“Impacts of Coffee Consumption on Sleep Quality and Health”

Abstract:

Ever wondered about the true impact of your daily coffee ritual? Our research explores coffee's complex relationship with health, revealing how this beloved beverage can be both a energizer and a potential sleep disruptor. Using advanced machine learning, we've developed a tool that helps individuals understand their unique coffee consumption patterns. Our goal is simple: empower people to enjoy their coffee while maintaining optimal health and sleep quality, transforming a daily habit into a mindful experience. The goal isn't to demonize coffee, but to help you become more aware. We're offering a roadmap to enjoying your favorite drink while keeping your health in perfect balance. It's about finding that sweet spot where you can savor your coffee and still get the restorative sleep your body craves.

Introduction:

Coffee is one of the most widely consumed beverages globally, with billions of cups enjoyed daily across cultures and demographics. At the heart of its popularity is caffeine, a powerful stimulant known to enhance alertness, concentration, and energy levels. However, its influence extends far beyond mere wakefulness. Scientific evidence highlights caffeine's complex interactions with human biology, particularly its effects on sleep quality, duration, and health. These dualities make caffeine consumption a fascinating and essential area for investigation, especially in an era where sleep disorders and lifestyle-related health challenges are on the rise.

This project seeks to understand and analyze the relationships between caffeine consumption, sleep quality, and overall health by leveraging data-driven approaches. By analyzing demographic and lifestyle data, the study aims to identify patterns and insights that could help individuals make informed decisions about their coffee habits. Modern machine learning techniques are employed to explore these dynamics, allowing for sophisticated predictions and actionable recommendations. Factors such as age, occupation, exercise frequency, and co-consumption of substances like alcohol and tobacco are integrated into the analysis to paint a comprehensive picture.

Through the integration of exploratory data analysis (EDA) and advanced machine learning models, this project not only aims to enhance our understanding of caffeine’s role in sleep and health but also to offer practical tools for lifestyle management. A key outcome is the development of a user-friendly web application that helps individuals track their coffee consumption and receive tailored recommendations to optimize their sleep and overall well-being.

Aim and Objectives:

The primary aim of this project is to analyze caffeine consumption's impact on sleep quality and overall health across diverse demographics. Specifically, the objectives include:

- Identifying safe daily caffeine consumption levels by age group to minimize adverse effects on sleep and health.
- Exploring correlations between lifestyle factors (e.g., exercise, occupation) and caffeine's effects on sleep efficiency and duration.
- Developing predictive models to analyze and forecast the consequences of caffeine consumption on individual lifestyles.
- Creating a web-based application that empowers users to monitor their habits and receive insights for improved well-being.

Methodology:

The project methodology for analyzing the impacts of coffee consumption on sleep quality and health is outlined below, following the process depicted in the provided flowchart. This structured approach ensures comprehensive and actionable insights.

1. **Data Collection**

   The initial phase involved collecting data on individuals' coffee consumption patterns, sleep habits, and other lifestyle factors such as exercise, alcohol intake, and smoking status. The dataset incorporated a wide range of variables, including sleep efficiency, REM sleep percentage, and health-related metrics. This step ensured a robust foundation for subsequent analysis.

2. **Data Processing**

   Raw data was processed to ensure consistency and usability. This phase included encoding categorical variables, handling missing values, and formatting the dataset for analysis. Relevant metrics were extracted and organized for seamless integration into the analytical pipeline.

3. **Data Cleaning**

   The data cleaning phase addressed outliers, noise, and inconsistencies to ensure the reliability of results. This step involved validating the dataset, ensuring it accurately captured relationships between coffee consumption and sleep quality.

4. **Exploratory Data Analysis (EDA)**

   EDA was performed to uncover patterns, correlations, and trends related to coffee consumption and its impact on sleep quality. Key visualizations, such as histograms, scatter plots, and correlation matrices, highlighted significant relationships between coffee intake, sleep efficiency, and overall health indicators.

