Health Data Analysis and Hypothesis Testing Report

1. Introduction

This project involves a comprehensive analysis of a health-related dataset containing 100,000 records and 49 features. The aim is to investigate the impact of various lifestyle and physiological factors on health status (target variable). The outcome of the project is to derive data-driven insights and present them through statistical tests and visual storytelling using Power BI.

2. Data Loading and Initial Overview

Loaded a CSV dataset with 100,000 entries and 49 features.
Features include physiological parameters (e.g., heart rate, glucose, BMI), lifestyle indicators (e.g., sleep, screen time, alcohol consumption), and the target variable (target) indicating health status (1 = healthy, 0 = unhealthy).

3. Data Preprocessing

Performed several critical steps to prepare the data for analysis:
- Handling Missing Values: Ensured all features had complete data by checking and imputing missing values if needed.
- Encoding Categorical Variables: Converted categorical features (like smoking_level, diet_type, etc.) into numerical format using Label Encoding.
- Data Scaling: Applied feature scaling using standardization (Z-score normalization) to ensure uniformity across features.
- Distribution Analysis: Plotted distributions of individual features to understand skewness, variance, and detect anomalies or outliers.
- Feature Importance: Trained a Random Forest model to rank features based on importance scores in predicting the target.
- Correlation Analysis: Used heatmaps and correlation matrices to analyze linear relationships between features and the health target.

4. Hypothesis Formulation and Statistical Testing

Health Data Analysis and Hypothesis Testing Report

Single Feature vs Target

Tested individual features against the health target using:

- T-Test (for continuous features like sugar intake, work hours, daily steps).
- Chi-Square Test (for categorical features like smoking level).

Results:

Feature	Test Used	P-value	Conclusion
sugar_intake	T-Test	0.8997	No significant relation
work_hours	T-Test	0.0006	Significant relation with health
daily_steps	T-Test	0.1263	No significant relation
smoking_level	Chi-Square	0.6406	No significant relation

5. Multi-Feature vs Target Testing

Grouped related lifestyle factors to test their joint significance with health:

- Sleep Hours, Screen Time, Stress Level (using logistic regression)
- Diet Quality & Physical Activity
- Substance Use (Alcohol + Smoking)
- All Lifestyle Factors Together

Used logistic regression models to test overall significance of each group:

- Group p-values and model summary showed that Sleep, Stress, Diet, and Physical Activity are significant predictors.
- Substance use alone did not show a statistically significant impact in isolation.

6. Conclusion

- Work hours had a statistically significant relationship with health.
- Sugar intake, smoking, and daily steps individually did not show significance, possibly due to complex interactions with other variables or insufficient sensitivity in the test.
- Combined lifestyle factors like sleep quality, stress level, and physical activity showed stronger predictive

power.

- Feature importance scores and correlation heatmaps revealed additional insights beyond statistical significance, supporting the value of including model-driven analysis in addition to hypothesis testing.