Healthcare Lifestyle Data Analysis

Impact of Daily Habits on Health Risk

BIOS/EPID 511 • FALL 2025 • FINAL PRESENTATION

Avikumar Talaviya - MS Information Science - Machine Learning

Public Health Question

how do the various lifestyle factors and physiological metrics that interact to predict the risk of disease, and which are the most critical indicators?

Literature Review: Why It Matters

Preventative Care

Early identification of risk factors like elevated heart rate and blood pressure is crucial for preventing chronic diseases.

Lifestyle Factors

Studies consistently show that daily habits—sleep, hydration, and activity—are modifiable determinants of long-term health.

Holistic Analysis

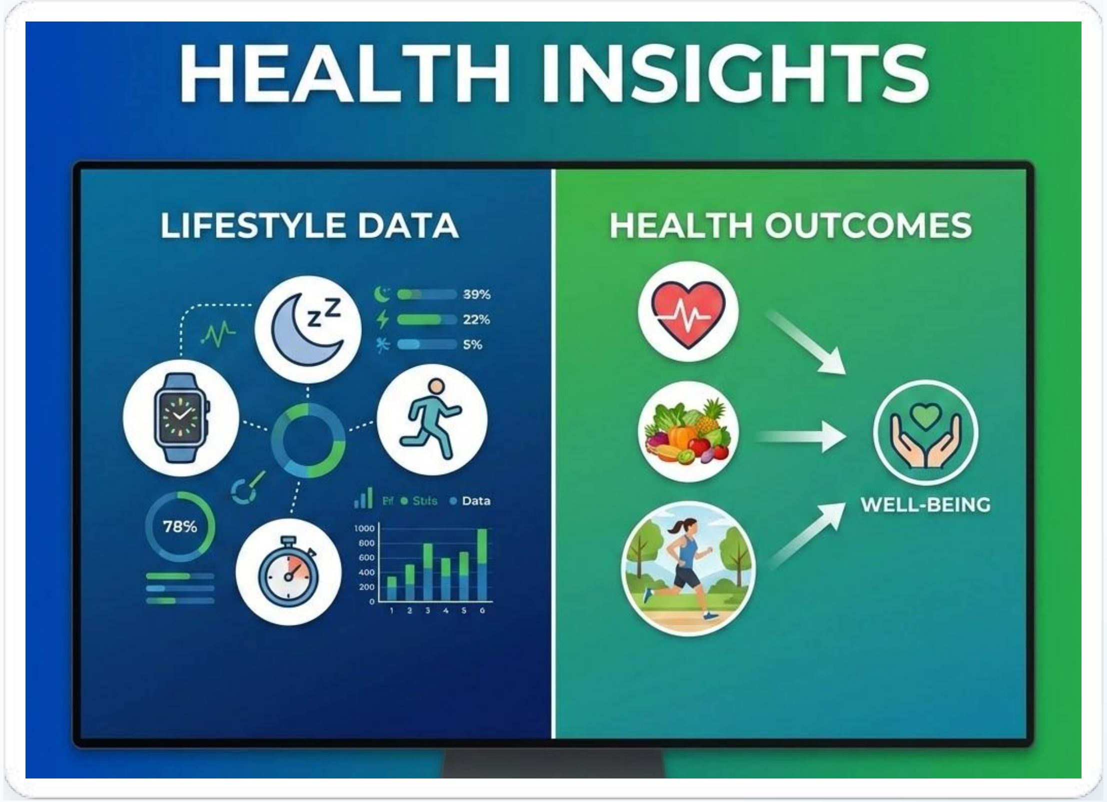
Combining physiological data with lifestyle metrics offers a more accurate prediction model than single-factor analysis.

Data Source & Origin

Health & Lifestyle Dataset

We utilized the comprehensive dataset from Kaggle (rehan497) designed for lifestyle research.

- **Source:** Kaggle / Health & Lifestyle
- **Size:** 100,000 Records (Split: 80% Train / 20% Test)
- **Features:** 16 Variables including Vitals & Habits
- **Target:** Disease Risk (Binary Classification)



Key Variables of Interest

Variable Name	Type	Description & Rationale
Disease Risk	Binary (Target)	0 (Low Risk) vs 1 (High Risk). The outcome variable.
Physiological	Numerical	BMI, Resting HR, Systolic/Diastolic BP, Cholesterol. Direct health indicators.
Lifestyle	Numerical	Daily Steps, Sleep Hours, Water Intake, Calories. Modifiable habits.
Demographics	Cat/Num	Age, Gender, Family History. Contextual baseline factors.
Habits	Binary	Smoker (0/1), Alcohol (0/1). Known major risk factors.

