Master Computer Science

Type 2 Diabetes in Context

Predicting Type 2 Diabetes Prevalence Through Machine Learning: Correlations with Social Networks, Lifestyle, Socioeconomics and Living Environment

Name: Ir. Simone Frederika Smits
Student ID: s2710676

Date: 23/01/2025

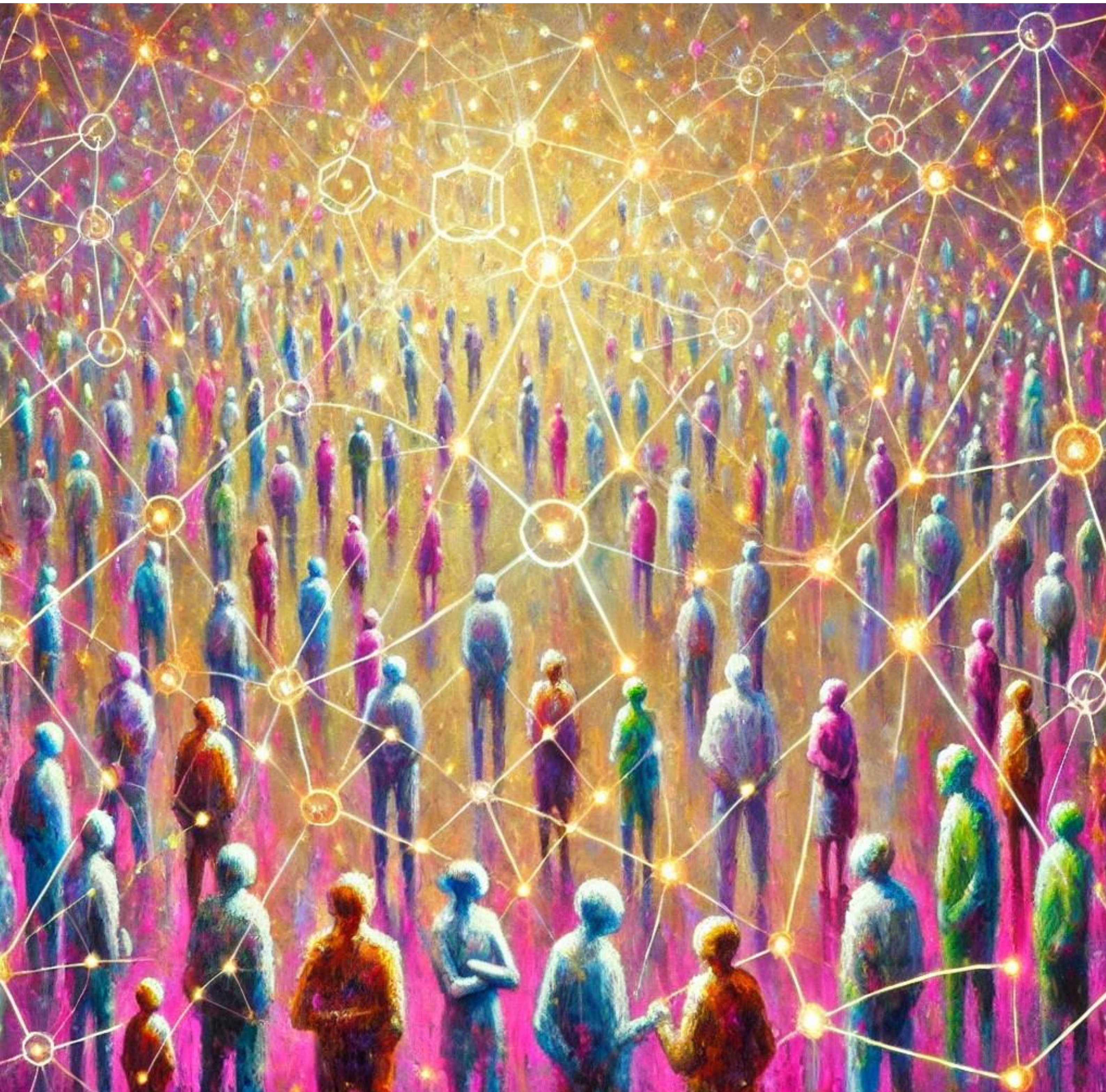
Specialisation: Bioinformatics

1st supervisor: Prof. dr. ir. Wessel Kraaij
2nd supervisor: Dr. Jeroen Pronk

Company Supervisors:
Prof. dr. Marjolijn Das (CBS)
Dr. ir. Edwin de Jonge (CBS)
Dr. Tanja Krone (TNO)

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands



Social Networks
Created using Dall E 3 – AI Image Generator

Acknowledgements

Simone Smits
Delft, January 2025

First and foremost, I would like to express my sincere gratitude to TNO and Centraal Bureau voor de Statistiek (CBS) for providing me with the opportunity to conduct this research within their organizations. I am super thankful that they gave me the opportunity to work with their project teams. It is thanks to the groundwork in the persoonsnetwerken dataset and the availability of all the other CBS datasets that this research was made possible. I would also like to acknowledge the expertise of TNO in the area of (preventive public) health, which provided me with a deeper understanding of the subject.

I would particularly like to thank Edwin de Jonge and Marjolijn Das from CBS and Jeroen Pronk and Tanja Krone from TNO, for their valuable supervision throughout my thesis project. I truly enjoyed the research process, and much of that enjoyment stemmed from your expertise and enthusiastic support. I am grateful for how you actively involved me in the project, shared your knowledge, and consistently took the time to review the (extensive) drafts of my thesis. Throughout it all, you remained encouraging and supportive, which kept me enthusiastic.

Also, I would like to thank everyone in the persoonsnetwerken group at CBS and the team at TNO for their input and support. It helped a lot to receive input from so many different angles.

A special thanks to Barteld Braaksma and Tanja Krone for their direct enthusiasm to help find an interesting and challenging thesis internship for me. You help was crucial in connecting me with the right people within the organizations.

Furthermore, I would like to thank Wessel Kraaij for his supervision from Universiteit Leiden. I have really enjoyed working together with you. Every thesis conversation was filled with enthusiasm, fresh insights, and valuable feedback. In addition to the valuable scientific insights, I also really appreciated the shared interest in leveraging research to have social impact.

Lastly, a big thank you to my friends, family, and housemates (my own personal networks ;)). It's always important to balance hard work with relaxation, and that was certainly possible thanks to you all.

