ST3189 MACHINE LEARNING

Coursework

Name: Xavier Chia De-Qing

Student Number: 220455051

Table of Contents

Unsupervised Learning

1.1 Substantive Issue

Sleep disorders are a growing global health concern, affecting millions of people across different demographics. According to the World Health Organization (WHO), up to 40% of the global population suffers from sleep disorders, with insomnia affecting nearly 10-30% of adults worldwide (WHO, 2023). Additionally, modern lifestyle factors—including high work stress, excessive screen time, and reduced physical activity—further contribute to declining sleep quality worldwide (CDC, 2023).

Having established the substantive issue of analysing and evaluating factors contributing to sleep disorders, we aim to implement unsupervised machine learning techniques to group individuals into clusters based on their sleep patterns and health metrics. By interpreting these clusters, we seek to identify distinct groups affected by sleep disorders and uncover potential lifestyle or health-related factors influencing their condition.

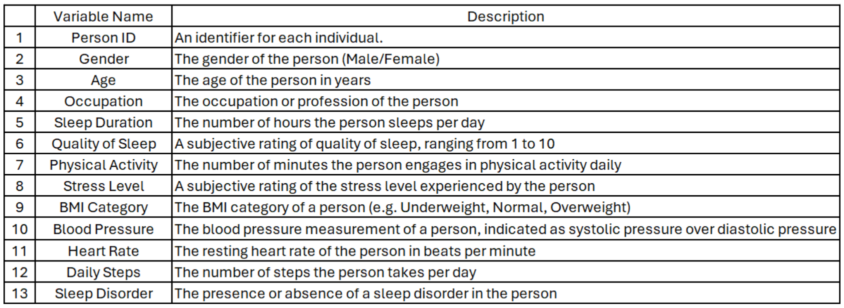
1.2 Research Questions

The research questions for the substantive issue are as follows:

- RQ1: What is the optimal number of clusters for grouping individuals?
- RQ2: What distinct groups can be identified based on sleep and lifestyle related factors?

1.3 Dataset & Variables

The dataset used for this unsupervised learning task is titled as "Sleep Health and Lifestyle Dataset". This dataset was created by collecting information on various health and lifestyle factors that influence sleep quality and disorders. The dataset consists of 13 variables and 374 rows, representing individuals, and multiple variables related to sleep patterns, physical activity, stress levels, and overall health metrics. A summary of the information collected is presented in the table below:



	Variable Name	Description
1	Person ID	An identifier for each individual.
2	Gender	The gender of the person (Male/Female)
3	Age	The age of the person in years
4	Occupation	The occupation or profession of the person
5	Sleep Duration	The number of hours the person sleeps per day
6	Quality of Sleep	A subjective rating of quality of sleep, ranging from 1 to 10
7	Physical Activity	The number of minutes the person engages in physical activity daily
8	Stress Level	A subjective rating of the stress level experienced by the person
9	BMI Category	The BMI category of a person (e.g. Underweight, Normal, Overweight)
10	Blood Pressure	The blood pressure measurement of a person, indicated as systolic pressure over diastolic pressure
11	Heart Rate	The resting heart rate of the person in beats per minute
12	Daily Steps	The number of steps the person takes per day
13	Sleep Disorder	The presence or absence of a sleep disorder in the person

Table 1: Description of Variables in Sleep Health and Lifestyle Dataset

We will remove "Person ID" variable as it is a unique identifier assigned to each individual and does not contribute any meaningful patterns for clustering.

1.4 Methodology

The unsupervised learning task focuses on uncovering patterns, structures, or relationships within the dataset without prior knowledge or labelled examples. The goal is to group similar data points into clusters based on their inherent similarities or differences, helping to identify meaningful subgroups within the data.

To achieve this, unsupervised learning techniques such as dimensionality reduction methods like Principal Component Analysis (PCA) and clustering algorithms like K-Means and Hierarchical Clustering will be applied. These techniques will help simplify the dataset, making it easier to analyse and visualise. The resulting clusters will then be examined and interpreted to gain insights into the underlying patterns within the data.

1.5 Analysis

1.5.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms a large set of variables into a smaller one, called principal components, while retaining most of the original data's variation. It does this by finding new, uncorrelated axes (components) that capture the maximum variance in the data, making it easier to analyse, visualize, and reduce noise in high-dimensional datasets.

This R code prepares a dataset for Principal Component Analysis (PCA) by first loading and cleaning the data. We first remove the “Person ID” column, converts categorical variables into factors, ensuring only numeric columns remain. The dataset is then standardized, which normalizes all variables, making it ready for PCA analysis.

Figure 1 visualizes the variance explained by each principal component in PCA. The blue bars represent the proportion of variance explained by each component, while the red line shows the cumulative variance. Since the cumulative variance explained by the first four principal components is 82.5%, we can retain them for further analysis while reducing dimensionality and preserving most of the dataset’s information.

