	Use	Role	Level
Absenteeism_se	**No**	Target	Nominal
Month_of_abser	**No**	Sequence	Interval
Reason_for_abs	**Yes**	Target	Nominal
dataobs	**No**	ID	Interval
employee_ID	**Yes**	ID	Nominal
record_ID	**No**	ID	Nominal

The variables that are used are employee_ID and Reason_for_abseenteism. Both variables are nominal level.

Result:

Association Report

Relations	Expected Confidence (%)	Confidence (%)	Support (%)	Lift	Transaction Count	Rule	Left Hand of Rule	Right Hand of Rule
4	2.78	100.00	2.78	36.00	1.00	28 & 21 ==> 12 & 5	28 & 21	12 & 5
4	2.78	100.00	2.78	36.00	1.00	21 & 6 ==> 12 & 5	21 & 6	12 & 5
4	2.78	100.00	2.78	36.00	1.00	27 & 22 ==> 12 & 9	27 & 22	12 & 9
4	2.78	100.00	2.78	36.00	1.00	27 & 14 ==> 12 & 9	27 & 14	12 & 9
4	2.78	100.00	2.78	36.00	1.00	9 & 1 ==> 13 & 4	9 & 1	13 & 4
4	2.78	100.00	2.78	36.00	1.00	17 ==> 16 & 10 & 8	17	16 & 10 & 8
4	2.78	100.00	2.78	36.00	1.00	17 ==> 16 & 11 & 8	17	16 & 11 & 8
3	2.78	100.00	2.78	36.00	1.00	17 ==> 16 & 8	17	16 & 8
4	2.78	100.00	2.78	36.00	1.00	25 & 17 ==> 16 & 8	25 & 17	16 & 8
4	2.78	100.00	2.78	36.00	1.00	23 & 17 ==> 16 & 8	23 & 17	16 & 8
4	2.78	100.00	2.78	36.00	1.00	22 & 17 ==> 16 & 8	22 & 17	16 & 8
4	2.78	100.00	2.78	36.00	1.00	21 & 17 ==> 16 & 8	21 & 17	16 & 8
4	2.78	100.00	2.78	36.00	1.00	18 & 17 ==> 16 & 8	18 & 17	16 & 8
4	2.78	100.00	2.78	36.00	1.00	17 & 11 ==> 16 & 8	17 & 11	16 & 8
4	2.78	100.00	2.78	36.00	1.00	17 & 10 ==> 16 & 8	17 & 10	16 & 8
3	2.78	100.00	2.78	36.00	1.00	22 & 16 ==> 17	22 & 16	17
3	2.78	100.00	2.78	36.00	1.00	21 & 16 ==> 17	21 & 16	17
3	2.78	100.00	2.78	36.00	1.00	21 & 8 ==> 17	21 & 8	17
3	2.78	100.00	2.78	36.00	1.00	18 & 16 ==> 17	18 & 16	17
3	2.78	100.00	2.78	36.00	1.00	16 & 8 ==> 17	16 & 8	17
4	2.78	100.00	2.78	36.00	1.00	25 & 22 & 16 ==> 17	25 & 22 & 16	17
4	2.78	100.00	2.78	36.00	1.00	25 & 21 & 16 ==> 17	25 & 21 & 16	17
4	2.78	100.00	2.78	36.00	1.00	25 & 21 & 8 ==> 17	25 & 21 & 8	17
4	2.78	100.00	2.78	36.00	1.00	25 & 18 & 16 ==> 17	25 & 18 & 16	17
4	2.78	100.00	2.78	36.00	1.00	25 & 16 & 8 ==> 17	25 & 16 & 8	17

Rule 1: Often Missing Work Then Reasons 12 and 5.

An employee will likely not be working because of Reason 28 and 21 if they had Reason 12 & 5 before. (100% Confidence Expected)

Rule 2: Reason 21 and 6 Absences Predict Reason 12 and 5 Absences.

