Predicting Student Attendance Based on Stress, Mood, and Other Features

Kelvin

Executive Summary

The goal of this analysis is to forecast student attendance ("Present" vs. "Absent") based on several behavioral and psychological characteristics, such as mood score, anxiety level, stress level, and sleep duration. With an equal number of "Present" and "Absent" observations, the data is balanced. This project aims to investigate the relationships between these variables and attendance, as well as to develop predictive models using machine learning techniques. To predict attendance status, key techniques included data cleansing, exploratory data analysis, and fitting logistic regression and k-nearest neighbors models.

Data Description and Variables

The "Student Monitoring" dataset is available on Kaggle and can be downloaded from this link.

The dataset contains 15,000 records and has 6 fields/attributes, with "Attendance.Status" as the target variable. It includes several variables related to student behavior, such as:

- Student.ID: A unique identifier for each student.
- Attendance.Status: Indicates whether a student was "Present" or "Absent" for a particular class.
- Stress.Level..GSR.: The student's stress level based on Galvanic Skin Response (GSR).
- Sleep.Hours: The number of hours of sleep the student had.
- Anxiety.Level: The self-reported anxiety level of the student.
- Mood.Score: The self-reported mood score of the student. The goal of this analysis is to predict whether a student will attend class based on these features, using machine learning models.

Methods and Analysis

Data Cleaning and Transformation

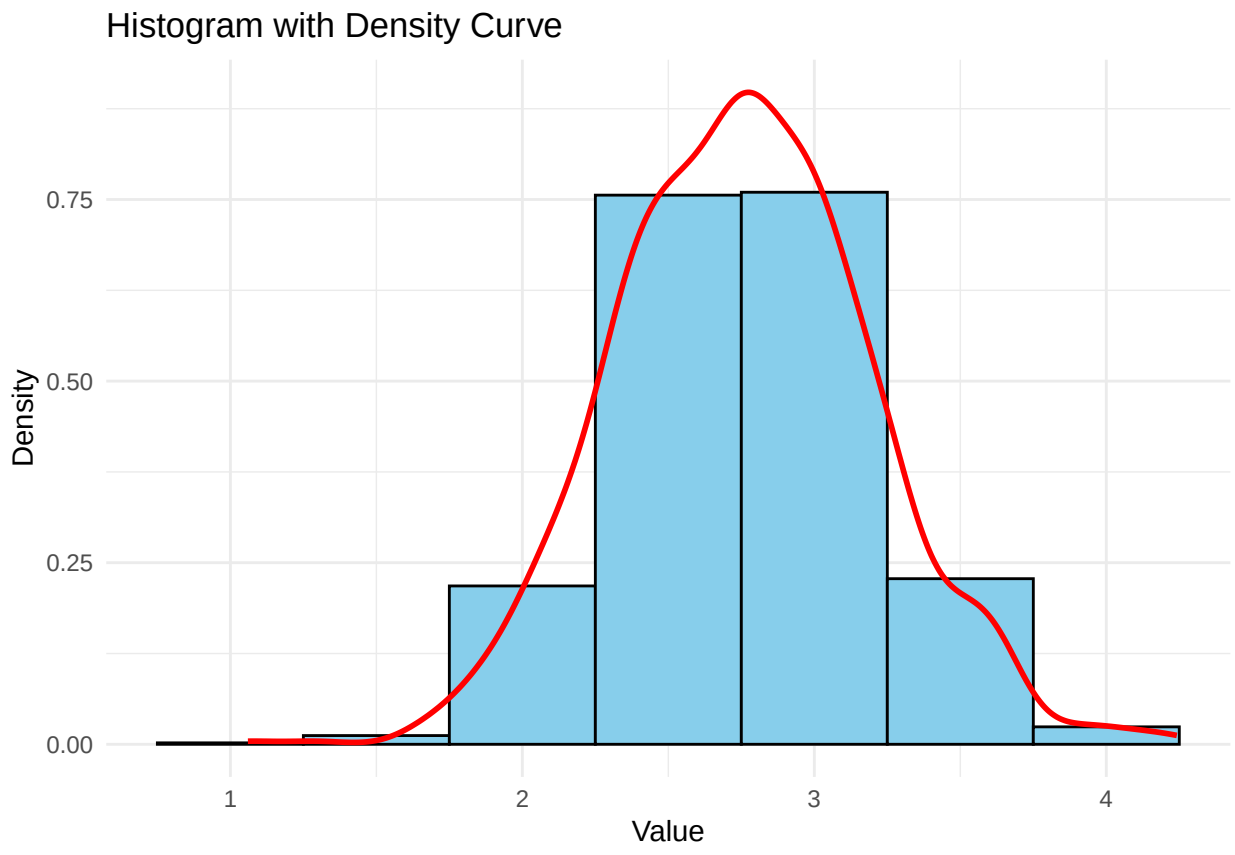
First, we ensure that the `Attendance.Status` and `Risk.Level` are properly converted to factors. The data is summarized by student ID, where we compute the mean values for stress, anxiety, sleep hours, and mood scores. After this transformation, we exclude any students who were marked as "Late," as we focus on "Present" vs. "Absent" classification.

