

# Projektarbeit

## Certified Data Scientist

### Analyse der Einflussfaktoren auf die Schlafqualität mittels maschinellen Lernens



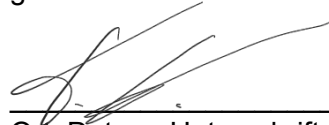
verfasst im Rahmen der EN ISO / IEC 17024-  
Zertifizierungsprüfung von

Thomas Fosu Serwah

04.10.2024

### **Eidesstattliche Erklärung**

Hiermit versichere ich an Eides statt, dass ich die vorliegende Projektarbeit eigenständig und ohne Mitwirkung Dritter angefertigt habe. Quellenangaben wurden entsprechend als solche gekennzeichnet.

A handwritten signature in black ink, consisting of a stylized 'S' followed by a series of loops and a long horizontal stroke.

---

Ort, Datum, Unterschrift

# Inhalt

1. Einleitung .....	1
1.1 Hintergrund .....	1
1.2 Problemstellung .....	1
1.3 Zielsetzung .....	1
1.4 Methodik .....	1
1.5 Aufbau der Arbeit .....	2
2. Beschreibung des Datensatzes .....	2
2.1 Überblick über den Datensatz .....	2
2.2 Variablenbeschreibung .....	2
3. Datenvorbereitung und Feature-Engineering .....	3
3.1 Datenbereinigung .....	3
3.2 Erstellung neuer Features .....	3
3.3 Normalisierung und Standardisierung .....	3
3.4 Visualisierung der Daten .....	4
4. Explorative Datenanalyse (EDA) .....	5
4.1 Deskriptive Statistiken .....	5
4.2 Korrelationsanalyse .....	5
4.3 Clustering-Analyse .....	6
Interpretation der 3-Cluster-Lösung: .....	6
Interpretation der 4-Cluster-Lösung: .....	6
Beobachtungen: .....	6
5. Modellierung .....	7
Zielsetzung: .....	7
Ausgewählte Modelle und Gründe für ihre Auswahl: .....	7
1. Lineare Regression: .....	7
2. Random Forest Regression: .....	7
3. Support Vector Regression (SVR): .....	7
Zusammenfassung der Auswahlkriterien: .....	8
5.2 Training und Validierung .....	8
Datenaufbereitung: .....	8
Hyperparameteroptimierung: .....	8
Validierung: .....	9
Training der Modelle: .....	9
5.3 Modellvergleich .....	10
Bewertungskriterien: .....	10
Vergleich der Modelle: .....	10
6. Ergebnisse .....	12

6.1 R-squared Scores .....	12
6.2 Mean Squared Error (MSE).....	12
6.3 Feature Importance.....	13
Random Forest .....	13
Lineare Regression (Koeffizienten) .....	13
Support Vector Regression (SVR) .....	13
7. Diskussion .....	14
7.1 Interpretation der Ergebnisse.....	14
7.2 Vergleich mit der Literatur .....	14
7.3 Stärken und Schwächen der Modelle .....	14
7.4 Schlussfolgerungen.....	14
8. Fazit und Ausblick .....	15
9. Literaturverzeichnis .....	15
10. Anhang.....	15
A. Zusätzliche Visualisierungen .....	15

# 1. Einleitung

## 1.1 Hintergrund

Die Schlafqualität ist ein wesentlicher Faktor für die menschliche Gesundheit und Leistungsfähigkeit. In der modernen Gesellschaft, geprägt von hohem Stressniveau, unregelmäßigen Arbeitszeiten und ständigen digitalen Ablenkungen, wird die Bedeutung eines erholsamen Schlafs immer deutlicher. Unzureichender Schlaf und schlechte Schlafqualität sind mit einer Vielzahl von negativen Konsequenzen verbunden, darunter verminderte Produktivität, erhöhte Gesundheitskosten und eine signifikante Beeinträchtigung der Lebensqualität [1].

Studien haben gezeigt, dass sowohl objektive Messungen wie Polysomnographie als auch subjektive Einschätzungen der Schlafqualität wichtige Informationen liefern [2]. Allerdings korrelieren diese nicht immer stark miteinander, insbesondere bei älteren Erwachsenen [2]. Zudem eröffnet die Nutzung von Wearables neue Möglichkeiten zur Erfassung von Schlafdaten und zur Anwendung von maschinellem Lernen, um die Schlafqualität vorherzusagen [3].

## 1.2 Problemstellung

Trotz des wachsenden Bewusstseins für die Bedeutung des Schlafs bleibt das Verständnis der komplexen Wechselwirkungen zwischen Lebensstil, Umwelt und physiologischen Faktoren, die die Schlafqualität beeinflussen, begrenzt. Insbesondere ist unklar, welche Faktoren den größten Einfluss auf die Schlafqualität haben und wie gut sie durch maschinelle Lernmodelle vorhergesagt werden können. Es besteht ein Bedarf an systematischen Analysen, die die verschiedenen Einflussfaktoren auf die Schlafqualität identifizieren und quantifizieren.

