Analysis of lifestyle factors and sleep health as predictors of chronic disease risk.

(COMP3125 Individual Project)

Brian Morillo
School of Computing and Data Science
Wentworth Institute of Technology

***Abstract*—**This study analyzes the relationship between lifestyle habits, sleep health, and chronic disease risk using two public health datasets. The primary objective was to develop a predictive model for chronic disease classification and to investigate occupational impacts on sleep quality. Two machine learning algorithms were implemented and compared: a Logistic Regression model with balanced class weights and a Random Forest Classifier. The results demonstrated a significant "Accuracy Paradox" inherent in imbalanced medical datasets. While the Random Forest model achieved a superior accuracy of 75.1%, it failed to identify any high-risk individuals (Recall of 0.00). In contrast, the Logistic Regression model prioritized sensitivity, achieving a Recall of 0.49 despite a lower overall accuracy of 49.9%. Furthermore, exploratory data analysis revealed that psychological stress is a stronger predictor of reduced sleep duration than physical activity. Occupational analysis identified distinct high-risk groups, with Nurses exhibiting an 83.6% prevalence of Sleep Apnea and Salespersons showing a 90.6% prevalence of Insomnia. These findings underscore the importance of selecting appropriate evaluation metrics beyond accuracy in health data science and highlight specific target populations for workplace health interventions.

Keywords—Chronic Disease Prediction, Sleep Health, Random Forest, Class Imbalance

I. Introduction (*Heading 1*)

The relationship between daily lifestyle factors, sleep health, and the risk of developing chronic disease represents a concern in public health. Understanding these complex interactions is crucial because lifestyle behaviors are often modifiable targets for intervention, potentially reducing the burden of long-term illness. This project undertakes the analysis of lifestyle factors and sleep health as points of prediction for chronic disease risk using comprehensive datasets that capture various physiological and behavioral metrics.

A significant area of investigation involves how occupational demands and corresponding stress levels impact overall well-being. This analysis seeks to explore whether specific professional roles, such as Doctors, Nurses, Accountants, or Sales Representatives, are associated with a higher frequency of diagnosed sleep problems, including Insomnia or Sleep Apnea. Furthermore, the study aims to uncover the relationship between an individual's self-reported stress levels and their engagement in physical activity, evaluating how these factors collectively influence the quality and total duration of their sleep.

Building on these explorations, the objective of this research is to leverage quantitative analysis to determine if a comprehensive health and habit profile can accurately predict a person's chronic disease risk (categorized as high or low). This profile incorporates specific metrics such as Body Mass Index (BMI) category, measured blood pressure, average daily steps taken, and documented sleep hours. Initial examination of relevant datasets indicates a diverse population ranging in age from 18 to 79 years. This existing data shows that approximately 25% of the population under consideration is already classified as being at risk for chronic disease. This established data context sets the stage for the utilization of the classification methodology, the appropriate technique for predicting categorical outcomes, to generate a predictive model for chronic disease risk.

II. Datasets

A. Source of dataset (Heading 2)

The data used for this project was primarily sourced from credible public data repositories on Kaggle. This research is based on two distinct datasets relating to health, lifestyle, and sleep disorders. The "Health & Lifestyle Dataset" dataset was created by Rehan Liaqat on the 4th October of 2025 and the "Sleep Disorder Diagnosis Dataset" dataset was created by Sultanul Ovi on the 7th of September of 2025.

B. Character of the datasets

The primary data source, the Health & Lifestyle Dataset, is a large, structured compilation of quantitative and categorical health information. This dataset contains a total of 100,000 records (rows), each representing an individual subject, and includes 16 distinct features (columns). The dataset is structured to support classification methodology, as it incorporates both predictor variables (lifestyle habits and physiological measurements) and a clear binary target variable. The features within the Health & Lifestyle Dataset encompass detailed numerical and categorical information essential for modeling chronic disease risk.

Key features:

Parameter/Feature	Character Attributes
id	Unique identifier
age	18 – 79 years
gender	Male/Female
bmi (Body Mass Index)	18 – 40
daily_steps	1,000 – 19,999
sleep_hours	3 – 10 hrs
cholesterol	150 – 299 mg/dL
disease_risk	0 (Low), 1 (High)

979-8-3315-0993-4/25/$31.00 ©2025 IEEE

The second dataset is the “Sleep Disorder Diagnosis Dataset”, which focuses specifically on the interaction between occupation, lifestyle, and documented sleep problems. This smaller, yet highly detailed, dataset contains 374 records (rows) and 13 features (columns). The dataset captures essential demographic (e.g., Age, Gender), behavioral (e.g., Occupation, Stress Level), and physiological data (e.g., Heart Rate) for each individual.

