REPORT

PREDICTING BMI CATEGORY BASED ON LIFESTYLE FACTORS

(Predictive Modeling · Data Analysis · R Programming)

NOVEMBER 2024

BY EWAITONDE SHEILA HOUMEY

GitHub softlyshebuilds

Sheila Houmey

SECTION I
IDEA AND DATASET DESCRIPTION

PROJECT IDEA

The goal of this project is to develop a machine learning model predict Body Mass Index (BMI) categories (Normal Weight, Overweight, or Obese)**,** based on three key lifestyle factors: **Daily Steps**, **Sleep Duration**, and **Stress Level**. Our hypothesis is that these factors can be used to effectively predict BMI categories.

The classification of individuals based on their lifestyle factors follows these rules:

- **Normal Weight:** Individuals who get 7-9 hours of sleep, take more than 10,000 steps per day, and have low stress level ≤ 5 .
- **Overweight:** Individuals who have moderate activity levels (between 5,000 and 10,000 steps), experience moderate to high stress $>= 5$, and may not fall within the optimal sleep range.
- **Obese:** Individuals with low activity levels (fewer than 5,000 steps), who either sleep too little or too much (less than 7 or more than 9 hours), and report high stress levels > 5 .

DATASET SOURCE AND DESCRIPTION

The dataset used in this project was provided during a professional data exploration internship led by a collaborator at Fitbit. It contains **anonymized lifestyle data from real users (sleep, steps, stress levels)**. **As it was shared in a professional context, the dataset is not publicly available and cannot be redistributed here.**

It contains 374 records of individuals, containing data about their daily lifestyle habits and their BMI category. The attributes in the dataset are:

- **PersonID:** Unique identifier for each individual (Nominal categorical).
- **Gender:** Gender of the individual (Nominal categorical).
- **Occupation:** The occupation of the individual (Nominal categorical).
- **Age:** Age of the individual (Ratio numerical).
- **Sleep Duration:** Average number of hours of sleep per night (Numerical).
- **Quality of Sleep:** A rating of the quality of sleep, ranging from 1 to 10. (Ordinal categorical).
- **Physical Activity Level:** The number of minutes the person engages in physical activity daily. (Numerical).
- **Stress Level:** Self-reported stress level on a scale from 1 to 10 (Ordinal categorical).
- **BMI Category:** The BMI category of the person (Class label)

Our dataset has **216 normal weight** tuples**, 148 Overweight** tuples **and 10 Obese** tuples. We consider the data imbalanced particularly for the Obese category because its underrepresented which could impact the model's ability to correctly predict this class.

SECTION II

DATA ANALYSIS AND/OR PREPARATION

DESCRIPTIVE STATISTICS ANALYSIS

DAILY STEPS ANALYSIS:

- **Mean Daily Steps:** 6,816.85 steps
- **Median Daily Steps:** 7,000 steps
- **Standard Deviation:** 1,617.92 steps

Since median daily steps is **7,000** we assume that most individuals fall into the moderately active category(**5000 -10000** steps daily) but the high standard deviation indicates some individuals are super active or inactive in contrast to the rest

Conclusion: We expect a high portion of individuals to be categorized as normal weight.

SLEEP DURATION ANALYSIS:

- **Mean Sleep Duration:** 7.13 hours
- **Median Sleep Duration:** 7.2 hours
- **Standard Deviation:** 0.80 hours

Notice that the median value of sleep is **7.2 hours** indicating that half of the population is above the mean. The low standard deviation shows that sleep hours don't vary much from person to person.

Conclusion: We assume that BMI categories do not determine sleep hours but sleep hours are determined by health choices.

STRESS LEVEL ANALYSIS

- **Mean Stress Level:** 5.39
- **Median Stress Level:** 5
- **Standard Deviation:** 1.77

The median is **5**, the data shows **moderate stress** on individuals. A standard deviation of **1.77** makes us expect some individual to experience significant amounts of stress in their day to day.

Conclusion: Food is the leading factor of obesity , but stress oftentimes contributes to food intake

OUTLIER DETECTION

As part of the data preparation process, we performed an outlier detection analysis using the Interquartile Range (IQR) method for the variables **Daily Steps**, **Sleep Duration** and **Stress Level**. Outliers, or extreme values, can potentially distort model performance.
To address this, we applied the **Interquartile Range (IQR) method**, a widely used technique for detecting outliers in numerical data.

IQR CALCULATION AND OUTLIER DETECTION IN DAILY STEPS 👣

The IQR method identifies outliers based on the range between the **25th percentile (Q1)** and the **75th percentile (Q3)**. Values falling outside **1.5 times** the IQR below Q1 or above Q3 are considered outliers.