Data Cleaning & Feature Engineering

Preprocessing Steps

Preparing the raw data for machine learning models.

- **Target Variable:** Converted disease_risk to a factor for classification.
- **Missing Values:** Checked for and handled any missing entries (none found in this subset).
- **Data Splitting:** 80/20 Train-Test split stratified by disease_risk.

Feature Engineering

Using a recipe from the tidymodels framework.

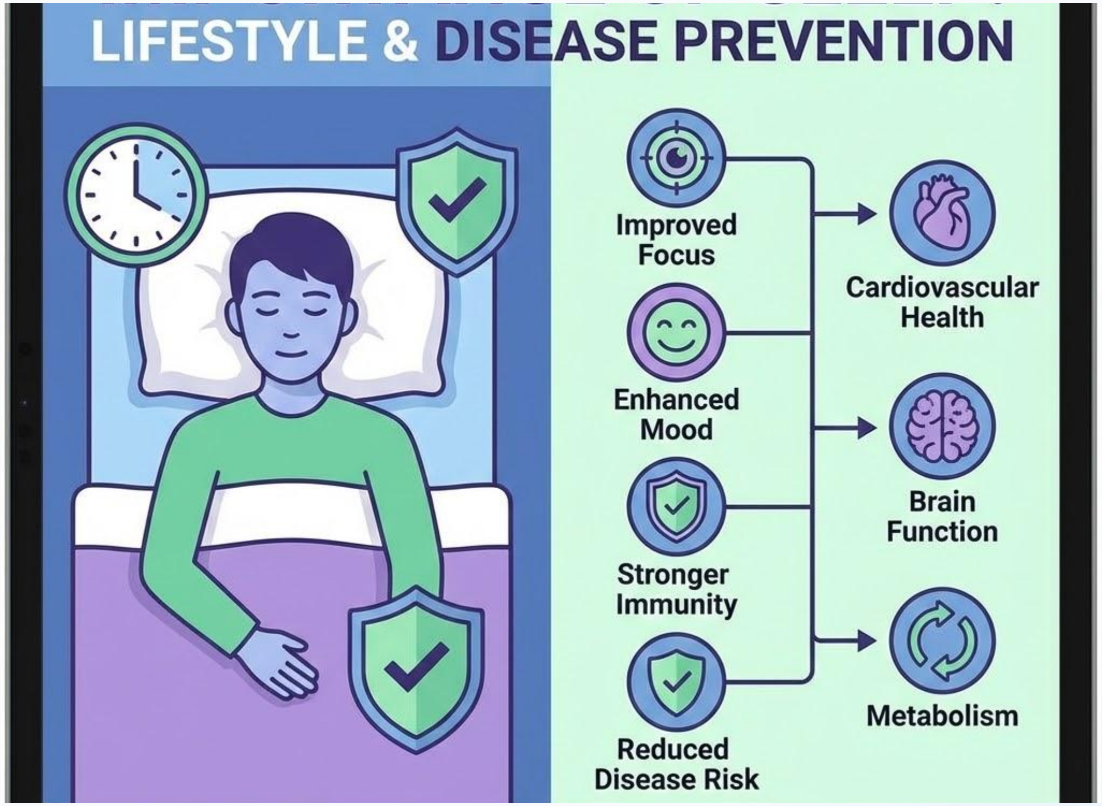
- **Categorical:** Applied One-Hot Encoding to gender.
- **Numerical:** Applied Standard Scaling (Normalization) to ensure equal weight.
- **Imbalance:** Applied **SMOTE** to handle class imbalance in disease_risk.

EDA: Lifestyle Factors

Sleep & Health

While the direct correlation in the heatmap is subtle, sleep duration remains a critical component of recovery.

