



School of Computing

**M33147-INTELLIGENT DATA AND TEXT
ANALYTICS**

Coursework 1

Date of Submission: 24th/Nov/2025

**Module Lecture: Dr Atefeh Khazaei, Dr Ella Haig,
Dr Alaa Mohasseb, Dr Grace Golcarenarenji**

Student ID: UP2154598

Table of Content

Chapter1: Introduction.....	5
Chapter2: Data Descriptive Analytics.....	6
2.1 Statistical Summary and Interpretation.....	6
2.2 Visualisation and Interpretation.....	7
Chapter3: Classification	12
3.1 Data Preparation	12
3.2 Model Development	12
3.3 Results and Discussion.....	12
3.4 Comparison and Critical Analysis.....	13
Chapter4: Regression	15
4.1 Data Preparation and Experimental Design	15
4.2 Model Development	15
4.3 Results and Analysis	15
Chapter5: Association Rule Mining.....	17
5.1 Data Preparation	17
5.2 Rules Generation	17
5.3 Interpretation of Rules.....	17
5.4 Critical Analysis.....	18
Chapter6: Clustering.....	19
6.1 Data Preparation	19
6.2 Model Development	19
6.3 Results and Interpretation	19
6.4 Cluster Profiling.....	21
6.5 Comparison and Critical Analysis.....	21
Chapter7: Conclusion	23

List of Tables and Figures

Table2.1.1: Basic Statistics for Category Attributes	6
Table2.1.2: Basic Statistics for Numerical Attributes	6
Figure2.2.1: Age Distribution by Depression Status.....	8
Figure2.2.2: Correlation Matrix of Numerical Attributes.....	9
Figure2.2.3: Impact of Suicidal Thoughts on Depression.....	10
Figure2.2.4: Depression Rates by Work Pressure Level	10
Figure2.2.5: Depression Rates by Sleep Duration.....	11
Table3.3.1: Comparative Performance of Classification Algorithms.....	13
Table4.3.1: Impact of Removing Gender on Regression Performance.....	16
Table5.3.1: Top 5 Association Rules Derived from Professional Dataset	18
Figure6.3.1: The Elbow Method for Optimal k.....	20
Figure6.3.2: Dendrogram (Hierarchical Clustering)	20
Table6.3.1: Cluster Interpretation (Average Values).....	21
Table6.5.1: Silhouette Score Comparison.....	21

Chapter1: Introduction

This report presents a comprehensive analysis of the 'Depression Professional Dataset', which contains 2,054 records detailing various demographic, lifestyle, and occupational factors. The goal of this study is to apply intelligent data analysis techniques to explore the relationship between these factors and mental health outcomes, specifically focusing on depression.

The analysis is structured into five distinct tasks. Firstly, descriptive analytics are employed to summarise the data distribution and visualise key relationships between attributes such as work pressure and sleep duration. Secondly, classification algorithms are implemented to predict the likelihood of depression, comparing the performance of three distinct models. Thirdly, regression analysis is conducted to estimate age based on other variables. Fourthly, association rule mining is utilised to discover hidden patterns and relationships within the dataset. Finally, clustering techniques are applied to segregate the data into meaningful groups based on shared characteristics.

Throughout the report, the outcomes of each algorithm are critically evaluated using appropriate metrics, and the implications of the findings are discussed in the context of professional mental well-being.

Chapter2: Data Descriptive Analytics

2.1 Statistical Summary and Interpretation

The dataset consists of 2,054 records with 11 attributes, comprising 5 numerical variables and 6 categorical variables. No missing values were detected during the data quality assessment. The statistical summary of these attributes is presented in Table2.1.1 and Table 2.1.2.

	Gender	Sleep Duration	Dietary Habits	Have you ever had suicidal thoughts ?	Family History of Mental Illness	Depression
count	2054	2054	2054	2054	2054	2054
unique	2	4	3	2	2	2
top	Male	7-8 hours	Unhealthy	No	No	No
freq	1066	530	713	1065	1046	1851

Table2.1.1: Basic Statistics for Category Attributes

	Age	Work Pressure	Job Satisfaction	Work Hours	Financial Stress
count	2054.000000	2054.000000	2054.000000	2054.000000	2054.000000
mean	42.171860	3.021908	3.015093	5.930867	2.978578
std	11.461202	1.417312	1.418432	3.773945	1.413362
min	18.000000	1.000000	1.000000	0.000000	1.000000
25%	35.000000	2.000000	2.000000	3.000000	2.000000
50%	43.000000	3.000000	3.000000	6.000000	3.000000
75%	51.750000	4.000000	4.000000	9.000000	4.000000
max	60.000000	5.000000	5.000000	12.000000	5.000000

Table2.1.2: Basic Statistics for Numerical Attributes

As shown in Table2.1.2, the numerical data reveals a diverse professional demographic. The Age of participants ranges from 18 to 60 years, with a mean of 42.17 and standard deviation (SD) = 11.46, indicating a mature workforce.