- **Q1 (25th percentile):** 5600 steps (25% of the individuals in the dataset take fewer than 5600 steps daily)
- **Q3 (75th percentile):** 8000 steps (75% of the individuals take fewer than 8000 steps daily)
- **IQR (Interquartile Range): The difference between Q3 and Q1** is 2400 steps, representing the spread of the central 50% of the data

We calculated the outlier boundaries as follows:

- **Lower Bound:** 2000 steps **(calculated as Q1 - 1.5 * IQR)**
- **Upper Bound:** 11600 steps **(calculated as Q3 + 1.5 * IQR)**

Any values outside this range (below 2000 or above 11600 steps) would be considered outliers.

Based on these calculations, no outliers were detected in the dataset, as all the values for daily steps fell within the defined range.

IQR CALCULATION AND OUTLIER DETECTION FOR SLEEP DURATION

- **Q1 (25th percentile)**: 6.4 hours
- **Q3 (75th percentile):** 7.8 hours
- **IQR**: 1.4 hours
- **Lower Bound:** 4.3 hours; **Upper Bound:** 9.9 hours

Based on these calculations, no outliers were detected in the **Sleep Duration** variable, indicating a clean and consistent dataset for this factor.

IQR CALCULATION AND OUTLIER DETECTION FOR STRESS LEVEL

- **Q1 (25th percentile)**: 4
- **Q3 (75th percentile):** 7
- **IQR**: 3
- **Lower Bound:** -0.5; **Upper Bound:** 11.5

Similarly, no outliers were detected for the **Stress Level** variable, ensuring that this data is reliable for further analysis.

There are no **outliers for daily steps, stress level and sleep duration**. So we consider our dataset to be clean and reliable. There aren't any misssing values either

SECTION III

ALGORITHM DESCRIPTION

CHOICE OF ALGORITHM: DECISION TREE

We used a **decision tree classification algorithm** to predict BMI category **(Normal Weight, Overweight, Obese)**.
Decision trees are good for this because they are easy to understand and can handle data with different types of attributes. This lets us see how the attributes affect the BMI classification.

Decision trees are good for projects that want to find links between things like stress level, sleep duration, and daily steps and a target class. They can handle lots of different types of data **(categorical and numerical values),** which is important for interpreting the different lifestyles in the dataset.

DATA PREPARATION AND CLEANING

We prepared the data before training the model to ensure a clean and consistent dataset:

- **Standardizing BMI Category Labels:** To harmonize labels, all instances of "Normal" in the **BMI.Category** column were standardized to **"Normal Weight."**

- **Handling Missing Values and Outliers:** Extreme values were identified using the **IQR method** for the attributes **Daily Steps, Sleep Duration, and Stress Level.** Observations within the defined bounds **(Q1 - 1.5 * IQR to Q3 + 1.5 * IQR)** were retained for model training, ensuring consistent lifestyle representation.

- **Data Splitting:** The dataset was divided into **training (80%)** and **testing (20%)** sets to evaluate model performance objectively.

DECISION TREE DESCRIPTION

The decision tree algorithm relies on a hierarchical structure, where each node represents a test on an attribute **(e.g., Daily Steps > 5000)**, and each branch represents the result of this test. The leaves of the tree indicate the predicted **BMI category (Normal Weight, Overweight, Obese)** for an individual. The model was trained using R's **rpart** library, set to maximize classification accuracy by categorizing individuals based on lifestyle habits.

In our script, the model uses three primary attributes:

- **Daily Steps:** Average daily steps, indicating the level of physical activity.
- **Sleep Duration:** Average sleep duration, essential for assessing rest.
- **Stress Level:** Self-reported stress level, connected to mental and physical health.

The decision tree is built to separate BMI classes as distinctly as possible. The number of steps you take a day and how long you sleep for are important. Research shows that sleep and exercise affect your health.

EXECUTION AND PREDICTION

Following the decision tree creation, the model was applied to predict BMI categories in the test set. The following code summarizes the steps for training and prediction:

```
# Train the decision tree model
model <- rpart(`BMI.Category` ~ Daily.Steps + Sleep.Duration + Stress.Level,
               data = trainData, method = "class")

# Make predictions on the test set
predictions <- predict(model, testData, type = "class")
```

The **predictions** are then compared to the actual **BMI.Category values** in the test set to evaluate model performance.