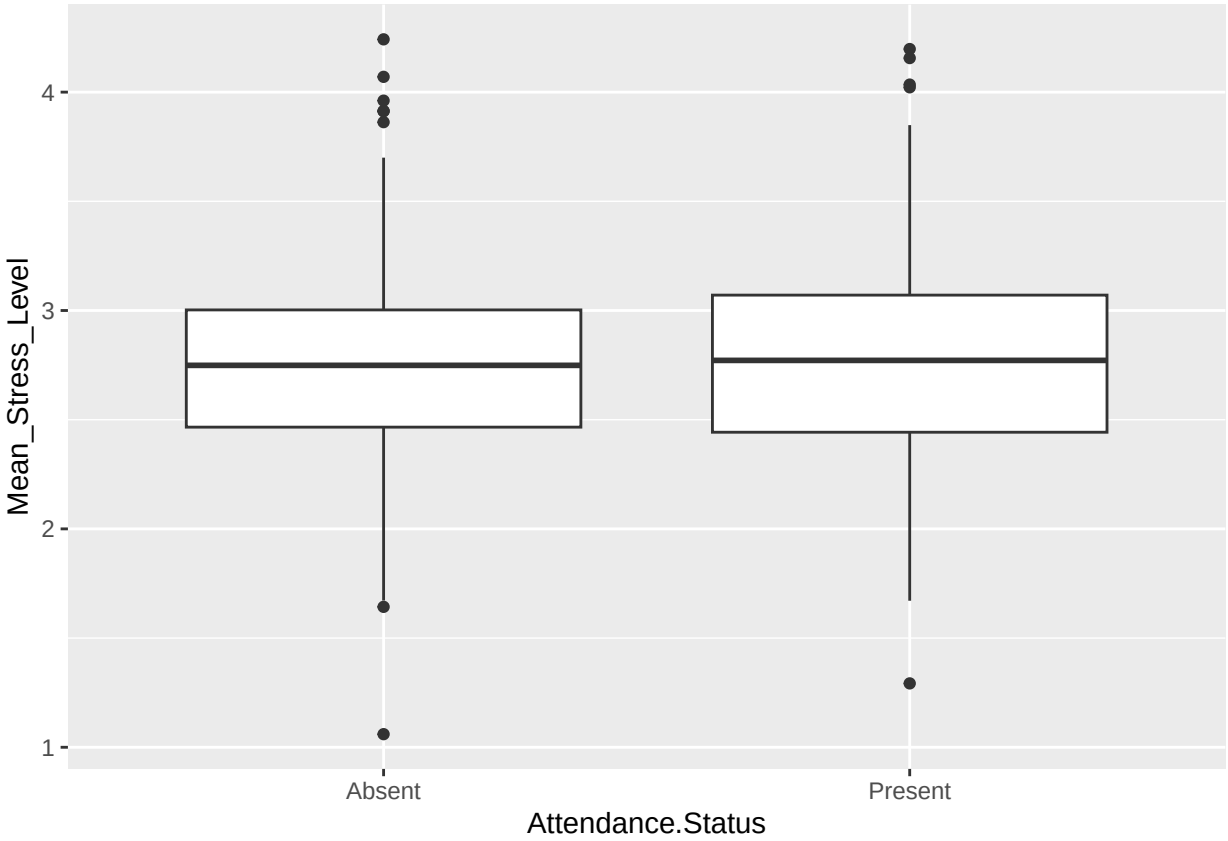
```
## 'data.frame':    1000 obs. of  6 variables:
##  $ Student.ID        : int  1 1 2 2 3 3 4 4 5 5 ...
##  $ Attendance.Status : Factor w/ 2 levels "Absent","Present": 1 2 1 2 1 2 1 2 1 2 ...
##  $ Mean_Stress_Level : num  3.3 3.1 2.72 2.37 3.08 ...
##  $ Mean_Sleep_Hours  : num  7.2 6.61 6.98 7.17 7.06 ...
##  $ Mean_Anxiety_Level: num  7 4.91 6.7 5 5.25 ...
##  $ Mean_Mood_Score   : num  5.88 6.55 6.4 6.78 4.38 ...
```

Exploratory Data Analysis

We visualized each feature to explore its distribution and relationship with attendance status. Here are the key findings:

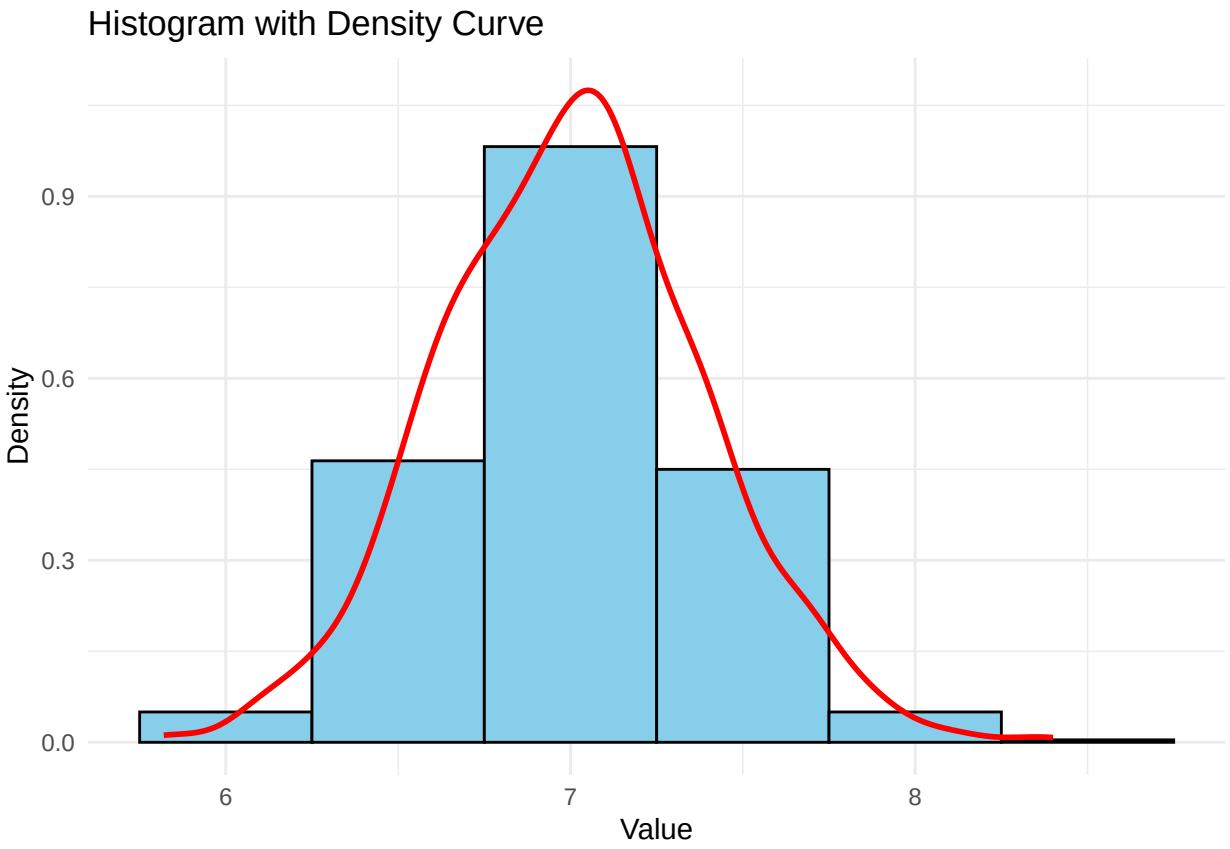
stress level is normally distributed as can be seen from the shape of a bell curve below