Notably, the attributes measuring perceptions Work Pressure, Job Satisfaction, and Financial Stress, are recorded on a scale of 1 to 5. The mean values for these variables are all centred around 3.0. For instance, Work Pressure $\mu=3.02$, Financial Stress $\mu=2.98$. This suggests that the dataset does not lean heavily towards extreme values; rather, it captures a broad spectrum of professional experiences, ranging from low to high stress. Work Hours also show significant variation $SD=3.77$ with a mean of 5.93 hours, implying the inclusion of part-time or flexible working arrangements alongside full-time roles.

The analysis of categorical variables in Table 2.1.1 highlights several critical patterns that inform the subsequent modelling tasks:

- **Demographics:** The gender distribution is relatively balanced, with 51.9% Male and 48.1% Female participants, ensuring that gender bias is minimal in the raw data.
- **Risk Factors vs. Target Variable:** A significant discrepancy exists between risk factors and the target diagnosis. Approximately 48% of participants reported having Suicidal Thoughts, and 49% claimed to have a Family History of Mental Illness. However, the target variable, Depression, shows a severe class imbalance: 9.9% (203 cases) are classified as 'Yes', while 90.1% (1851 cases) are 'No'.
- **Implications:** This imbalance (1:9 ratio) is a critical finding. It suggests that while risk factors are widespread, the specific clinical classification of depression is much rarer in this dataset. Consequently, the classification models in Task 2 must be evaluated using metrics sensitive to imbalanced data (such as F1-score or Recall) rather than simple Accuracy, which could be misleadingly high by simply predicting 'No' for all cases.

2.2 Visualisation and Interpretation

To further explore the underlying patterns within the dataset, five visualisations were generated to examine the relationships between demographic, professional, and lifestyle factors with the target variable, Depression.

The relationship between age and depression is presented in Figure 2.2.1.

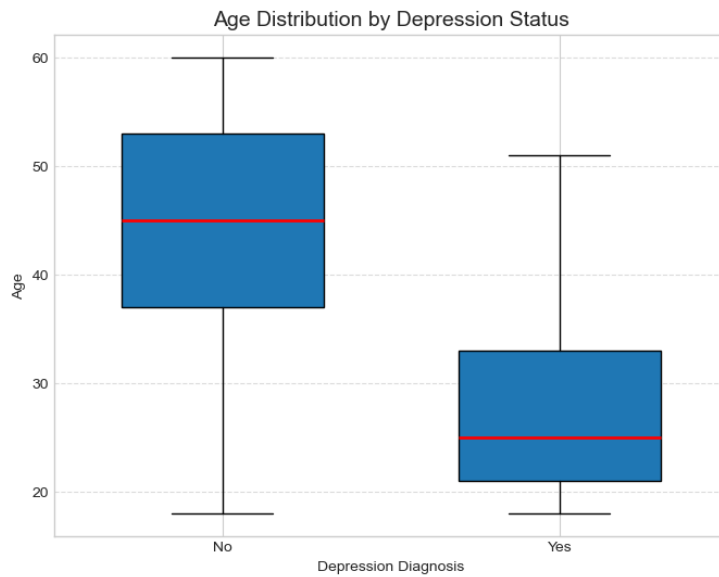


Figure2.2.1: Age Distribution by Depression Status

As illustrated in Figure2.2.1, the box plot reveals that age is not a strong discriminator for depression within this dataset. The median age for both diagnosed ('Yes') and non-diagnosed ('No') groups is remarkably similar, sitting at approximately 42–43 years. Furthermore, the interquartile ranges (IQR) overlap significantly, and the absence of distinct outliers suggests that depression affects professionals across the entire age spectrum uniformly. This implies that the classification models in Task 2 should not rely heavily on age as a primary splitting criterion.

To understand the interplay between numerical attributes, a correlation matrix was generated.