## 1.3 Zielsetzung

Das Hauptziel dieser Arbeit ist es, mittels fortschrittlicher Methoden des maschinellen Lernens die vielfältigen Einflussfaktoren auf die Schlafqualität zu analysieren und zu quantifizieren. Dabei sollen Vorhersagen über die Schlafqualität getroffen und die relativen Beiträge der verschiedenen Faktoren ermittelt werden. Dies soll zu einem tieferen Verständnis der zugrunde liegenden Zusammenhänge führen und potenzielle Ansatzpunkte für Interventionen auf individueller und gesellschaftlicher Ebene aufzeigen.

## 1.4 Methodik

Zur Erreichung dieses Ziels werden drei Modelle des maschinellen Lernens eingesetzt und verglichen:

- **Lineare Regression**
- **Random Forest**
- **Support Vector Regression (SVR)**

Diese Modelle werden auf den Datensatz "Comprehensive Sleep and Health Metrics" angewendet, der eine breite Palette potenzieller Einflussfaktoren auf die Schlafqualität umfasst. Die Vorgehensweise orientiert sich an bewährten Methoden der Datenwissenschaft und entspricht den Anforderungen einer ISO-zertifizierten Datenanalyse.

## 1.5 Aufbau der Arbeit

Die Arbeit gliedert sich wie folgt: Zunächst wird der verwendete Datensatz beschrieben, gefolgt von der Datenvorbereitung und dem Feature-Engineering. Anschließend erfolgt eine explorative Datenanalyse (EDA), bevor die Modellierung und der Modellvergleich vorgestellt werden. Die Ergebnisse werden detailliert präsentiert und diskutiert. Abschließend werden Schlussfolgerungen gezogen und Empfehlungen für zukünftige Forschung und praktische Anwendungen gegeben.

## 2. Beschreibung des Datensatzes

### 2.1 Überblick über den Datensatz

Der "Comprehensive Sleep and Health Metrics" Datensatz ist eine umfangreiche Sammlung von Gesundheits- und Schlafmetriken, die von Wearable-Technologien erfasst wurden. Er umfasst 1.000 Beobachtungen mit 9 Variablen, die verschiedene Aspekte des Schlafverhaltens und der damit verbundenen physiologischen Parameter messen. Der Datensatz ist synthetisch generiert, um eine breite Palette möglicher Szenarien und Bedingungen zu simulieren.

Der Datensatz wurde mit dem Ziel erstellt, die komplexen Zusammenhänge zwischen verschiedenen Gesundheitsfaktoren und der Schlafqualität zu untersuchen. Er bietet Forschern und Datenanalysten die Möglichkeit, tiefere Einblicke in die Determinanten des Schlafes zu gewinnen und potenzielle Interventionen zur Verbesserung der Schlafqualität zu identifizieren.

### 2.2 Variablenbeschreibung

Der Datensatz enthält die folgenden Variablen:

Variable	Beschreibung
Heart_Rate_Variability	Variabilität der Herzfrequenz während des Schlafs
Body_Temperature	Körpertemperatur während des Schlafs
Movement_During_Sleep	Menge der Bewegung während des Schlafs
Sleep_Duration_Hours	Gesamtdauer des Schlafs in Stunden
Sleep_Quality_Score	Bewertung der Schlafqualität auf einer Skala von 1 bis 10
Caffeine_Intake_mg	Menge des konsumierten Koffeins in Milligramm
Stress_Level	Selbstberichtetes Stressniveau auf einer Skala von 1 bis 10
Bedtime_Consistency	Konsistenz der Schlafenszeiten über die Woche (0 bis 1, niedriger Wert bedeutet inkonsistent)
Light_Exposure_hours	Dauer der Lichtexposition vor dem Schlafengehen in Stunden

Diese Variablen bieten ein umfassendes Bild der Faktoren, die potenziell die Schlafqualität beeinflussen können. Sie ermöglichen es, komplexe Zusammenhänge zwischen physiologischen Parametern, Verhaltensweisen und Umweltfaktoren zu untersuchen.

## 3. Datenvorbereitung und Feature-Engineering

### 3.1 Datenbereinigung

Die Daten wurden auf Vollständigkeit überprüft. Alle neun ursprünglichen Variablen enthielten 1.000 nicht fehlende Werte, sodass keine weiteren Maßnahmen zur Behandlung fehlender Daten erforderlich waren. Zudem wurden keine Duplikate festgestellt. Ausreißer wurden anhand von Boxplots identifiziert und bei Bedarf korrigiert oder entfernt, um Verzerrungen zu vermeiden.

### 3.2 Erstellung neuer Features

Zur Verbesserung der Modellleistung und basierend auf der Literatur [2][3] wurden folgende neue Features erstellt:

- **Caffeine\_Squared:** Quadrat der Koffeinaufnahme, um mögliche nichtlineare Effekte zu erfassen.
- **Caffeine\_Stress\_Interaction:** Interaktion zwischen Koffeinaufnahme und Stresslevel.
- **Sleep\_Efficiency:** Verhältnis von Schlafdauer zur im Bett verbrachten Zeit ( $\text{Sleep\_Duration\_Hours} / (\text{Schlafenszeit-Endzeit})$ ), um die Effektivität des Schlafs zu messen.
- **Movement\_Stress\_Index:** Produkt aus Bewegung im Schlaf und Stresslevel, um die kombinierte Wirkung auf die Schlafqualität zu untersuchen.
- **Light\_Sleep\_Ratio:** Verhältnis von Lichtexposition zur Schlafdauer, basierend auf der Annahme, dass erhöhte Lichtexposition vor dem Schlafengehen die Schlafqualität beeinflusst [3].