Key Features:

Parameter/Feature	Character Attributes
id	Unique identifier
gender	Categorical
age	Age in years
occupation	Profession or job category
sleep Duration	Hours slept per day (e.g., 6.1, 7.8)
stress_level	Subjective stress rating from 1 (low) to 10 (high)
blood_pressure	Measured as systolic/diastolic
sleep_disorder	Presence of disorder: None, Insomnia, or Sleep Apnea

III. Methodology

A. Classification

In order to predict if someone is at high or low risk of developing a chronic disease based on their health and habits (like BMI, blood pressure, daily steps, and hours of sleep), the appropriate methodology is Classification. This method is chosen because it is designed to predict a categorical outcome. The goal is to train a model that accurately maps inputs to a binary output: either 'High Risk' (1) or 'Low Risk' (0).

Classification is a type of supervised machine learning method used to predict which category an observed data point belongs to. In this project, the classification model will learn the relationship between input variables (e.g., bmi, sleep_hours, daily_steps, systolic_bp, diastolic_bp, cholesterol) and the target variable, disease_risk.

Assumptions of this method/model:

- Relevance of Features: The selected health and habit metrics (BMI, blood pressure, steps, sleep) are assumed to have a meaningful statistical relationship with the chronic disease risk outcome.
- Target Variable Definition: The disease risk variable must accurately and reliably categorize individuals into high or low-risk groups.
- Independence: The observations (individuals/rows) within the dataset are assumed to be independent of one another.

Advantages:

• Direct Prediction: Classification models provide a direct prediction of a categorical result (High or Low Risk), which aligns with the research question.

Disadvantages:

• Complexity and Interpretation: Depending on the specific algorithm chosen, the model may function as a "black box," making it challenging to interpret exactly why a prediction was made.

B. Model Selection and Implementation

To predict chronic disease risk, two distinct supervised learning algorithms were implemented using the Python Scikit-learn library:

1. *Logistic Regression:* A linear model was selected as a baseline to test for direct linear relationships between lifestyle variables and disease risk. To address the dataset's class imbalance (where "Low Risk" cases significantly outnumber "High Risk" cases), the model was configured with class_weight='balanced', which assigns a higher penalty to misclassifying the minority class.
2. *Random Forest Classifier:* An ensemble learning method constructing 100 decision trees was employed to capture non-linear interactions between features.

C. Evaluation Metrics

Given the imbalanced nature of the dataset, relying solely on Accuracy can be misleading. Therefore, the models were evaluated using a comprehensive set of metrics defined as follows:

1. *Accuracy:* The ratio of correctly predicted observations to the total observations.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2. *Recall (Sensitivity):* The ratio of correctly predicted positive observations (High Risk) to all observations in actual class. This is critical in medical diagnosis to ensure sick patients are not overlooked.

$$Recall = \frac{TP}{TP + FN}$$

3. *Precision:* The ratio of correctly predicted positive observations to the total predicted positive observations.

$$Precision = \frac{TP}{TP + FP}$$

Where: TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives.

D. Analytical Techniques

To address the research questions regarding lifestyle correlations, Grouping and Aggregation methods were

employed. Rather than relying solely on global correlation coefficients, the data was grouped by specific attributes (e.g., Occupation, Stress Level) to calculate mean health metrics. This approach allowed for the identification of specific non-linear patterns and "high-risk" categories within the population.

IV. Results

A. Prediction of Chronic Disease Risk

Table 1 Model Performance Comparison

Model	Accuracy	Training Time	Class	Precision	Recall	F1-Score
Logistic Regression	49.9%	0.069s	Low Risk	0.75	0.50	0.60
			High Risk	0.25	0.49	0.33
Random Forest	75.1%	18.63s	Low Risk	0.75	1	0.86
			High Risk	0.25	0	0

The comparative analysis of the two predictive models, summarized in Table 1, revealed a significant discrepancy between overall accuracy and clinical utility (Recall).