5. **Machine Learning Models and Statistical Analysis**

   After EDA, machine learning models and statistical techniques were applied to predict the effects of coffee consumption on sleep quality and health. Regression and classification algorithms were utilized to derive actionable insights. A predictive model was developed to analyze the impacts of varying coffee consumption levels on lifestyle factors.

6. **Building Data Products**

   The insights from the analysis were integrated into a user-friendly web application. This tool allows individuals to track and monitor their coffee consumption and receive personalized recommendations based on the developed predictive model. The application aims to help users understand and improve their sleep quality and overall health.

Requirements:

1. Hardware Requirements

- Laptop/PC (Intel Core i5, 8 GB RAM, 256 GB SSD, internet connectivity)
- Optional cloud resources (AWS, GCP, Firebase).

2. Software Requirements

- **Programming Language:** Python 3.8+
- **Libraries:** Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, Streamlit, Firebase, Requests
- **IDE:** Jupyter Notebook, VS Code, or PyCharm
- **Version Control:** Git/GitHub
- **Hosting Platforms:** Streamlit Cloud, Firebase Hosting

3. Data Requirements

- Dataset with coffee consumption, sleep metrics, and lifestyle factors.
- Real-time data integration via Firebase.

4. Functional Requirements

- Data processing (cleaning, encoding, handling missing values).
- Machine learning models for prediction and analysis.
- Streamlit-based web application for user interaction.
- Real-time data updates using Firebase.

Phase 1:

In Phase 1 of the project, the focus was on analyzing caffeine consumption and its effects on sleep quality and overall health across various age groups. The aim was to determine daily coffee consumption levels for each age group that would minimize interference with sleep quality and health.

Data Preparation and Cleaning:

1. **Cleaning Process**:
    a. Negative values in the Caffeine consumption column were replaced with NaN and adjusted using the median value for positive consumption.
    b. Missing or erroneous values for Bedtime and Wakeup time were estimated using Sleep duration or default wakeup times. Rows with invalid or missing data in critical columns were dropped.
    c. Sleep duration values were constrained to a realistic range (0–10 hours).
2. **Categorization**:
    a. Age groups were created using bins (e.g., <20, 20-30, etc.), and new metrics like wakefulness_time were calculated for further analysis.

Exploratory Data Analysis (EDA):

The EDA aimed to uncover relationships between caffeine consumption and lifestyle or health variables:

1. **Sleep Duration**:
    a. Younger individuals (under 30) experienced shorter sleep durations with higher caffeine consumption.
    b. Older adults maintained more consistent sleep patterns regardless of caffeine intake.
2. **Sleep Efficiency**:
    a. Low to moderate caffeine consumption (0–50 mg) correlated with higher sleep efficiency, while high consumption (200–1000 mg) significantly reduced sleep quality.
3. **Age and Efficiency**:
    a. Younger groups maintained higher sleep efficiency with moderate caffeine intake. Older adults (51+) showed sharper declines in efficiency with high caffeine consumption.
4. **Occupational Influence**:
    a. High-stress occupations (e.g., Engineers, Teachers, Lawyers) had higher average caffeine consumption compared to healthcare workers like Doctors and Nurses.

5. **Exercise and Sleep**:
    a. Regular exercisers consuming low caffeine (0–50 mg) showed similar sleep durations to non-caffeine consumers, with 7–8 hours of sleep being common.
6. **Alcohol and Smoking Correlation**:
    a. Combined caffeine and alcohol consumption exacerbated sleep efficiency issues. Smokers tended to consume more caffeine, with wider variability in intake.
7. **Wakefulness Time**:
    a. Younger individuals exhibited slightly higher wakefulness times, but caffeine's influence was minor across age groups, especially for older adults.

Hypotheses:

Insights from EDA supported several hypotheses:

- Younger individuals are more sensitive to caffeine's effects on sleep duration and efficiency.
- Moderate caffeine consumption has limited adverse effects, but high levels disrupt sleep quality significantly.
- Lifestyle factors such as occupation, smoking, and alcohol consumption influence caffeine intake and its impacts on health.

