A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Denisse Garcia Vivolo |
| *Student Number* | sbs24050 |
| *Module Title* | Machine Learning |
| *Assessment Title* | CA1 Project |
| *Assessment Due Date* | 28/04/2024 |
| *Date of Submission* | 28/04/2024 |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

**Influence of Lifestyle Factors on Sleep Disorders and Quality**

Contents

[Introduction 1](#_Toc165145277)

[Motivation 1](#_Toc165145278)

[Problem Statement 1](#_Toc165145279)

[Objectives 1](#_Toc165145280)

[Data Characterization and Pre-processing 1](#_Toc165145281)

[Hyperparameter Tuning 2](#_Toc165145282)

[Results Interpretation and Model Rationale 2](#_Toc165145283)

[References 5](#_Toc165145284)

[NOTE: The table of contents above has been included for your convenience. To refresh the table, simply click on it, then select 'Update Table' using the mouse. You can choose to update either the page numbers exclusively or the entire table as needed.]

The student would need to consider the following instructions (a - d) during the development of this

project.

a) Logical justification based on the reasoning for the specific choice of machine learning approaches.

b) Multiple machine learning models (at least two) using hyperparameters and a comparison between the chosen modelling approaches.

c) Visualise your comparison of ML modelling outcomes. You may use a statistical approach to argue that one feature is more important than other features.

d) Cross-validation methods should be used to justify the authenticity of your ML results.

You will present the findings and defend the results in the report (MS Doc) by highlighting your work. Your

report should capture the following aspects that are relevant to your project investigations.

1. A clear introduction, motivation, a description of the problem domain, and an explanation of how the project's goals are justified using Prediction / Classification algorithms.

(20 marks)

2. Characterization of data, pre-processing, explanation and description of techniques used for the

variation in the accuracy across three training splits (20%, 25% and 30%) using cross validation

techniques.

(30 marks)

3. What is the primary purpose of hyperparameter tuning in machine learning? Could you elaborate on

specific hyperparameter tuning techniques (e.g., GridSearchCV) applied to machine learning models

to find optimal parameters?

(25 marks)

4. Interpret and explain the results obtained, discuss overfitting / underfitting / generalisation, provide a

rationale for the chosen models and use visualisations to support your findings. Comments in Python

code, conclusions of the project should be specified at the end of the report. Harvard Style must be

used for citations and references.

(25 marks)

# Introduction

Sleep is a crucial component of overall health, influencing not only physical well-being but also mental and emotional health. Adequate sleep has been linked to improved attention, behaviour, learning, memory, emotional regulation, quality of life, and mental and physical health. In contrast, insufficient sleep can lead to a wide range of health consequences including obesity, diabetes, cardiovascular disease, and depression.

Recent studies have shown that various lifestyle factors such as stress levels, physical activity, diet, and occupational pressures significantly affect sleep quality and duration (Knutson et al.). However, the complex interplay between these factors and sleep health is not fully understood. By leveraging data-driven approaches, this CA aims to uncover deeper insights into how different lifestyle choices impact sleep. This understanding can inform targeted interventions to improve sleep health, thereby enhancing overall well-being.

In order to understand the sleep health that factor that might impact, we will exploring the application of machine learning techniques to the Sleep Health and Lifestyle Dataset. The dataset contains information on various aspects of individual lifestyles and their sleep patterns, including age, gender, occupation, sleep duration, and health-related metrics.

The goal is to uncover insights into sleep patterns and disorders, which can inform health interventions and lifestyle adjustments.

# Problem Statement

Sleep health is a critical component of overall well-being, influencing physical health, mental health, and quality of life. Sleep deficiency poses significant challenges across various aspects of life, including work, school, driving, and social interactions. It can impact cognitive functions such as learning, focusing, and emotional recognition, leading to frustration and difficulty in social situations.

Furthermore, sleep deficiency is associated with numerous chronic health conditions like heart disease, high blood pressure, diabetes, and obesity, as well as mental health issues like depression. It also increases the risk of accidents and injuries, including car crashes and falls, across all age groups. Tragic accidents, such as nuclear reactor meltdowns and plane crashes, have been linked to sleep deficiency (National Heart, Lung, and Blood Institute).