EVALUATION METRICS

We created a confusion matrix to evaluate the model's quality. It shows the accuracy, sensitivity, and specificity for each BMI category. The calculated metrics are:

- **Overall Accuracy:** The number of correct predictions for all BMI categories.
- **Sensitivity:** True positive rate for each category, indicating if the model correctly identified each BMI class (normal weight, overweight, obese).
- **Specificity:** True negative rate, representing the model's ability to avoid false classifications.

```
# Generate the confusion matrix
conf_matrix <- confusionMatrix(predictions, testData$`BMI.Category`)

# Print the confusion matrix
print(conf_matrix)

# Extract and display overall accuracy
accuracy <- conf_matrix$overall['Accuracy']
cat("Overall Accuracy: ", accuracy, "\n")
```

LIMITATIONS AND FUTURE IMPROVEMENTS

While the decision tree is an effective and interpretable model, it has limitations when dealing with imbalanced classes, such as the **underrepresented Obese category** in this dataset.

In the future, techniques like **resampling** or **class weighting** could be implemented to improve classification performance in this minority class.

SECTION IV

RESULTS AND ANALYSIS (INSIGHTS!)

The confusion matrix analysis revealed that the dataset's predictions were highly accurate for both the normal weight and overweight categories. With a total **216 normal weight instances** and **148 overweight instances**.
The model exhibited a sensitivity of **85% for the normal weight** class and **92% for the overweight** class.
Hence the model accurately classified a significant majority correctly.

When we conducted our measure of central tendencies on lifestyle factors such as sleep duration, daily step count and stress level the results were normally distributed.
The mean values were **7,000 steps with a 14.29% standard deviation**, **7 hours of sleep with a 10.53% standard deviation**, and a common **stress level rating of 5 with a 32.83% standard deviation**.

We know that the majority of the population are correctly identified therefore we conclude that individuals who follow good lifestyle are mainly classified as normal weight. This enforced our prediction that BMI categories plays a part on lifestyle factors.

The charts and plots attached below visualizes the distribution of lifestyle factors. Figure 1-4

We observe that the values fall within expected ranges, indicating that there are no unusual or outlier values that might suggest errors in the dataset.

The specificity of normal weight is **92%** ,overweight is **89%** and obese is **97%** From these results we acknowledge that our model is fairly accurate because it posses a good ability to identify negative instances, **The negative predicted values** across our data attributes for BMI categories range **from 70% to 90% percent** and **the positive predicted values range from 82% to 90%** across BMI categories **(Normal Weight, Overweight)**.

Obese is not expressed as having neither a positive predicted value and a negative predicted value because it is underrepresented in the dataset, considering that there are a total of only **10 instances** of Obese in our data.

This is unfortunate because it reduces the credibility of our prediction but the positive skewness of data in the variables (daily step , sleep duration and stress level) supports the categories with sufficient data.

Normal weight attribute has a balanced accuracy of **86%** and the **overweight** attribute has a balanced accuracy of **90%** . This indicate that model performs well in classifying individuals in both categories.

SUMMARY

While the model does face some challenges due to underrepresentation of the obese category.
It proves to be fairly accurate for normal weight and obese instances. The high performance of out data being able to accurately identify them on our data enhances the the reliability of our prediction.
Based on the results, the statistical analysis most effectively communicated that BMI categories are influenced by lifestyle factors.

VISUALIZING THE DISTRIBUTION FOR DAILY STEPS

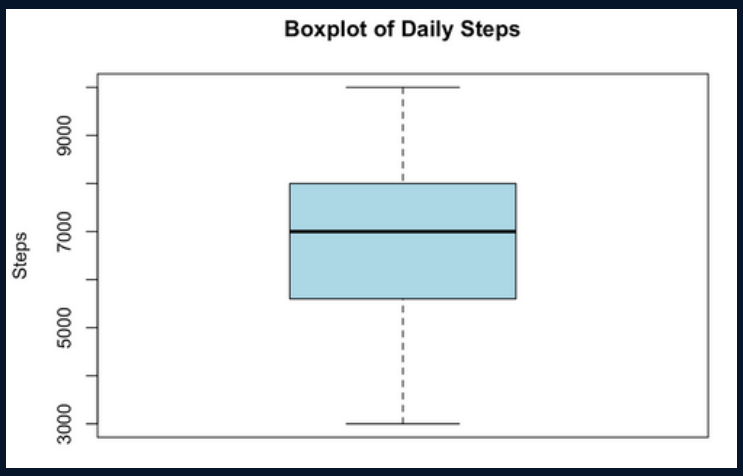

Figure 1: The boxplot illustrates the spread of the data, with the central **50%** (the interquartile range) lying between 5600 and 8000 steps. **No outliers are visible** outside the whiskers of the boxplot, confirming that all values are within the acceptable range.