This groundwork laid the foundation for developing a machine learning model in Phase 2 to analyze and predict the effects of caffeine on individual lifestyles, integrated with a web application for user tracking and analysis.

Phase 2:

Phase 2 expanded on the findings from Phase 1 by implementing advanced machine learning models and further exploratory data analysis to understand the relationships between caffeine consumption, sleep quality, and overall health.

Data Preparation:

1. **Data Cleaning**:
    a. Similar to Phase 1, missing values were handled through imputation techniques, with additional outlier removal using interquartile ranges and percentiles.
    b. Non-numeric columns like Occupation were converted into dummy variables, and features were standardized for improved model performance.
2. **Feature Engineering**:
    a. New features such as Caffeine_Age_Interaction, logarithmic transformations, and polynomial features were introduced to capture complex relationships.
    b. Interaction terms between key variables like caffeine consumption, age, and sleep quality were explored.

Machine Learning Models and Key Analyses:

1. **K-Means Clustering**:
    a. Applied to group individuals based on Age, Caffeine consumption, and Sleep duration.
    b. The elbow method determined the optimal cluster count (k=3), validated by silhouette scores.
    c. Clustering identified distinct patterns, such as similar sleep durations within certain age groups, revealing caffeine consumption trends.
2. **Logistic Regression**:
    a. Used to classify high vs. low sleep efficiency.
    b. Polynomial features enhanced the model's ability to capture non-linear patterns.
    c. GridSearchCV for hyperparameter tuning achieved an accuracy of 74%, with an AUC score of 0.75, indicating a reliable classification.
3. **Random Forest Regressor**:
    a. Predicted caffeine consumption based on demographic and lifestyle factors like occupation, sleep patterns, and exercise frequency.
    b. Achieved near-perfect performance with an MSE of 0.0005 and R^2 of 0.999, effectively modeling complex relationships.

4. **LightGBM Regressor**:
    a. Predicted Sleep duration using multiple features, including Caffeine consumption, Alcohol consumption, and Exercise frequency.
    b. Delivered strong results with an MSE of 0.11 and R^2 of 0.84, demonstrating effective handling of feature interactions.
5. **SVM Classifier**:
    a. Classified caffeine consumption as high or low based on Sleep duration, Age, and Exercise frequency.
    b. Achieved an accuracy of 84.5%, supported by a confusion matrix to evaluate predictions.
6. **XGBoost Regressor**:
    a. Modeled Sleep duration with hyperparameter tuning for optimal performance.
    b. Produced an MSE of 0.255 and R^2 of 0.64, highlighting its ability to generalize predictions effectively.
7. **Decision Tree Classifier**:
    a. Classified individuals into high or low awakening frequency groups based on lifestyle factors.
    b. Delivered a 96% accuracy, effectively segmenting individuals by sleep disruption frequency.
8. **KNN Regressor**:
    a. Predicted sleep efficiency by age group, yielding an MSE of 0.026.
    b. Results indicated that age alone was insufficient for strong predictive power of sleep efficiency.

Insights:

1. **Caffeine and Sleep**:
    a. High caffeine consumption significantly reduced sleep efficiency and duration, with pronounced effects in younger age groups.
    b. Moderate caffeine intake (50–75 mg) had minimal negative impacts.
2. **Lifestyle Factors**:
    a. Stressful occupations correlated with higher caffeine consumption, while exercise frequency moderated some adverse effects.
    b. Alcohol and smoking amplified caffeine's disruptive effects on sleep.
3. **Model Effectiveness**:
    a. Advanced models captured intricate relationships between caffeine, sleep, and health, offering valuable predictions and actionable insights for lifestyle interventions.

Phase 2 concluded with the development of a web application incorporating these machine learning models, enabling users to track their caffeine consumption, predict sleep impacts, and receive personalized recommendations to improve their health and well-being.