Abstract

Background: Type 2 diabetes (T2D) is a significant public health issue with multifactorial influences, including personal and environmental factors. Understanding these risk and protective factors is crucial for effective prevention and intervention strategies.
Objective: This study investigates what the correlations and polarities are of an individual's social network, personal lifestyle, socioeconomic status, and living environment (specifically in terms of food and physical activity opportunities) with the prevalence of T2D among adults in the Netherlands.
Methods: Using a random forest and a logistic regression model as binary classifiers, this research predicts diabetes medication use based on various variables derived from government registration data, national health registers, and the national health monitor survey, while also examining the relative contributions of these variables. The study sample consists of over 290,000 individuals aged 40 and older who participated in the Dutch health monitor survey in 2016. Both models are optimized for the average precision on the diabetes medication group. Shapley values are employed to assess the impact of various factors.
Results: Both models demonstrate similar performance in terms of average precision on unseen data, namely an average precision of 0.29 for the random forest model and 0.28 for the logistic regression model. The analysis examined four categories of risk factors to assess their association with T2D. For the first, social networks, the findings reveal that T2D is more prevalent among individuals whose social networks have a high prevalence of T2D and lower education level. Family networks have the highest correlation followed by workplace and neighborhood same-gender networks. Remarkably, the exposure within family networks is highly predictive, even more than the amount of time someone exercises. Regarding the second group of risk factors, namely socioeconomic status, individuals with lower socioeconomic status are more likely to have T2D. In terms of the third about lifestyle, BMI and exercise engagement are as expected very predictive for the prevalence of T2D. For the fourth, the living environment, it appears that having exercise environments, for example parks and public green space, very close by (approximately 1 km) reduces the risk of T2D, while further away has no effect. However, no clear association was found between T2D prevalence and (limited) access to healthy food.
Conclusion: While existing research indicates that lifestyle behavior is a major determinant of T2D, our research shows that also an individual's social network greatly associates with T2D. This research provides quantitative evidence for the importance of identifying and understanding social networks where T2D is either very prevalent or almost absent. The existence of those healthy and unhealthy social networks seems to go hand in hand with high- and low-educated social networks. Those findings imply that prevention and intervention strategies in the Netherlands should focus not only on individuals, but could be more effective by implementing group interventions tailored to specific risk groups. The random forest model and logistic regression model turn out to have similar performance on unseen data, so although the expectation was that the random forest perform better because of its ability to capture non-linear relationships, this was not clearly the case.
Limitations & Future Research: It is crucial to recognize the limitations of the data used in this study. For example, there is noise from individuals with T1D and of those with T2D not on medication. Additionally, the sample population may lack representativeness and bias may be introduced through the

handling of missing values. Also, the study reflects correlations, not causality, and predicts current, not future, T2D cases. Future research could focus on the development of a robust prognostic model, alongside an in-depth evaluation of its performance across different minority and demographic groups, allowing for a more nuanced understanding of the factors influencing T2D within specific at-risk groups. Additionally, the methodologies and approaches used in this research could be applied to investigate other health outcomes, such as depression.

Contents

Nomenclature

Abbreviation	Definition	Dutch Translation
BMI	Body Mass Index	
CBS	Central Bureau of Statistics	Centraal Bureau voor de Statistiek
T1D	Type 1 Diabetes	
T2D	Type 2 Diabetes	
SES	Socioeconomic status	

1 Introduction

In this chapter an introduction to the problem context and literature studies about the research topic will be given followed by the research questions and unique characteristics of this research.

1.1. The Growing Burden of Type 2 Diabetes

Type 2 Diabetes (T2D) is a chronic disease with lots of personal and societal burden and is currently on the rise worldwide [1–4] In the Netherlands currently more than 1 million people are diagnosed with T2D [5, 6]. There is an even bigger group of people in the Netherlands, namely 1.4 million, who have prediabetes [7] , which is a preliminary stage of T2D. Additionally, there are also people with T2D who are undiagnosed, it is unclear how big this group is [6, 7]. As prognosis for the coming years, it is expected that the number of people with T2D in the Netherlands will increase to over 1.3 million based on the expected demographic development of the Dutch population [5]. When looking to the current adult Dutch population aged 45 and older, 1 in 3 is expected to develop T2D in the future [7].

When looking at the societal level, the burden of diabetes also translates into high costs. In the Netherlands diabetes (type 1 and 2) and its complications are on number 7 of the list of most expensive diseases [8]. This comes down to 1.3 billion euros in 2019, which is 1.4% of the health expenses which is only the lower limit as many costs for diabetes complications are not included due to limitations in administration [8, 9]. These costs thus include type 1 diabetes (T1D) and T2D, however T1D is not preventable and much less common; of the people with diabetes in the Netherlands, 9 out of 10 have T2D [7].

T2D is more prevalent among men than women [10] and is more common in individuals with a migration background [7]. Additionally, as T2D is age-related, the likelihood of developing the disease increases with age [11].

T2D is caused by impaired insulin secretion and insulin resistance in tissues, leading to elevated blood sugar levels [12]. Unlike T1D, which results from an autoimmune response, T2D develops gradually due to dysfunctions in insulin regulation [13]. This persistent high blood sugar damages blood vessels, affecting organs like the heart, kidneys, and eyes, and can lead to serious complications such as coronary heart disease, stroke, and nerve damage [2, 14–16]. T2D rarely occurs alone, often coexisting with other conditions such as cardiovascular disease, hypertension, and kidney disease [12, 16, 17]. In the Netherlands the people with T2D have heart disease (12.1%) as most common comorbidity [16].

1.2. Type 2 Diabetes Prevention Efforts Have Fallen Short

It has long been recognized that T2D is a growing public health issue in the Netherlands. In the 2006 prevention report 'Kiezen voor gezond leven' (Choosing for a Healthy Life), T2D was already identified as an emerging challenge [18]. The number of people diagnosed with diabetes (both T1D and T2D) in 2007 was around 750,000 [18]. In the reports it is also predicted that without significant policy changes, the number of people diagnosed with diabetes would double to over 1.3 million by 2025, with half of the cases being preventable [18]. This preventable portion was attributed to the growing number of people with obesity and other risk factors for T2D [18], which could and should be addressed through national prevention efforts according to the report. The remaining growth was expected to result from population aging and improved diagnostic capabilities [18], which is thus not preventable.

