Norwegian University of Science and Technology

TDT4259 - Applied Data Science

Group Assignment

A Data-Driven Approach to Lifestyle Risk Prediction Using Clinical Biomarkers: Smoking and Drinking Analysis

Authors:

Marco Prosperi - 151613
Andrea Richichi - 151790
Tizita Belachew Tamirat - 128549
Gianluigi Vazzoler - 152698

Fall, 2025

Contents

1 Introduction

1.1 Background and Context

Lifestyle-related health risks, particularly smoking and alcohol consumption, represent significant public health challenges worldwide. These behaviors are associated with numerous chronic diseases, including cardiovascular disease, cancer, liver disease, and respiratory conditions. Traditional approaches to identifying at-risk individuals often rely on self-reported surveys, which can be subject to underreporting due to social desirability bias or lack of awareness.

Recent advances in preventive medicine have highlighted the potential of using objective biomedical data to identify individuals with risky lifestyles before serious health complications arise. Clinical biomarkers—measurable indicators derived from routine health examinations—provide objective evidence of physiological changes associated with smoking and drinking behaviors. These markers include liver function enzymes (AST, ALT), cholesterol levels, blood pressure measurements, and anthropometric data.

The ability to predict lifestyle risks from biomedical data has important implications for healthcare providers. Early identification of at-risk individuals enables targeted interventions, personalized health counseling, and preventive care strategies. This proactive approach can reduce the burden of lifestyle-related diseases on healthcare systems and improve patient outcomes.

1.2 Problem Definition

Despite the availability of extensive biomedical data from routine health screenings, healthcare providers often lack effective tools to systematically identify individuals with risky lifestyle behaviors. The challenge lies in developing predictive models that can:

1. Accurately classify individuals based on smoking and alcohol consumption patterns using objective clinical measurements
2. Provide interpretable predictions that healthcare professionals can act upon (NOTE THAT THIS PART IS MISSING IN THE CODE, WE NEED TO ADD IT)
3. Identify the most relevant biomarkers that signal lifestyle-related health risks

The primary research question addressed in this project is: can machine learning models effectively predict smoking status and alcohol consumption levels using clinical biomarkers, and which biomarkers are most predictive of these lifestyle behaviors?

1.3 Motivation

The motivation for this project stems from multiple factors. First, the global burden of lifestyle-related diseases continues to rise, with smoking and excessive alcohol consumption being among the leading preventable causes of mortality. According to the World Health Organization, tobacco use kills more than 8 million people annually, while harmful use of alcohol results in approximately 3 million deaths each year.

Second, traditional screening methods based on self-reporting have well-documented limitations in accuracy and reliability. Individuals may underreport their consumption due to stigma, denial, or simple forgetfulness. An objective, data-driven approach using readily available biomedical measurements could overcome these limitations.

Third, routine health examinations already collect extensive biomedical data, but this information is often underutilized for predictive purposes. By developing machine learning models that leverage this existing data, healthcare providers can implement risk prediction systems without requiring additional invasive tests or costly procedures.

Finally, early identification of at-risk individuals creates opportunities for preventive interventions that are more effective and less costly than treating advanced disease. This aligns with the broader shift in healthcare toward preventive and personalized medicine

1.4 Team and Roles

1.5 Report Outline

This report is structured to provide a comprehensive overview of the predictive modeling approach for lifestyle risk identification. The organization follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, ensuring systematic progression from problem understanding to deployment recommendations.
Section 2: Project Objectives and Scope defines the specific goals of the analysis, project limitations, and the broader context within preventive healthcare.
Section 3: Data Strategy and Management describes the dataset sources, preprocessing procedures, quality assurance measures, and the tools employed for analysis.
Section 4: Methods and Modeling details the analytical approach, including feature engineering techniques, algorithm selection, model training procedures, and interpretability considerations.
Section 5: Analysis and Results presents the exploratory data analysis, model performance metrics, feature importance analysis, and interpretation of findings.
Section 6: Recommendations and Deployment provides actionable recommendations for healthcare providers, discusses deployment strategies, and outlines an implementation roadmap.
Section 7: Monitoring and Maintenance establishes key performance indicators, monitoring procedures, and risk management strategies for maintaining model effectiveness over time.
Section 8: Conclusion summarizes the key findings, discusses limitations, and suggests directions for future research.

2 Project Objectives and Scope

2.1 Project Goals

2.2 Scope and Limitations

2.3 Broader Context and Relevance

3 Data Strategy and Management

3.1 Data Sources and Description

3.2 Data Collection and Preprocessing

3.3 Data Quality and Cleaning

3.4 Data Strategy Framework

3.5 Tools and Technologies Used

4 Methods and Modeling

4.1 Analytical Approach

4.2 Modeling Techniques and Algorithms

4.3 Model Training and Validation