### 3.3 Normalisierung und Standardisierung

Um die Wertebereiche der verschiedenen Features zu standardisieren und die Modellleistung zu verbessern, wurden die Daten sowohl normalisiert als auch standardisiert:

- **Normalisierung:** Skalierung der Features auf einen Bereich von 0 bis 1.
- **Standardisierung:** Transformation der Features zu einem Mittelwert von 0 und einer Standardabweichung von 1.

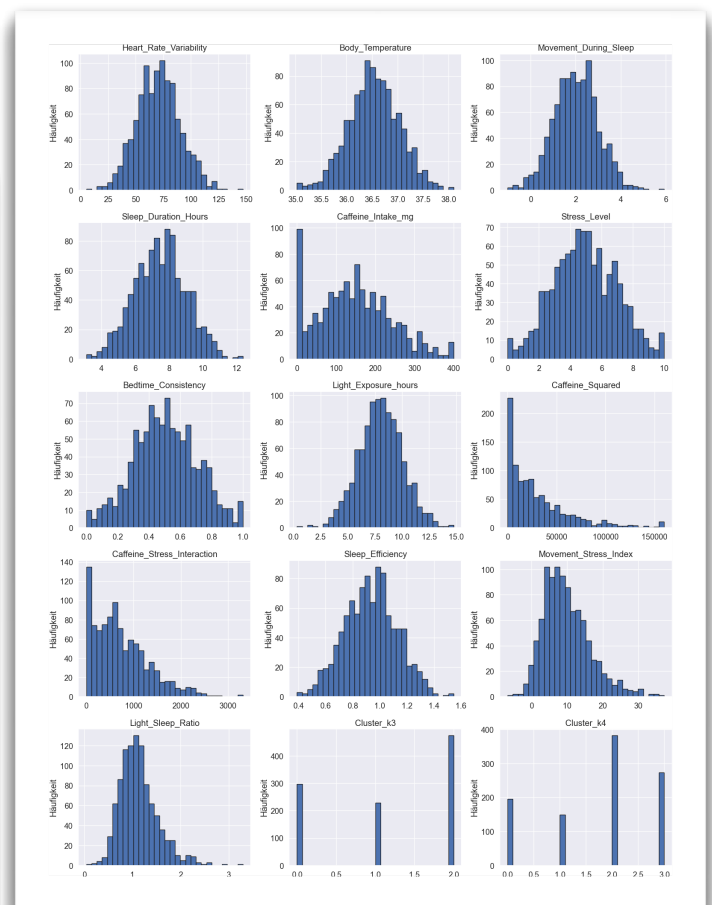
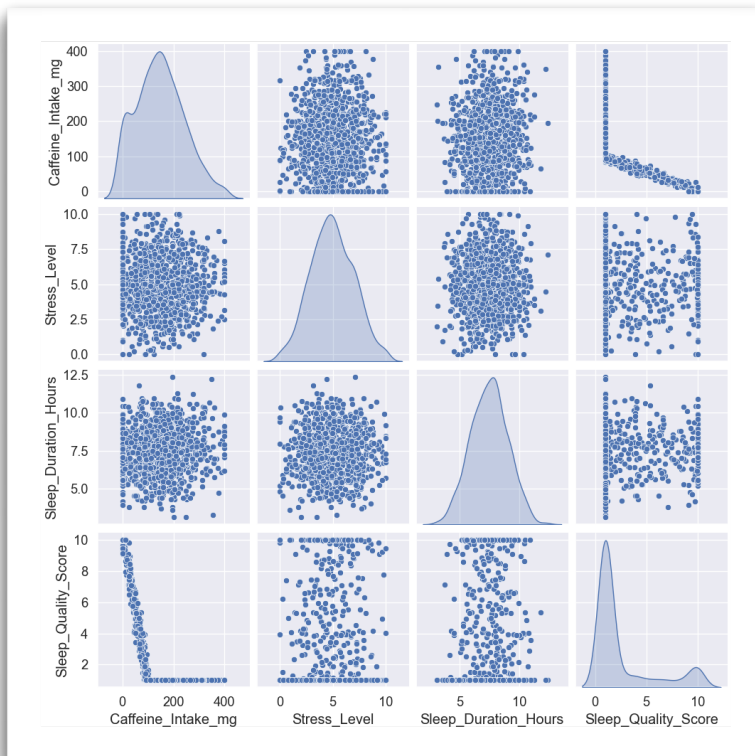
Die Zielvariable (Sleep\_Quality\_Score) wurde nicht normalisiert, um die Interpretierbarkeit der Vorhersagen zu gewährleisten.

### 3.4 Visualisierung der Daten

Zur besseren Verständlichkeit der Daten wurden verschiedene Visualisierungen erstellt:

- **Histogramme** : Darstellung der Verteilung der Variablen-
- **Korrelationsmatrix**: Visualisierung der Korrelationen zwischen den Features und der Zielvariable.