An employee who missed work because of Reason 21 and 6 will likely miss work because of Reason 12 and 5. (100% Confidence Expected)

Rule 3: Missing Reason 27 and 22 Causes Missing Reason 12 and 9

An employee will likely miss work because of 12 and 9 if they missed work because of Reason 27 and 22. (100% Confidence Expected)

Rule 4: Reason 27 and 14 Absences Signify Days 12 and 9 Absences

An employee will miss work on Reason 12 and 9 if they had faced Reason 27 and 14. (100% Confidence Expected)

Rule 5: Reason 9 and Reason 1 caused Absences Correlate with Reason 13 and Reason 4

An employee will miss work because of Reason 13 and 4 if they missed work on a day because of Reason 9 and 1. (100% Confidence Expected)

Rule 6: Missing any part of day and the reason is Reason 17 will also miss parts of days because of Reason 16, 10, & 8

It is very likely that an employee who missed work because of 17 will miss work because Reason 16, 10, and 8. (100% Confidence Expected)

2. **Target: severe_of_absenteeism**

Name	Use	Role	Level
Absenteeism_se	**Yes**	Target	Nominal
Month_of_abser	**No**	Sequence	Interval
Reason_for_abs	**No**	Target	Nominal
dataobs	**No**	ID	Interval
employee_ID	**Yes**	ID	Nominal
record_ID	**No**	ID	Nominal

The variables used are ‘Absenteeism_severe’ and ‘employee_ID’, target role, and ID role, respectively. Both variables are Nominal Levels.

Results:

Association Report

Relations	Expected Confidence (%)	Confidence (%)	Support (%)	Lift	Transaction Count	Rule	Left Hand of Rule	Right Hand of Rule	Rule Item 1
2	94.44	93.75	83.33	0.99	30.00	1 ==> 0	1	0	1
2	88.89	88.24	83.33	0.99	30.00	0 ==> 1	0	1	0

If an employee has a Mild Severity (Severity 1), there is a 94.44% chance that the employee will be present the next day. Conversely, if an employee does not have any absent issue (Severity 0), there is an 88.89% chance that he will have a mild severity (Severity 1).

5.3 Sequence Analysis

One data mining technique used to find patterns and relationships in sequential data sets is sequence analysis. It entails looking for significant trends and dependencies in the ordered sequences of absence episodes within the framework of our absenteeism study. Sequence analysis highlights the chronological order of occurrences, in contrast to standard research methods that concentrate on individual data points. This enables us to spot recurring trends in employees' absenteeism behaviour over time.

Sequence analysis aids in understanding the chronological order of various absence types in the context of workplace absenteeism and may provide light on the causes of employee absenteeism. We want to uncover common sequences, patterns, and dependencies that contribute to a more thorough knowledge of absenteeism trends by examining the temporal links between absences. This analysis can provide valuable information for human resource management strategies and policy formulation, allowing organisations to tailor interventions and support systems to address specific absenteeism patterns effectively.

1. **Target: Reasons_of_absenteeism**

Name	Use	Role	Level
Absenteeism_se	**No**	Target	Nominal
Month_of_abser	**Yes**	Sequence	Interval
Reason_for_abs	**Yes**	Target	Nominal
dataobs	**No**	ID	Interval
employee_ID	**Yes**	ID	Nominal
record_ID	**No**	ID	Nominal

In the association rules analysis, we explore the interplay between employee_ID and Reason_for_absenteeism, both of which are nominal-level variables. We also set the ‘Month_of_absenteeism’ as a sequence role.

