Prediction of Diabetes Risk according to Social & Health Indicators

Supervised Regression

Alexandru Profir[1]
[1]Babeș-Bolyai University

October 31, 2025

1 Problem definition

The purpose of this particular approach is to develop a framework capable of predicting the *diabetes risk score* based on a set of individual health indicators. This predictive capability can assist in preventive healthcare and early detection, helping individuals and practitioners make informed medical decisions.

2 Problem specification

This section presents the specification of the problem, by analyzing the input and output features, as well as pre and post conditions applied to them.

2.1 Input Data

The input data consists of $100,000$ patient profiles with features based on demographics, lifestyle habits, family history, and clinical measurements that are well-established indicators of diabetes risk. The data is generated using statistical distributions from real-world medical research, ensuring *privacy preservation* while reflecting realistic health patterns.

Table 1 presents a description of all input features used in this problem. This table showcases the name of the feature, the type of the feature, a brief description, and the domain (*value range*) of the presented feature. As confirmed by the creators of this dataset, all values fall within realistic medical ranges, matching realistic population health patterns, being a viable option when exploring lifestyle and clinical health patterns.

Feature	Type	Description	Domain
patient_id	Integer	Unique patient identifier	$\mathbb{N} \cap [1, 100000]$
age	Integer	Age of patient in years	$\mathbb{N} \cap [18, 90]$
gender	Categorical	Patient gender	{Male, Female, Other}
ethnicity	Categorical	Ethnic background	{White, Hispanic, Black, Asian, Other}
education_level	Categorical	Highest completed education	{No formal, Highschool, Graduate, Postgraduate}
income_level	Categorical	Income category	{Low, Medium, High}
employment_status	Categorical	Employment type	{Employed, Unemployed, Retired, Student}
smoking_status	Categorical	Smoking behavior	{Never, Former, Current}
alcohol_consumption_per_week	Float	Drinks consumed per week	$\mathbb{R}_+ \cap [0, 30]$
physical_activity_minutes_per_week	Integer	Physical activity (weekly minutes)	$\mathbb{N} \cap [1, 100000]$
diet_score	Integer	Diet quality (higher = healthier)	$\mathbb{N} \cap [1, 10]$
sleep_hours_per_day	Float	Average daily sleep hours	$\mathbb{R}_+ \cap [3, 12]$
screen_time_hours_per_day	Float	Average daily screen time hours	$\mathbb{R}_+ \cap [0, 12]$
family_history_diabetes	Integer	Family history of diabetes	$\mathbb{N} \cap \{0, 1\}$
hypertension_history	Integer	Hypertension history	$\mathbb{N} \cap \{0, 1\}$
cardiovascular_history	Integer	Cardiovascular history	$\mathbb{N} \cap \{0, 1\}$
bmi	Float	Body Mass Index (kg/m^2)	$\mathbb{R}_+ \cap [15, 45]$
waist_to_hip_ratio	Float	Waist-to-hip ratio	$\mathbb{R}_+ \cap [0.7, 1.2]$
systolic_bp	Integer	Systolic blood pressure ($mmHg$)	$\mathbb{N} \cap [90, 180]$
diastolic_bp	Integer	Diastolic blood pressure ($mmHg$)	$\mathbb{N} \cap [60, 120]$
heart_rate	Integer	Resting heart rate (bpm)	$\mathbb{N} \cap [50, 120]$
cholesterol_total	Float	Total cholesterol (mg/dL)	$\mathbb{R}_+ \cap [120, 30]$
hdl_cholesterol	Float	HDL cholesterol (mg/dL)	$\mathbb{R}_+ \cap [20, 100]$
ldl_cholesterol	Float	LDL cholesterol (mg/dL)	$\mathbb{R}_+ \cap [50, 200]$
triglycerides	Float	Triglycerides (mg/dL)	$\mathbb{R}_+ \cap [50, 500]$
glucose_fasting	Float	Fasting glucose (mg/dL)	$\mathbb{R}_+ \cap [70, 250]$
glucose_postprandial	Float	Post-meal glucose (mg/dL)	$\mathbb{R}_+ \cap [90, 350]$
insulin_level	Float	Blood insulin level ($\mu U/mL$)	$\mathbb{R}_+ \cap [2, 50]$
hba1c	Float	HbA1c (%)	$\mathbb{R}_+ \cap [4, 14]$

Table 1: Input data description

2.2 Output Data

This problem proposes to solve the issue of predicting the diabetes risk score, based on social & health indicators. The *diabetes risk score* is a computed integer value based on the proposed social & health indicators, ranging between $[0, 100]$. Based on the proposed features, the solution will provide a *diabetes risk score* within the same range, indicating if a patient has a risk to diabetes or not.

3 Learning Task

This section presents the formalization of the learning task regarding the presented problem, accounting the features proposed to solve the problem and the goal provided by the solution. The specification of the learning task is presented below:

- **Task**: prediction of the diabetes risk score;
- **Performance**: quality of the resulted prediction, represented by how accurate is the prediction with respect to the real value;
- **Experience**: prediction of the diabetes risk score on a large set of labeled instances;

The above described task will be solved through an empirical analysis, driven by data patterns rather than explicit programming.