By 2024, these projections have largely proven accurate. The number of diabetes patients (T1D and T2D) has reached 1.2 million, with an additional 52,000 new cases each year [7]. Furthermore, the number of individuals with prediabetes has increased significantly, from approximately 1.1 million in 2018 to 1.4 million in 2024 [7, 19]. These trends show that efforts such as the 2018 National Prevention Agreement, aimed at reducing obesity and preventing T2D, have not been sufficient to curb the growth [20].

Thus, while T2D prevention is firmly on the national agenda, tangible effects remain elusive. The current approach falls short in mitigating the rise of T2D, with the disease burden continuing to increase alongside unsustainable healthcare costs [20]. The Diabetes Fund has called for stronger government intervention [19], advocating for earlier identification of people at high risk of T2D and supporting them with personalized lifestyle advice [19].

1.3. Lifestyle plays a major role in the development of T2D, but can not be viewed in isolation

Lifestyle factors play a major role in the development of T2D [12, 21–27]. It appears however that lifestyle factors, which can lead to T2D, can not just be viewed in isolation but rather as part of interconnected social, economic, and environmental factors in a population or community [28]. There thus seems to be clustering of pre-existing health, social network (see section 1.3.3), socioeconomic status and environmental conditions and the onset of T2D. This clustering can come forth of persistent social and economic inequalities [29]. It is therefore important to shift the focus from individual lifestyles (micro-level) to the wider social and environmental context (macro-level) in which people live [30]. The influence of lifestyle and the associations between T2D and socioeconomic status, social network and living environment are discussed below.

1.3.1. Lifestyle

T2D arises from a complex interplay of genetic, metabolic and environmental factors [12]. While factors like ethnicity and family history contribute due to genetic predisposition, epidemiological evidence highlights that many T2D cases can be prevented by addressing the main modifiable risk factors, which include obesity, low physical activity, and unhealthy diet [12, 21, 25–27]. Dietary patterns high in processed foods, saturated fats, and sugars, coupled with low intake of fruits, vegetables, and whole grains, correlate strongly with increased T2D risk [12, 22–24]. Besides those main risk factors, other lifestyle factors like smoking habits [31, 32], alcohol drinking [33, 34] and sleep patterns [35, 36] also play a critical role in T2D development. Furthermore, sedentary behavior and inadequate sleep duration disrupt metabolic homeostasis, exacerbating insulin resistance and glucose intolerance. Concurrently, tobacco smoking amplifies systemic inflammation and oxidative stress, accelerating pancreatic β-cell dysfunction and insulin resistance [31, 32, 37, 38].

1.3.2. Socioeconomic Status

Socioeconomic status (SES), including education and income, is strongly linked to health outcomes, including the prevalence of T2D. In the Netherlands, individuals with only primary education are significantly more affected by T2D (11%) compared to those with university degrees (2.3%) [39]. Additionally, people with lower levels of education and income experience significant disparities in life expectancy and health. For example, men in the highest wealth groups live 25 years longer in good health than those in the lowest wealth groups, while for women, this difference is 23 years [40].

Lower SES groups are more likely to face lifestyle-related risk factors, such as poor diet, lack of exercise, and higher stress levels, which contribute to these disparities [41]. Poverty and low literacy further restrict access to healthy food and health information, worsening health outcomes [40]. These differences are also influenced by factors such as financial stress, poor working and living conditions, and limited health literacy, all of which increase the risk on poor health outcomes, including T2D [42–44].

1.3.3. Social Network

Studies investigating social network characteristics related to T2D are relatively scarce [45], but the research that does exist reveals several key factors. Additionally literature about obesity and social network characteristics is also considered as this could also be relevant to understanding T2D, given that both conditions are closely related metabolic diseases, sharing common disease pathways [46], and given the fact that obesity serves as a common risk factor for T2D [47].

There is Clustering of Obesity in Social Networks

There is evidence of clustering in social networks among people with obesity [48–50]. Clustering means that a person with obesity has more contacts with obesity within their social network than would be expected by chance. We hypothesize that T2D will also cluster as T2D and obesity are closely related metabolic diseases [46].

The literature and logic reasoning offers multiple explanations why there can be clustering of obesity and hypothetically also clustering of T2D in social networks:

- **Social Contagion Hypothesis**

The first explanation is the social contagion hypothesis that posits that individuals tend to adopt behaviors and norms observed within their social circles [45, 51]. These social circles can exert both positive and negative influences on health behavior through shared norms and behavioral observations.

- **Shared Background Characteristics**
  The second explanation is that obesity and potentially T2D cluster in social networks due to shared background characteristics, which can be either observed (such as socioeconomic status and living environment) or unobserved (such as genetics, shared meals, norms, values, and unobserved life events).

  Socioeconomic Status & Living Environment
  The associations between socioeconomic status and living environment with the prevalence of T2D is discussed in sections 1.3.2 and 1.3.4, respectively. Given those associations, it is noteworthy that education level, a measure of SES, tends to cluster within work, neighborhood, family, and household networks [52, 53], which implies that T2D, which also associates with education level, might cluster in those networks. Additionally, when T2D is linked to the living environment, it implies that T2D automatically clusters within neighborhood and household networks.

  Genetic Predisposition
  There is a genetic predisposition to T2D, though it is typically not the deciding factor [12]. Nevertheless, genetic predisposition can contribute to the clustering of T2D within family networks, as members share similar genetic backgrounds.

  Shared Meals within households
  Dietary habits significantly impact T2D risk, and families or households often share meals, leading to clustering due to common dietary practices rather than social contagion [12, 21, 25–27].

  Norms, Values, Life Events
  Clustering may also be influenced by unaccounted variables, such as shared norms, life events, or experiences within social networks. These factors can act as confounders, creating the appearance that social ties drive T2D development, when other unobserved factors may be responsible.

The Type of Social Network Connection Matters for Clustering of Obesity

Earlier research on obesity indicates that clustering varies depending on the type of social network connection, which refers to the nature of the relationships between individuals. Friends have a stronger influence on the risk of obesity than family or spouses [49]. The research of Christakis & Fowler [49] hypothesized that your social norm regarding the acceptance of obesity is adjusted when you have contact with an obese friend. Additionally, in the same research, neighbors appeared to not increase chances of becoming obese [49].

There Seem to Be Gender Effects in Clustering of Obesity

The research of Christakis & Fowler [49] highlights the role of gender in social networks. In same-sex friendships, particularly male-male, the risk of developing obesity is associated with a friend's obesity status, whereas no significant effect is observed in opposite-sex friendships. Similarly, same-sex sibling pairs are more likely to have correlated weight changes compared to opposite-sex siblings. Spouses (of opposite gender) also show associations in their weight status, although friends and siblings exhibit stronger correlations with obesity risk [49].