Result:

Sequence Report

Chain Length	Transaction Count	Support (%)	Confidence (%)	Pseudo Lift	Rule	Chain Item 1	Chain Item 2	Chain Item 3	Rule Index	Left Hand of Rule	Right Hand of Rule
2	14	38.89	77.78	1.22	13 ==> 23	13	23		1	13	23
2	14	38.89	60.87	0.95	23 ==> 23	23	23		2	23	23
2	12	33.33	52.17	1.04	23 ==> 13	23	13		3	23	13
2	12	33.33	52.17	0.82	0 ==> 23	0	23		4	0	23
2	11	30.56	47.83	0.75	0 ==> 0	0	0		5	0	0
2	11	30.56	61.11	0.96	13 ==> 0	13	0		6	13	0
2	11	30.56	47.83	0.75	23 ==> 0	23	0		7	23	0
2	11	30.56	64.71	1.01	19 ==> 23	19	23		8	19	23
2	10	27.78	43.48	1.20	23 ==> 10	23	10		9	23	10
2	10	27.78	58.82	1.18	19 ==> 13	19	13		10	19	13
2	10	27.78	55.56	1.18	13 ==> 19	13	19		11	13	19
2	10	27.78	76.92	1.20	25 ==> 23	25	23		12	25	23
2	10	27.78	43.48	1.20	23 ==> 25	23	25		13	23	25
3	10	27.78	71.43	1.12	13 ==> 23 ==> 23	13	23	23	14	13 ==> 23	23
2	9	25.00	52.94	0.83	19 ==> 0	19	0		15	19	0
2	9	25.00	50.00	1.38	13 ==> 10	13	10		16	13	10
2	9	25.00	50.00	1.00	13 ==> 13	13	13		17	13	13
2	9	25.00	75.00	1.17	18 ==> 23	18	23		18	18	23
2	9	25.00	81.82	1.28	26 ==> 23	26	23		19	26	23
2	9	25.00	50.00	1.64	13 ==> 26	13	26		20	13	26
2	9	25.00	39.13	1.28	23 ==> 26	23	26		21	23	26
2	9	25.00	39.13	0.94	0 ==> 28	0	28		22	0	28
2	9	25.00	39.13	0.94	23 ==> 28	23	28		23	23	28
2	9	25.00	60.00	1.44	28 ==> 28	28	28		24	28	28
3	9	25.00	81.82	1.28	13 ==> 0 ==> 23	13	0	23	25	13 ==> 0	23

Rule 1: Absence Reason 13 followed by Absence Reason 23

If an employee is absent due to reason 13, there is a high likelihood (Support: 38.89%, Confidence: 77.78%) they will also be absent due to reason 23. (Lift: 1.22)

Rule 2: Absence Reason 23 followed by Another Absence Reason 23

If an employee is absent due to reason 23, there is a significant probability (Support: 38.89%, Confidence: 60.87%) they will have another absence due to reason 23. (Lift: 0.95)

Rule 3: Absence Reason 23 preceding Absence Reason 13

If an employee is absent due to reason 23, there is a moderate chance (Support: 33.33%, Confidence: 52.17%) they had a preceding absence due to reason 13. (Lift: 1.04)

Rule 4: Absence (Reason 0) followed by Absence Reason 23

If an employee is absent (Reason 0), there is a notable possibility (Support: 33.33%, Confidence: 52.17%) they will be absent due to reason 23. (Lift: 0.82)

Rule 5: Absence (Reason 0) followed by another Absence (Reason 0)

If an employee is absent (Reason 0), there is a substantial likelihood (Support: 30.56%, Confidence: 47.83%) they will continue being absent. (Lift: 0.75)

Rule 6: Absence Reason 13 preceding Absence (Reason 0)

If an employee is absent due to reason 13, there is a reliable chance (Support: 30.56%, Confidence: 61.11%) they will be absent on the following day. (Lift: 0.96)

Rule 7: Absence Reason 23 preceding Absence (Reason 0)

If an employee is absent due to reason 23, there is a substantial likelihood (Support: 30.56%, Confidence: 47.83%) they will be absent on the following day. (Lift: 0.75)

Rule 8: Absence Reason 19 preceding Absence Reason 23

If an employee is absent due to reason 19, there is a significant chance (Support: 30.56%, Confidence: 64.71%) they will be absent due to reason 23. (Lift: 1.01)