This research aims to examine how the sleep quality is affected by the new pace of life and other factors that may cause people to sleep less.

## Goals

1. **Identify Key Predictors**: Determine which factors most significantly impact sleep quality and duration.
2. **Develop Predictive Models:** Utilize machine learning models to predict sleep quality based on observable lifestyle factors.
3. **Provide Recommendations:** Based on the model findings, offer actionable recommendations to improve sleep health.

# Objectives

The primary objective is to use prediction and classification algorithms to:

1. Predict the quality of sleep based on lifestyle factors.
2. Classify individuals with potential sleep disorders.

These goals are pursued to enhance the understanding and management of sleep health.

Key Questions

1. **Gender Differences in Sleep Health:** What is the impact of gender on sleep quality? Are males or females more likely to report poor sleep quality?
   1. How does BMI influence sleep quality across genders? For example, you can analyze whether an increase in BMI category (from normal to overweight or obese) affects sleep quality differently in males vs. females.
2. Age and Sleep Duration: How does age affect sleep duration and quality? Is there a correlation between increasing age and changes in sleep duration or quality?
3. Occupational Impact on Sleep: Does occupation type correlate with sleep quality or duration? For example, comparing high-stress jobs (like doctors) vs. lower-stress jobs.
4. Physical Activity and Sleep Health: How does the level of physical activity affect sleep quality? Is higher physical activity associated with better sleep quality?
5. Stress and Sleep Disorders: What is the relationship between reported stress levels and the presence of sleep disorders? Can high stress levels predict the occurrence of specific sleep disorders like Sleep Apnea?
6. BMI, Blood Pressure, and Heart Rate as Predictors: Can BMI, blood pressure, and heart rate predict sleep quality or the presence of sleep disorders? This can be an exploration of how physiological measures correlate with sleep health.

# Data Characterization and Pre-processing

\*\*Data Characterization\*\*

The dataset includes 374 entries across 13 attributes such as age, sleep duration, physical activity, and more. Each attribute's role is briefly explained with statistical summaries.

\*\*Pre-processing Steps\*\*

Data cleaning involved removing duplicates and handling missing values. The dataset was randomly sampled to create different subsets for training (20%, 25%, 30% splits) to evaluate the models under varying conditions.

\*\*Cross Validation\*\*

Cross-validation techniques were used to assess the robustness of the model predictions. The variance in model accuracy across different splits is discussed to illustrate the stability of the predictive models.

# Hyperparameter Tuning

\*\*Purpose of Hyperparameter Tuning\*\*

Hyperparameter tuning is essential to optimize the performance of machine learning models. It involves adjusting parameters that govern the model training process.

\*\*Tuning Techniques\*\*

GridSearchCV was utilized to systematically work through multiple combinations of parameter tunes, cross-validating as it goes to determine which tune gives the best performance.

# Results Interpretation and Model Rationale

\*\*Model Interpretation\*\*

The outcomes are analyzed, focusing on the accuracy and robustness of predictions. The performance metrics are illustrated through confusion matrices and ROC curves.

\*\*Discussion on Model Fit\*\*

The analysis discusses potential overfitting or underfitting, highlighting how the models generalize to new data.

\*\*Rationale for Model Choice\*\*

The choice of models, including decision trees and SVM, is justified based on their performance and suitability for the dataset's characteristics.

Given the variables available in your "Sleep Health and Lifestyle Dataset," there are several interesting questions and analyses you can explore. Here are some potential questions based on the data:

1. \*\*Gender Differences in Sleep Health\*\*:

- What is the impact of gender on sleep quality? Are males or females more likely to report poor sleep quality?

- How does BMI influence sleep quality across genders? For example, you can analyze whether an increase in BMI category (from normal to overweight or obese) affects sleep quality differently in males vs. females.

2. \*\*Age and Sleep Duration\*\*:

- How does age affect sleep duration and quality? Is there a correlation between increasing age and changes in sleep duration or quality?

3. \*\*Occupational Impact on Sleep\*\*:

- Does occupation type correlate with sleep quality or duration? For example, comparing high-stress jobs (like doctors) vs. lower-stress jobs.

4. \*\*Physical Activity and Sleep Health\*\*:

- How does the level of physical activity affect sleep quality? Is higher physical activity associated with better sleep quality?