Living Alone is a Risk Factor for T2D

There is evidence in the literature indicating that living alone is associated with a higher risk of developing T2D, comparable to the risk posed by obesity or high blood pressure [45, 54]. There is a difference between genders here as living alone increases the likelihood of T2D more for men than for women [54].

Loneliness is a Risk Factor for T2D

Research has consistently shown that individuals who experience loneliness have a significantly higher likelihood of developing T2D—nearly double the risk compared to those who do not feel lonely [55–57].

A possible explanation that living alone as well as loneliness are a risk factor for T2D is given by the stress-buffering or stress-exacerbating hypothesis which suggests that one's social network or the lack thereof can either mitigate or intensify stress, which in turn affects biological processes [58]. This means that according to the stress-buffering or stress exacerbating hypothesis, living alone as well as loneliness can give stress, which can lead to unhealthy lifestyle behaviors and physiological effects of chronic stress [55–57].

1.3.4. Living Environment

Food Environment

There is substantial scientific evidence suggesting that the food supply in an environment is linked to the health of those who live or work there [59, 60].

From a public health perspective, the food environment encompasses the accessibility, affordability, promotion, quality, and sustainability of food and beverages [61]. Food consumption is influenced by more than personal choice; the environment plays a significant role. More than half of the food choiceson a day are being impulsive or unconscious [62]. In an obesogenic environment, unhealthy choices are often the default, making it harder for individuals to opt for healthier alternatives [63].

In the Netherlands, Government-commissioned research indicated that nearly 79% of products in major supermarket chains fell outside the Dutch dietary guideline [64]. Furthermore, 91% of products offered at 21 examined out-of-home food service chains were not conducive to healthy eating [65].

Epidemiological studies in the Netherlands provide insight into how the Dutch food environment impacts health. It was found for a study involving more than 100.000 residents in the Northern Netherlands that those living within 1 km of a fast-food outlet had higher Body Mass Index (BMI) scores [66]. National surveys also indicated that greater exposure to fast-food outlets correlated with higher incidences of T2D and cardiovascular disease [67, 68]. However, smaller studies have shown more nuanced results. For example, a Dutch study involving more than 8000 participants revealed no significant link between the number of nearby fast-food outlets and unhealthy dietary habits or obesity [69]. Similarly, research with more than 4000 Amsterdam residents found no significant association between the healthiness of the food environment and diet quality [70]. Also it is found by Hoenink *et al.* [71] that people often buy food outside their residential neighborhoods, complicating measurement.

Exercise Environment

When considering the living environment in relation to physical activity, evidence suggests that the structure of the built environment can influence energy balance and levels of physical activity [72]. Neighborhoods with a higher degree of 'walkability,' including features such as sidewalks and good connectivity, are thought to promote active transportation like walking and cycling, which may help reduce the risk of obesity [73–75]. Access to recreational facilities is another crucial factor. Studies have shown that proximity to parks and sports facilities is associated with meeting recommended levels of physical activity, contributing to a more active lifestyle and lowering the likelihood of obesity [75, 76].

In the Netherlands, it has been found that the presence of green spaces in and around cities positively influences physical activity and health outcomes [77]. Statistical analyses reveal that neighborhoods with sufficient green space have significantly lower rates of childhood overweight compared to similar neighborhoods without such greenery. The research suggests that green spaces encourage children to engage in more physical activity, which helps reduce overweight. Although factors such as ethnicity and socioeconomic status also play a role in overweight prevalence, green space was found to independently contribute to promoting physical activity and overall health [77].

1.4. Research Questions

T2D imposes a significant disease burden due to its numerous side effects and the substantial healthcare costs associated with it, all while its prevalence continues to rise. Addressing T2D is complex, as it cannot be seen merely as an individual lifestyle issue. Instead, T2D is linked to socioeconomic factors, social networks, and the living environment. Therefore, it is essential to look beyond the individual (micro-level) and their lifestyle, and instead focus on the broader context in which individuals live (macro-level). Current policies have proven insufficient in curbing the rise of T2D, indicating that stronger, more comprehensive interventions are necessary. It is crucial for policymakers to better understand the factors associated with an increased risk of T2D in order to design more effective and innovative policies. Therefore, this research will explore the associations between micro- and macro-level factors and the prevalence of T2D, aiming to identify key leverage points for effective intervention. Therefore the first research question guiding this study is:

1. How do social networks, lifestyle, socioeconomic status, and living environment contribute to the prediction of Type 2 Diabetes prevalence among adults in the Netherlands?

In order to address this research question statistical models will be used to predict the prevalence of T2D using social network, lifestyle, socioeconomic status and living environment factors. The models that will be used are a random forest and logistic regression model. It will be examined how the models predict, so which variables are important for the prediction and what the polarities of those variables are.

In addition to investigating the substantive aspects that can inform policy recommendations, the study also undertakes a methodological comparison between the two statistical models. The random forest model is expected to be particularly useful, as the onset of diseases, like T2D, often occurs due to a culmination of factors, leading to a sudden tipping point. While logistic regression may struggle to capture such complexity due to its linear assumptions, random forest is capable of identifying both linear and non-linear patterns [78]. This dual capability is assumed to enhance the understanding of the relationships between the independent variables and T2D prevalence. Therefore the second research question guiding this study is:

2. How does the performance of a random forest model compare to that of a logistic regression model in the prediction of Type 2 Diabetes prevalence among adults in the Netherlands based on variables about an individual's social network, lifestyle, socioeconomic status, and living environment?

In addressing this research question, there will be an examination of which of the two statistical models is more effective in predicting the prevalence of T2D.

1.5. Hypotheses Regarding the Research Questions

Associations with T2D and Social Network

- **H1** As there is clustering of obesity in social networks [48, 49] and T2D is a closely related metabolic disease [46], it is hypothesized that more T2D in someone's network is positively associated with one's own T2D status.
- **H2** Since the association between social networks and obesity varies by network type [49], and obesity and T2D are closely related metabolic diseases [46], it is hypothesized that the association of social networks and T2D also differs depending on the type of network.
- **H3** As there is a stronger association between same-gender social contacts and the risks of obesity compared to opposite-gender contacts [49] and obesity and T2D are closely related metabolic diseases [46], it is hypothesized for T2D that especially the prevalence of T2D among same-gender contacts within someone's network is associated with one's own T2D status.
- **H4** A low education level of an individual is associated with a higher risk of T2D [39, 40]. It is hypothesized that besides someone's own education level, the education level of someone's network also associated with someone's T2D status. Where a higher education level of someone's network is negatively associated with T2D prevalence.
- **H5** As living alone is found to be a risk factor for T2D by Schram *et al.* [45] and Brinkhues *et al.* [54], it is hypothesized that living alone will be positively associated with T2D prevalence.
- **H6** As loneliness is found to be a risk factor for T2D [55–57], it is expected that loneliness is positively associated with T2D prevalence.