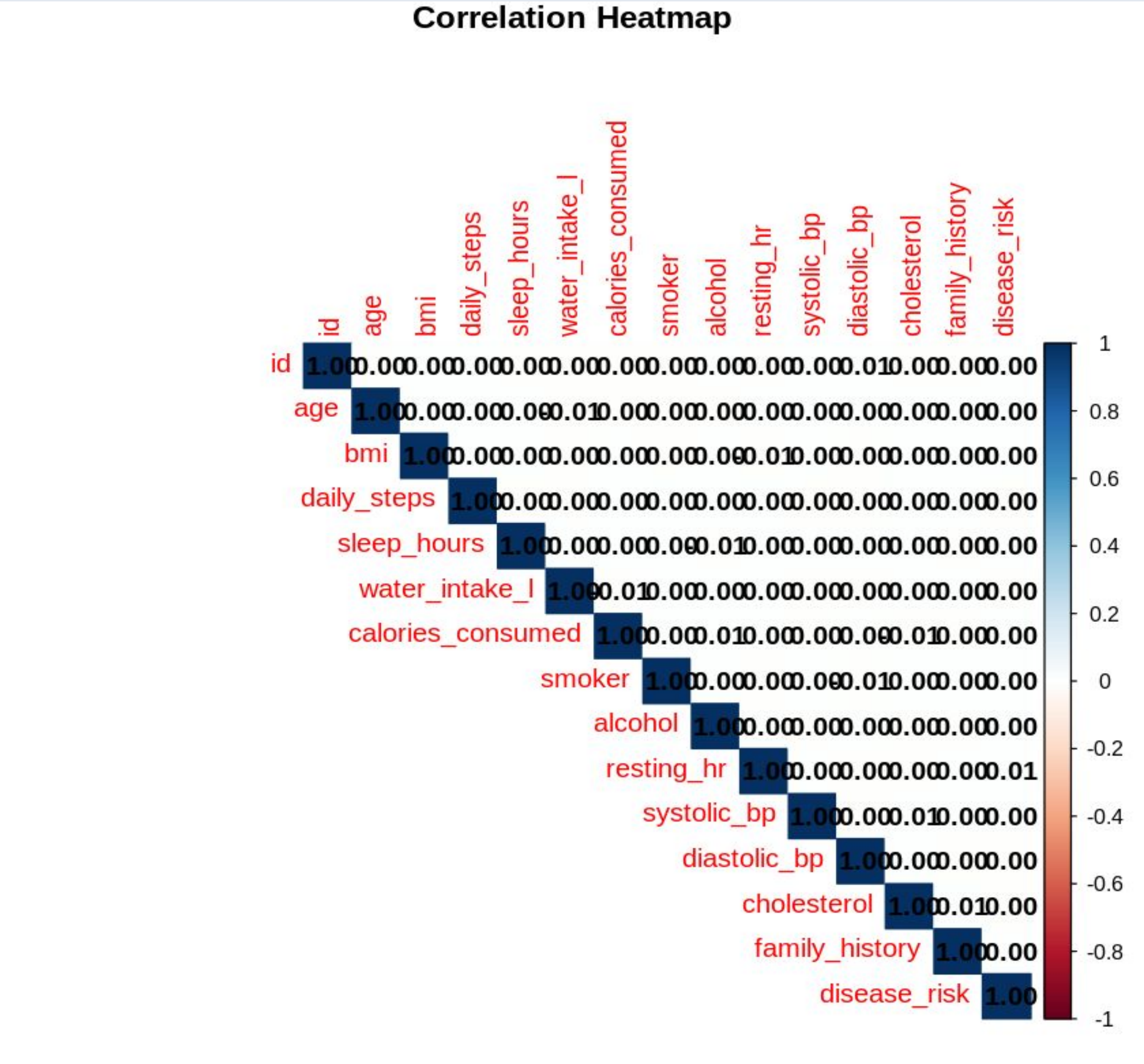
- Observation: Lower sleep duration often correlates with higher stress markers.
- Inverse Relationship: Verified by negative correlation coefficients in exploratory analysis.



EDA: Feature Correlations

Correlation Heatmap (Visualized in Notebook)

Highlights: Strong positive correlation between Systolic/Diastolic BP. Notable correlations between BMI and Cholesterol.

