Diabetes Determinator

Project Proposal

Mihail Kenarov

Table of Contents

[1. Introduction 3](#_Toc169476728)

[‘What?’ 3](#_Toc169476729)

[‘Why?’ 3](#_Toc169476730)

[‘Who?’ 3](#_Toc169476731)

[‘When?’ 3](#_Toc169476732)

[‘How?’ 3](#_Toc169476733)

[2. Domain Understanding 3](#_Toc169476734)

[3. Data sourcing 3](#_Toc169476735)

[3.1 Objective 3](#_Toc169476736)

[3.2 Data requirements 4](#_Toc169476737)

[3.3 Data sources 4](#_Toc169476738)

[3.4 Data Legality and Ethics 4](#_Toc169476739)

[3.5 Data Diversity 4](#_Toc169476740)

[3.6 Version Control 7](#_Toc169476741)

[3.7 Iterative Process 7](#_Toc169476742)

[4. Analytic Approach 8](#_Toc169476743)

[5. Bibliography 8](#_Toc169476744)

# 1. Introduction

## ‘What?’

## ‘Why?’

## ‘Who?’

## ‘When?’

## ‘How?’

# 2. Domain Understanding

# 3. Data sourcing

## 3.1 Objective

The goal of this project is to determine if a certain person, based on different kinds of data is sustainable to developing either kind of diabetes. This will be done with

## 3.2 Data requirements

To achieve such a system, we will need different kinds of mostly medical data, related to the person. Not only that be we would probably need some kind of lifestyle data since it is something that should be taken into consideration, when trying to predict something of this sort. General data would also be nice, something like the gender and age as an example to start with. The BMI also could also be useful since it can be affected by/affect diabetes. Sport activities may be somehow related, since there is research that shows that people who do more physical activity are less likely to develop the illness. Also, blood sugar levels should be of great help when it comes to determining the possibly of development.

* ***Age – Number***
* ***Gender – Male or Female – 0/1***
* ***BMI – Number***
* ***Smoking – yes/no – 1/0***
* ***Activity – yes/No – 1/0***
* ***Blood Sugat Levels - Number***

## 3.3 Data sources

Publicly available dataset:

<https://www.kaggle.com/datasets/rabieelkharoua/diabetes-health-dataset-analysis>

This dataset comprises extensive health information for 1,879 patients, each uniquely identified with IDs ranging from 6000 to 7878. It includes demographic details, lifestyle factors, medical history, clinical measurements, medication usage, symptoms, quality of life scores, environmental exposures, and health behaviors. Each patient is linked to a confidential primary care doctor, ensuring privacy and confidentiality.

## 3.4 Data Legality and Ethics

The data provenance for this dataset is confidential to protect the identities and privacy of the patients involved.

## 3.5 Data Diversity

***Patient Information – All the types of information we will be using.***

* Patient ID
* Demographic Details
* Lifestyle Factors
* Medical History
* Clinical Measurements
* Medications
* Symptoms and Quality of Life
* Environmental and Occupational Exposures
* Health Behaviors
* Diagnosis Information

***Patient Information***

Patient ID

* PatientID: A unique identifier assigned to each patient (6000 to 7878).

Demographic Details

* Age: The age of the patients ranges from 20 to 90 years.
* Gender: Gender of the patients, where 0 represents Male and 1 represents Female.
* Ethnicity: The ethnicity of the patients, coded as follows:
  + 0: Caucasian
  + 1: African American
  + 2: Asian
  + 3: Other
* SocioeconomicStatus: The socioeconomic status of the patients, coded as follows:
  + 0: Low
  + 1: Middle
  + 2: High
* EducationLevel: The education level of the patients, coded as follows:
  + 0: None
  + 1: High School
  + 2: Bachelor's
  + 3: Higher

***Lifestyle Factors***

* BMI: Body Mass Index of the patients, ranging from 15 to 40.
* Smoking: Smoking status, where 0 indicates No and 1 indicates Yes.
* AlcoholConsumption: Weekly alcohol consumption in units, ranging from 0 to 20.
* PhysicalActivity: Weekly physical activity in hours, ranging from 0 to 10.
* DietQuality: Diet quality score, ranging from 0 to 10.
* SleepQuality: Sleep quality score, ranging from 4 to 10.