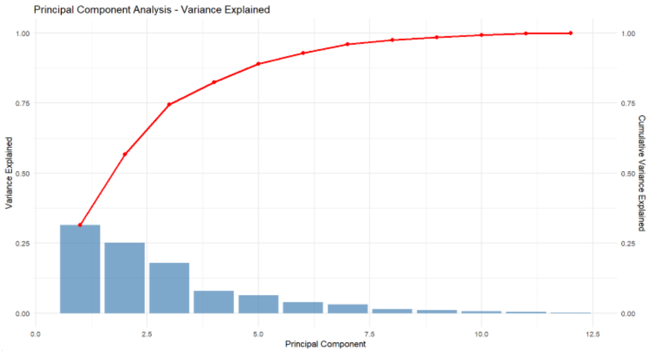


Figure 1: Variance Explained by Principal Component

1.5.2 K-Means Clustering

K-Means clustering is an unsupervised machine learning algorithm used to group data into distinct clusters based on similarity. It partitions a dataset into k clusters, where each data point belongs to the cluster with the nearest mean (centroid). The algorithm iteratively updates the centroids by minimizing the variance within each cluster. The optimal number of clusters can be determined using methods like the elbow method, which examines the within-cluster sum of squares, and the silhouette method, which measures how well data points fit into their assigned clusters.

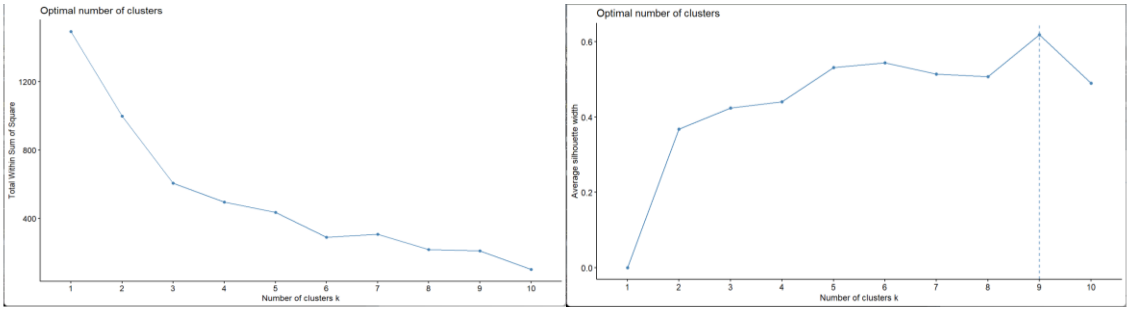


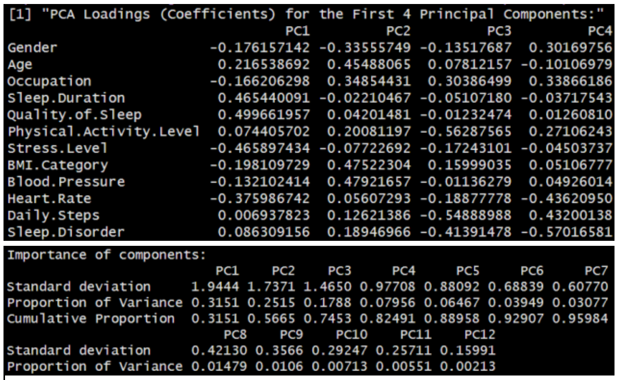
Figure 2: Determining Optimal Number of Clusters by Elbow Method and Silhouette Method

From Figure 2, choosing 6 clusters strikes a balance between the elbow method and silhouette method results. The elbow method suggests 3 clusters, which may be too simplistic and fail to capture finer details in the data. On the other hand, the silhouette method indicates 9 clusters, which could overcomplicate the model and risk overfitting. At 6 clusters, the rate of decrease in the within-cluster sum of squares (elbow curve) begins to level off, indicating a good trade-off between reducing variance and maintaining model simplicity. Additionally, 6 clusters has the second-highest average silhouette width of almost 0.6, which is a strong indicator of well-defined and distinct clusters. This choice ensures meaningful groupings while avoiding excessive fragmentation, making it a practical and interpretable solution for clustering analysis.

1.5.3 Hierarchical Clustering

Hierarchical clustering is an unsupervised learning technique that groups data into a hierarchy without predefining the number of clusters. Ward's method minimizes within-cluster variance, merging clusters in a way that results in the smallest increase in total variance. The final output is a dendrogram, a tree-like structure that visually represents the clustering process. Here, we use the 6 clusters identified by *K-Means clustering* to create a dendrogram, allowing us to compare both methods. This helps visualize hierarchical relationships between data points while ensuring consistency with K-Means.