Phase 3:

This project leverages Firebase for real-time data integration and Streamlit for web application development. The web application aims to analyze sleep-related metrics such as sleep duration and sleep efficiency, utilizing machine learning (ML) models, including Logistic Regression, KNN, SVM, and LightGBM. Users can interact with the app through four distinct pages:

1. **Home Page**: The homepage introduces the application, offering an overview of its purpose. Using a typewriter animation built with HTML, CSS, and JavaScript, the page highlights the key features of the app, including sleep analysis, data entry, and visualizations.
2. **Data Entry Page**: This page allows users to input their personal and lifestyle data. Key inputs include age, gender, occupation, sleep duration, sleep efficiency, caffeine consumption, alcohol consumption, exercise frequency, and other relevant factors. Once submitted, the data is stored in the Firebase database in real-time. This functionality is facilitated by a form, where users can provide data such as sleep-related habits and health factors.
3. **View Data Page**: The view data page provides users the ability to view all previously submitted records. Users can filter the data by age range using an interactive slider. This page dynamically retrieves data from Firebase and displays it in a tabular format, allowing for easy analysis and management of the entered information.
4. **Visualizations Page**: The final page showcases various machine learning model predictions and visualizations. The models, which include LightGBM, Logistic Regression, KNN, and SVM, predict sleep duration, sleep efficiency, caffeine consumption, and more. Visualizations include prediction vs. actual graphs, classification metrics such as accuracy, ROC curves, and confusion matrices. These visualizations provide insights into the relationship between lifestyle factors and sleep quality.

Code Overview:

In this code, we have integrated machine learning models for predicting sleep-related outcomes. Below are the key components:

- **Data Integration**: Firebase is used to store and retrieve user data in real time. The requests module communicates with the Firebase database to fetch and add data as JSON objects.
- **Machine Learning Models**:
    - **LightGBM**: Used to predict sleep duration based on various lifestyle factors. The model is trained on features like light sleep percentage, alcohol consumption, caffeine intake, and exercise frequency.
    - **Logistic Regression**: Employed to predict sleep efficiency, with a binary classification approach that categorizes individuals based on whether their sleep efficiency is above or below the median.

- **SVM**: Applied to predict caffeine consumption levels, classifying individuals as high or low caffeine consumers.
- **KNN Regressor**: Used to predict sleep efficiency, providing a regression-based approach to assess the impact of various lifestyle factors on sleep quality.

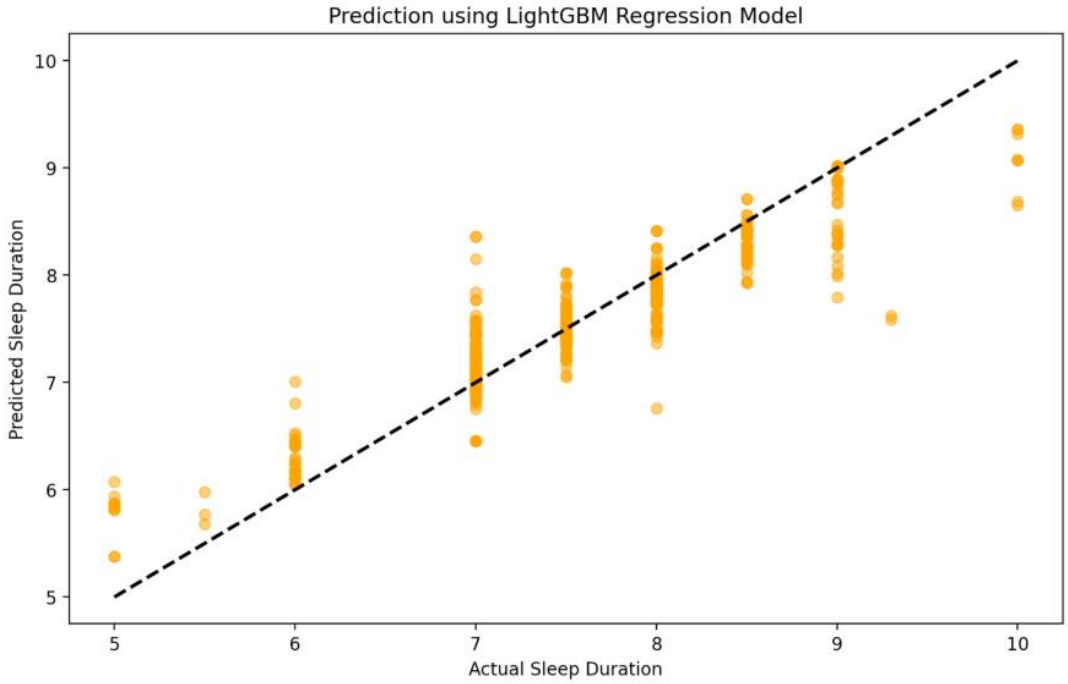
Performance Metrics: Each model's performance is evaluated using metrics such as Mean Squared Error (MSE), R-squared (R^2), Accuracy, and AUC (Area Under the Curve). The visualizations present a clear comparison between actual vs. predicted values, helping users understand how well the models perform.