Associations with T2D and Living Environment

- **H7** As an unhealthy food environment increases the likelihood of higher BMI and the risk of T2D among residents [66–68], it is hypothesized that an unhealthy food environment is positively associated with T2D prevalence.
- **H8** As proximity to parks, recreational facilities and green spaces promotes physical activity and decreases the likelihood of obesity [75, 76], therefore it is hypothesized that proximity to those places is negatively associated with T2D prevalence.

Random Forest Model vs. Logistic Regression Model

- **H9** We hypothesize that a random forest algorithm, requiring fewer assumptions and capable of capturing not only linear but also non-linear patterns [78], will outperform regression methods in predicting relations between independent variables and T2D.

1.6. Main Contributions of this Study

This research distinguishes itself from prior studies mainly in two ways, which will be discussed below.

1.6.1. Innovative Dataset

This research utilizes multiple datasets, namely data on diabetes medication use, population network, lifestyle, living environment information and social-demographic information. All these datasets are linked on an individual-level.

This research is very innovative as it utilizes an unprecedented and comprehensive population network dataset to investigate the relationship between individual-level health outcomes, specifically T2D, and network-level health. The population network dataset is constructed by 'Centraal Bureau voor de Statistiek' (CBS) and is a unique dataset. It encompasses the entire population of the Netherlands using administrative data, providing a person-level of detail [79]. This dataset, sourced from official governmental registers, includes includes family relationships, addresses, employment details, and educational enrollment. The administrative data enables the mapping of connections between family members, neighbors, coworkers, household members, and classmates. While analyzing administrative data and employing network science techniques in sociology have been done before, the creation of a complete nation-wide integral population network is completely new. Only Denmark has also recently (following the work of the CBS) built a similar network for their country [80]. Having nation-wide population networks enables social scientists to use network science methods to analyze detailed, individually-linked administrative data on a completely new and much larger scale [79].

The study by de Zoete [81] is at the best of my knowledge the only published study where the Dutch population network data is linked to health data. In that research the data is linked at the community level and utilized as measure for a social capital outcome.

1.6.2. Model Complexity

This research employs a random forest model to capture non-linear relationships between T2D and the independent variables. This while previous research using the population network dataset [81–83] and previous research about the relationship between social network characteristics and T2D [49, 54, 84] uses regression techniques.

In the remainder of this thesis we will focus on developing a framework of social network factors and other factors that may be related to the prevalence of T2D (chapter 2). Following an explanation of the methods used and a description of the dataset, we will evaluate the performance of both the random forest and logistic regression models, examine the associations between factors and T2D and discuss the findings (chapter 3). The implications of these results will be explored, with a focus on informing future research and providing recommendations for policy changes (chapter 4).

2 Methods

This chapter provides a detailed overview of the study design and data preparation process. It covers the study population and sample, data sources, operationalization of T2D, and summary statistics of the dataset. Additionally, the creation and pre-processing of variables, the exclusion of certain variables, the handling of missing values and the training and interpretation of the models using Shapley are explained.

2.1. Research Flow

The two research questions are:

1. How do social networks, lifestyle, socioeconomic status, and living environment contribute to the prediction of Type 2 Diabetes prevalence among adults in the Netherlands?
2. How does the performance of a random forest model compare to that of a logistic regression model in the prediction of Type 2 Diabetes prevalence among adults in the Netherlands based on social network, lifestyle, socioeconomic status, and living environment?

In order to answer the two research questions, we will:

- Select a study sample and data
- Create social network variables (exposure scores for T2D)
- Pre-process data (exclusion of variables, handling missing data, handle categorical and numerical data)
- Train a random forest and a logistic regression model.
- Evaluate the performance of those models.
- Examine the influence (magnitude and polarity) of each variable on the model outputs using Shapley values.
- Compare the influence of variables between the random forest and logistic regression model.

2.2. Study Population

The study population are individuals living in the Netherlands aged 40 and older. This age group was selected because T2D is very rare for people under 40 [11].

2.3. Data Sources

The input data for this research was sourced from multiple datasets, primarily provided by the Centraal Bureau voor de Statistiek (CBS). These datasets include government registration data on social networks, socioeconomic characteristics, and living environment factors, along with health-related data from the Gezondheidsmonitor [85]. The data and its sources are:

- Background variables: demographic variables, socioeconomic status and statistics about the living environment from the CBS.
- Dutch Health Monitor [85] of 2016 from the RIVM on a person level.
- Person network data of the CBS for family, work, colleague and neighbor networks.
- Diabetes medication use for 2016 and 2022 on a person level from health insurance data retrieved by the CBS.
- Exposure values for the education level in one's network for 2016 from the CBS on a person level.
- Exposure values to the use of diabetes medication in one's network for 2016 from the CBS on a person level.

2.4. Operationalization of Type 2 Diabetes

To determine if an individual has T2D, health insurance data on diabetes medication use is utilized. An individual is classified as having T2D if they use insulin and analogs (ATC code A101A), blood glucose-lowering agents excluding insulin (ATC code A10B), or other diabetes medications (ATC code A10X). However, this method does not differentiate between T1D and T2D, meaning that the data includes noise from T1D cases (approximately 10% of individuals with diabetes are estimated to have T1D [7]). Additionally, this approach fails to capture individuals with T2D who do not use medication.

2.5. Study Sample

The study sample is the group of individuals aged 40 and older who participated in the Dutch Health Monitor [85] of 2016. In figure 2.1 the age distribution and diabetes medication use per age group of the study sample is shown. It can be seen that there is a marked increase in diabetes medication use starting at age 40, indicating that within the Health Monitor [85] study population the onset of T2D for the first persons is likely starting around 40, thereby reinforcing the choice of the age group of 40+.