1.6 Results



```
[1] "PCA Loadings (Coefficients) for the First 4 Principal Components:"
                                PC1         PC2         PC3         PC4
Gender                 -0.176157142 -0.33555749 -0.13517687  0.30169756
Age                     0.216538692  0.45488065  0.07812157 -0.10106979
Occupation             -0.166206298  0.34854431  0.30386499  0.33866186
Sleep.Duration          0.465440091 -0.02210467 -0.05107180 -0.03717543
Quality.of.Sleep        0.499661957  0.04201481 -0.01232474  0.01260810
Physical.Activity.Level  0.074405702  0.20081197 -0.56287565  0.27106243
Stress.Level           -0.465897434 -0.07722692 -0.17243101 -0.04503737
BMI.Category           -0.198109729  0.47522304  0.15999035  0.05106777
Blood.Pressure         -0.132102414  0.47921657 -0.01136279  0.04926014
Heart.Rate             -0.375986742  0.05607293 -0.18877778 -0.43620950
Daily.Steps             0.006937823  0.12621386 -0.54888988  0.43200138
Sleep.Disorder          0.086309156  0.18946966 -0.41391478 -0.57016581
```

```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7
Standard deviation     1.9444 1.7371 1.4650 0.97708 0.88092 0.68839 0.60770
Proportion of Variance 0.3151 0.2515 0.1788 0.07956 0.06467 0.03949 0.03077
Cumulative Proportion  0.3151 0.5665 0.7453 0.82491 0.88958 0.92907 0.95984
                           PC8    PC9    PC10    PC11    PC12
Standard deviation     0.42130 0.3566 0.29247 0.25711 0.15991
Proportion of Variance 0.01479 0.0106 0.00713 0.00551 0.00213
```

Figure 3: Principal Component Analysis

Based on Figure 3, the principal components can be named based on their heavily loaded variables: PC1: Sleep and Stress Health Index (Sleep Duration, Quality of Sleep, Stress Level), PC2: Biometric Health Profile (Age, BMI Category, Blood Pressure), PC3: Activity and Sleep Disorder Impact (Physical Activity Level, Daily Steps, Sleep Disorder), and PC4: Cardiovascular and Activity Link (Heart Rate, Daily Steps, Sleep Disorder). These names reflect the dominant variables that contribute significantly to each component, capturing key patterns such as sleep health, biometric metrics, physical activity, and cardiovascular interactions. The heavy loadings indicate these variables are central to the variance explained by each principal component. These 4 principal components account for 82.5% of the total variance.

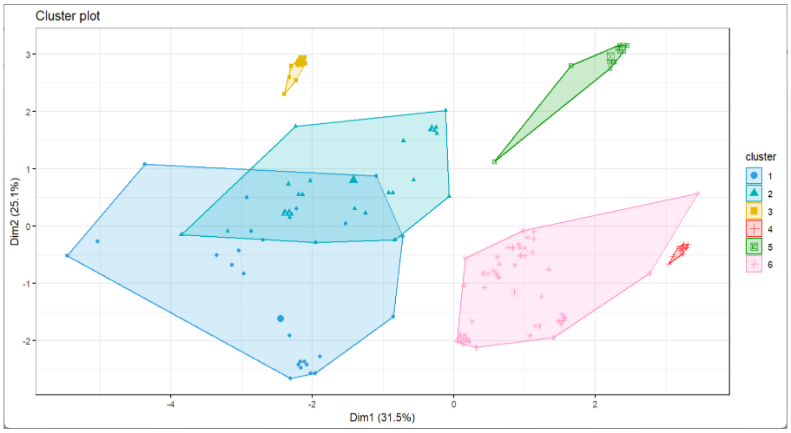


Figure 4: K-Means Cluster Plot

Figure 4 shows a K-Means Cluster Plot of 6 distinct clusters, as determined by the K-Means analysis. This visualization highlights the separation and grouping of data points into meaningful clusters, providing insights into the underlying structure of the dataset.

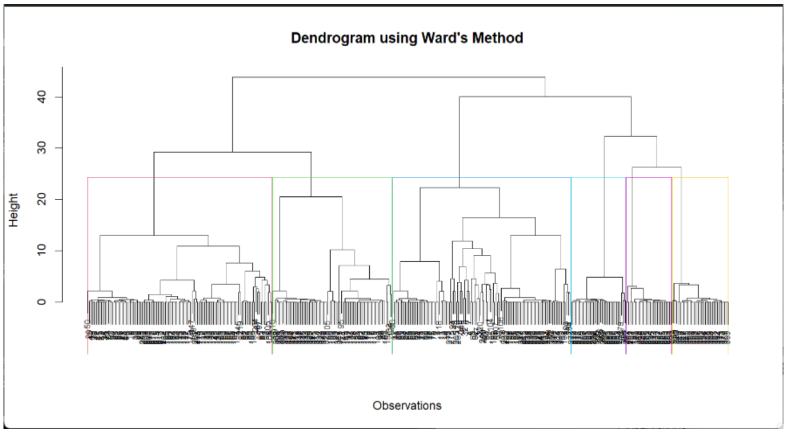


Figure 5: Hierarchical Cluster Plot

Figure 5 shows a dendrogram generated by Ward's method, which also depicts 6 distinct clusters. The dendrogram illustrates the hierarchical relationships and distances between observations, further validating the clustering results.

Thus, we addressed the research questions by grouping individuals into 6 optimal clusters using K-means and hierarchical clustering. The dendrogram validated the cluster structure, while PCA helped interpret the clusters by identifying key variables driving the patterns. This combined approach ensured meaningful and interpretable groupings, supported by robust statistical analysis.