- *Random Forest Performance*: The Random Forest model achieved a high overall accuracy of 75.1%. However, analysis of the Confusion Matrix (Fig. 1, Bottom) and Classification Report reveals that this accuracy is driven entirely by the majority class. The model yielded a Recall of 0.00 for the "High Risk" class, effectively classifying every patient as "Low Risk" to maximize its accuracy score.

- *Logistic Regression Performance*: The Logistic Regression model, forced to balance class weights, achieved a lower accuracy of 49.9%. However, it demonstrated a significantly higher Recall of 0.49, successfully identifying nearly half of the at-risk population, though at the cost of numerous False Positives.

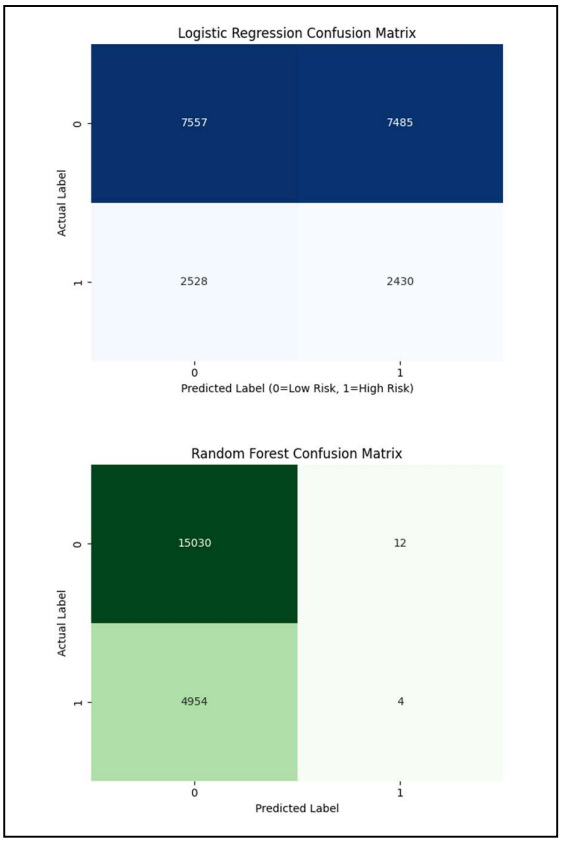

Fig. 1. Comparison of Confusion Matrices. The Logistic Regression model (Top) attempts to identify high-risk individuals (Recall 0.49), resulting in higher error rates. The Random Forest model (Bottom) maximizes accuracy by predicting only the majority class, failing to identify any high-risk cases.

B. Impact of Stress on Sleep and Activity

The analysis of lifestyle factors indicated a strong correlation between psychological stress and sleep health, but there was no direct link to physical activity.

- *Stress and Sleep*: A clear negative trend was observed. As self-reported stress levels increased from 3 to 8, average sleep duration dropped consistently from 8.2 hours to 6.0 hours.

- *Stress and Activity*: Physical activity levels showed high variance with no linear correlation to stress, suggesting that exercise habits are independent of perceived stress levels in this population.

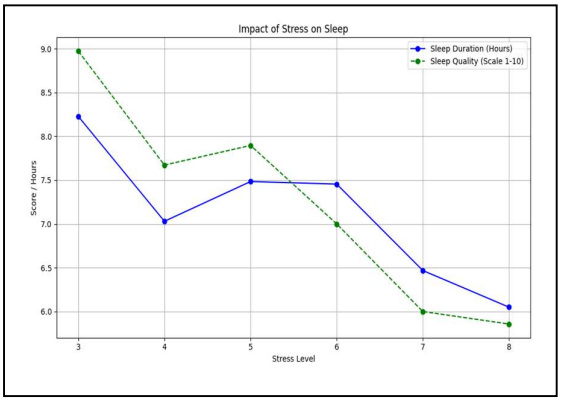

Fig. 2. Impact of Stress Level on Sleep Duration and Quality. Higher stress levels correlate strongly with reduced sleep duration.

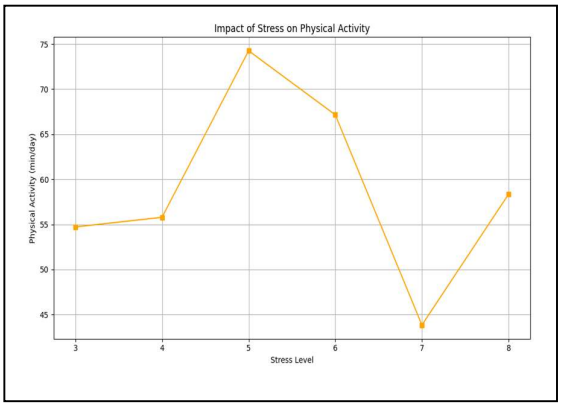

Fig. 3. Impact of Stress Level on Physical Activity. No significant linear correlation was observed between stress levels and daily activity.