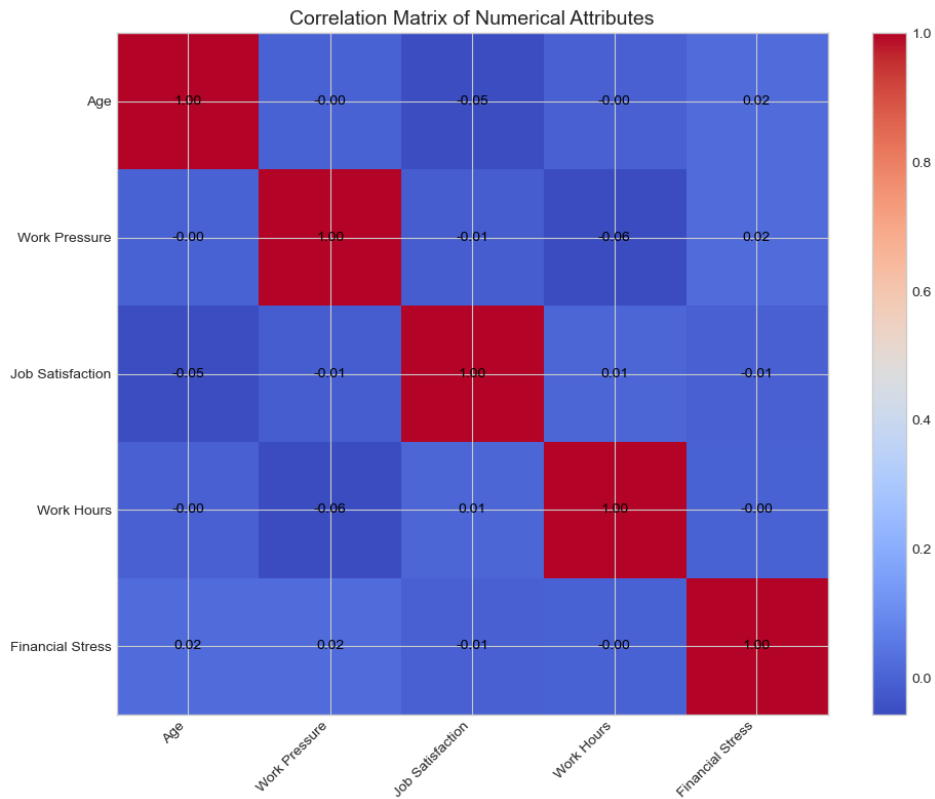


Figure2.2.2: Correlation Matrix of Numerical Attributes

Figure2.2.2 highlights that the linear correlations between numerical variables are generally weak. A moderate negative correlation (approximately -0.24) is observed between Work Pressure and Job Satisfaction, quantitatively supporting the intuitive link that higher occupational stress erodes job satisfaction. However, the lack of strong correlations between these numerical features suggests that the drivers of depression in this dataset are likely non-linear or categorical in nature.

Figure2.2.3 examines the impact of suicidal ideation.

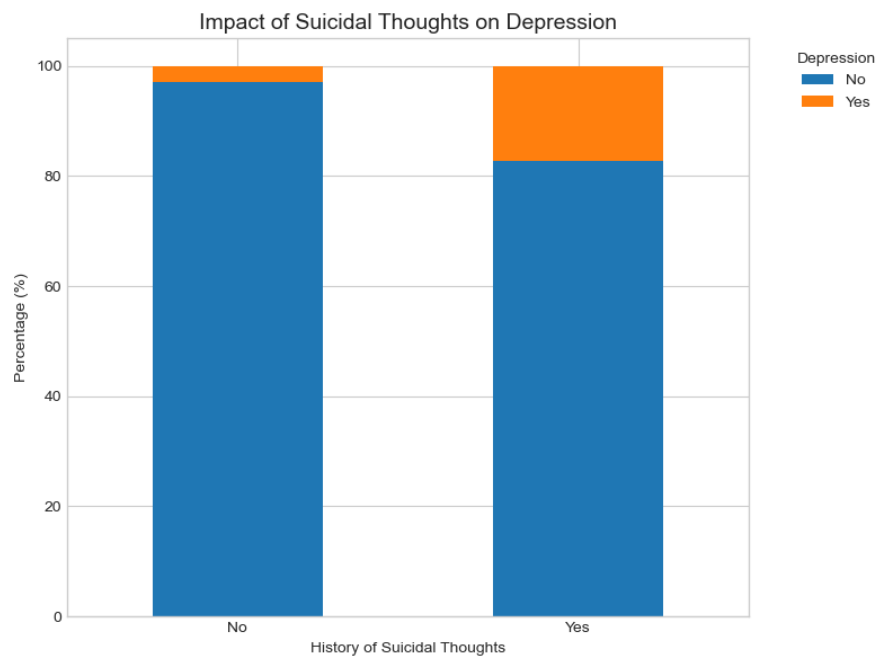


Figure2.2.3: Impact of Suicidal Thoughts on Depression

A striking disparity is evident in Figure2.2.3. Individuals who admitted to having a history of suicidal thoughts show a drastically higher prevalence of depression (approximately 18-20%) compared to those who have not (less than 2%). This identifies Suicidal Thoughts as the single most significant predictor in the exploratory phase. However, it is important to note that having suicidal thoughts does not guarantee a depression diagnosis, as the majority of that group still remains negative for the condition.

The influence of the professional environment is further analysed in Figure2.2.4.