Regression

2.1 Substantive Issue

A key challenge in medical insurance pricing is balancing affordability with risk assessment, as factors like age and health history can make premiums costly for high-risk individuals. This raises concerns about fairness and bias in pricing models. Regression analysis helps identify relationships between health issues and premium costs, allowing insurers to predict fair pricing based on historical data while minimizing bias and ensuring financial stability.

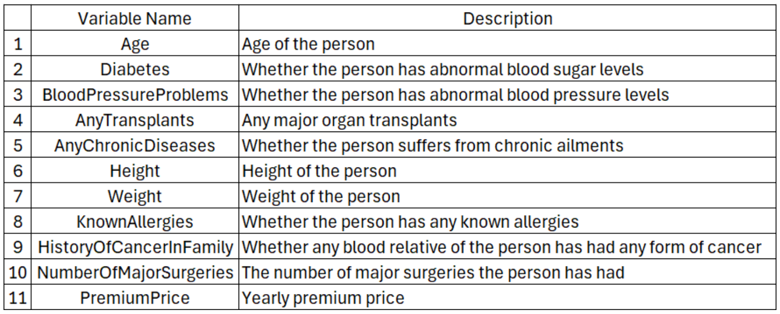
2.2 Research Questions

The research questions for the substantive issue are as follows:

- RQ1: How do demographic and health factors affect medical insurance premiums?
- RQ2: Can regression models help reduce biases in premium pricing for fairness?

2.3 Dataset & Variables

The dataset used for this regression analysis task is titled as “Medical Insurance Premium Prediction”. The dataset was created by collecting health-related information of almost 1000 customers. The dataset consists of 11 variables and 986 rows, representing individuals and their health-related parameters. A summary of the information collected is presented in the table below:



	Variable Name	Description
1	Age	Age of the person
2	Diabetes	Whether the person has abnormal blood sugar levels
3	BloodPressureProblems	Whether the person has abnormal blood pressure levels
4	AnyTransplants	Any major organ transplants
5	AnyChronicDiseases	Whether the person suffers from chronic ailments
6	Height	Height of the person
7	Weight	Weight of the person
8	KnownAllergies	Whether the person has any known allergies
9	HistoryOfCancerInFamily	Whether any blood relative of the person has had any form of cancer
10	NumberOfMajorSurgeries	The number of major surgeries the person has had
11	PremiumPrice	Yearly premium price

Table 2: Description of Variables in Medical Insurance Premium Dataset

We calculate BMI from height and weight, then remove the original height and weight columns to reduce multicollinearity and simplify the regression model.

2.4 Methodology

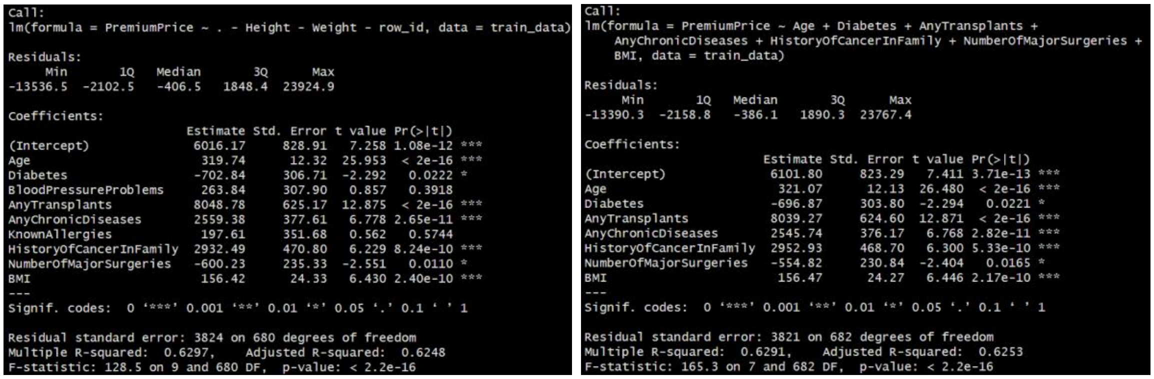
This regression task predicts medical insurance premiums using Linear Regression, CART, and Random Forest. Models will be evaluated using R^2, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) to assess accuracy and error. The dataset will be split into 70% training and 30% testing. R^2 measures variance explained, MAE calculates average prediction error, and RMSE highlights large errors. Comparing these metrics will help identify the best-performing model.

2.5 Analysis

2.5.1 Linear Regression

Linear regression models the relationship between a dependent variable and one or more independent variables by fitting a linear equation. It predicts outcomes by minimizing errors, helping identify trends and the impact of factors on the target variable.