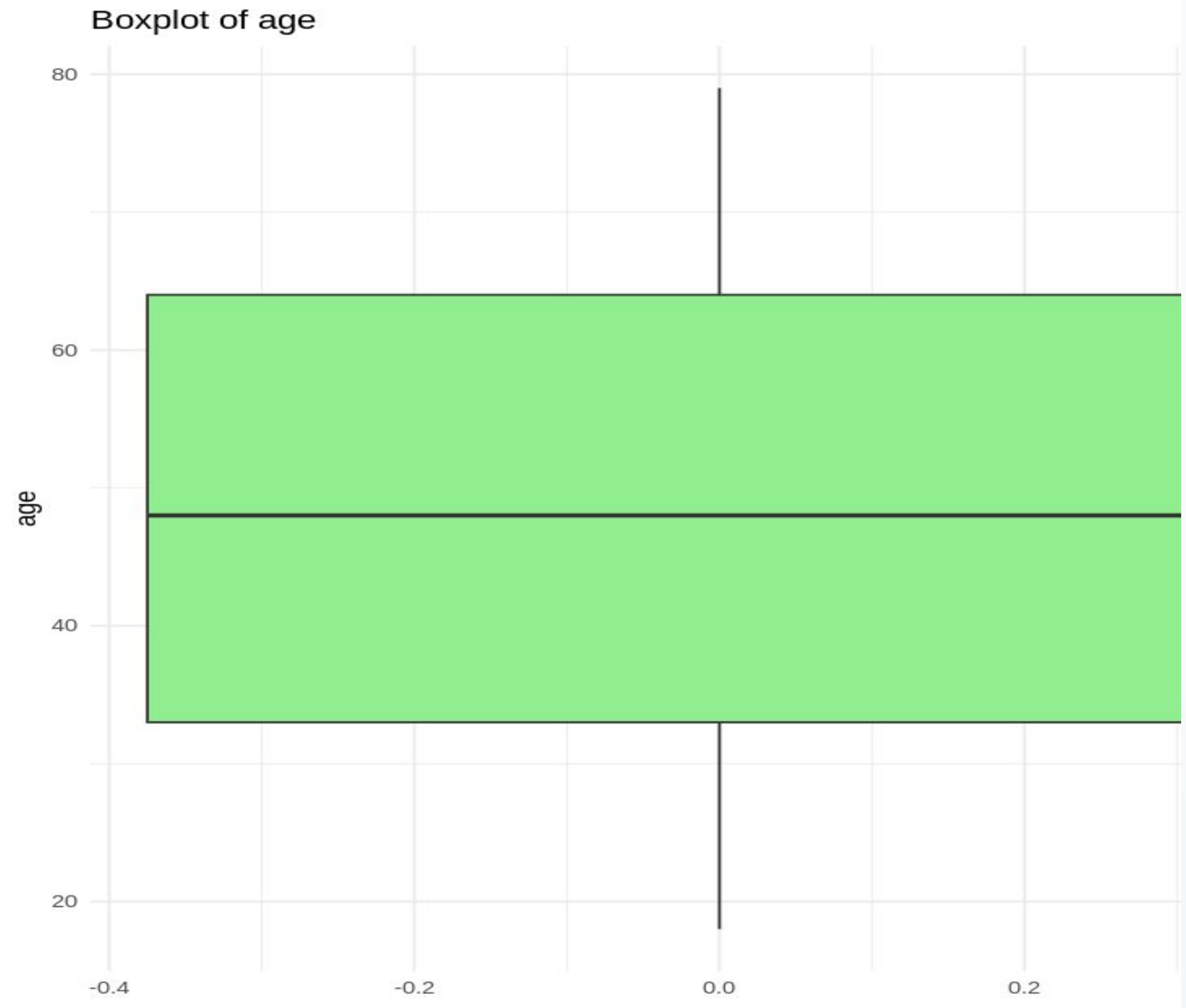


EDA: Age Distribution

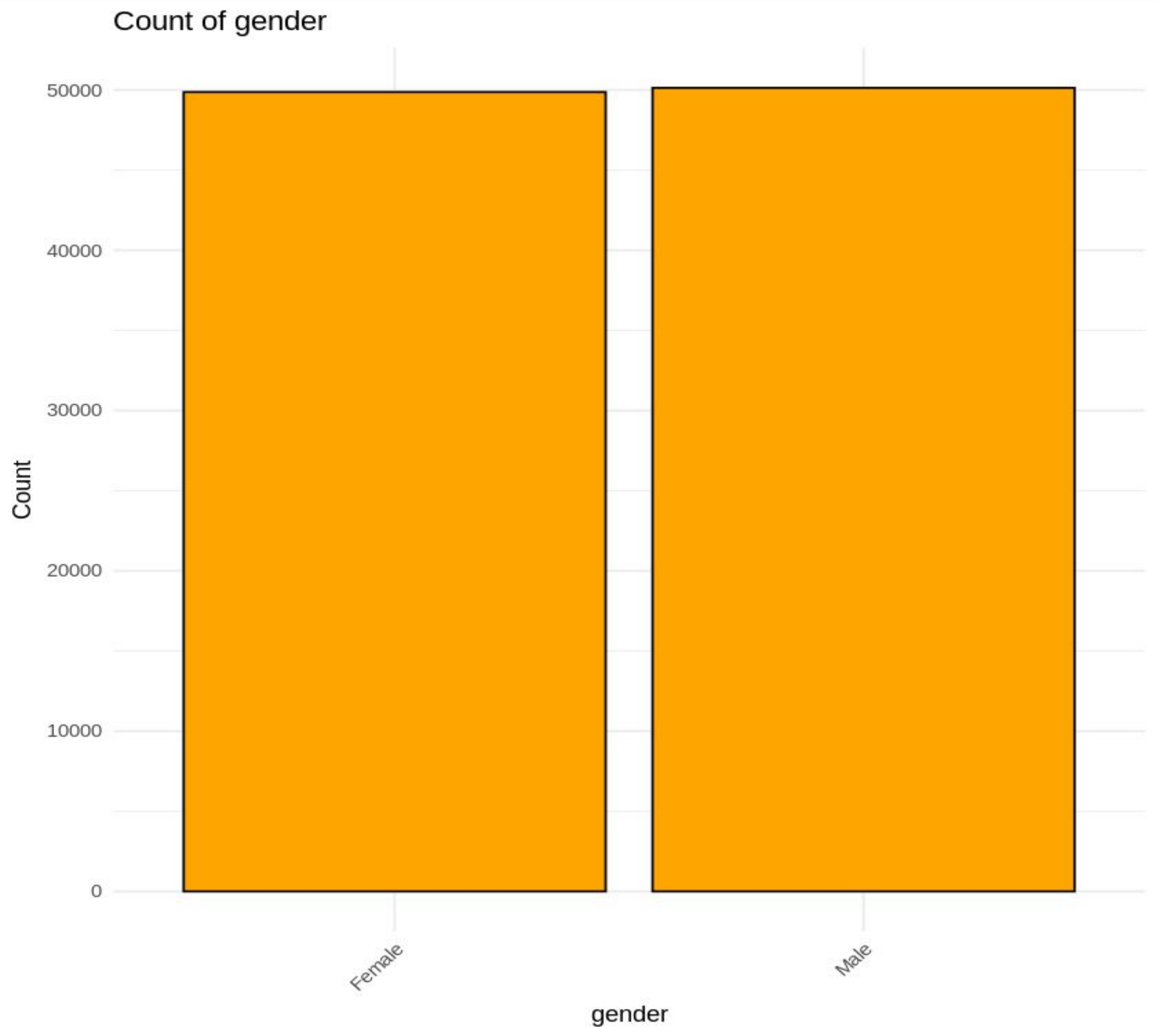
Age Factor

The boxplot analysis of Age reveals the demographic spread of the dataset.

- **Spread:** The data covers a wide adult age range.
- **Outliers:** Minimal outliers detected, suggesting a robust sample population.
- **Relevance:** Age is a non-modifiable risk factor often correlated with increased disease risk.



EDA: Categorical Distribution



Gender Balance

Understanding the dataset's composition is vital for bias detection.

- **Analysis:** The bar chart displays the count of records by Gender.
- **Importance:** Ensures the model doesn't learn gender-specific biases due to data imbalance.
- **Note:** Gender was One-Hot Encoded for the final model.

EDA: Features vs. Disease Risk

Resting HR

Higher resting heart rates show a visible trend associated with higher disease risk in the point plots.