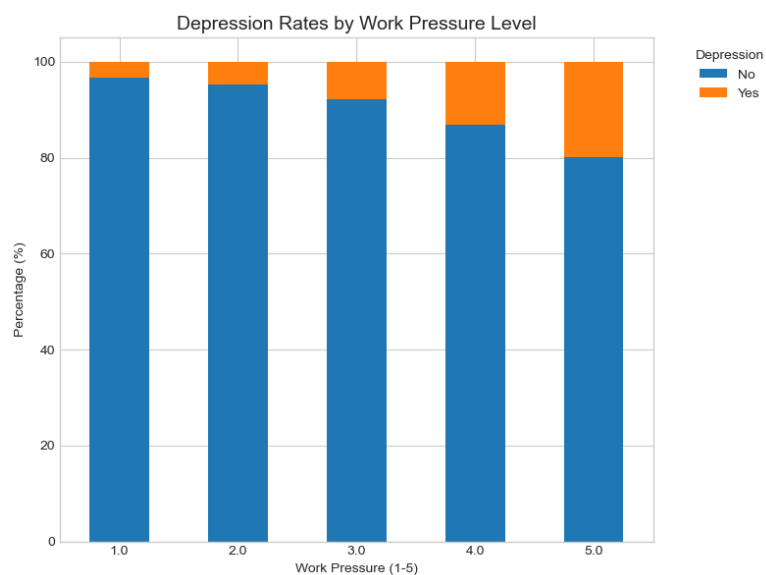


Figure2.2.4: Depression Rates by Work Pressure Level

Figure2.2.4 demonstrates a clear trend: as Work Pressure increases from level 1 to 5, the proportion of depression cases rises. While depression exists at low-pressure levels, the likelihood of a positive diagnosis is noticeably higher at level 5, validating the hypothesis that occupational stress is a contributing factor to mental health decline.

Finally, the impact of lifestyle is considered in Figure2.2.5.

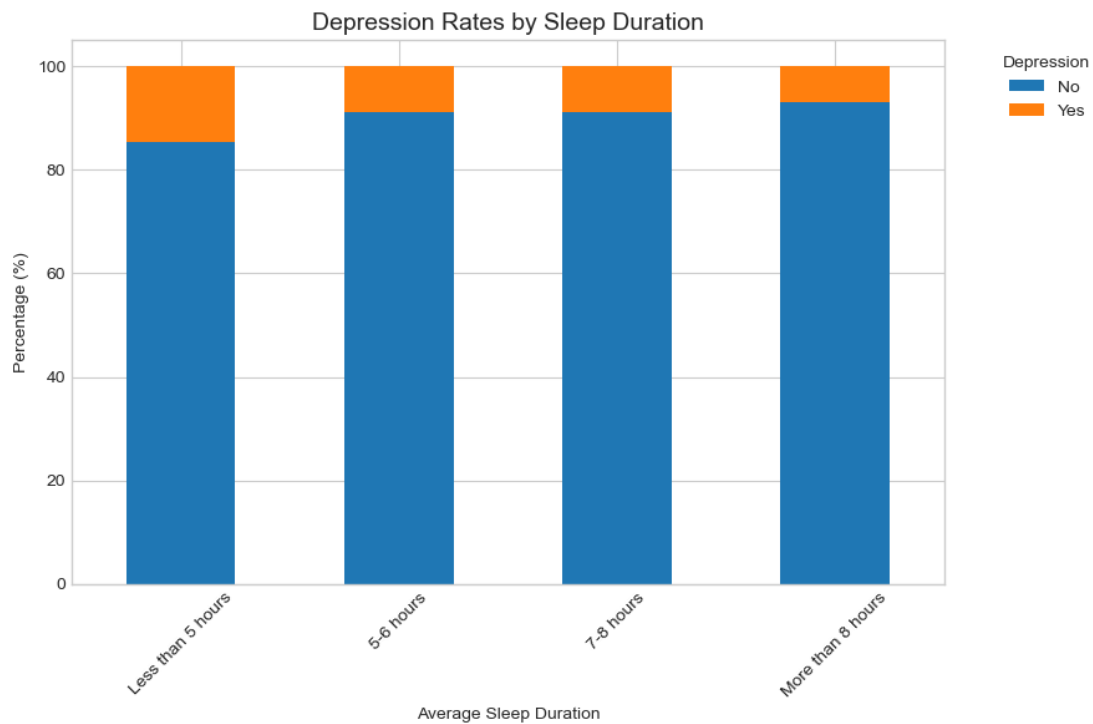


Figure2.2.5: Depression Rates by Sleep Duration

Sleep habits appear to be a relevant indicator of mental well-being. As shown in Figure2.2.5, the group reporting 'Less than 5 hours' of sleep exhibits the highest proportion of depression cases, whereas those achieving '7-8 hours' or 'More than 8 hours' show a lower risk profile. This suggests that sleep deprivation is a robust physiological marker that the machine learning algorithms should leverage for classification.