We apply stepwise regression which automatically selects the optimal set of predictors by iteratively removing variables based on Akaike Information Criterion (AIC). This process improves the model by eliminating irrelevant or redundant predictors, enhancing efficiency and accuracy.



```
Call:
lm(formula = PremiumPrice ~ . - Height - Weight - row_id, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-13536.5  -2102.5   -406.5   1848.4  23924.9 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)              6016.17     828.91   7.258 1.08e-12 ***
Age                       319.74      12.32  25.953  < 2e-16 ***
Diabetes                 -702.84     306.71  -2.292   0.0222 *  
BloodPressureProblems     263.84     307.90   0.857   0.3918    
AnyTransplants           8048.78     625.17  12.875  < 2e-16 ***
AnyChronicDiseases       2559.38     377.61   6.778 2.65e-11 ***
KnownAllergies            197.61     351.68   0.562   0.5744    
HistoryOfCancerInFamily  2932.49     470.80   6.229 8.24e-10 ***
NumberOfMajorSurgeries   -600.23     235.33  -2.551   0.0110 *  
BMI                       156.42      24.33   6.430 2.40e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3824 on 680 degrees of freedom
Multiple R-squared:  0.6297,	Adjusted R-squared:  0.6248 
F-statistic: 128.5 on 9 and 680 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = PremiumPrice ~ Age + Diabetes + AnyTransplants + 
    AnyChronicDiseases + HistoryOfCancerInFamily + NumberOfMajorSurgeries + 
    BMI, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-13390.3  -2158.8   -386.1   1890.3  23767.4 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)              6101.80     823.29   7.411 3.71e-13 ***
Age                       321.07      12.13  26.480  < 2e-16 ***
Diabetes                 -696.87     303.80  -2.294   0.0221 *  
AnyTransplants           8039.27     624.60  12.871  < 2e-16 ***
AnyChronicDiseases       2545.74     376.17   6.768 2.82e-11 ***
HistoryOfCancerInFamily  2952.93     468.70   6.300 5.33e-10 ***
NumberOfMajorSurgeries   -554.82     230.84  -2.404   0.0165 *  
BMI                       156.47      24.27   6.446 2.17e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3821 on 682 degrees of freedom
Multiple R-squared:  0.6291,	Adjusted R-squared:  0.6253 
F-statistic: 165.3 on 7 and 682 DF,  p-value: < 2.2e-16
```

Figure 6: Coefficients before (left) and after (right) Stepwise Regression

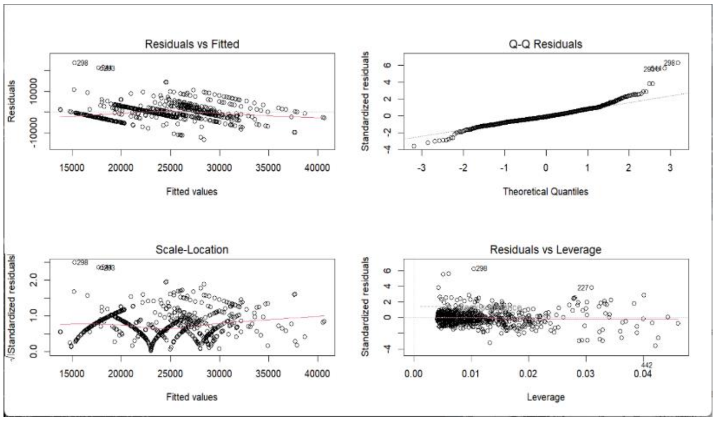


Figure 7: Diagnostic Plots

We then generate diagnostic plots to assess model validity. **Residuals vs. Fitted** reveals a slight curve, suggesting non-linearity in the model. **Q-Q Plot** shows deviations in the tails, indicating potential outliers affecting normality. **Scale-Location Plot** suggests heteroscedasticity, as residuals show uneven variance. **Residuals vs. Leverage** highlights influential points that could distort model predictions.

2.5.2 Classification and Regression Trees (CART)

CART is a decision tree algorithm for classification and regression. It splits data into subsets based on feature values, creating a tree-like structure where each leaf node represents a predicted outcome. In regression, CART predicts continuous values by dividing the data to minimize variance within each subset.

The tree splits data based on key health-related factors. Each node represents a decision rule, leading to branches that further refine predictions. The final leaf nodes indicate predicted premium values, with darker shades representing higher premiums. The tree structure helps insurers assess risk and determine pricing based on historical data.

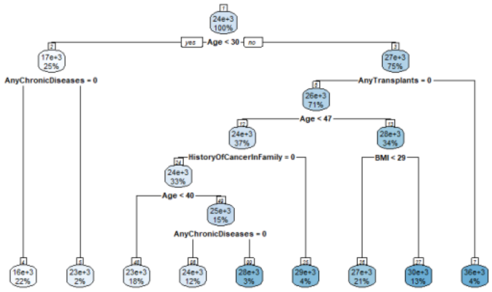


Figure 8: Optimal Trees

2.5.3 Random Forest

A regression tree is a decision tree used for predicting continuous numerical values. It splits data into subsets based on input features, minimizing variance within each group. Each leaf node represents a predicted value, making it useful for modelling complex relationships and identifying key factors influencing outcomes.