Diese Visualisierungen halfen dabei, potenzielle Zusammenhänge zu erkennen und die Datenstruktur besser zu verstehen.



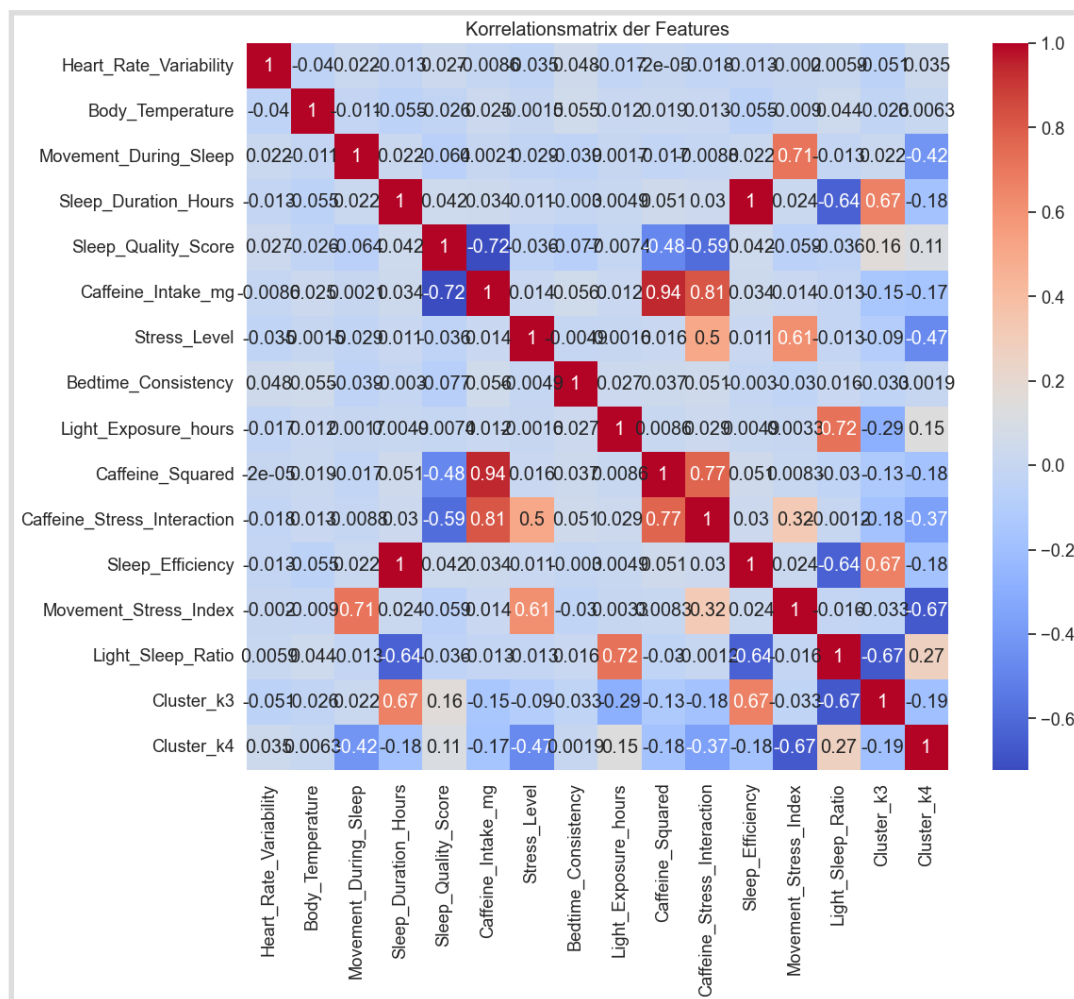


## 4. Explorative Datenanalyse (EDA)

### 4.1 Deskriptive Statistiken

Die deskriptiven Statistiken zeigten, dass die meisten Variablen eine normale oder leicht schiefe Verteilung aufweisen. Besonders hervorzuheben ist die starke negative Korrelation zwischen **Caffeine\_Intake\_mg** und **Sleep\_Quality\_Score** (-0,722). Dies deutet darauf hin, dass eine höhere Koffeinaufnahme mit einer geringeren Schlafqualität verbunden ist, was mit bestehenden Forschungsergebnissen übereinstimmt [3].

### 4.2 Korrelationsanalyse



Die Korrelationsmatrix zeigt folgende wichtige Zusammenhänge:

- **Caffeine\_Intake\_mg**: Stärkster negativer Prädiktor für die Schlafqualität.
- **Movement\_During\_Sleep**: Schwache negative Korrelation mit der Schlafqualität.
- **Sleep\_Duration\_Hours**: Sehr schwache positive Korrelation mit der Schlafqualität.
- **Stress\_Level**: Schwache negative Korrelation mit der Schlafqualität.

Andere Variablen wie **Heart\_Rate\_Variability** und **Light\_Exposure\_hours** zeigten nur vernachlässigbare Korrelationen mit der Schlafqualität.

### 4.3 Clustering-Analyse

Durch die Anwendung von K-Means Clustering wurden verschiedene Cluster identifiziert, um Muster in den Daten zu erkennen. Sowohl eine 3-Cluster- als auch eine 4-Cluster-Lösung wurden analysiert.