Rule 9: Absence Reason 23 preceding Absence Reason 10

If an employee is absent due to reason 23, there is a notable probability (Support: 27.78%, Confidence: 43.48%) that they will be absent due to reason 10. (Lift: 1.20)

Rule 10: Absence Reason 19 preceding Absence Reason 13

If an employee is absent due to reason 19, there is a considerable likelihood (Support: 27.78%, Confidence: 58.82%) they will be absent due to reason 13. (Lift: 1.18)

2. **Target: Absenteeism_severe**

Name	Use	Role	Level
Absenteeism_se	**Yes**	Target	Nominal
Month_of_abser	**Yes**	Sequence	Interval
Reason_for_abs	**No**	Target	Nominal
dataobs	**No**	ID	Interval
employee_ID	**Yes**	ID	Nominal
record_ID	**No**	ID	Nominal

In the analysis of association rules, we delve into the interaction between employee identification (employee_ID) and absenteeism severity (Absenteeism_severe). Besides that, we also set month (Month_of_absenteeism) as the sequence. Both variables are classified at the nominal level, signifying unique identifiers for employees and specific categories for reasons behind their absenteeism.

Result:

Sequence Report

Chain Length	Transaction Count	Support (%)	Confidence (%)	Pseudo Lift	Rule	Chain Item 1	Chain Item 2	Chain Item 3	Rule Index	Left Hand of Rule	Right Hand of Rule
2	29	80.56	85.29	0.90	0 ==> 0	0	0		1	0	0
2	28	77.78	87.50	0.93	1 ==> 0	1	0		2	1	0
2	26	72.22	81.25	0.91	1 ==> 1	1	1		3	1	1
2	25	69.44	73.53	0.83	0 ==> 1	0	1		4	0	1
3	24	66.67	82.76	0.88	0 ==> 0 ==> 0	0	0	0	5	0 ==> 0	0
3	23	63.89	92.00	0.97	0 ==> 1 ==> 0	0	1	0	6	0 ==> 1	0
3	23	63.89	92.00	1.04	0 ==> 1 ==> 1	0	1	1	7	0 ==> 1	1
2	22	61.11	95.65	1.01	0 & 1 ==> 0	0 & 1	0		8	0 & 1	0
2	22	61.11	64.71	1.01	0 ==> 0 & 1	0	0 & 1		9	0	0 & 1
3	22	61.11	84.62	0.95	1 ==> 1 ==> 1	1	1	1	10	1 ==> 1	1
2	21	58.33	91.30	1.03	0 & 1 ==> 1	0 & 1	1		11	0 & 1	1
3	21	58.33	75.00	0.79	1 ==> 0 ==> 0	1	0	0	12	1 ==> 0	0
3	21	58.33	95.45	1.01	0 ==> 0 & 1 ==> 0	0	0 & 1	0	13	0 ==> 0 & 1	0
3	21	58.33	72.41	0.81	0 ==> 0 ==> 1	0	0	1	14	0 ==> 0	1
3	21	58.33	75.00	0.84	1 ==> 0 ==> 1	1	0	1	15	1 ==> 0	1
3	20	55.56	76.92	0.81	1 ==> 1 ==> 0	1	1	0	16	1 ==> 1	0
3	20	55.56	90.91	1.02	0 & 1 ==> 0 ==> 1	0 & 1	0	1	17	0 & 1 ==> 0	1
3	20	55.56	90.91	1.02	0 ==> 0 & 1 ==> 1	0	0 & 1	1	18	0 ==> 0 & 1	1
2	19	52.78	59.38	0.93	1 ==> 0 & 1	1	0 & 1		19	1	0 & 1
3	19	52.78	76.00	1.19	0 ==> 1 ==> 0 & 1	0	1	0 & 1	20	0 ==> 1	0 & 1
2	18	50.00	78.26	1.22	0 & 1 ==> 0 & 1	0 & 1	0 & 1		21	0 & 1	0 & 1
3	18	50.00	85.71	0.91	0 & 1 ==> 1 ==> 0	0 & 1	1	0	22	0 & 1 ==> 1	0
3	18	50.00	62.07	0.97	0 ==> 0 ==> 0 & 1	0	0	0 & 1	23	0 ==> 0	0 & 1
3	18	50.00	64.29	1.01	1 ==> 0 ==> 0 & 1	1	0	0 & 1	24	1 ==> 0	0 & 1
3	18	50.00	81.82	1.28	0 ==> 0 & 1 ==> 0 & 1	0	0 & 1	0 & 1	25	0 ==> 0 & 1	0 & 1