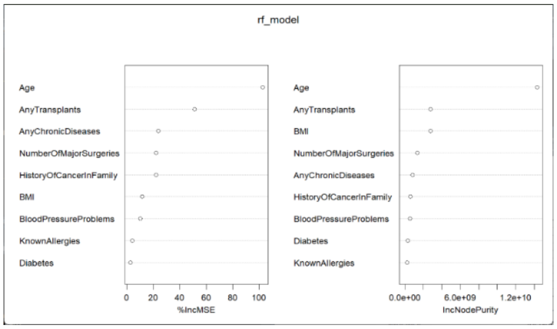


Figure 9: Variable Importance in RF Model

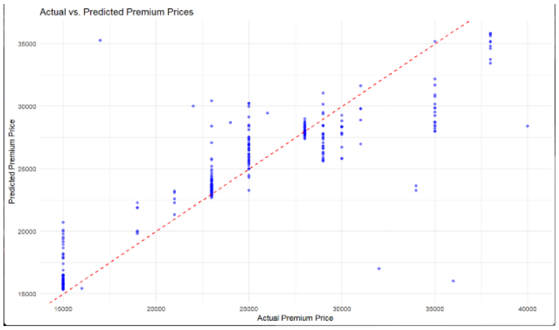
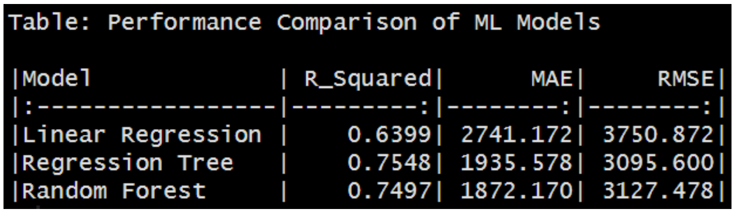


Figure 10: Actual vs Predicted Plot

The variable importance plot highlights Age and AnyTransplants as key predictors, while KnownAllergies and BloodPressureProblems have minimal impact, aligning with the linear regression model. The Actual vs. Predicted Plot shows a general trend match but with some deviations, suggesting room for model improvement.

2.6 Results



```
Table: Performance Comparison of ML Models

|Model              | R_Squared|      MAE|     RMSE|
|:------------------|---------:|--------:|--------:|
|Linear Regression  |    0.6399| 2741.172| 3750.872|
|Regression Tree    |    0.7548| 1935.578| 3095.600|
|Random Forest      |    0.7497| 1872.170| 3127.478|
```

Figure 11: Models Performance Results

Based on the performance comparison, the Regression Tree is the best model, achieving the highest R^2 (75.48%) and lowest RMSE (3095.600), indicating superior accuracy and consistency. While the Random Forest has a slightly lower MAE (1872.170 vs. 1935.578), the difference is marginal, and the Regression Tree's stronger overall metrics make it the optimal choice. Thus, the Regression Tree outperforms both Linear Regression and Random Forest for this task.

Classification

3.1 Substantive Issue

A major global issue related to this dataset is educational inequality, where factors like socioeconomic status, access to resources, and parental support create disparities in student performance. By analysing this dataset, researchers can identify patterns and predictors of academic success, helping to develop targeted interventions, improve educational policies, and promote equal learning opportunities for all students.

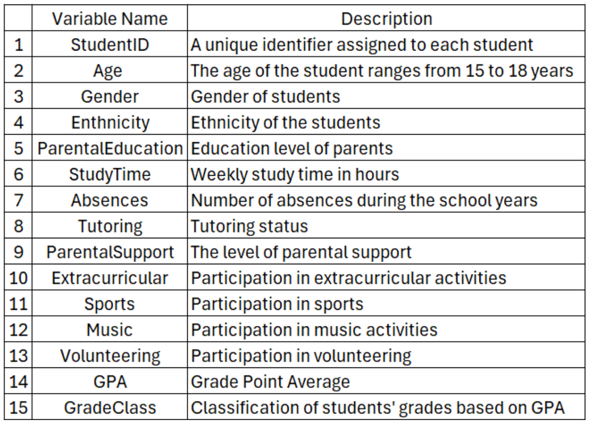
3.2 Research Questions

The research questions for the substantive issue are as follows:

- RQ1: What factors most influence students' grade classification?
- RQ2: Can machine learning accurately predict grade classification?

3.3 Dataset & Variables

The dataset used for this classification analysis task is titled as “Students Performance Dataset”. The dataset was created by collecting information of over 2000 students. The dataset consists of 15 variables and 2392 rows, including various factors influencing their academic performance. A summary of the information collected is presented in the table below:



	Variable Name	Description
1	StudentID	A unique identifier assigned to each student
2	Age	The age of the student ranges from 15 to 18 years
3	Gender	Gender of students
4	Enthnicity	Ethnicity of the students
5	ParentalEducation	Education level of parents
6	StudyTime	Weekly study time in hours
7	Absences	Number of absences during the school years
8	Tutoring	Tutoring status
9	ParentalSupport	The level of parental support
10	Extracurricular	Participation in extracurricular activities
11	Sports	Participation in sports
12	Music	Participation in music activities
13	Volunteering	Participation in volunteering
14	GPA	Grade Point Average
15	GradeClass	Classification of students' grades based on GPA

Table 3: Description of Variables in Students Performance Dataset

We will remove the “StudentID” column as it does not contribute to predicting student performance.