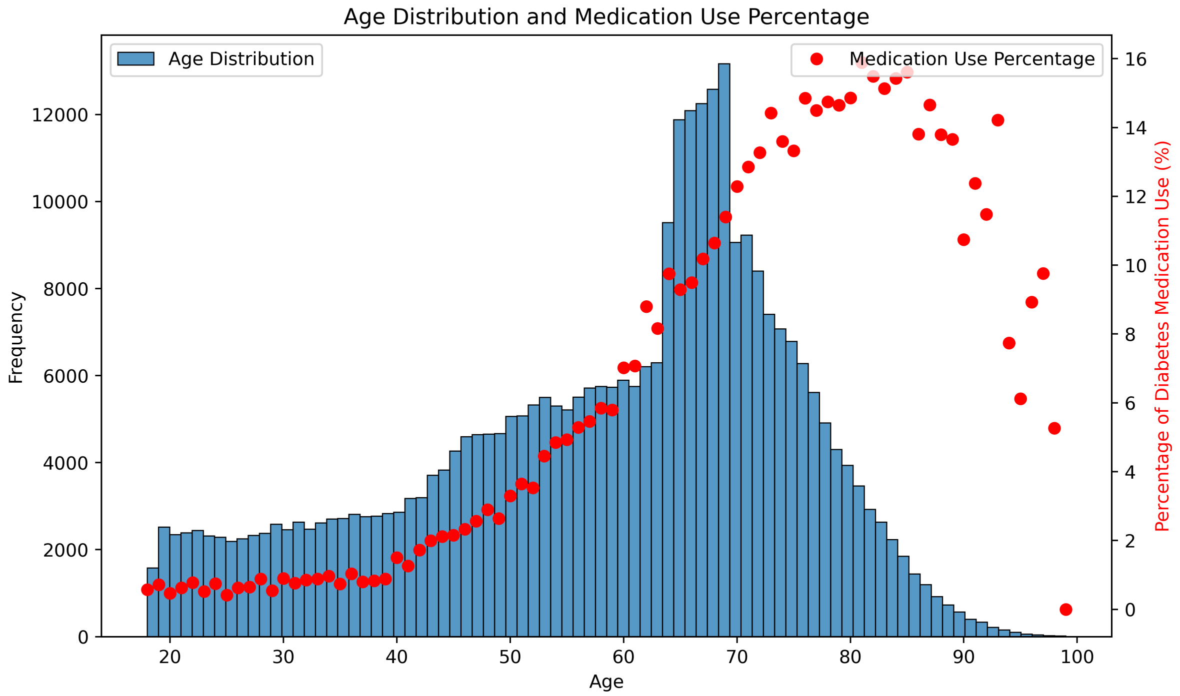


Figure 2.1: The Age Distribution of the GEMON dataset and the percentage of diabetes medication use. (Except for anyone over 100, due to privacy rules)

2.6. Dataset

All variables used to train the model, except for certain demographic variables, are categorized into four main groups: social network, lifestyle, socioeconomic status, and living environment. These variables serve as independent variables in the model. The dependent variable is 'diabetes medication use,' which indicates whether an individual uses diabetes medication or not. Below, we discuss all variable groups and present their summary statistics for all individuals included in the training of the models. The summary statistics are expressed either as the mean of the variable (e.g., the average age) or as the percentage of individuals belonging to a specific group (e.g., 51.04% belong to the 'women' group). These statistics are shown for two categories:

- The **Overall** group, which includes all individuals—both those who do not use diabetes medication and those who do.
- The **Using Diabetes Medication** group, which is about the individuals using diabetes medication. So for example in table 2.1 it can be seen that the average age of peole using diabetes medication is 69.57. Additionally it can be seen that 10.76% of men are using diabetes medication, while only 7.62% of women are.

All those individuals in the summary statistics are 40+ and participated in the Health Monitor of 2016 (see the description of the study sample in section 2.5. There are also left out individuals because they have missing data (which is discussed in section 2.13), however those are not included in the summary statistics and have their own summary statistics shown in appendix B.

For the study sample 90.85% (263,802 people) do not use diabetes medication and 9.15% (26,576 people) do use diabetes medication.

2.7. Demographic Variables

The summary statistics about the demographics variables are shown in table 2.1. Those variables are: age, gender and origin. The age variable can be any (discrete) number. The gender variable is either man or woman. For origin there is a division in 10 categories, namely: Dutch, Other European, Turkish, Moroccan, Surinamese, Dutch Caribbean, Indonesian, Other African, Other Asian and Other American and Oceanian.

Category	**Overall**		**Using Diabetes Medication**	
Average age	64.20 ± 11.56 years			69.57 ± 9.36 years
Men	48.78%	(141638)	10.76%	(15240)
Women	51.22%	(148740)	7.62%	(11336)
Dutch	87.82%	(255000)	8.76%	(22348)
Other European	6.20%	(17994)	9.93%	(1786)
Turkish	0.39%	(1123)	19.15%	(215)
Moroccan	0.28%	(824)	24.03%	(198)
Surinamese	0.78%	(2264)	23.76%	(538)
Dutch Caribbean	0.29%	(839)	15.26%	(128)
Indonesian	2.80%	(8128)	11.17%	(908)
Other African	0.32%	(935)	13.58%	(127)
Other Asian	0.71%	(2052)	12.72%	(261)
Other American & Oceanian	0.42%	(1219)	5.50%	(67)

Table 2.1: Demographic summary statistics of the study sample with a comparison between the overall study sample and individuals using diabetes medication. Either the average of the variable value or the percentage of people that comply with that variable is shown. Besides the percentages, also the total number of people that belong to the percentage are shown.

2.8. Social Network Variables

The summary statistics about the social network variables are shown in table 2.3. These variables can be grouped into four subcategories: household, loneliness, exposure to education and exposure to diabetes medication use. The household subcategory addresses the type of household in which an individual lives, such as living alone, with a partner, with children, etc. The loneliness variable is a variable from the Dutch health monitor [85] of 2016, the higher the more lonely someone is.The subcategory, exposure to education levels, measures the direct and indirect exposure to different education levels within an individual's network across all layers, including master's, bachelor's, secondary, and low education and is a pre-calculated variable by the CBS (see section 2.8.2). The exposure to diabetes subcategory quantifies the proportion of people in an individual's social network who use diabetes medication. Those diabetes exposure scores are calculated during this research using social network data from the CBS, see below for the definitions of the networks and see section 2.8.2 for the creation of those scores.