Rule 1: No absence (Severity 0) followed by another no absence (Severity 0)

If an employee has no absence (Severity 0), there is a high probability (Support: 80.56%, Confidence: 85.29%) that they will continue to have no absence. (Lift: 0.90)

Rule 2: Mild Absence (Severity 1) followed by No Absence (Severity 0)

If an employee has a mild absence (Severity 1), there is a substantial likelihood (Support: 77.78%, Confidence: 87.50%) that they will have no absence in the subsequent period. (Lift: 0.93)

Rule 3: Mild Absence (Severity 1) Followed by Another Mild Absence (Severity 1)

If an employee has a mild absence (Severity 1), there is a notable probability (Support: 72.22%, Confidence: 81.25%) that they will have another mild absence. (Lift: 0.91)

Rule 4: No Absence (Severity 0) Followed by Mild Absence (Severity 1)

If an employee has no absence (Severity 0), there is a significant chance (Support: 69.44%, Confidence: 73.53%) they will have a mild absence. (Lift: 0.83)

Rule 5: No Absence (Severity 0) Followed by No Absence (Severity 0) Followed by No Absence (Severity 0)

If an employee has no absence (Severity 0) for three consecutive periods, there is a strong likelihood (Support: 66.67%, Confidence: 82.76%) they will continue to have no absence. (Lift: 0.88)

Rule 6: No Absence (Severity 0) Followed by Mild Absence (Severity 1) Followed by No Absence (Severity 0)

If an employee has no absence (Severity 0) followed by a mild absence (Severity 1), there is a high probability (Support: 63.89%, Confidence: 92.00%) they will return to having no absence. (Lift: 0.97)

Rule 7: No Absence (Severity 0) Followed by Mild Absence (Severity 1) Followed by Mild Absence (Severity 1)

If an employee has no absence (Severity 0) followed by two consecutive mild absences (Severity 1), there is a high probability (Support: 63.89%, Confidence: 92.00%) they will continue to have mild absences. (Lift: 1.04)

Rule 8: No Absence (Severity 0) & Mild Absence (Severity 1) Followed by No Absence (Severity 0)

If an employee has no absence (Severity 0) and a mild absence (Severity 1), there is a high likelihood (Support: 61.11%, Confidence: 95.65%) they will have no absence in the subsequent period. (Lift: 1.01)

Rule 9: No Absence (Severity 0) followed by No Absence (Severity 0) & Mild Absence (Severity 1)

If an employee has no absence (Severity 0) followed by no absence (Severity 0) and a mild absence (Severity 1), there is a moderate chance (Support: 61.11%, Confidence: 64.71%) they will continue to have no absence. (Lift: 1.01)

Rule 10: Mild Absence (Severity 1) followed by Another Mild Absence (Severity 1) Followed by Another Mild Absence (Severity 1)

If an employee has a mild absence (Severity 1) for three consecutive periods, there is a significant probability (Support: 58.33%, Confidence: 84.62%) they will continue to have mild absences. (Lift: 0.95)

The examination of association rules reveals significant trends in employee absenteeism at work, highlighting the complex connection between employee_ID and Reason_for_absenteeism. Particular causes of absence show unique patterns that provide important information for comprehending and forecasting absenteeism trends. To successfully address certain patterns, interventions and policies must be tailored with this information in mind.