Model Evaluation (LightGBM)

Mean Squared Error (MSE): 0.12312166245026282

R-squared (R2): 0.8336191319461749

Prediction vs Actual Values (LightGBM)



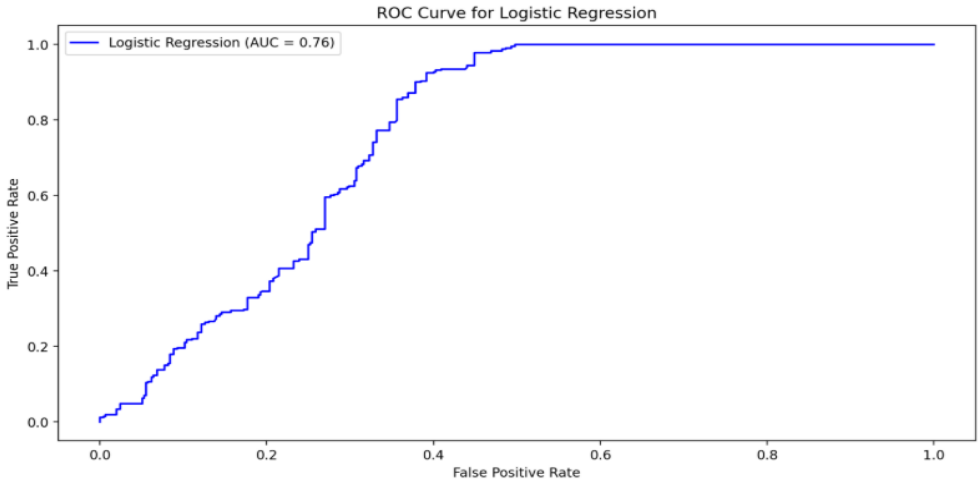
Logistic Regression Results

Accuracy: 0.7410404624277457

Classification Report: precision recall f1-score support

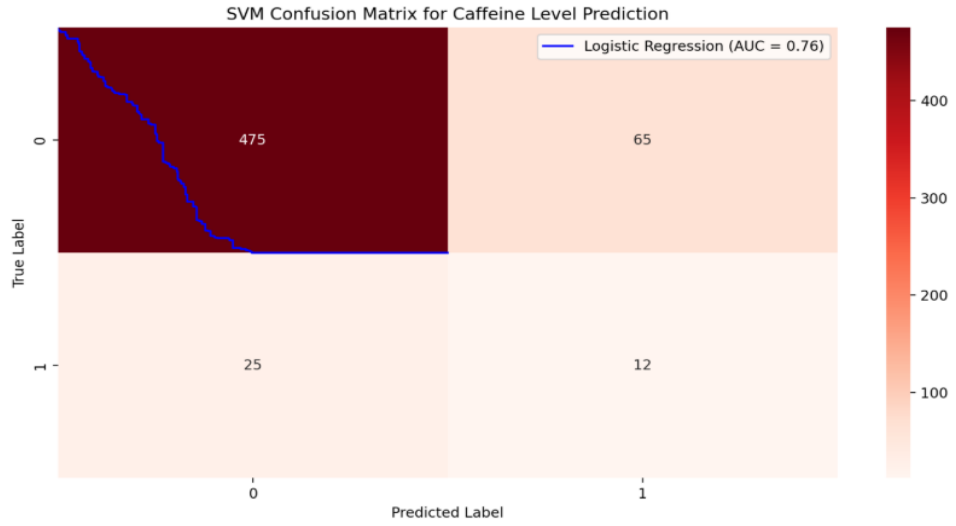
```
0       0.98      0.52      0.68       452
        1       0.65      0.99      0.78       413

accuracy                            0.74       865
```