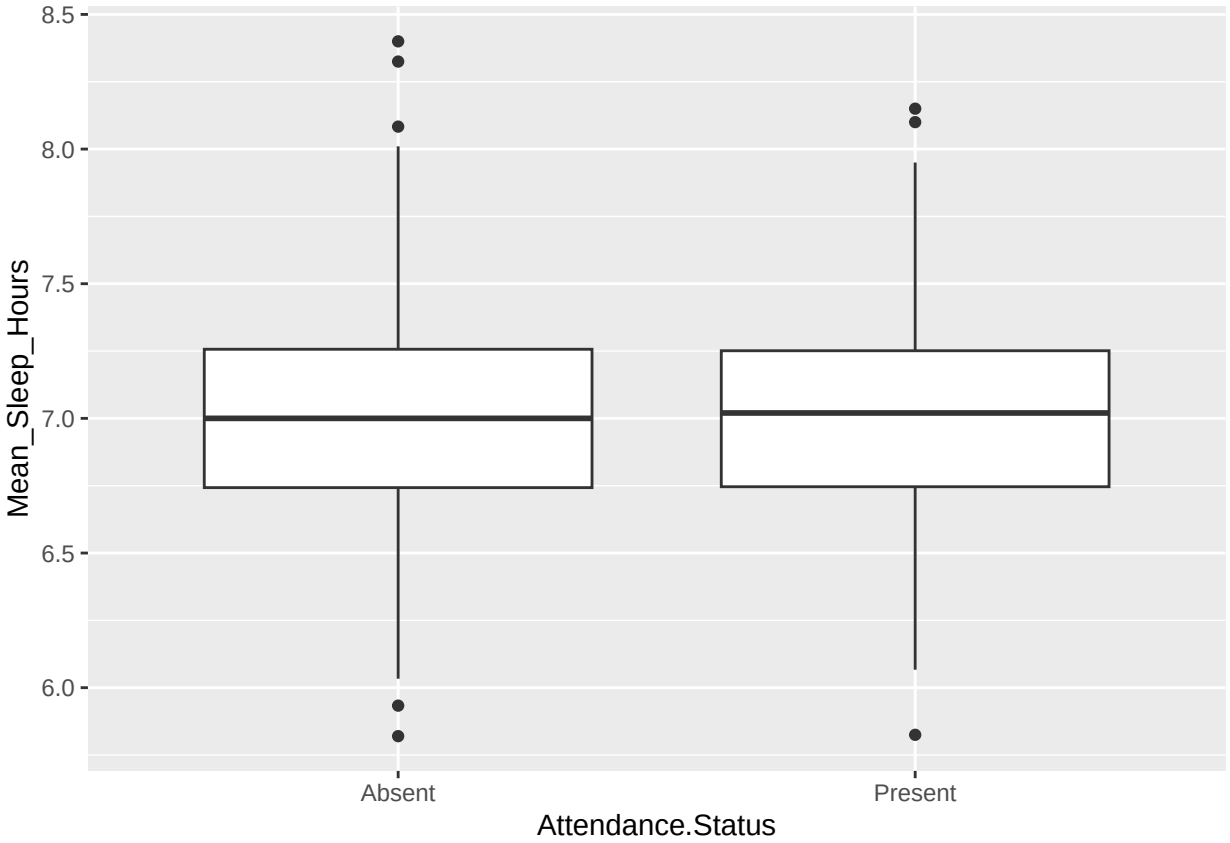




Stress Level: Students who are present tend to have slightly higher stress levels than those who are absent. When students are present, they experience slightly higher stress than when they are absent, which could indicate accumulated stress from additional responsibilities at school. In contrast, when students are absent, they may not be participating in these stressful situations, and as a result, their stress levels may not reflect the same intensity.