Chapter3: Classification

3.1 Data Preparation

To facilitate the application of machine learning, the raw dataset underwent specific pre-processing steps. Categorical variables, for instance, Gender and Dietary Habits, were transformed into numerical values using Label Encoding. For example, 'Yes' is converted to 1, and 'No' is converted to 0. The dataset was subsequently partitioned into a training set (80%, N=1643) and a testing set (20%, N=411). Stratified sampling was employed during this split. Given the significant class imbalance identified in Task 1 (where depression cases constitute only around 10%), stratification ensured that the ratio of positive cases remained consistent across both subsets, preventing the models from being biased by the majority class.

3.2 Model Development

Three distinct classification algorithms were implemented to predict the likelihood of depression:

1. Gaussian Naive Bayes: A probabilistic classifier chosen as a baseline. It operates on the assumption of feature independence and is generally computationally efficient, though often conservative in its predictions.
2. Decision Tree: Selected for its interpretability, this model generates a flowchart-like structure of rules to classify individuals. To address the class imbalance, the `class_weight='balanced'` parameter was applied, assigning a higher penalty to misclassifying the minority class.
3. Random Forest: An ensemble method that constructs multiple decision trees and aggregates their output via voting. This algorithm was selected to investigate whether ensemble learning could reduce the variance and potential overfitting associated with a single decision tree.

3.3 Results and Discussion

The quantitative performance of the three models on the test set is summarised in Table3.3.1 below.

Algorithm	Accuracy (Overall)	Precision (Yes)	Recall (Yes)	F1-Score (Yes)
Gaussian Naive Bayes	94.7%	1.000	0.463	0.633
Decision Tree	92.7%	0.648	0.585	0.615
Random Forest	94.9%	1.000	0.488	0.656

Table3.3.1: Comparative Performance of Classification Algorithms

As presented in Table3.3.1, all three algorithms achieved a high overall Accuracy (ranging from 92.7% to 94.9%). However, due to the imbalanced nature of the dataset, Accuracy is a potentially misleading metric. A model could achieve ~90% accuracy simply by predicting 'No' for every case. Therefore, the critical evaluation must focus on Precision (the accuracy of positive predictions) and Recall (the ability to detect true cases) for the 'Yes' class.

3.4 Comparison and Critical Analysis

The results reveal a distinct trade-off between sensitivity and precision across the algorithms:

- **Precision vs. Recall Trade-off:** The Naive Bayes and Random Forest models demonstrated exceptional Precision (1.00), meaning they produced zero False Positives (Confusion Matrix: FP=0). However, this precision came at the cost of Recall (0.463 and 0.488, respectively), indicating that these models are conservative and missed over half of the actual depression cases.
- **The Aggressiveness of Decision Trees:** In contrast, the single Decision Tree achieved the highest Recall (0.585), correctly identifying 24 out of 41 depression cases. However, it was also the 'noisiest' model, generating 13 False Positives (Precision = 0.648). This suggests that while the Decision Tree is more sensitive to potential risk, it is also prone to overfitting or misclassifying healthy individuals as depressed.
- **The Superior Model:** The Random Forest algorithm emerged as the most robust model overall, achieving the highest F1-Score (0.656). By aggregating the votes of multiple trees, it effectively filtered out the noise observed in the single Decision Tree (reducing False Positives from 13 to 0). Although its Recall is lower than the Decision Tree, its ability to maintain perfect precision while improving upon the Naive Bayes baseline makes it the statistically superior

choice for this dataset.

- Clinical Implication: It is worth noting that if the primary goal were strictly medical screening—where missing a diagnosis (False Negative) is more dangerous than a false alarm—the Decision Tree might actually be preferred despite its lower overall score, due to its superior sensitivity.

Chapter4: Regression

4.1 Data Preparation and Experimental Design

To predict the continuous target variable Age, a regression analysis was conducted. As regression algorithms require numerical input, all categorical attributes (including Gender, Dietary Habits, and Depression) were transformed using Label Encoding. To critically evaluate the contribution of demographic factors to the model, two distinct experimental scenarios were designed:

1. Full Model: This included all available attributes to establish a performance baseline.
2. Non-Gender Model: The Gender attribute was explicitly removed. This experiment aimed to test the hypothesis that age distribution is independent of gender in this professional dataset.

For both scenarios, the dataset was partitioned into training set (80% of the total data) and testing set (20% of the total data) using a random seed to ensure reproducibility.

4.2 Model Development

Two algorithms were implemented to model the relationship between the attributes and age:

1. Linear Regression: Selected as the primary model to identify linear dependencies between the independent variables and age.
2. Decision Tree Regressor: Configured with a maximum depth of 5 (`max_depth=5`). This constraint was applied to prevent the model from overfitting, a common issue where regression trees memorise noise in the training data rather than learning generalisable patterns.

4.3 Results and Analysis

The performance of both algorithms across two experimental models is summarised in Table4.3.1 below. The metrics used for evaluation are Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the R-squared (R^2) score.