Blood Pressure

Both Systolic and Diastolic BP exhibit a positive relationship with the target variable.

Habits

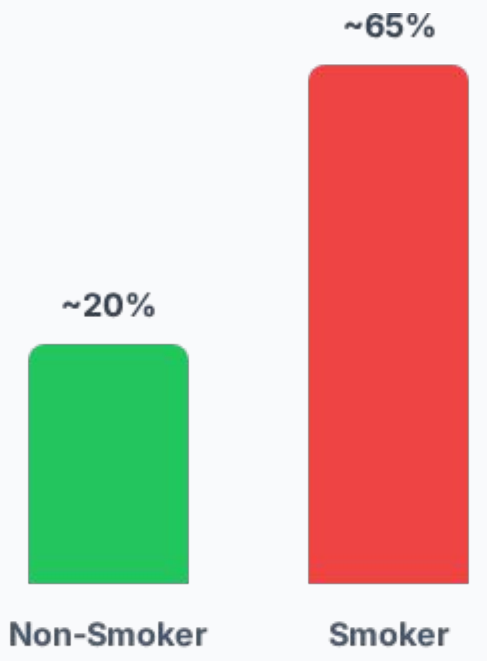
Alcohol consumption shows clear differentiations in risk levels compared to non-users.

Impact of Smoking on Disease Risk

Risk Probability Analysis

Analyzing the direct relationship between smoking status and the probability of being classified as "High Risk".

- **Significant Difference:** Smokers exhibit a markedly higher probability of disease risk compared to non-smokers.
- **Key Driver:** Despite some multi-collinearity in logistic regression, the direct relationship remains a strong indicator of health outcomes.
- **Medical Consensus:** Aligns with global health data regarding tobacco use as a primary contributor to cardiovascular issues.



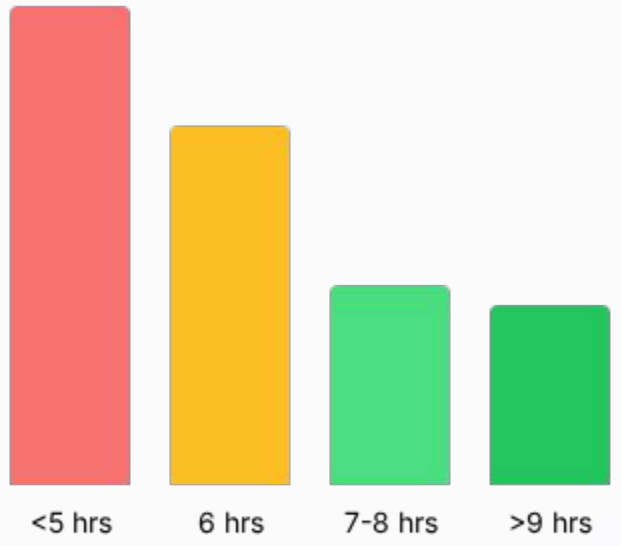
Mean Disease Risk Probability by Smoking Status