C. *Occupational Sleep Health*

The investigation into occupation-specific health risks identified distinct clusters of sleep disorders among specific professions.

- *High Risk Groups*: Nurses exhibited the highest rate of Sleep Apnea (83.6%), likely associated with shift work and high-stress environments. Salespersons (90.6%) and Teachers (67.5%) showed the highest prevalence of Insomnia.
- *Low Risk Groups*: Doctors, Engineers, and Lawyers reported the highest sleep health, with approximately 90% of individuals in these roles reporting no sleep disorders.

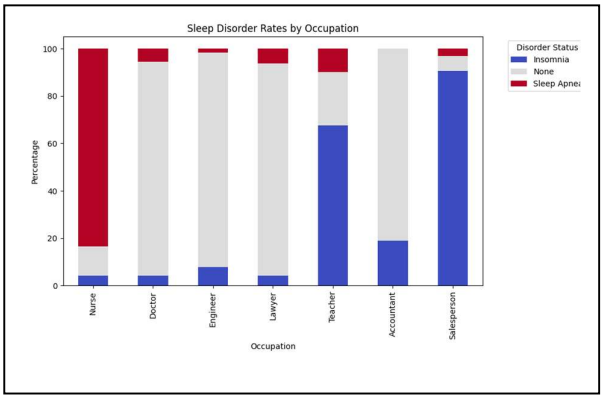

Fig. 4. Prevalence of Sleep Disorders by Occupation. Nurses show extreme rates of Sleep Apnea, while Salespersons are prone to Insomnia.

V. Discussion

A. *The Accuracy Paradox in Disease Prediction*

The results demonstrate the "Accuracy Paradox" often encountered in medical data science. The Random Forest model achieved 75% accuracy simply by exploiting the class imbalance, predicting the majority outcome ("Low Risk") for every case. While statistically accurate, this model is clinically useless as it fails to detect the disease. Conversely, the Logistic Regression model provided a more balanced, albeit less accurate, approach by prioritizing sensitivity (Recall). This highlights that for disease screening, metrics like Recall must take precedence over raw Accuracy, and data other augmentation techniques are necessary to train effective models on imbalanced datasets.

B. *Occupational Health Implications*

The clustering of Sleep Apnea among Nurses and Insomnia among Salespersons suggests that sleep health is heavily influenced by professional demands. The high rate of apnea in nurses is particularly concerning and warrants further investigation into the correlation with BMI and shift-work schedules. These findings suggest that public health interventions should be targeted: stress-management programs for Teachers/Salespersons and physical health screenings for Nurses.

C. *Limitations*

The primary limitation of this study is the dataset size (374 records for sleep analysis) and the significant class imbalance in the disease dataset. Additionally, "Stress Level" is a subjective, self-reported metric, which may introduce bias compared to physiological measurements like cortisol levels.

VI. Conclusion

This project analyzed the relationship between lifestyle habits, occupation, and chronic disease risk. The study confirms that Stress is a dominant predictor of poor sleep health, overshadowing the benefits of physical activity in high-stress populations. Furthermore, specific occupations (Nursing, Sales, Teaching) were identified as high-risk categories for sleep disorders.

From a predictive modeling perspective, the research highlights the dangers of relying on Accuracy as a sole performance metric. The superior accuracy of the Random

Forest model masked its complete failure to identify at-risk patients, underscoring the need for careful model evaluation using Recall and Precision in healthcare applications. Future work should focus on addressing data imbalance to create predictive models that are both accurate and clinically sensitive.

VII. References

[1] R. Liaqat, "Health & Lifestyle Dataset," Kaggle.com, 2025. [Online]. Available: https://www.kaggle.com/datasets/rehan497/health-lifestyle-dataset.

[2] Sultanul Ovi, "Sleep Disorder Diagnosis Dataset," Kaggle.com, 2025. [Online]. Available: https://www.kaggle.com/datasets/mdsultanulislamovi/sleep-disorder-diagnosis-dataset.