Simultaneously, sequence analysis reveals patterns in the evolution of absence reasons over time by highlighting the chronological order of absenteeism occurrences. Workers with certain motivations typically follow predictable patterns, which makes it easier to spot recurrent absenteeism trends. Concurrently, the analysis of absenteeism severity offers perceptions into the shift between severity levels, enabling preventative actions and guaranteeing a more seamless return to work following absences.

6. Modify

The modification that has been done to our dataset is imputing the missing value and variable transformation, which was done using SAS Enterprise Miner. The below picture shows the picture of the SAS Enterprise Miner Impt Node and Transform Variables Node that was used for imputing missing values and transformation variable.

For imputing missing values, the Impt Node was used to fill the missing value using the Tree Surrogate method. Tree Surrogates are used to identify surrogate variables that are highly associated with the missing variables. The algorithm uses the value of the chosen surrogate variable to fill in the missing value. Below are the parameters used for the Impt node.

General	
Node ID	Impt
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Nonmissing Variables	No
Missing Cutoff	50.0
Class Variables	
Default Input Method	Tree Surrogate
Default Target Method	None
Normalize Values	Yes

In our Dataset, Absenteeism_severity has a few missing values. To impute the missing value, we used the Tree Surrogate method. Below shows how the missing values are being imputed where the Absenteeism_severity is depends on the Absent_time_in_hours. (Time in hours => 8, then it counts as severe.)

Absenteeism_time_in_hours	Absenteeism_severity	Imputed: ...
4	0	0
0	.	0
2	0	0
4	.	0
2	0	0
2	0	0
8	1	1
4	.	0
40	1	1
8	1	1
8	1	1
8	1	1
8	1	1
1	.	0
4	.	0
8	.	1

For transformation variables, the Transform Variables Node was used to bring variables onto a common scale (0 until 1) which will be used for modelling. Below are the one of the examples of variables in our dataset that are continuous.

Absenteeism_time_in_hours	Service_time
4	13
0	18
2	18
4	14
2	13
2	18
8	3
4	11
40	14
8	14
8	11
8	11
8	11
1	18
4	18
8	16
2	18
8	18

After the variables are transformed into a range of 0 until 1, the variables are changed as per picture below.

Transformed: Absenteeism ...	Transformed: Servic...
0.033333	0.428571
0	0.607143
0.016667	0.607143
0.033333	0.464286
0.016667	0.428571
0.016667	0.607143
0.066667	0.071429
0.033333	0.357143
0.333333	0.464286
0.066667	0.464286
0.066667	0.357143
0.066667	0.357143
0.066667	0.357143
0.008333	0.607143
0.033333	0.607143
0.066667	0.535714
0.016667	0.607143

7. Model & Assess

The first process of modelling is to split the dataset into training and validation. The node Data partition is used to split the dataset before feeding it into the model. To observe the characteristics of underfitting, a Logistic regression model was used. Underfitting happens when the model performs poorly on training data. The problem of underfitting can be resolved by parameter tuning. By finding the right parameter, the model can perform better.

Data Set Allocations	
Training	60.0
Validation	40.0
Test	0.0

The figure above shows the data set allocation for Logistic model 1. The figure above shows that only 60% of the dataset are used for training. This leads to underfitting as the model does not have enough training data to make predictions.

Data Set Allocations	
Training	80.0
Validation	20.0
Test	0.0

The figure above shows the data set allocation for Logistic model 2. After parameter tuning, 80% of the dataset is used for training instead of the 60%. To prove that Logistic model 2 would perform better, the node Model comparison is used.

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label
Y	Reg2	Reg2	Regression	Absenteeis...	Absenteeis...
	Reg	Reg	Regression	Absenteeis...	Absenteeis...

The figure above shows that Logistic model 2 performs better than Logistic model 1. The table below shows the evaluation metric obtained from both models.