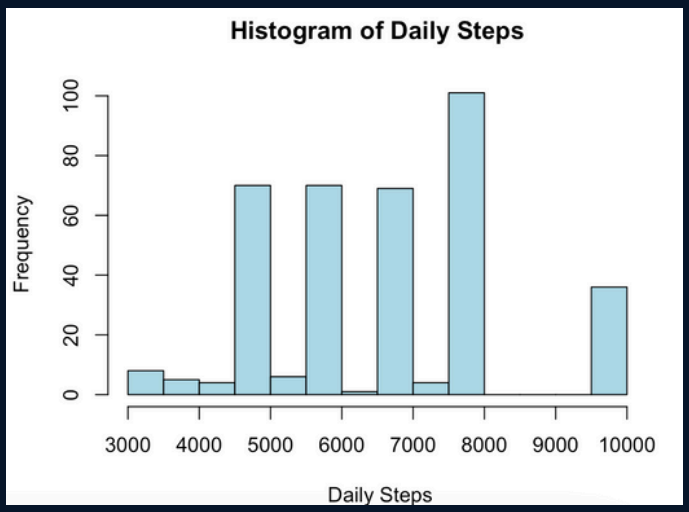

Figure 2: The histogram provides a frequency distribution of daily steps. Most individuals take between 5000 and 8000 steps per day, with no unusually high or low values observed.

VISUALIZING THE DISTRIBUTION FOR STRESS LEVEL

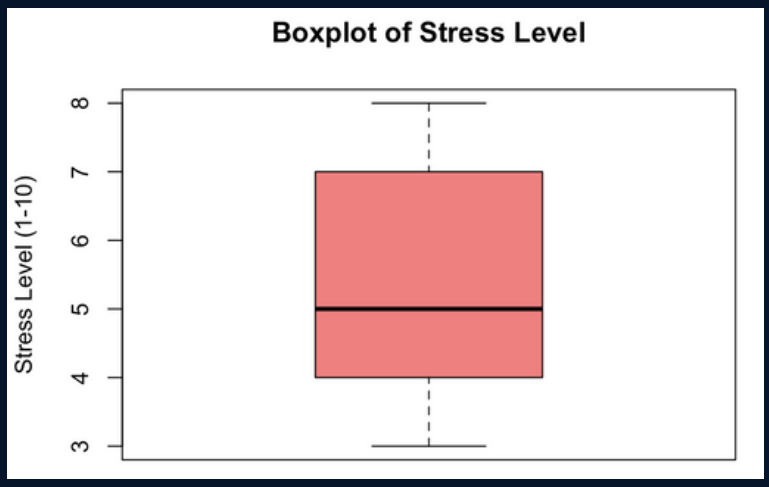

Figure 3: The boxplot for stress level shows the interquartile range between 4 and 7 on a 10-point scale. There are no outliers, indicating that the dataset represents a consistent distribution of stress levels among the participants.

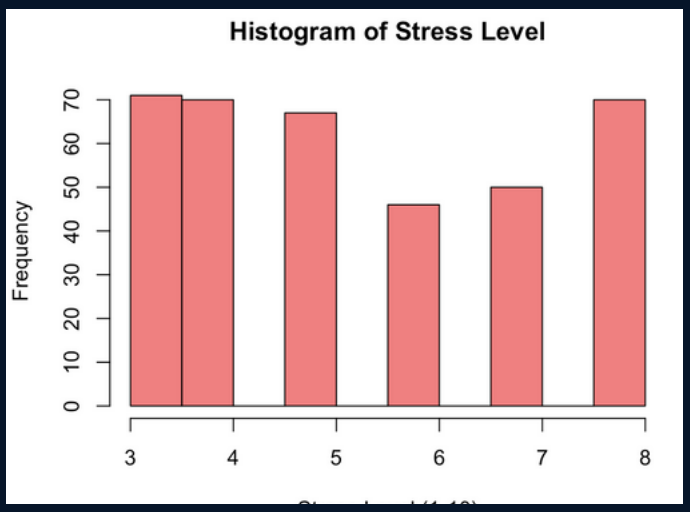

Figure 4: The histogram for stress level supports the notion that most individuals have a moderate stress level, with no extreme high or low values. This clean distribution will allow accurate modeling when predicting BMI based on stress levels.

