Healthcare Risk Factors Prediction Project

Leveraging machine learning for early identification of high-risk individuals in healthcare.

The Challenge: Manual Risk Assessment

Healthcare professionals face challenges in quickly evaluating multiple risk factors for conditions like diabetes, hypertension, and obesity.

- Time-consuming
- Prone to human error
- Does not scale for large patient volumes

Our objective is to build a data-driven approach to predict health risks using available patient attributes.

Project Objectives

01

Data Analysis

Analyse large healthcare datasets with demographic, clinical, and lifestyle features.

02

Preprocessing

Perform thorough data cleaning, preprocessing, and exploratory analysis.

03

ML Model Training

Train an ML model to classify patient medical conditions.

04

Insight Generation

Identify influential predictors and derive actionable insights for clinical decisions.

Dataset Overview

Our dataset comprises ~30,000 records with over 20 attributes, covering various patient aspects.

Demographics	Age, Gender
Clinical	Glucose, Blood Pressure, HbA1c, BMI, Cholesterol, Oxygen Saturation
Lifestyle	Diet Score, Physical Activity, Smoking, Alcohol Intake, Sleep Hours, Stress Level
Other	Length of Stay, Family History
Target Variable	Medical Condition (multiple classes)

Data Cleaning & Preprocessing

Key Steps Performed:

- Handled missing values (median/mode imputation)
- Removed noise columns (e.g., random_notes)
- Corrected data types

Encoding & Scaling:

- Label encoding for target variable
- One-hot encoding for categorical features
- Standard scaling for numeric values

Exploratory Data Analysis: Key Insights

Demographic Trends

Older individuals show higher blood pressure and hypertension risk. Age correlates with stress and sleep variations.

Clinical Findings

High HbA1c and Glucose strongly indicate diabetes. Elevated BMI aligns with obesity and cholesterol issues. Blood Pressure links to hypertension.

Lifestyle Insights

Low physical activity links to multiple chronic conditions. Poor diet score associates with higher triglyceride levels.

Modeling Approach: Random Forest Classifier

Selected for its excellent performance with heterogeneous datasets, ability to capture non-linear patterns, and built-in interpretability.

Training Details:

- 80/20 train–test split
- Preprocessing pipeline: Scaling + One-Hot Encoding
- Evaluation metrics: Accuracy, F1 Score, ROC AUC

Model Performance & Top Features

Performance Metrics:

- Accuracy: ~78%
- Macro F1 Score: ~0.73
- ROC AUC (macro): ~0.92

Best-performing classes: Diabetes, Hypertension, Healthy.

Top Feature Importances:

- HbA1c – 0.125
- Glucose – 0.108
- Length of Stay – 0.107
- Physical Activity – 0.083
- Blood Pressure – 0.083
- Age – 0.081
- BMI – 0.073

Key Findings: Drivers of Health Risks

Metabolic Markers

HbA1c and Glucose are primary predictors of diabetes.

Clinical Indicators

BMI, BP, and Age are critical for hypertension and obesity.

Lifestyle Choices

Significantly influence health risks.

Sleep & Stress

Contribute more than gender-based differences.

Business Impact & Final Summary

Impact:

- Early Detection: Identify at-risk patients sooner.
- Operational Efficiency: Reduce clinical workload.
- Financial Impact: Aid insurers in risk evaluation.
- Improved Outcomes: Support preventive interventions.

Summary:

Clinical measurements (HbA1c, Glucose, BP, BMI, Oxygen Saturation) are primary drivers. Lifestyle patterns (Activity, Diet, Sleep, Stress) are secondary risk indicators. Gender plays a minor role.

These insights help healthcare providers prioritize metabolic screening, design preventive interventions, and implement patient triage models.