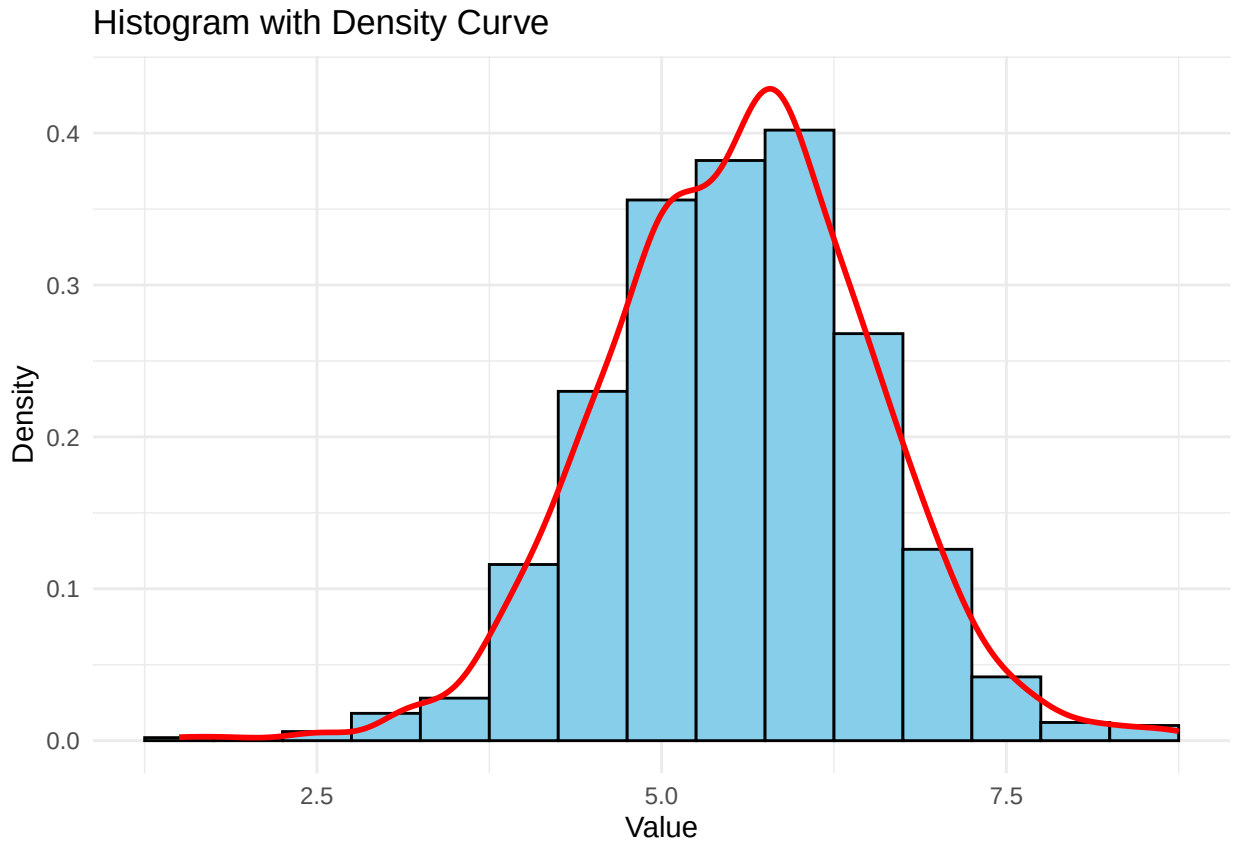
Sleep hours are normally distributed, as can be seen from the bell curve shape below.

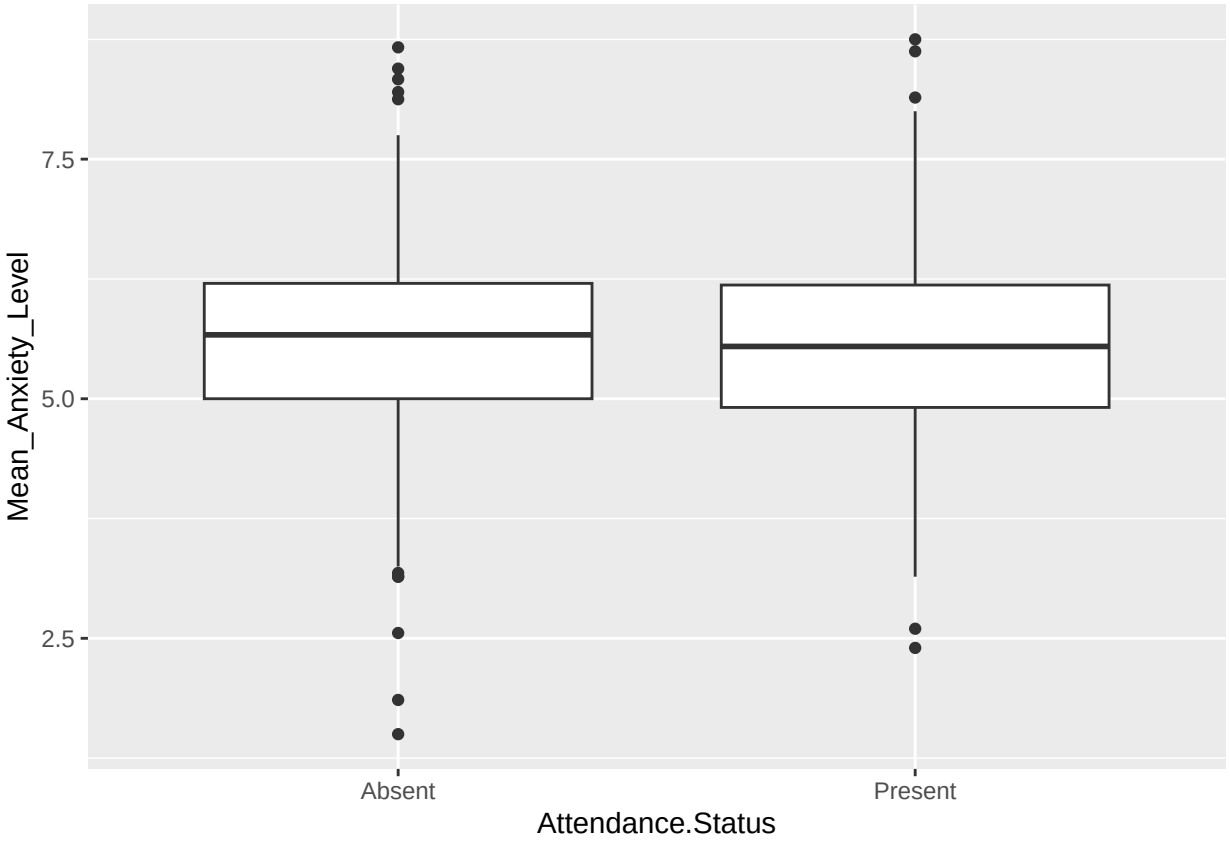

Histogram with Density Curve
Density
0.9
0.6
0.3
0.0
6
7
8
Value



Sleep Hours: Present students tend to have slightly better sleep compared to those who are absent.