5. \*\*Stress and Sleep Disorders\*\*:

- What is the relationship between reported stress levels and the presence of sleep disorders? Can high stress levels predict the occurrence of specific sleep disorders like Sleep Apnea?

6. \*\*BMI, Blood Pressure, and Heart Rate as Predictors\*\*:

- Can BMI, blood pressure, and heart rate predict sleep quality or the presence of sleep disorders? This can be an exploration of how physiological measures correlate with sleep health.

### Example Analysis: Impact of BMI on Sleep Quality by Gender

To address the question about the impact of BMI on sleep quality and how it varies by gender, you could perform the following analysis:

- \*\*Statistical Analysis\*\*: Use statistical tests to compare sleep quality across different BMI categories within each gender. This might involve using ANOVA or Kruskal-Wallis tests if the sleep quality is ordinal or not normally distributed.

- \*\*Visualization\*\*: Create boxplots or bar charts to visualize the differences in sleep quality across BMI categories for each gender.

- \*\*Predictive Modeling\*\*: Use logistic regression or decision trees to predict sleep quality based on BMI, gender, and their interaction.

Let's start with a statistical summary and a visualization to explore the relationship between BMI, gender, and sleep quality. I will generate the necessary code for this analysis. If this aligns with your needs, we can proceed!

\*\*Visualizations\*\*

Graphs and charts are included to visualize the data distributions and model performances.

\*\*Python Code Comments\*\*

Commented Python code is provided to show the implementation of data processing, model training, and evaluation.

### Conclusion

The report concludes with a summary of findings, implications for sleep health management, and suggestions for further research.

---

# References

Draft -

### Problem Domain

The "Sleep Health and Lifestyle Dataset" provides a unique opportunity to explore this area. It includes data on several variables such as age, gender, occupation, sleep duration, sleep quality, physical activity level, stress level, and more. These variables are indicators of lifestyle and health, making it possible to study their effects on sleep.

### Project Goals and Justification Using Machine Learning

#### Goals:

1. \*\*Identify Key Predictors\*\*: Determine which factors most significantly impact sleep quality and duration.

2. \*\*Develop Predictive Models\*\*: Utilize machine learning models to predict sleep quality based on observable lifestyle factors.

3. \*\*Provide Recommendations\*\*: Based on the model findings, offer actionable recommendations to improve sleep health.

#### Justification for Using Prediction/Classification Algorithms:

- \*\*Predictive Analysis\*\*: Machine learning models can analyze complex datasets and identify patterns that may not be apparent through traditional statistical methods. Predictive models like regression or sophisticated algorithms like Random Forest can quantify the impact of various factors on sleep health.

- \*\*Classification\*\*: Classification algorithms can be used to categorize individuals into different levels of sleep quality based on their lifestyle choices. This can help in targeting specific groups for interventions.

- \*\*Algorithm Selection\*\*:

- \*\*Regression Models\*\* (Linear, Logistic): To predict outcomes like sleep duration or categorize sleep quality (good, average, poor).

- \*\*Tree-Based Models\*\* (Decision Trees, Random Forest): These models are beneficial for handling nonlinear relationships and interactions between multiple variables.

- \*\*Neural Networks\*\*: For capturing complex patterns in the data, which might be crucial given the multifaceted nature of lifestyle data.

### Implementation Strategy

1. \*\*Data Preprocessing\*\*: Clean and prepare data, handle missing values, encode categorical variables.

2. \*\*Exploratory Data Analysis\*\*: Gain insights into the data distribution and initial correlations.

3. \*\*Model Building and Evaluation\*\*: Build multiple models to identify the best performer based on metrics such as accuracy, precision, recall, and F1-score.

4. \*\*Feature Importance Analysis\*\*: Determine which features most strongly impact sleep quality using techniques like feature importance in ensemble models.

5. \*\*Validation\*\*: Use techniques like cross-validation to ensure the models are robust and generalize well to unseen data.

This structured approach not only addresses the complex nature of the sleep health problem but also leverages advanced analytical techniques to derive meaningful outcomes that can drive health interventions. Through this project, we aim to contribute to the growing field of health informatics by providing data-backed insights and recommendations to improve sleep health.