***Medical History***

* FamilyHistoryDiabetes: Family history of diabetes, where 0 indicates No and 1 indicates Yes.
* GestationalDiabetes: History of gestational diabetes, where 0 indicates No and 1 indicates Yes.
* PolycysticOvarySyndrome: Presence of polycystic ovary syndrome, where 0 indicates No and 1 indicates Yes.
* PreviousPreDiabetes: History of previous pre-diabetes, where 0 indicates No and 1 indicates Yes.
* Hypertension: Presence of hypertension, where 0 indicates No and 1 indicates Yes.

***Clinical Measurements***

* SystolicBP: Systolic blood pressure, ranging from 90 to 180 mmHg.
* DiastolicBP: Diastolic blood pressure, ranging from 60 to 120 mmHg.
* FastingBloodSugar: Fasting blood sugar levels, ranging from 70 to 200 mg/dL.
* HbA1c: Hemoglobin A1c levels, ranging from 4.0% to 10.0%.
* SerumCreatinine: Serum creatinine levels, ranging from 0.5 to 5.0 mg/dL.
* BUNLevels: Blood Urea Nitrogen levels, ranging from 5 to 50 mg/dL.
* CholesterolTotal: Total cholesterol levels, ranging from 150 to 300 mg/dL.
* CholesterolLDL: Low-density lipoprotein cholesterol levels, ranging from 50 to 200 mg/dL.
* CholesterolHDL: High-density lipoprotein cholesterol levels, ranging from 20 to 100 mg/dL.
* CholesterolTriglycerides: Triglycerides levels, ranging from 50 to 400 mg/dL.

***Medications***

* AntihypertensiveMedications: Use of antihypertensive medications, where 0 indicates No and 1 indicates Yes.
* Statins: Use of statins, where 0 indicates No and 1 indicates Yes.
* AntidiabeticMedications: Use of antidiabetic medications, where 0 indicates No and 1 indicates Yes.

***Symptoms and Quality of Life***

* FrequentUrination: Presence of frequent urination, where 0 indicates No and 1 indicates Yes.
* ExcessiveThirst: Presence of excessive thirst, where 0 indicates No and 1 indicates Yes.
* UnexplainedWeightLoss: Presence of unexplained weight loss, where 0 indicates No and 1 indicates Yes.
* FatigueLevels: Fatigue levels, ranging from 0 to 10.
* BlurredVision: Presence of blurred vision, where 0 indicates No and 1 indicates Yes.
* SlowHealingSores: Presence of slow-healing sores, where 0 indicates No and 1 indicates Yes.
* TinglingHandsFeet: Presence of tingling in hands or feet, where 0 indicates No and 1 indicates Yes.
* QualityOfLifeScore: Quality of life score, ranging from 0 to 100.

***Environmental and Occupational Exposures***

* HeavyMetalsExposure: Exposure to heavy metals, where 0 indicates No and 1 indicates Yes.
* OccupationalExposureChemicals: Occupational exposure to harmful chemicals, where 0 indicates No and 1 indicates Yes.
* WaterQuality: Quality of water, where 0 indicates Good and 1 indicates Poor.

***Health Behaviors***

* MedicalCheckupsFrequency: Frequency of medical check-ups per year, ranging from 0 to 4.
* MedicationAdherence: Medication adherence score, ranging from 0 to 10.
* HealthLiteracy: Health literacy score, ranging from 0 to 10.

***Diagnosis Information (Target Variable)***

* Diagnosis: Diagnosis status for Diabetes, where 0 indicates No and 1 indicates Yes.

## 3.6 Version Control

A history of the project and the processing of the data will be kept on a Git repository in case of an incident and as a way for a version control.

The link to the git repository:

* <https://git.fhict.nl/I509460/video-game-reommendation.git>

## 3.7 Iterative Process

The model will be check for its performance after the iterations. Depending on the results, if they are found are satisfactory or not, changes will be made either to the model or the preparation of the data.

# 4. Analytic Approach

Currently this is a

# 5. Bibliography