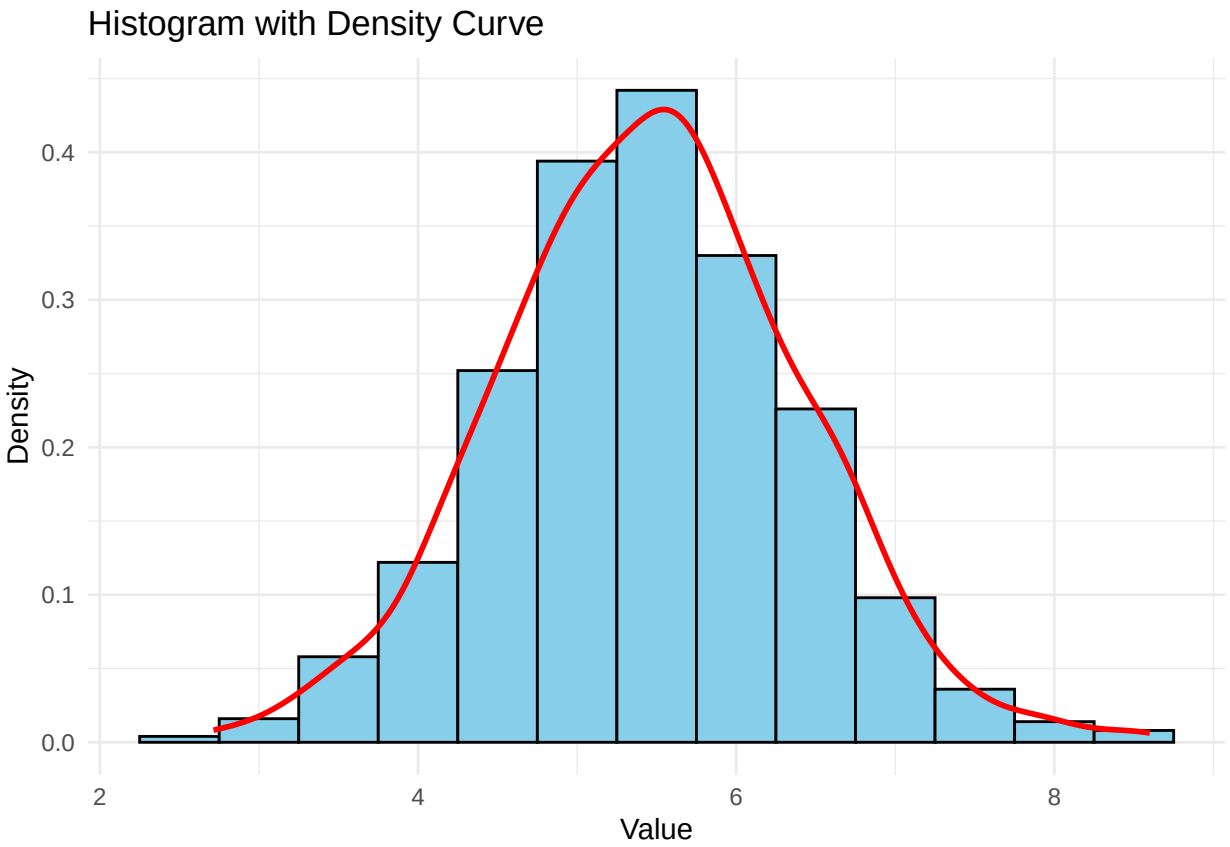
Anxiety level is normally distributed as can be seen from the shape of a bell curve shape below