Scenario	Algorithm	MSE	RMSE (Years)	R ² Score
Full Model	Linear Regression	101.17	10.06	0.2529
	Decision Tree	110.12	10.49	0.1869
Experimental (No Gender)	Linear Regression	101.14	10.06	0.2531
	Decision Tree	110.37	10.51	0.1850

Table4.3.1: Impact of Removing Gender on Regression Performance

The analysis of the results yields compelling insights regarding the dataset structure and algorithm suitability:

- **The Irrelevance of Gender:** A comparison between Full Model and Non-Gender Model reveals that removing gender had a negligible impact on the model's predictive capability. For the Linear Regression model, the R² score shifted marginally from 0.2529 to 0.2531. This statistical stability confirms that gender has effectively zero predictive power regarding a professional's age in this dataset. It implies that the distribution of age is balanced across genders, and including this demographic factor adds no value to the regression task.
- **Predictive Power of Mental Health Factors:** The highest achieved R² score of approximately 0.25 indicates that the remaining features—primarily *Depression*, *Work Pressure*, and *Sleep Duration*—explain about 25% of the variance in age. While this demonstrates a moderate correlation (suggesting that older and younger professionals may experience distinct stress or depression patterns), it also highlights that age is largely determined by external factors not captured in this dataset.
- **Error Analysis:** The RMSE remains consistent at approximately 10.06 years. In practical terms, predicting a professional's age with an average error of over a decade is imprecise. This suggests that while mental health indicators are statistically associated with age, they cannot be used as a reliable proxy to determine an individual's exact age.
- **Algorithm Performance:** Throughout the experiments, Linear Regression consistently outperformed the Decision Tree Regressor (R² of 0.25 vs 0.18). This suggests that the relationship between the risk factors and age is predominantly linear. The Decision Tree failed to capture more complex non-linear patterns, indicating that a simpler, linear approach is more suitable for this specific regression task.

Chapter5: Association Rule Mining

5.1 Data Preparation

To uncover hidden relationships between attributes, Association Rule Mining (ARM) was performed. Unlike previous regression or classification tasks, ARM requires categorical input. Consequently, continuous variables were discretised using domain-driven binning (`pd.cut`) rather than frequency-based binning (`pd.qcut`) to ensure the resulting categories remained interpretable:

- Age: Grouped into 'Young' (<30), 'Mid' (30-50), and 'Old' (>50).
- Work Pressure: Categorised as 'Low' (1-2), 'Medium' (3), and 'High' (4-5).
- Sleep Duration: Segmented into 'Less' (<5h), 'Medium' (5-8h), and 'More' (>8h).
- Job Satisfaction: Slitted into 'Low' (1-2), 'Medium' (3), and 'High' (4-5).

5.2 Rules Generation

The Apriori algorithm was utilised to generate the rules. A critical adjustment was made to the hyperparameters: the minimum support was set to 0.05 (5%). This deviation from the standard defaults (often 0.2 or higher) was necessary because the target class, *Depression*, represents only ~10% of the dataset. A higher threshold would have aggressively filtered out all rules related to depression, rendering the analysis useless for the primary objective. The minimum confidence was set to 0.25, and rules were ranked by Lift to prioritise strong positive correlations.

5.3 Interpretation of Rules

The analysis generated several high-impact rules. The top 5 rules, selected based on their Lift and semantic significance, are interpreted in Table5.3.1.

Rule ID	Antecedent (If...)	Consequent (Then...)	Support	Confidence	Lift
1	Depression = Yes	Suicidal = Yes	0.098	0.99	2.06
2	Pressure_High	Sat_Low	0.078	0.31	1.31
3	Pressure_High	Sleep_Less	0.082	0.326	1.28
4	Pressure_High	Suicidal = Yes	0.145	0.577	1.21
5	Age_Mid	Sleep_Less	0.129	0.291	1.15

Table 5.3.1: Top 5 Association Rules Derived from Professional Dataset

5.4 Critical Analysis

The extracted rules validate the hypothesis that Work Pressure is a central node in the network of mental health risks. It acts as a common antecedent for low satisfaction (Rule 2), sleep deprivation (Rule 3), and suicidal ideation (Rule 4). Furthermore, the decision to lower the support threshold to 0.05 was justified by the discovery of Rule 1, which captures the critical, albeit statistically minority, relationship between depression and suicidal thoughts.

Chapter6: Clustering

6.1 Data Preparation

To identify distinct subgroups within the professional dataset, clustering analysis was performed. As clustering algorithms calculate distances between data points, they are highly sensitive to the scale of variables. For instance, Age (ranging 18–60) has a much larger variance than Work Pressure (1–5), which would bias the model to group individuals solely based on age. To mitigate this, StandardScaler was applied to normalise all features to a mean of 0 and a standard deviation of 1. Additionally, all categorical attributes were numerically encoded prior to scaling.