2.8.1. Definitions of the Social Network Layers

The definitions of the network layers are outlined in files provided by CBS [86, 87]. For this research, the updated version of the definitions from [86] was used. Details of these updates are documented in [87], which is used to construct the network layers. For understanding the definitions it is good to know that an ego denotes the person of interest and its associated network.

Family Network Layer

The family network consists of the following relationships:

- **Core family**: Individuals who share a biological or adoptive familial bond. The core family includes the following relationships: parents, grandparents, grandchildren, (half-)siblings, co-parents (individuals who share a child but may not cohabitate, as the cohabitation relationship is captured by the 'partner' relation), aunts/uncles, nieces/nephews, and cousins.
- **Partners**: Individuals residing at the same address who are married, in a registered partnership, or in a cohabiting relationship.
- **Stepfamily**: The partner of an ego's parent (who is not themselves the ego's parent) with whom the parent cohabits, as well as any children of this partner, regardless of their place of residence.
- **In-laws**: The family members of an ego's partner.

Household Network Layer

- The household network layer consists of individuals residing at the same address as the ego.

Colleague Network Layer

- The work network includes individuals who worked at the same company as the ego for at least one month during the reference year. If the ego had more than 100 colleagues during the year, only the 100 colleagues residing closest to the ego's home address are selected.

Neighbor Network Layer

- All residents of the ten closest addresses to the ego.
- Up to 20 randomly selected neighbors. A neighbor is defined as an individual residing within a 200-meter radius of the ego's address. For additional details, refer to [82, 86, 87].

Deduplication and removal of relationships

It is important to note several aspects of relationship overlap and to note deduplication in the construction of network layers: First, it is possible for individuals to appear in multiple network layers of an ego. For example, a married couple with children who cohabitate would share three distinct relationships: partner (married), co-parent, and household member. Second, for the calculation of exposure scores (see section 2.8.2), relationships are deduplicated within each network layer. Each individual can only have one relationship per layer. Additionally, deduplication occurs across network layers, where only one type of relationship between an ego and another individual is retained. For instance, pairs of individuals with a relationship in the household layer are removed from the family layer. In the example above, the three relationships would be consolidated, retaining only the relationship in the household layer. Finally, it is important to understand the implications of changes in relationship status. For example, when an ego divorces their partner, the partner relationship ceases. However, other relationships may persist, such as the co-parent relationship if they share a child. This relationship remains even after divorce and regardless of cohabitation status.

2.8.2. Creation of Exposure Scores

Creation of Indirect Exposure Scores for Education Level (by the CBS)

The creation of the indirect exposure scores was done by Van der Laan *et al.* [79] of the CBS and the methods are described in the paper [79]. Exposure scores are between 0 and 1 and quantify how much a person is exposed to a certain education level. A score of 0 means no exposure, while a score of 1 is maximum exposure (for example: everyone in one's network has a master). These scores are calculated by aggregating the characteristics of a person's network contacts and adjusting by the size of the network. In smaller networks, each person's traits have a larger impact on the exposure score, while in larger networks, the influence is spread out. For first-degree connections, network layers such as family, work, neighbors, household members and classmates are considered separately, with each layer contributing equally to the exposure score. This ensures that no one layer is overly dominant, and each social context is weighted the same. Additionally, those closer in the network (e.g., direct family, colleagues, neighbors, household members and classmates) contribute more to the score than those farther away. This score is called indirect exposure score as it also takes into account more than one degree of separation, so not only direct, but also indirect contacts. In the final exposure score, no distinction can be made between the different network layers or between the influence of direct and indirect contacts. This means that all layers and contacts are combined and weighted together, without separately identifying their individual contributions.

Creation of Direct Exposure Scores for T2D

The creation of direct exposure scores was done specifically for this research using the population network data [79] of the CBS and the diabetes medication data. For the creation of direct exposure scores only one degree of separation is considered and only one network layer per time is considered and only people in someone's network who are 40+ are considered (for the same reasons as described in section 2.2). This makes it possible to distinguish between the influence of different network layers. Besides creating a direct exposure score to T2D per network layer, there are also two additional direct exposure scores per network layer, namely exposure to people with the same gender and exposure to people with the opposite gender, this as from previous studies it becomes evident that gender can matter in the spread of obesity within a network [49] and it is then possible to test if it plays a role for T2D as well.

For those not represented in a particular network layer, the variables associated with that layer are filled with a default value of '0'. For example, if a person is not employed and therefore not included in the colleague network layer, all variables pertaining to this layer will be assigned a value of '0' for that individual. In table 2.2 it can be seen how many (of the in total 290,378 individuals) lack a certain network type. Filling those missings with zeros is not ideal especially for the colleague layer as this applies to 67.9% of the individuals. However, post hoc analyses are done for only the working people (see section 3.3 in chapter 3). Also this is further discussed in the discussion, see chapter 4.

Network layer	Number of individuals	percentage of total population
Family	7,295	2.51%
Household	70,536	24.3%
Colleague	197,074	67.9%
Neighbor	36	0.01%

Table 2.2: The number and percentages of individuals relative to the total included population (290,378) that are lacking a network type.

Category	Overall (N)		Using Diabetes Medication	
Living at parents home	0.36%	(1053)	5.41%	(57)
Living alone	20.89%	(60651)	12.16%	(7376)
Partner in unmarried couple without children living at home	4.42%	(12845)	7.33%	(941)
Partner in married couple without children living at home	50.62%	(146987)	10.20%	(14987)
Partner in unmarried couple with children living at home	2.53%	(7336)	2.70%	(198)
Partner in married couple with children living at home	17.72%	(51469)	4.20%	(2161)
Parent in single-parent household	2.51%	(7280)	7.24%	(527)
Reference person in other household	0.12%	(349)	11.75%	(41)
Other household member	0.76%	(2193)	11.99%	(263)
Member of institutional household	0.07%	(215)	11.63%	(25)
Not lonely	56.69%	(164627)	7.71%	(12693)
Slightly lonely	34.88%	(101282)	10.53%	(10666)
Lonely	5.47%	(15892)	12.72%	(2022)
Very lonely	2.95%	(8577)	13.93%	(1195)
Average loneliness **	0.55 ± 0.73		0.69 ± 0.80	
Indirect exposure to people * across all network layers	0.05 ± 0.06		0.07 ± 0.07	
Exposure to family members *	0.06 ± 0.11		0.10 ± 0.14	
Exposure to family members of the same gender *	0.06 ± 0.13		0.08 ± 0.16	
Exposure to family members of a different gender *	0.07 ± 0.14		0.10 ± 0.18	
Exposure to household members *	0.06 ± 0.24		0.11 ± 0.32	
Exposure to household members of the same gender *	0.00 ± 0.05		0.00 ± 0.06	
Exposure to household members of a different gender *	0.06 ± 0.24		0.11 ± 0.31	
Exposure to neighbors *	0.08 ± 0.07		0.10 ± 0.08	
Exposure to neighbors of the same gender *	0.08 ± 0.09		0.10 ± 0.10	
Exposure to neighbors of a different gender *	0.08 ± 0.09		0.10 ± 0.10	
Exposure to colleagues *	0.01 ± 0.04		0.01 ± 0.05	
Exposure to colleagues of the same gender *	0.01 ± 0.04		0.01 ± 0.06	
Exposure to colleagues of a different gender *	0.01 ± 0.05		0.01 ± 0.04	
Exposure to colleagues * (***)	0.03 ± 0.05		0.05 ± 0.09	
Exposure to colleagues * of the same gender (***)	0.03 ± 0.06		0.06 ± 0.11	
Exposure to colleagues * of a different gender (***)	0.03 ± 0.08		0.03 ± 0.08	
Exposure to people with master education	0.14 ± 0.12		0.11 ± 0.11	
Exposure to people with bachelor education	0.24 ± 0.12		0.22 ± 0.12	
Exposure to people with middle education	0.45 ± 0.16		0.46 ± 0.15	
Exposure to people with low education	0.17 ± 0.13		0.20 ± 0.15	