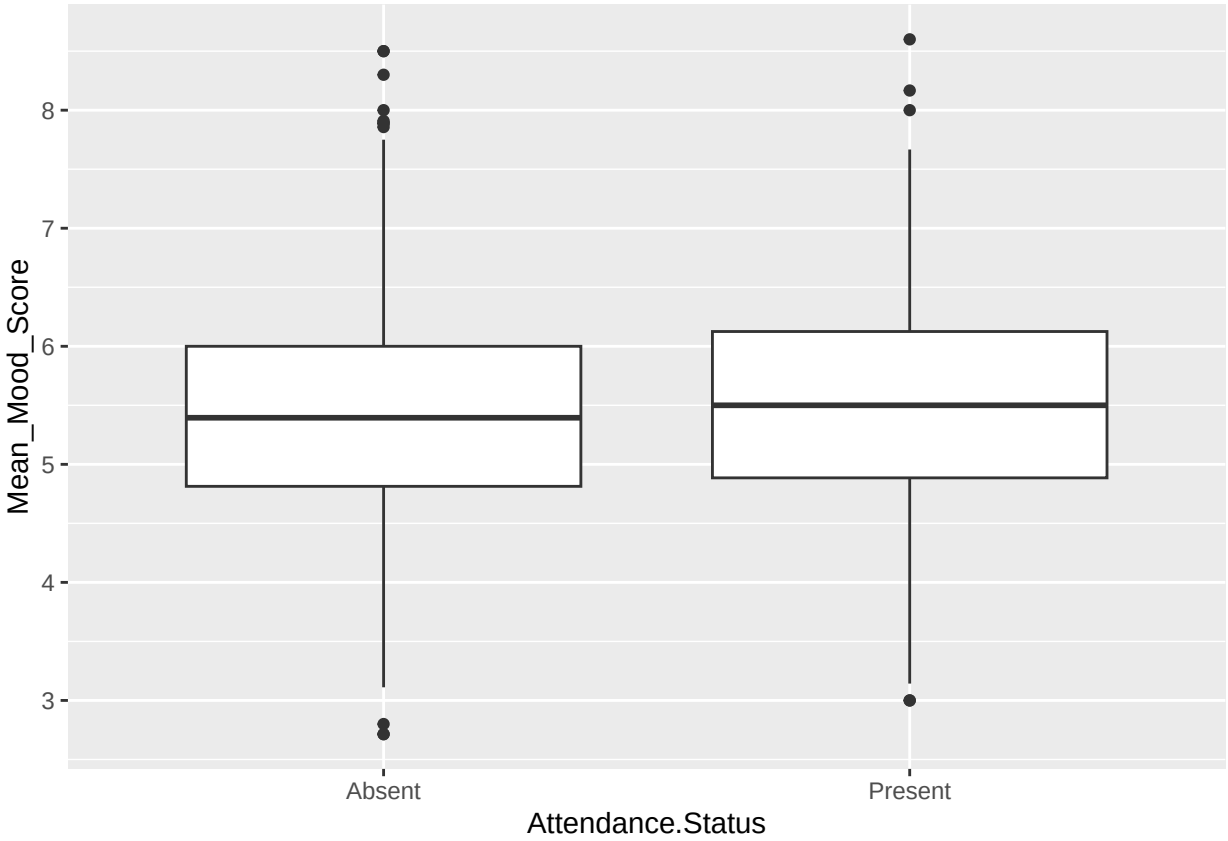




Anxiety Level: Students who are absent report higher anxiety levels. When students are present, they have lower anxiety levels than when they are absent, suggesting that absence might allow students to avoid situations that trigger their anxiety. Their absence could reflect anxiety about school, with being absent providing temporary relief. On the other hand, when students are present, they may be actively confronting and managing their anxiety through exposure, social interaction, and academic engagement, all of which can help reduce anxiety.

Mood score is normally distributed as can be seen from the shape of a bell curve shape below


Histogram with Density Curve
Density
0.0
0.1
0.2
0.3
0.4
2
4
6
8
Value



Mood Score: Present students have higher mood scores compared to absent students. The higher mood in students when they are present, compared to when they are absent, suggests that school attendance is associated with a more positive emotional state. This is likely due to social engagement, routine, academic participation, and the opportunity for positive emotional experiences at school. In contrast, when students are absent, they may experience a lower mood due to isolation, avoidance, or struggles with emotional or psychological challenges, which could prevent them from benefiting from the mood-regulating advantages of school life.

###Correlation between the features

```
##                    Mean_Stress_Level Mean_Sleep_Hours Mean_Anxiety_Level
## Mean_Stress_Level         1.000000000       0.01783065       -0.004980809
## Mean_Sleep_Hours          0.017830647       1.00000000       -0.025283639
## Mean_Anxiety_Level       -0.004980809      -0.02528364        1.000000000
## Mean_Mood_Score           0.041163250      -0.02708605       -0.029388165
##                    Mean_Mood_Score
## Mean_Stress_Level       0.04116325
## Mean_Sleep_Hours       -0.02708605
## Mean_Anxiety_Level     -0.02938817
## Mean_Mood_Score         1.00000000
```

The correlations between these features are generally weak, indicating that stress levels, sleep hours, anxiety levels, and mood scores do not show strong relationships with one another. These variables may be influenced by other factors not captured in this dataset, or there may be complexities (e.g., non-linear relationships) that aren't captured by simple correlation analysis.

Mood shows a slight positive correlation with stress (correlation of 0.0412). Anxiety and sleep hours exhibit

very weak negative relationships with mood and anxiety, respectively, but these trends are not strong enough to be considered highly significant

###correlation of features with attendance status

```
cor(data_summary %>% mutate(Attendance.Status = ifelse(Attendance.Status == "Absent", 0, 1)) %>%
                           select(-Student.ID))
```

```
##                    Attendance.Status Mean_Stress_Level Mean_Sleep_Hours
## Attendance.Status         1.000000000       0.020574693      0.005113742
## Mean_Stress_Level         0.020574693       1.000000000      0.017830647
## Mean_Sleep_Hours          0.005113742       0.017830647      1.000000000
## Mean_Anxiety_Level       -0.029031537      -0.004980809     -0.025283639
## Mean_Mood_Score           0.042221663       0.041163250     -0.027086052
##                    Mean_Anxiety_Level Mean_Mood_Score
## Attendance.Status        -0.029031537      0.04222166
## Mean_Stress_Level        -0.004980809      0.04116325
## Mean_Sleep_Hours         -0.025283639     -0.02708605
## Mean_Anxiety_Level        1.000000000     -0.02938817
## Mean_Mood_Score          -0.029388165      1.00000000
```

There is a slight tendency for attendance to be positively correlated with mood and stress, but these correlations are minimal.

Machine learning

Guessing model

We will build a guessing model to compare it against the other two models we will create, namely the logistic regression model and k-Nearest Neighbors with cross-validation. First, we will create a training and test set using the createDataPartition function. Then, we will simulate random guessing and evaluate its performance.

```
## [1] 0.5166667
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Absent Present
##    Absent      77      72
##    Present     73      78
##
##                Accuracy : 0.5167
##                  95% CI : (0.4585, 0.5745)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : 0.3017
##
##                   Kappa : 0.0333
##
##  Mcnemar's Test P-Value : 1.0000
##
##             Sensitivity : 0.5133
##             Specificity : 0.5200
##          Pos Pred Value : 0.5168
##          Neg Pred Value : 0.5166
```

```
##              Prevalence : 0.5000
##          Detection Rate : 0.2567
##    Detection Prevalence : 0.4967
##       Balanced Accuracy : 0.5167
##
##        'Positive' Class : Absent
##
```

```
## RMSE  0.6952218
```

rmse is far from zero meaning the predicted values are not close to the true values which is expected since we are guessing

The next step is understanding which feature can generate the highest accuracy in predicting attendance status

```
## 'data.frame':    700 obs. of  6 variables:
##  $ Student.ID        : int  1 2 4 5 6 7 7 8 8 9 ...
##  $ Attendance.Status : Factor w/ 2 levels "Absent","Present": 2 1 2 2 2 1 2 1 2 1 ...
##  $ Mean_Stress_Level : num  3.1 2.72 2.8 2.59 2.09 ...
##  $ Mean_Sleep_Hours  : num  6.61 6.98 7.08 6.86 6.3 ...
##  $ Mean_Anxiety_Level: num  4.91 6.7 5.17 4.92 6.57 ...
##  $ Mean_Mood_Score   : num  6.55 6.4 4.42 4.25 4.71 ...
```

```
##  Mean_Stress_Level   Mean_Sleep_Hours Mean_Anxiety_Level    Mean_Mood_Score
##          0.5200000          0.5171429          0.5328571          0.5500000
```

There is no feature that predicts better than random guessing, as their accuracy falls within the confidence interval of the guessing model, which ranges from 0.46 to 0.57. The best feature among all the features is the mood score.