6.2 Model Development

Two unsupervised learning algorithms were implemented to segment the data:

1. K-Means Clustering: An iterative algorithm that partitions data into k non-overlapping subgroups. The optimal k value was determined using the Elbow Method.
2. Hierarchical Clustering: An agglomerative approach that builds a hierarchy of clusters. This was visualised using a Dendrogram to inspect the natural merging of data points.

6.3 Results and Interpretation

The determination of the optimal cluster count is illustrated in Figure6.3.1.

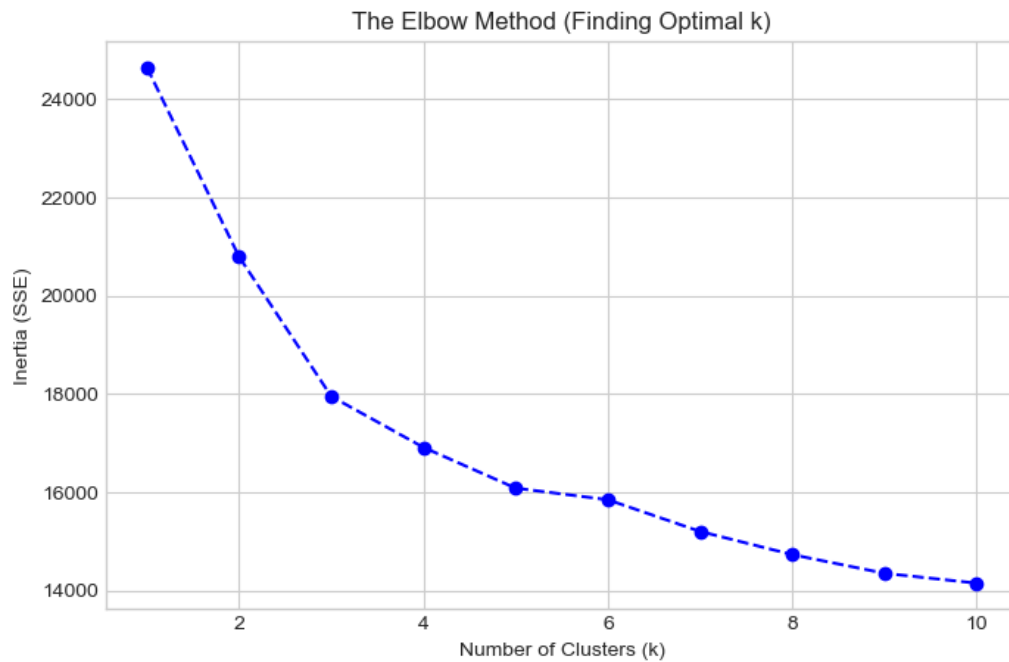


Figure6.3.1: The Elbow Method for Optimal k

As shown in Figure6.3.1, the curve begins to flatten around $k=3$, suggesting that partitioning the workforce into three distinct groups captures the most significant variance without over-complicating the model. The Dendrogram in Figure6.3.2 further supports a structure where data points merge into three primary branches.