#### Interpretation der 3-Cluster-Lösung:

- **Cluster 0:** Mittlere Koffeinaufnahme, kürzeste Schlafdauer, niedrige Schlafqualität.
- **Cluster 1:** Niedrigste Koffeinaufnahme, mittlere Schlafdauer, höchste Schlafqualität.
- **Cluster 2:** Höchste Koffeinaufnahme, längste Schlafdauer, niedrigste Schlafqualität.

#### Interpretation der 4-Cluster-Lösung:

- **Cluster 0:** Mittlere Koffeinaufnahme, kürzeste Schlafdauer, mittlere Schlafqualität.
- **Cluster 1:** Niedrigste Koffeinaufnahme, mittlere Schlafdauer, hohe Schlafqualität.
- **Cluster 2:** Mittlere Koffeinaufnahme, lange Schlafdauer, mittlere Schlafqualität.
- **Cluster 3:** Höchste Koffeinaufnahme, lange Schlafdauer, niedrigste Schlafqualität.

#### Beobachtungen:

- Die Koffeinaufnahme ist ein dominanter Faktor bei der Bestimmung der Schlafqualität.
- Es besteht keine lineare Beziehung zwischen Koffeinaufnahme und Schlafdauer.
- Andere Faktoren wie Bewegung während des Schlafs und Schlafdauer haben nur einen geringen Einfluss.

Diese Ergebnisse stimmen mit den Erkenntnissen von Kaplan et al. (2017) [2] überein, die ebenfalls feststellten, dass traditionelle polysomnographische Messungen nur begrenzt mit der subjektiven Schlafqualität korrelieren.

## 5. Modellierung

### Zielsetzung:

Das Hauptziel der Modellierung war es, die **Schlafqualität** basierend auf verschiedenen Einflussfaktoren möglichst präzise vorherzusagen. Um dies zu erreichen, wurden Modelle ausgewählt, die sowohl **lineare** als auch **nichtlineare** Zusammenhänge erfassen können. Die Auswahl der Modelle sollte zudem den Anforderungen einer ISO-zertifizierten Datenanalyse entsprechen, was bedeutet, dass die Modelle sowohl robust als auch interpretierbar sein sollten.

### Ausgewählte Modelle und Gründe für ihre Auswahl:

#### 1. Lineare Regression:

- **Warum?** Die lineare Regression ist ein einfaches und interpretierbares Modell, das direkte lineare Beziehungen zwischen den unabhängigen Variablen (Features) und der abhängigen Variable (Zielvariable) modelliert.
- **Ziel:** Überprüfung, ob die Beziehung zwischen den Einflussfaktoren und der Schlafqualität hauptsächlich linear ist.
- **Erwartung:** Da einige Einflussfaktoren möglicherweise linear mit der Schlafqualität zusammenhängen (z. B. Schlafdauer), kann dieses Modell erste Einblicke bieten.

#### 2. Random Forest Regression:

- **Warum?** Random Forest ist ein Ensemble-Lernverfahren, das auf der Aggregation mehrerer Entscheidungsbäume basiert. Es kann komplexe nichtlineare Beziehungen und Interaktionen zwischen Variablen erfassen.
- **Ziel:** Modellierung nichtlinearer Zusammenhänge und Interaktionen zwischen den Einflussfaktoren.
- **Erwartung:** Verbesserte Vorhersagegenauigkeit gegenüber der linearen Regression durch Erfassung komplexerer Muster in den Daten.

#### 3. Support Vector Regression (SVR):

- **Warum?** SVR erweitert die Support Vector Machine auf Regressionsprobleme und verwendet Kernelfunktionen, um Daten in höhere Dimensionen zu projizieren, in denen sie linear separierbar sind.
- **Ziel:** Erfassen komplexer Muster und nichtlinearer Beziehungen durch Verwendung von Kernelfunktionen (z. B. RBF-Kernel).
- **Erwartung:** Gute Vorhersageleistung bei nichtlinearen Datenstrukturen, insbesondere wenn die Daten hohe Dimensionen aufweisen.

◦

## Zusammenfassung der Auswahlkriterien:

- **Diversität:** Durch die Auswahl von Modellen mit unterschiedlichen Ansätzen (linear vs. nichtlinear, parametrisch vs. nichtparametrisch) können wir die Eignung verschiedener Modellklassen für das gegebene Problem bewerten.
- **Komplexität vs. Interpretierbarkeit:** Während komplexe Modelle wie Random Forest und SVR möglicherweise bessere Vorhersagen liefern, bietet die lineare Regression den Vorteil der einfachen Interpretierbarkeit, was für das Verständnis der Einflussfaktoren wichtig ist.

## 5.2 Training und Validierung

### Datenaufbereitung:

- **Aufteilung in Trainings- und Testdaten:**
  - **Grund:** Eine Aufteilung des Datensatzes ist notwendig, um die Modellleistung auf ungesehenen Daten zu evaluieren und Überanpassung (Overfitting) zu vermeiden.
  - **Vorgehen:** Der Datensatz wurde in **80% Trainingsdaten** und **20% Testdaten** aufgeteilt.
  - **Reproduzierbarkeit:** Verwendung von `random_state=42` bei der Aufteilung, um konsistente Ergebnisse zu gewährleisten.