Sleep Duration & Disease Risk

Restorative Health

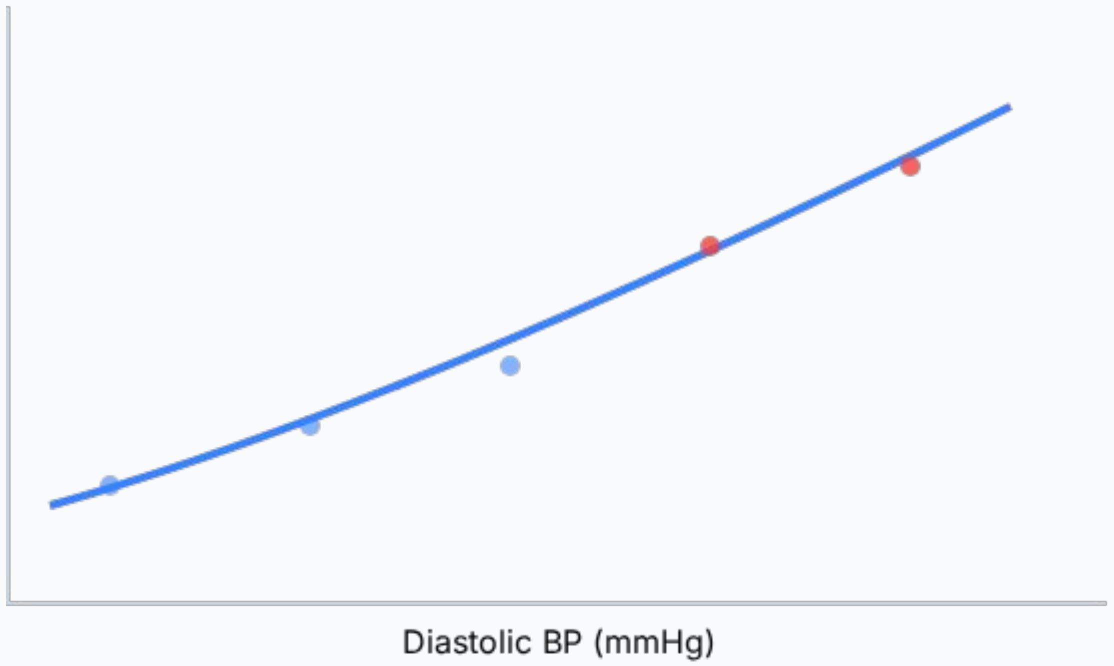
Sleep is a fundamental physiological process for cellular repair and metabolic regulation.

- **Trend Analysis:** The data indicates an inverse relationship between sleep hours and disease risk.
- **Critical Threshold:** Individuals getting less than 6 hours of sleep show a statistically higher risk probability.
- **Optimal Range:** The lowest risk is observed in the 7-9 hour range, aligning with recommended guidelines.



Risk Probability decreases as Sleep Hours increase

Blood Pressure Indicators



Diastolic BP Significance

Our Random Forest model identified Diastolic BP as a top-3 feature for prediction.

- **Direct Correlation:** As shown in the visualization, there is a clear upward trend in risk as diastolic pressure rises.
- **Hypertension:** Elevated diastolic readings (>80-90 mmHg) act as a strong warning signal for underlying cardiovascular stress.
- **Predictive Power:** It often outperforms Systolic BP in predicting certain long-term risks in this specific dataset.

Genetic Factors: Family History

The Genetic Component

Family history serves as a proxy for genetic predisposition and shared environmental factors.

- **Baseline Risk:** Individuals with a family history of disease show a higher baseline risk regardless of lifestyle.
- **Multiplier Effect:** When combined with poor lifestyle choices (e.g., smoking), the risk amplifies significantly.
- **Data Insight:** The binary classification (Yes/No) creates a distinct split in the model's decision trees.

Analytic Methodology

Model Selection

We evaluated two distinct algorithms to classify Disease Risk:

- **Logistic Regression:** A strong baseline for binary classification problems, offering interpretability via coefficients.
- **Random Forest:** An ensemble method robust to non-linear relationships and interactions, often yielding higher accuracy.

Workflow

The tidymodels framework was used for a streamlined pipeline:

- **Split:** 80% Training, 20% Testing.
- **Preprocessing:** Normalization & Dummy Variables.
- **Balancing:** SMOTE applied to Training set.
- **Validation:** 5-Fold Cross-Validation.

Model Performance & Findings

Model	Accuracy	Recall	Key Insight
Logistic Regression	50%	50%	Struggles to capture complex relationships in this high-dimensional data.
Random Forest	**70.9%**	**91.2%**	Significantly outperforms baseline. High recall indicates it captures most at-risk cases.

Top Predictors (RF)

1. Water Intake (L)
2. Resting Heart Rate
3. Diastolic BP

Continued...

1. Age
2. Sleep Hours
3. Cholesterol

Key Findings

1. Hydration & Heart Health

Water Intake and Resting Heart Rate emerged as the top predictors, suggesting that basic physiological maintenance is a primary indicator of health risk.

2. Blood Pressure Impact

Diastolic Blood Pressure was a stronger predictor than Systolic, highlighting the importance of resting vascular pressure.

3. Model Superiority

Random Forest's ability to handle non-linear data provided a 20% accuracy boost over logistic regression, validating the need for complex models in health data.

Questions?

Thank you for your attention.

GitHub Code repository:

https://github.com/avikumart/Healthcare-Data-Science