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

(COMP3125 Individual Project)

Brian Morillo
dept. name of organization

Abstract*—This electronic document is a “live” template and already defines the components of your paper [title, text, heads, etc.] in its style sheet. **CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract. (*provide a short abstract*)

Keywords*—*example1, example2, example3, example 4, example 5 (provide 3-5 keywords)

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)

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.