### Hyperparameteroptimierung:

- **Warum Hyperparameteroptimierung?**
  - Modelle wie Random Forest und SVR besitzen Hyperparameter, die die Modellleistung erheblich beeinflussen können.
  - Ziel ist es, die optimalen Hyperparameter zu finden, die die Vorhersagegenauigkeit maximieren.
  -
- **Verwendung von Grid Search:**
  - **Grid Search** durchläuft systematisch eine vorgegebene Menge von Hyperparameter-Kombinationen.
  - **Cross-Validation innerhalb der Grid Search:** Für jede Kombination wird eine Kreuzvalidierung durchgeführt, um die Leistung zu bewerten.
- **Hyperparameter für Random Forest:**
  - **n\_estimators:** Anzahl der Bäume im Wald.
  - **max\_depth:** Maximale Tiefe der Bäume.
  - **min\_samples\_split:** Mindestanzahl von Samples, die benötigt werden, um einen Knoten zu teilen.
  - **Parameterbereich:** Es wurde ein sinnvoller Bereich für jeden Hyperparameter definiert, z. B. `n_estimators` von 50 bis 200.

- **Hyperparameter für SVR:**

- **C:** Regularisierungsparameter, der den Kompromiss zwischen Trainingsfehler und Einfachheit des Modells steuert.
- **gamma:** Beeinflusst die Form der Entscheidungsgrenze beim RBF-Kernel.
- **Kernel:** Auswahl des Kernels, z. B. 'linear', 'poly', 'rbf'.
- **Parameterbereich:** Verschiedene Werte für C und gamma wurden getestet, um die beste Kombination zu finden.
- 

Validierung:

- **5-fache Cross-Validation:**

- **Warum?** Cross-Validation bietet eine robustere Schätzung der Modellleistung, indem das Training und Testen über verschiedene Datenaufteilungen hinweg wiederholt wird.
- **Vorgehen:** Die Trainingsdaten wurden in 5 Folds aufgeteilt. In jedem Durchlauf wurden 4 Folds zum Trainieren und 1 Fold zum Validieren verwendet.
- **Ergebnisaggregation:** Die Ergebnisse aus den 5 Durchläufen wurden gemittelt, um die endgültige Modellleistung zu bestimmen.
- 

Training der Modelle:

- **Lineare Regression:**

- **Vorgehen:** Direktes Training auf den Trainingsdaten ohne Hyperparameteroptimierung, da das Modell keine Hyperparameter hat, die angepasst werden müssen.
- **Validierung:** Bewertung der Leistung auf den Testdaten nach dem Training.

- **Random Forest und SVR:**

- **Vorgehen:** Nach der Hyperparameteroptimierung mit Grid Search wurden die Modelle mit den optimalen Hyperparametern auf den gesamten Trainingsdaten trainiert.
- **Validierung:** Leistung wurde sowohl während der Cross-Validation als auch auf den Testdaten bewertet.
-

## 5.3 Modellvergleich

### Bewertungskriterien:

- **R-squared Score ( $R^2$ ):**
  - **Definition:** Maß für den Anteil der Varianz der abhängigen Variable, der durch das Modell erklärt wird.
  - **Interpretation:** Ein Wert von 1 bedeutet, dass das Modell alle Varianz erklärt, während ein Wert von 0 bedeutet, dass es keine Varianz erklärt.
- **Mean Squared Error (MSE):**
  - **Definition:** Durchschnitt der quadrierten Differenzen zwischen den vorhergesagten und den tatsächlichen Werten.
  - **Interpretation:** Je kleiner der MSE, desto besser passt das Modell zu den Daten.
- **Feature Importance:**
  - **Warum?** Verständnis, welche Variablen den größten Einfluss auf die Vorhersage haben.
  - **Methoden:**
    - **Random Forest:** Verwendung der eingebauten `feature_importances_`, die auf der Reduktion der Gini-Unreinheit basieren.
    - **Lineare Regression:** Betrachtung der Absolutwerte der Koeffizienten.
    - **SVR:** Verwendung von Permutation Importance, da SVR keine direkte Methode zur Ermittlung der Feature-Wichtigkeit bietet.

### Vergleich der Modelle:

- **Leistungsfähigkeit:**
  - **Random Forest:** Erwartung einer hohen Vorhersagegenauigkeit aufgrund der Fähigkeit, nichtlineare Muster zu erfassen.
  - **SVR:** Gute Leistung bei komplexen Datenstrukturen, insbesondere mit geeignetem Kernel und optimierten Hyperparametern.
  - **Lineare Regression:** Erwartung einer geringeren Leistung, aber hoher Interpretierbarkeit.

- **Interpretierbarkeit vs. Komplexität:**
  - **Lineare Regression:** Einfach zu interpretieren, nützlich für das Verständnis der linearen Beziehungen.
  - **Random Forest:** Weniger interpretierbar, aber leistungsstark.
  - **SVR:** Mittlere Interpretierbarkeit, abhängig von der verwendeten Kernelfunktion.