VISUALIZING THE DISTRIBUTION FOR SLEEP DURATION

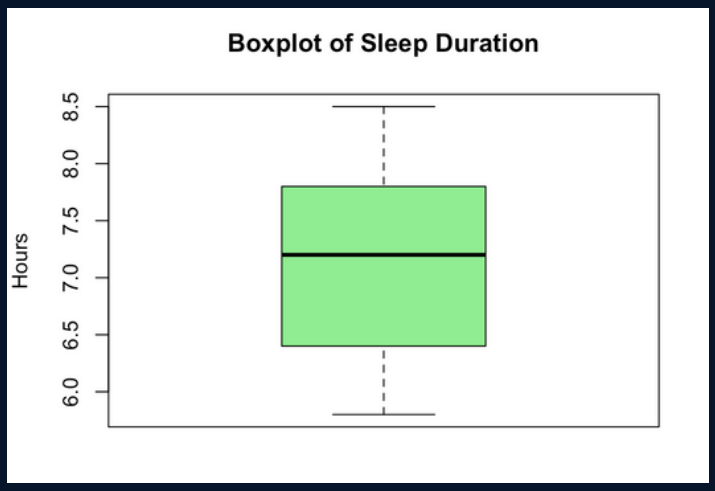

Figure 5: The boxplot for sleep duration shows that the **interquartile range is between 6.4 and 7.8 hours**, with no extreme outliers. This is crucial for ensuring that the dataset is consistent, and the majority of individuals in the dataset seem to fall within the recommended sleep duration for healthy living.

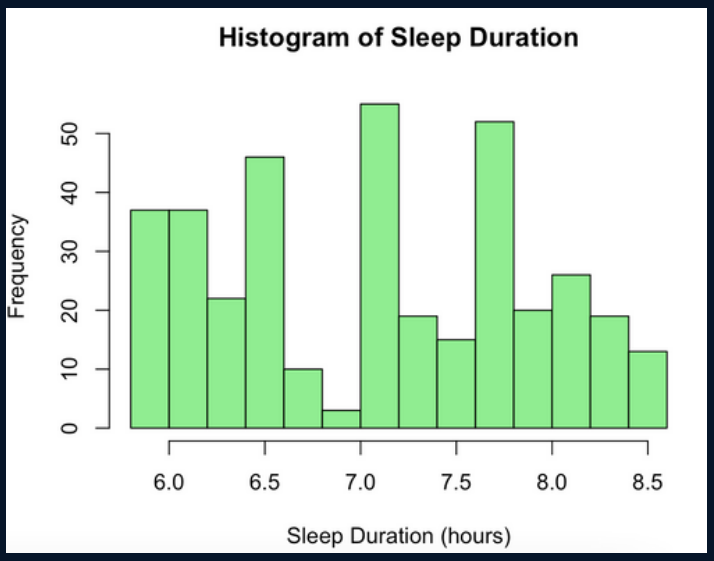

Figure 6: The histogram illustrates that **most individuals in the dataset sleep between 6 and 8 hours**, which aligns with standard health guidelines. The absence of outliers ensures that the dataset is robust for model training.

CONFUSION MATRIX

Confusion Matrix and Statistics

	predictions		
	Normal Weight	Obese	Overweight
Normal Weight	41	0	2
Obese	2	0	0
Overweight	5	0	24

Figure 7

	Class: Normal Weight	Class: Obese	Class: Overweight
Sensitivity	0.8542	NA	0.9231
Specificity	0.9231	0.97297	0.8958
Pos Pred Value	0.9535	NA	0.8276
Neg Pred Value	0.7742	NA	0.9556
Prevalence	0.6486	0.00000	0.3514
Detection Rate	0.5541	0.00000	0.3243
Detection Prevalence	0.5811	0.02703	0.3919
Balanced Accuracy	0.8886	NA	0.9095

Figure 8

BIBLIOGRAPHY

OUTLIER DETECTION WITH IQR:

https://www.datacamp.com/community/tutorials/statistics-in-R-IQR

https://stat.ethz.ch/R-manual/R-devel/library/stats/html/quantile.html

CONFUSION MATRIX & MODEL EVALUATION:

https://www.geeksforgeeks.org/confusion-matrix-in-r/

https://cran.r-project.org/web/packages/caret/vignettes/caret.html

DECISION TREES WITH RPART:

https://www.rdocumentation.org/packages/rpart/versions/4.1-15/topics/rpart

DESCRIPTIVE STATISTICS & VISUALIZATION

https://www.geeksforgeeks.org/how-to-add-mean-and-median-to-histogram-in-r/