First model will be logistic regression

The features have extremely weak correlations with the attendance status but the model might still find some patterns in the data.

first step will be without using the training function in the caret package

```
##
## Call:
## glm(formula = Attendance.Status ~ Mean_Mood_Score + Mean_Anxiety_Level +
##     Mean_Stress_Level + Mean_Sleep_Hours, family = binomial,
##     data = train_set)
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.912113   1.609196  -0.567   0.5708
## Mean_Mood_Score     0.164951   0.080390   2.052   0.0402 *
## Mean_Anxiety_Level -0.038487   0.079117  -0.486   0.6266
## Mean_Stress_Level   0.062694   0.171039   0.367   0.7140
## Mean_Sleep_Hours    0.007644   0.199533   0.038   0.9694
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 970.41  on 699  degrees of freedom
## Residual deviance: 965.67  on 695  degrees of freedom
## AIC: 975.67
##
## Number of Fisher Scoring iterations: 3
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Absent Present
##    Absent      75      72
##    Present     75      78
##
##                Accuracy : 0.51
##                  95% CI : (0.4519, 0.5679)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : 0.3864
##
##                   Kappa : 0.02
##
##  Mcnemar's Test P-Value : 0.8690
##
##             Sensitivity : 0.5000
##             Specificity : 0.5200
##          Pos Pred Value : 0.5102
##          Neg Pred Value : 0.5098
##              Prevalence : 0.5000
##          Detection Rate : 0.2500
##    Detection Prevalence : 0.4900
##       Balanced Accuracy : 0.5100
##
##        'Positive' Class : Absent
##
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Absent Present
##    Absent      77      72
##    Present     73      78
##
##                Accuracy : 0.5167
##                  95% CI : (0.4585, 0.5745)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : 0.3017
##
##                   Kappa : 0.0333
##
##  Mcnemar's Test P-Value : 1.0000
##
##             Sensitivity : 0.5133
##             Specificity : 0.5200
```

```
##          Pos Pred Value : 0.5168
##          Neg Pred Value : 0.5166
##              Prevalence : 0.5000
##          Detection Rate : 0.2567
##    Detection Prevalence : 0.4967
##       Balanced Accuracy : 0.5167
##
##        'Positive' Class : Absent
##

## using all predictors 0.7
```

The summary of the model shows that only mood score is statistically significant, as its p-value is less than 0.05. Therefore, we will use mood score separately after attempting to predict using all the other predictors in the models to test if the model improves.

Here are the confusion matrix results of the various variations of the model

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Absent Present
##    Absent     132     134
##    Present     18      16
##
##                Accuracy : 0.4933
##                  95% CI : (0.4354, 0.5514)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : 0.6136
##
##                   Kappa : -0.0133
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.8800
##             Specificity : 0.1067
##          Pos Pred Value : 0.4962
##          Neg Pred Value : 0.4706
##              Prevalence : 0.5000
##          Detection Rate : 0.4400
##    Detection Prevalence : 0.8867
##       Balanced Accuracy : 0.4933
##
##        'Positive' Class : Absent
##

## Confusion Matrix and Statistics
##
##           Reference
## Prediction Absent Present
##    Absent      75      72
##    Present     75      78
##
##                Accuracy : 0.51
```

```
##                  95% CI : (0.4519, 0.5679)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : 0.3864
##
##                   Kappa : 0.02
##
##  Mcnemar's Test P-Value : 0.8690
##
##             Sensitivity : 0.5000
##             Specificity : 0.5200
##          Pos Pred Value : 0.5102
##          Neg Pred Value : 0.5098
##              Prevalence : 0.5000
##          Detection Rate : 0.2500
##    Detection Prevalence : 0.4900
##       Balanced Accuracy : 0.5100
##
##        'Positive' Class : Absent
##
```

```
## RMSE logmodel 0.7
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Absent Present
##    Absent     132     134
##    Present     18      16
##
##                Accuracy : 0.4933
##                  95% CI : (0.4354, 0.5514)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : 0.6136
##
##                   Kappa : -0.0133
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.8800
##             Specificity : 0.1067
##          Pos Pred Value : 0.4962
##          Neg Pred Value : 0.4706
##              Prevalence : 0.5000
##          Detection Rate : 0.4400
##    Detection Prevalence : 0.8867
##       Balanced Accuracy : 0.4933
##
##        'Positive' Class : Absent
##
```

```
## RMSE logmodelmood 0.7118052
```

```
## Confusion Matrix and Statistics
```

```
##
##           Reference
## Prediction Absent Present
##    Absent      75      72
##    Present     75      78
##
##                Accuracy : 0.51
##                  95% CI : (0.4519, 0.5679)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : 0.3864
##
##                   Kappa : 0.02
##
##  Mcnemar's Test P-Value : 0.8690
##
##             Sensitivity : 0.5000
##             Specificity : 0.5200
##          Pos Pred Value : 0.5102
##          Neg Pred Value : 0.5098
##              Prevalence : 0.5000
##          Detection Rate : 0.2500
##    Detection Prevalence : 0.4900
##       Balanced Accuracy : 0.5100
##
##        'Positive' Class : Absent
##

## RMSE_logmodeltraining 0.7

## Confusion Matrix and Statistics
##
##           Reference
## Prediction Absent Present
##    Absent      75      78
##    Present     75      72
##
##                Accuracy : 0.49
##                  95% CI : (0.4321, 0.5481)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : 0.6569
##
##                   Kappa : -0.02
##
##  Mcnemar's Test P-Value : 0.8715
##
##             Sensitivity : 0.5000
##             Specificity : 0.4800
##          Pos Pred Value : 0.4902
##          Neg Pred Value : 0.4898
##              Prevalence : 0.5000
##          Detection Rate : 0.2500
##    Detection Prevalence : 0.5100
##       Balanced Accuracy : 0.4900
##
```