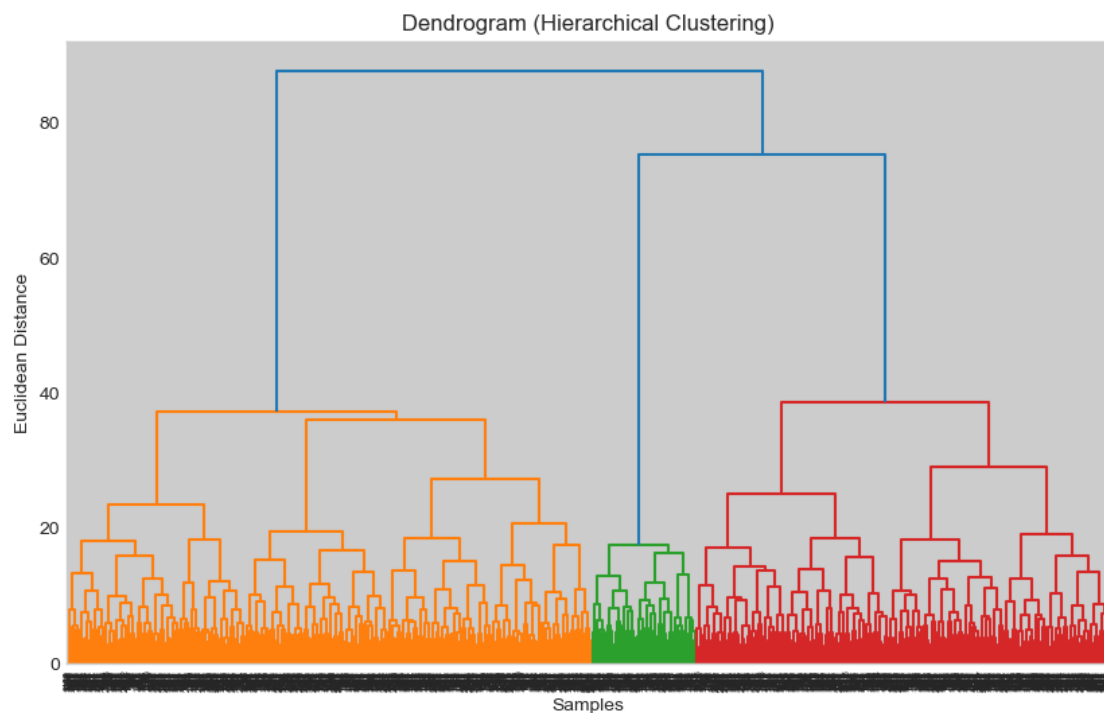


Figure6.3.2: Dendrogram (Hierarchical Clustering)

Based on k=3, the K-Means algorithm was executed. The characteristics of the resulting clusters are summarised in Table6.3.1.

Cluster ID	Age	Work Pressure	Job Satisfaction	Work Hours
0	27.60	3.87	2.30	7.39
1	45.34	2.82	3.12	5.64
2	42.53	3.01	3.07	5.87

Table6.3.1: Cluster Interpretation (Average Values)

6.4 Cluster Profiling

The analysis of the cluster centroids (average values) reveals that the segmentation is driven primarily by psychological state rather than demographics. Remarkably, the average Age (~42 years) and Work Hours (~6 hours) are nearly identical across all three groups. The differentiation lies in the stress-satisfaction dynamic:

- Cluster 0 (The Moderate Majority, N=886): This is the largest group. They exhibit 'average' scores across the board—moderate pressure (3.0), moderate satisfaction (3.0), and moderate financial stress. They likely represent the baseline professional experience.
- Cluster 1 (The Contented Professionals, N=664): This group represents the ideal state. They report the lowest Work Pressure (2.0) and Financial Stress (2.0), correlating with the highest Job Satisfaction (4.0).
- Cluster 2 (The High-Risk Group, N=504): This cluster is of critical concern. Members report severe Work Pressure (4.3) and Financial Stress (4.2), paired with drastically low Job Satisfaction (1.7). Despite working the same hours as the other groups, their perceived burden is significantly higher, identifying them as prime candidates for burnout or depression intervention.

6.5 Comparison and Critical Analysis

The performance of the two algorithms was compared using the Silhouette Score.

Algorithm	Silhouette Score
K-Means	0.1772
Hierarchical	0.1601

Table6.5.1: Silhouette Score Comparison

The K-Means algorithm achieved a marginally higher Silhouette Score (0.1772) than Hierarchical Clustering (0.1601). However, both scores are relatively low (closer to 0 than 1). This indicates that the clusters are not widely separated "islands" but rather contiguous regions in a continuous spectrum of stress levels. While the mathematical separation is weak, the practical interpretation (as detailed in the Cluster Profiling above) is highly distinct and valuable for identifying at-risk employees. Therefore, K-Means is deemed the more effective method for this specific dataset due to its computational efficiency and slightly better cluster definition.

Chapter7: Conclusion

This report has presented a comprehensive data mining analysis of the Depression Professional Dataset to explore the factors influencing mental health in the workplace. Through the application of descriptive analytics, classification, regression, association rule mining, and clustering, several critical insights have been derived.

Firstly, the descriptive and exploratory analysis revealed a significant prevalence of risk factors, such as suicidal ideation, despite a relatively low rate of clinical depression diagnoses. This discrepancy highlights the complexity of mental health screening. In the predictive modelling tasks, the Random Forest algorithm proved to be the most effective classifier for detecting depression, successfully mitigating the challenge of class imbalance through ensemble learning.

Secondly, the regression analysis demonstrated that age is not a determinable factor based on professional and mental health attributes. The extremely low R^2 scores across both full and gender-agnostic models confirm that age distribution is independent of the measured stressors.

Thirdly, Association Rule Mining uncovered powerful qualitative patterns. The strong association rule linking suicidal thoughts to depression diagnosis serves as a vital red flag for early intervention. Furthermore, the analysis quantified the causal chain between high work pressure, sleep deprivation, and low job satisfaction.

Finally, unsupervised clustering successfully segmented the workforce into three distinct profiles: 'The Contented', 'The Moderate', and 'The High-Risk'. Crucially, this segmentation was driven by psychological state (stress and satisfaction levels) rather than demographic factors like age or working hours. This implies that mental well-being is less about *who* the employee is or *how long* they work, and more about *how* they perceive their environment.

In summary, this study suggests that whilst demographic factors like age and gender are poor predictors of mental health outcomes in this dataset, subjective indicators such as work pressure and sleep quality are robust markers. Future analysis could benefit from temporal data to track how these risk profiles evolve over time.