3.4 Methodology

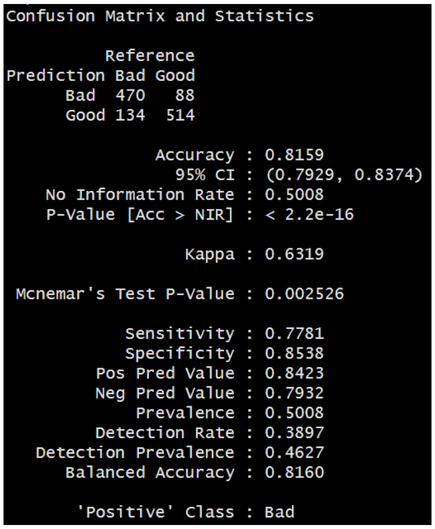
This classification task predicts student performance (GradeClass) using logistic regression, SVM, and neural networks, evaluated by confusion matrix, accuracy, and prediction error. We convert **GradeClass** into a binary variable ("Good" for 0 and 1, "Bad" otherwise) and change it to a factor. To address class imbalance, we apply oversampling with the ROSE package for a more balanced dataset. Finally, the data is split into a 70-30 train-test set for model training and evaluation.

3.5 Analysis

3.5.1 Logistic Regression

Logistic regression is a statistical method used for binary classification, where the goal is to predict the probability of an event occurring based on one or more predictor variables. It models the relationship between the input features and the binary outcome using a logistic function, producing values between 0 and 1.

The logistic regression model achieves 81.6% accuracy, correctly classifying 470 "Bad" and 514 "Good" cases. It shows stronger performance in identifying positive cases (85.4% specificity) than detecting negatives (77.8% sensitivity), with more false negatives (134) than false positives (88). The substantial Kappa score (0.632) and balanced accuracy (81.6%) demonstrate reliable performance, though the significant McNemar's p-value (0.0025) suggests uneven error distribution. This makes the model particularly suitable for applications where correctly identifying "Good" outcomes is prioritized, while still maintaining reasonable detection of "Bad" cases.



```
Confusion Matrix and Statistics

          Reference
Prediction Bad Good
      Bad  470   88
      Good 134  514

               Accuracy : 0.8159
                 95% CI : (0.7929, 0.8374)
    No Information Rate : 0.5008
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6319

 Mcnemar's Test P-Value : 0.002526

            Sensitivity : 0.7781
            Specificity : 0.8538
         Pos Pred Value : 0.8423
         Neg Pred Value : 0.7932
             Prevalence : 0.5008
         Detection Rate : 0.3897
   Detection Prevalence : 0.4627
      Balanced Accuracy : 0.8160

       'Positive' Class : Bad
```

Figure 12: Confusion Matrix and Statistics for Logistic Regression

3.5.2 Support Vector Machine

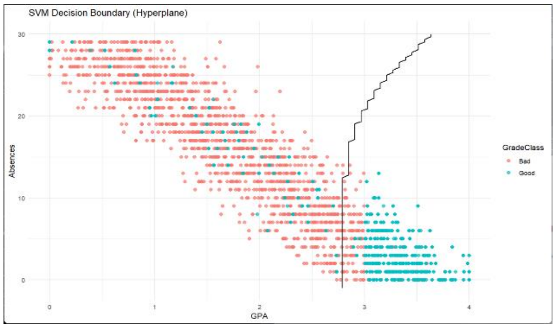
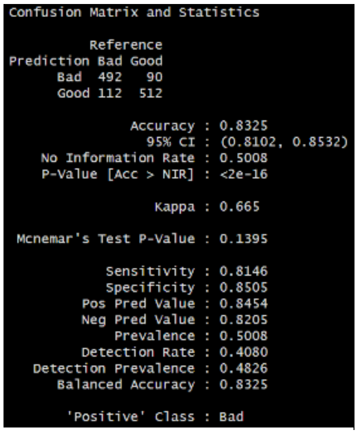


Figure 13: SVM Hyperplane

Support Vector Machine (SVM) is a supervised learning algorithm used for classification tasks. It finds the hyperplane that best separates data into different classes by maximizing the margin between them, making it effective for both linear and non-linear classification problems.

This plot shows an SVM decision boundary for predicting GradeClass using GPA and Absences, the most influential features. The boundary separates "Good" and "Bad" grades, highlighting that higher GPAs and fewer absences correlate with better performance. While some misclassifications exist, the model captures the overall trend well.

The SVM model achieves 83.25% accuracy, correctly identifying 492 "Bad" and 512 "Good" cases. It shows balanced performance with 81.5% sensitivity and 85.1% specificity, demonstrating consistent reliability across both classes. The model's strong precision (84.5%) for "Bad" predictions and substantial Kappa score (0.665) confirm its effectiveness for applications requiring unbiased classification of both outcomes.



```
Confusion Matrix and Statistics

          Reference
Prediction Bad Good
      Bad  492   90
      Good 112  512

               Accuracy : 0.8325
                 95% CI : (0.8102, 0.8532)
    No Information Rate : 0.5008
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.665

 Mcnemar's Test P-Value : 0.1395

            Sensitivity : 0.8146
            Specificity : 0.8505
         Pos Pred Value : 0.8454
         Neg Pred Value : 0.8205
             Prevalence : 0.5008
         Detection Rate : 0.4080
   Detection Prevalence : 0.4826
      Balanced Accuracy : 0.8325

       'Positive' Class : Bad
```

Figure 14: Confusion Matrix and Statistics for SVM

3.5.3 Neural Networks

Neural networks are computational models used for classification and regression tasks by learning complex patterns in data through training, adjusting weights based on error to improve predictions.