Evaluation metric	**Logistic model 1**	**Logistic model 2**
Root Mean Square Error (RMSE)	0.110524	0.017056
Average Squared Error (ASE)	0.012216	0.0002909
Misclassification Rate	0.016129	0

RMSE is a measure of the average magnitude of the errors between predicted and actual values. Lower RMSE values indicate better model performance. In this case, Model 2 has a significantly lower RMSE, suggesting it has better predictive accuracy.

ASE is the mean of the squared errors between predicted and actual values. Similar to RMSE, lower ASE values indicate better model performance. Again, Model 2 has a much lower ASE, indicating superior performance in terms of error magnitude.

The misclassification rate represents the proportion of incorrectly classified instances. A lower misclassification rate is better. Model 2 has a misclassification rate of 0, suggesting perfect classification with no misclassifications.

After finding out the optimised parameter to be used, now we will proceed with modelling. Four (4) models are chosen which are Logistic regression, gradient boosting, decision tree and autoneural. Each model's parameters are tuned and optimised to achieve its best performance before being compared. The parameters for each model are stated below:

Logistic regression

- Main Effects: Include main effects.
- Two-Factor Interactions: Include two-factor interactions.

- Polynomial Terms and Polynomial Degree: Include quadratic (degree = 2) polynomial terms.
- Selection Method: Use Stepwise selection**.**

Gradient boosting

- Number of Iterations (NIterations): 100
- Seed: 7458
- Max Branch: 2
- Max Depth: 4
- Reuse Variable: 1

Decision tree

- Significance Level: 0.05
- Max Branch: 2
- Max Depth:5
- Minimum Categorical Size:5
- Leaf Size:7
- Number of Rules:5

Autoneural

- Architecture:Single layer
- Max Iterations: 30
- Tolerance: Medium
- Total Time: 1 hour
- Number of Hidden Units: 10

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label
Y	Tree	Tree	Decision Tr...	Absenteeis...	Absenteeis...
	Boost	Boost	Gradient Bo...	Absenteeis...	Absenteeis...
	AutoNeural	AutoNeural	AutoNeural	Absenteeis...	Absenteeis...
	Reg2	Reg2	Regression	Absenteeis...	Absenteeis...

The figure above shows that the best performing model is Decision Tree. The evaluation metrics for each model are shown in the table below:

Evaluation metric	**Logistic regression**	**AutoNeural**	**Gradient boosting**	**Decision tree**
Root Mean Square Error (RMSE)	0.484296	0.176868	-	-
Average Squared Error (ASE)	0.234542	0.031282	0.124622	0
Misclassification Rate	0.375	0.03125	0	0

Root Mean Square Error (RMSE):

Logistic Regression: 0.484296 - A measure of the discrepancy between the predicted probabilities and the actual binary outcomes.

AutoNeural: 0.176868 - Represents the square root of the average squared differences between the predicted probabilities and the actual binary outcomes.

Gradient Boosting: No value.

Decision Tree: No value.

Average Squared Error (ASE):

Logistic Regression: 0.234542 - The average of squared errors between predicted and actual values.

AutoNeural: 0.031282 - The average of squared differences between predicted and actual values.

Gradient Boosting: 0.124622 - Similar to ASE for AutoNeural.

Decision Tree: 0 - ASE is 0, suggesting perfect predictions.

Misclassification Rate:

Logistic Regression: 0.375 - Represents the proportion of misclassified instances.

AutoNeural: 0.03125 - Indicates the proportion of misclassified instances.

Gradient Boosting: 0 - No misclassifications observed.

Decision Tree: 0 - No misclassifications observed.

Conclusion

In conclusion, employing the SEMMA methodology on a diverse UCI dataset has revealed intricate absenteeism patterns. Stratified sampling ensured a fair representation of absence types, while exploration explored association rules and sequence analysis.

Data modification, including handling missing values and normalisation, prepared the dataset for robust modelling. Classification models were pivotal in uncovering significant patterns. Thorough assessment validated model effectiveness and result reliability.

This systematic approach equips organisations with actionable insights for evidence-based strategies, fostering productivity, a healthy workplace, and informed HR policies.