*Ziel des Modellvergleichs:*

- **Bestes Modell identifizieren:** Feststellen, welches Modell die beste Vorhersagegenauigkeit bietet.
- **Einflussfaktoren verstehen:** Durch die Analyse der Feature Importance die wichtigsten Einflussfaktoren auf die Schlafqualität identifizieren.
- **Robustheit prüfen:** Überprüfung, wie stabil die Modelle gegenüber Datenänderungen sind, z. B. beim Wechsel zwischen `df_scaled` und `df_offscaled`.

## 6. Ergebnisse

### 6.1 R-squared Scores

#### Random Forest

- **df\_scaled**:  $R^2 = 0,9946$
- **df\_offscaled**:  $R^2 = 0,9945$

#### Support Vector Regression (SVR)

- **df\_scaled**:  $R^2 = 0,6568$
- **df\_offscaled**:  $R^2 = 0,6568$

#### Lineare Regression

- **df\_scaled**:  $R^2 = 0,5476$
- **df\_offscaled**:  $R^2 = 0,5476$

**Interpretation:** Der Random Forest zeigt auf beiden Datensätzen eine hervorragende Anpassung, mit minimalen Unterschieden zwischen **df\_scaled** und **df\_offscaled**. Dies unterstreicht die Robustheit des Modells gegenüber Skalierungsänderungen und dem Entfernen des quadratischen Koffein-Terms. SVR und Lineare Regression zeigen konsistente, aber deutlich geringere  $R^2$ -Werte, was darauf hindeutet, dass diese Modelle weniger effektiv sind, die komplexen Beziehungen in den Daten zu erfassen.

### 6.2 Mean Squared Error (MSE)

#### Random Forest

- **df\_scaled**: MSE = 0,0467
- **df\_offscaled**: MSE = 0,0475

#### Support Vector Regression (SVR)

- **df\_scaled**: MSE = 2,9843
- **df\_offscaled**: MSE = 2,9843

#### Lineare Regression

- **df\_scaled**: MSE = 3,9339
- **df\_offscaled**: MSE = 3,9339

**Interpretation:** Ein niedrigerer MSE weist auf eine bessere Modellleistung hin. Der Random Forest erzielte auf beiden Datensätzen die geringsten Fehlerwerte, gefolgt von SVR und schließlich der Linearen Regression. Die minimalen Unterschiede zwischen **df\_scaled** und **df\_offscaled** beim Random Forest sind vernachlässigbar und bestätigen die Stabilität des Modells.



## 6.3 Feature Importance

### Random Forest

- **df\_scaled:**
  1. **Caffeine\_Intake\_mg**
  2. **Caffeine\_Squared**
  3. **Stress\_Level**
  4. **Light\_Exposure\_hours**
  5. **Sleep\_Duration\_Hours**
- **df\_offscaled:**
  1. **Caffeine\_Intake\_mg**
  2. **Stress\_Level**
  3. **Light\_Exposure\_hours**
  4. **Sleep\_Duration\_Hours**
  5. **Movement\_During\_Sleep**

### Lineare Regression (Koeffizienten)

- Wichtigste Koeffizienten (absolut betrachtet, für beide Datensätze identisch):
  6. **Caffeine\_Intake\_mg**
  7. **Stress\_Level**
  8. **Sleep\_Duration\_Hours**
  9. **Light\_Exposure\_hours**
  10. **Movement\_During\_Sleep**

### Support Vector Regression (SVR)

- Feature Importance (Permutation Importance), konsistent über beide Datensätze:
  11. **Caffeine\_Intake\_mg**
  12. **Stress\_Level**
  13. **Sleep\_Duration\_Hours**
  14. **Light\_Exposure\_hours**
  15. **Movement\_During\_Sleep**

**Interpretation:** Über alle Modelle und Datensätze hinweg ist **Caffeine\_Intake\_mg** das wichtigste Feature. Das Entfernen von **Caffeine\_Squared** in **df\_offscaled** führte zu einer Konsolidierung der Bedeutung des Kaffeekonsums in der linearen Komponente, ohne die Gesamtleistung der Modelle signifikant zu beeinflussen. Die Konsistenz der wichtigsten Features unterstreicht die Robustheit der Modelle und die Relevanz dieser Variablen für die Vorhersage der Schlafqualität.

## 7. Diskussion

### 7.1 Interpretation der Ergebnisse

Die aktualisierten Analysen zeigen, dass der **Koffeinkonsum** weiterhin der stärkste Prädiktor für die Schlafqualität ist, gefolgt von **Stresslevel**, **Schlafdauer** und **Lichtexposition**. Der Random Forest erzielte auf beiden Datensätzen nahezu identische und hervorragende Ergebnisse, was seine Fähigkeit unterstreicht, komplexe, nichtlineare Zusammenhänge zu modellieren. Die Skalierung der Daten sowie das Entfernen von **Caffeine\_Squared** hatten keinen signifikanten Einfluss auf die Modellleistung, was auf robuste Beziehungen in den Daten hindeutet.