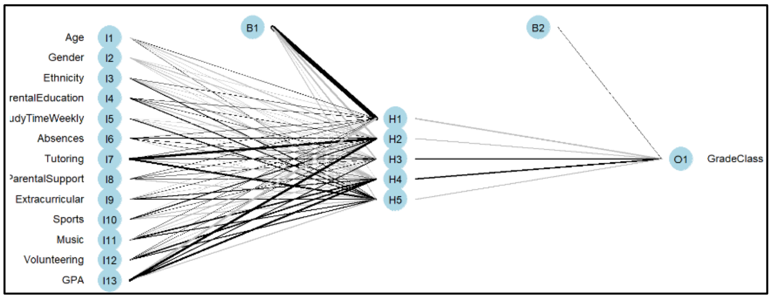
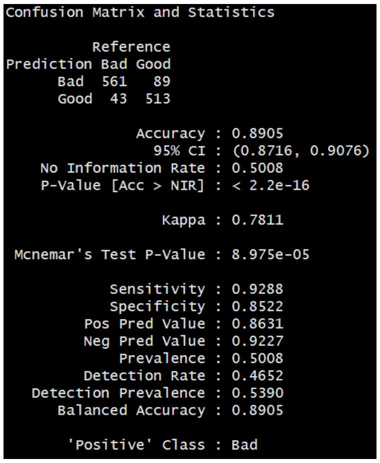


Figure 15: Neural Network Diagram

This neural network diagram is designed to predict student academic performance, as indicated by the GradeClass output node. The architecture consists of 13 input nodes representing various student attributes, feeding into two hidden layers: the first with 12 neurons (B1) and the second with 5 neurons (B2). This structure suggests a feedforward neural network, where data flows sequentially from input to output, with the hidden layers extracting patterns to map features to the target variable.

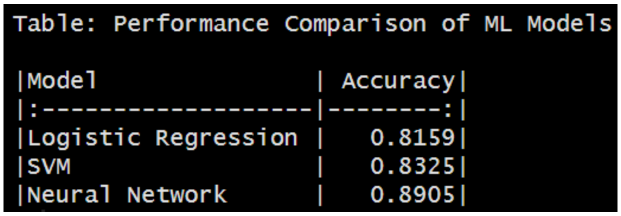


```
Confusion Matrix and Statistics

          Reference
Prediction Bad Good
      Bad  561   89
      Good  43  513

               Accuracy : 0.8905
                 95% CI : (0.8716, 0.9076)
    No Information Rate : 0.5008
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7811

 Mcnemar's Test P-Value : 8.975e-05

            Sensitivity : 0.9288
            Specificity : 0.8522
         Pos Pred Value : 0.8631
         Neg Pred Value : 0.9227
             Prevalence : 0.5008
         Detection Rate : 0.4652
   Detection Prevalence : 0.5390
      Balanced Accuracy : 0.8905

       'Positive' Class : Bad
```

Figure 16: Confusion Matrix and Statistics for NN

The neural network demonstrates excellent performance with 89.05% accuracy, correctly classifying 561 "Bad" cases and 513 "Good" cases. It shows particularly strong detection of negative cases (92.9% sensitivity) while maintaining solid identification of positives (85.2% specificity). The model makes marginally more false positives (89) than false negatives (43), indicating slightly better performance for "Bad" predictions. With a high Kappa score of 0.781 and statistically significant results (p < 2.2e-16), making it reliable for applications prioritizing "Bad" case detection.

3.6 Results



```
Table: Performance Comparison of ML Models

|Model               | Accuracy|
|:-------------------|--------:|
|Logistic Regression |   0.8159|
|SVM                 |   0.8325|
|Neural Network      |   0.8905|
```

Figure 17: Models Performance Results

For this classification task, the neural network is the best-performing model, achieving 89.05% accuracy—significantly higher than both SVM (83.25%) and logistic regression (81.59%). Its superior balance of sensitivity (92.9% for "Bad") and specificity (85.2% for "Good"), along with the fewest errors (89 FP, 43 FN) and highest Kappa score (0.781), demonstrates its robust ability to handle the dataset's complexity.

Appendix

1. Unsupervised Learning: Sleep Health and Lifestyle Dataset
   https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset
2. Regression: Medical Insurance Premium Prediction
   https://www.kaggle.com/datasets/tejashvi14/medical-insurance-premium-prediction
3. Classification: Students Performance Dataset
   https://www.kaggle.com/datasets/rabieelkharoua/students-performance-dataset