```
##        'Positive' Class : Absent
##
```

```
## RMSE_logmodelmoodtraining 0.7141428
```

The logistic regression model variations tested above do not outperform guessing. All models had lower accuracy compared to random guessing. The model that used only "Mean_Mood_Score" (without using the train function) showed a slight advantage in sensitivity. However, its accuracy was still below 50%, which means that logistic regression model performed poorly.

second model will be k-Nearest Neighbors with cross validation

we start by choosing the best k and testing the various variations of KNN to get the one with the best accuracy

These were the results of the various model variations of the knn model

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Absent Present
##    Absent      91      69
##    Present     59      81
##
##                Accuracy : 0.5733
##                  95% CI : (0.5152, 0.63)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : 0.006456
##
##                   Kappa : 0.1467
##
##  Mcnemar's Test P-Value : 0.426326
##
##             Sensitivity : 0.6067
##             Specificity : 0.5400
##          Pos Pred Value : 0.5687
##          Neg Pred Value : 0.5786
##              Prevalence : 0.5000
##          Detection Rate : 0.3033
##    Detection Prevalence : 0.5333
##       Balanced Accuracy : 0.5733
##
##        'Positive' Class : Absent
##
```

```
## RMSE_knn with only the mood score feature using train function 0.6531973
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Absent Present
##    Absent      78      61
##    Present     72      89
##
```

```
##                Accuracy : 0.5567
##                  95% CI : (0.4985, 0.6137)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : 0.02828
##
##                   Kappa : 0.1133
##
##  Mcnemar's Test P-Value : 0.38588
##
##             Sensitivity : 0.5200
##             Specificity : 0.5933
##          Pos Pred Value : 0.5612
##          Neg Pred Value : 0.5528
##              Prevalence : 0.5000
##          Detection Rate : 0.2600
##    Detection Prevalence : 0.4633
##       Balanced Accuracy : 0.5567
##
##        'Positive' Class : Absent
##

## RMSE_knn with all the features using the train function 0.6658328

## Confusion Matrix and Statistics
##
##           Reference
## Prediction Absent Present
##    Absent      91      70
##    Present     59      80
##
##                Accuracy : 0.57
##                  95% CI : (0.5119, 0.6268)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : 0.00889
##
##                   Kappa : 0.14
##
##  Mcnemar's Test P-Value : 0.37862
##
##             Sensitivity : 0.6067
##             Specificity : 0.5333
##          Pos Pred Value : 0.5652
##          Neg Pred Value : 0.5755
##              Prevalence : 0.5000
##          Detection Rate : 0.3033
##    Detection Prevalence : 0.5367
##       Balanced Accuracy : 0.5700
##
##        'Positive' Class : Absent
##

## RMSE_knn with only the mood score feature without using the train function 0.6658328

## Confusion Matrix and Statistics
```

```
##
##           Reference
## Prediction Absent Present
##    Absent      78      61
##    Present     72      89
##
##                Accuracy : 0.5567
##                  95% CI : (0.4985, 0.6137)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : 0.02828
##
##                   Kappa : 0.1133
##
##  Mcnemar's Test P-Value : 0.38588
##
##             Sensitivity : 0.5200
##             Specificity : 0.5933
##          Pos Pred Value : 0.5612
##          Neg Pred Value : 0.5528
##              Prevalence : 0.5000
##          Detection Rate : 0.2600
##    Detection Prevalence : 0.4633
##       Balanced Accuracy : 0.5567
##
##        'Positive' Class : Absent
##
```

```
## RMSE_knn with all the  features without using the train function 0.6658328
```

The k-NN model that used only mood score (train_knnmood) is the best of all the models tested, with the highest accuracy and sensitivity for 'Absent.' It performs better than random guessing, but it is still not a particularly strong model.

Conclusion\
This analysis aimed to predict student attendance by examining behavioral and psychological factors, such as stress levels, mood, anxiety, and sleep duration. Despite using various machine learning techniques like logistic regression and k-nearest neighbors (KNN), An accuracy of at least 80% could not be achieved with k-nearest neighbors (KNN) coming closest when using the mood score feature. This shows that there are more features that are not recorded that are more influential or the data taken is not as accurate as it should be or more advanced models are required i.e. decision trees, random forest and neural networks.

Key Insights: \
The exploratory analysis revealed that factors like stress, sleep, anxiety, and mood had weak correlations with attendance, although some notable trends were observed. For instance, students who attended classes had slightly higher stress levels and better mood scores, whereas absent students reported higher anxiety levels.

The machine learning models, including logistic regression and KNN, showed limited predictive power. The best result came from using just the "Mood Score" feature using the KNN model, which performed better than random guessing, but still fell short of ideal accuracy. Which shows mood plays a strong role which means additional information such as gender and school year could help more in increasing the robustness of the models.

Implications: \
This analysis provides an initial understanding of the link between student behavior and attendance, this will help in the collection of more features, making it possible to assist schools in identifying students at risk of absenteeism and offer targeted interventions.This could also help psychologists get to understand what psychological factors affect consistency leading to more advancement in the field.

Limitations:\
The dataset used may not encompass all the key factors influencing student attendance. For example, missing variables like academic performance, extracurricular activities, or personal issues may limit the model's ability to predict attendance accurately. The models also struggled with low accuracy, which could be attributed to weak feature-relationship correlations.

Future Directions:\
To improve predictive accuracy, future work could focus on:

Adding more features: Including additional behavioral, social, and academic data may enhance model performance. Exploring advanced models: Techniques like decision trees, random forests, or deep learning could better capture complex patterns and improve results. Expanding the variability of the dataset: more diverse datasets could provide a more comprehensive understanding of the factors influencing attendance.

In conclusion, while the models presented in this analysis offer some insights into the factors affecting student attendance, there is significant room for improvement to make these models more applicable in practical settings.

references\
This analysis was based on the dataset of ziya https://www.kaggle.com/ziya07.