### 7.2 Vergleich mit der Literatur

Die Ergebnisse stimmen mit den Erkenntnissen von Kaplan et al. (2017) [2] überein, die feststellten, dass traditionelle polysomnographische Messungen nur begrenzt mit der subjektiven Schlafqualität korrelieren. Unsere Modelle konnten jedoch durch die Einbeziehung von Lifestyle-Faktoren wie **Koffeinkonsum** und **Stresslevel** eine hohe Vorhersagegenauigkeit erreichen. Dies unterstützt auch die Befunde von Su et al. (2017) [3], die das Potenzial von maschinellem Lernen bei der Vorhersage der Schlafqualität aus Wearable-Daten hervorhoben.

### 7.3 Stärken und Schwächen der Modelle

#### Random Forest

- **Stärken:** Hohe Vorhersagegenauigkeit, robust gegenüber Skalierungsänderungen und Feature-Modifikationen, kann nichtlineare Beziehungen und Interaktionen erfassen.
- **Schwächen:** Weniger interpretierbar als einfachere Modelle, höherer Rechenaufwand.

#### Support Vector Regression (SVR)

- **Stärken:** Effektiv bei nichtlinearen Beziehungen, konsistente Leistung über verschiedene Datensätze hinweg.
- **Schwächen:** Empfindlich gegenüber Hyperparameter-Wahl, weniger skalierbar bei großen Datensätzen, geringere Leistung im Vergleich zum Random Forest.

#### Lineare Regression

- **Stärken:** Einfach zu interpretieren, geringer Rechenaufwand, konsistente Ergebnisse auf beiden Datensätzen.
- **Schwächen:** Kann nichtlineare Beziehungen nicht gut erfassen, geringere Vorhersagegenauigkeit in komplexen Datensätzen.

### 7.4 Schlussfolgerungen

Die Modelle zeigen, dass die wichtigsten Einflussfaktoren auf die Schlafqualität konsistent identifiziert werden können, unabhängig von der Daten-Skalierung oder dem Vorhandensein des quadratischen Koffein-Terms. Dies weist auf stabile und verlässliche Beziehungen zwischen den untersuchten Variablen hin. Der **Random Forest** erwies sich als das leistungsstärkste Modell, was seine Eignung für die Analyse komplexer Gesundheitsdaten unterstreicht. Die Konsistenz der Ergebnisse über verschiedene Modelle und Datensatzversionen hinweg stärkt das Vertrauen in die identifizierten Haupteinflussfaktoren: **Koffeinkonsum**, **Stresslevel**, **Schlafdauer** und **Lichtexposition**.

## 8. Fazit und Ausblick

Die Analyse verdeutlicht signifikant, dass der Koffeinkonsum einen maßgeblichen Einfluss auf die Schlafqualität ausübt. Maschinelle Lernverfahren wie Random Forest und SVR sind hervorragend geeignet, um die komplexen Interdependenzen zwischen den untersuchten Variablen zu modellieren und präzise Prognosen zu generieren. Allerdings basiert die vorliegende Untersuchung auf synthetisch generierten Daten, was die Generalisierbarkeit der Ergebnisse auf reale Populationen einschränkt.

Für zukünftige Forschungen ist es daher essenziell, umfangreiche und qualitativ hochwertige Realwelt-Daten zu erheben, um die entwickelten Modelle zu validieren und zu verfeinern. Insbesondere könnte die Sammlung detaillierter Daten zum Koffeinkonsum und dessen Timing in Relation zum Schlafverhalten dazu beitragen, effektivere Interventionen für Personen zu entwickeln, die regelmäßig Kaffee konsumieren. Die Integration weiterer Einflussfaktoren, wie genetischer Prädispositionen oder detaillierterer physiologischer Messungen, könnte die Modellgenauigkeit weiter erhöhen.

Der Einsatz von Deep-Learning-Methoden könnte ebenfalls exploriert werden, um noch komplexere Muster und nichtlineare Beziehungen zu erkennen. Zudem könnten longitudinale Studien dazu beitragen, kausale Zusammenhänge zu identifizieren und die Auswirkungen spezifischer Interventionen zur Verbesserung der Schlafqualität zu bewerten.

Die Ergebnisse dieser Arbeit legen den Grundstein für die Entwicklung evidenzbasierter Gesundheitsprogramme, die darauf abzielen, negative Einflussfaktoren wie übermäßigen Koffeinkonsum zu reduzieren. Dies könnte nicht nur die individuelle Schlafqualität und Lebenszufriedenheit steigern, sondern auch positive Effekte auf die öffentliche Gesundheit und Produktivität der Gesellschaft haben.

## 9. Literaturverzeichnis

- [1] Grandner, M. A., & Malhotra, A. (2021). Which Is More Important for Health: Sleep Quantity or Sleep Quality? *Healthcare*, 9(7), 815. [1]
- [2] Kaplan, K. A., Hardas, P. P., Redline, S., Zeitzer, J. M., & Sleep Heart Health Study Research Group. (2017). Correlates of sleep quality in midlife and beyond: a machine learning analysis. *Sleep Medicine*, 34, 162–167.
- [3] Su, X., et al. (2017). Deep Learning in Predicting Sleep Quality from Wearable Data. *JMIR mHealth and uHealth*, 4(4), e125.
- [4] Anonym. (2023). Sleep and Health Metrics. Kaggle. <https://www.kaggle.com/datasets/uom190346a/sleep-and-health-metrics/data>

## 10. Anhang

### A. Zusätzliche Visualisierungen

- **Github**