macro avg 0.81 0.75 0.73 865 weighted avg 0.82 0.74 0.73 865



SVM Model for Caffeine Level Prediction

Accuracy: 84.40207972270363%

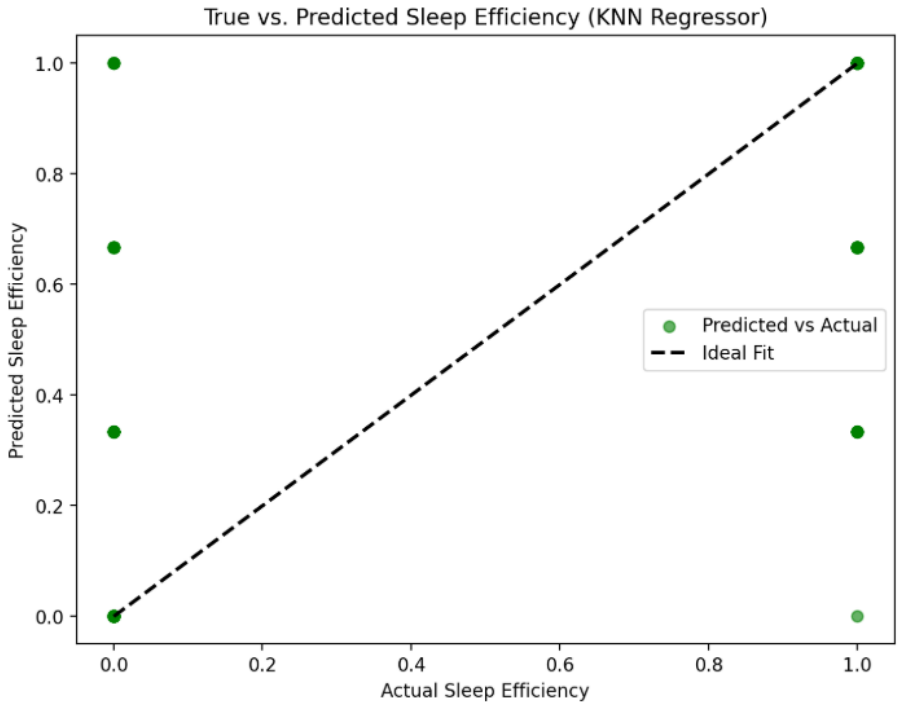


K-Nearest Neighbors Regressor Results for Sleep Efficiency Prediction

Mean Squared Error (MSE): 0.04

Mean Absolute Error (MAE): 0.07

R^2 Score: 0.85



For each model, predictions are compared against actual values, with plots showing the performance of the algorithms as shown in above figs. For classification models like Logistic Regression and SVM, confusion matrices, ROC curves, and classification reports are provided.

Conclusion:

This application provides users with an interactive and engaging way to track and analyze the impact of various lifestyle factors on sleep quality. By combining real-time data entry with machine learning models, it enables users to gain valuable insights and make informed decisions to improve their health and well-being. Through its intuitive interface, the app empowers individuals to understand how factors such as exercise, caffeine consumption, and sleep habits can influence their sleep duration and efficiency.

References:

https://docs.streamlit.io/

https://firebase.google.com/docs