Table 2.3: Social Network summary statistics of the study sample with a comparison between the overall study sample and individuals using diabetes medication. Either the average of the variable value or the percentage of people that comply with that variable is shown. Besides the percentages, also the absolute number of people (N) is shown. The '*' in the table stands for: using diabetes medication. **: The average loneliness ranges from 0 (not lonely) to 3 (very lonely). ***: Only the working population is included for those variables as only that group can have colleagues.

2.9. Lifestyle Variables

The summary statistics about the lifestyle variables are shown in table 2.4. The lifestyle variables are divided into the subcategories of exercise, smoking, alcohol, body mass index (BMI), and experienced health. The exercise subcategory includes variables related to the weekly minutes spent on low, moderate, and high-intensity activities and whether individuals meet the Dutch exercise guideline of the Dutch Health Council [88]. The smoking variable indicates whether a person has never smoked, is a former smoker, or is a current smoker. The alcohol subcategory covers whether someone is a drinker, a former drinker or a current drinker and the total number of drinks consumed during the week. Additionally, BMI is categorized into five distinct groups, and experienced health is also classified into five categories.

Category	Overall (N)		Using Diabetes Medication	
Minutes of light intensity exercise *	1407.19 ± 1069.38		1042.65 ± 974.71	
Minutes of middle intensity exercise *	851.57 ± 865.06		667.77 ± 819.39	
Minutes of high intensity exercise *	36.50 ± 111.85		15.67 ± 77.83	
Adherence to exercise guidelines	46.60%	(135306)	5.99%	(8110)
Never smoked	37.23%	(108103)	7.45%	(8054)
Ex-smoker	48.08%	(139618)	10.55%	(14734)
Smoker	14.69%	(42657)	8.88%	(3788)
Never drank alcohol	10.31%	(29938)	15.45%	(4626)
Alcohol drinker	83.06%	(241184)	7.62%	(18369)
Ex-alcohol drinker	6.63%	(19256)	18.60%	(3581)
Number of alcoholic drinks *	7.08 ± 9.12		5.40 ± 8.89	
Under weight (BMI: 18.5-)	0.94%	(2720)	3.27%	(89)
Normal weight (BMI: 18.5-20)	2.60%	(7555)	3.27%	(247)
Normal weight (20-25)	39.16%	(113717)	4.82%	(5483)
Overweight (BMI: 25-30)	41.27%	(119852)	9.51%	(11397)
Obese (BMI: 30+)	16.03%	(46534)	20.11%	(9360)
Very good experienced health	13.15%	(38175)	1.25%	(478)
Good experienced health	57.87%	(168044)	6.47%	(10876)
Moderate experienced health	24.35%	(70715)	17.07%	(12070)
Bad experienced health	4.11%	(11929)	23.34%	(2784)
Very bad experienced health	0.52%	(1515)	24.29%	(368)

Table 2.4: Lifestyle summary statistics of the study sample with a comparison between the overall study sample and individuals using diabetes medication. Either the average of the variable value or the percentage of people that comply with that variable is shown. Besides the percentages, also the absolute number of people (N) is shown. The * stands for: 'per week'.

2.10. Socioeconomic Variables

The summary statistics about the socioeconomic variables are shown in table 2.5. These variables are categorized into three subcategories: education level, socioeconomic category, and household income.

Category	Overall (N)		Using Diabetes Medication	
Low education (primary education)	6.73%	(19549)	18.98%	(3711)
Middle 1 education (Dutch: MAVO, LBO)	35.71%	(103702)	11.08%	(11495)
Middle 2 education (Dutch: HAVO, VWO, MBO)	29.20%	(84789)	8.00%	(6784)
High education (HBO, WO)	28.36%	(82338)	5.57%	(4586)
Unfit for work (Dutch: arbeidsongeschikt)	2.74%	(7953)	12.86%	(1023)
Social benefits (Dutch: bijstand)	0.99%	(2870)	13.52%	(388)
No income	4.20%	(12193)	5.82%	(710)
Retired	55.84%	(162154)	12.42%	(20140)
Social benefits (Dutch: sociale voorzieningen)	0.44%	(1264)	11.95%	(151)
Working	34.20%	(99304)	3.88%	(3850)
Using unemployment benefits (*)	1.60%	(4640)	6.77%	(314)
Average household income percentile	58.77 ± 25.80		48.84 ± 24.79	

Table 2.5: Socioeconomic summary statistics of the study sample with a comparison between the overall study sample and individuals using diabetes medication. Either the average of the variable value or the percentage of people that comply with that variable is shown. Besides the percentages, also the absolute number of people (N) is shown. *: Dutch: werkloosheidsuitkering

2.11. Living Environment Variables

The living environment variables are divided into three subcategories: urbanity, food environment and exercise environment. The urbanity level reflects the density of addresses within a neighborhood. The summary statistics about the urbanity levels and diabetes medication use are shown in table 2.6. The food environment measures the availability of food-related locations within a 3 km radius, including supermarkets, other daily grocery stores, cafés, cafeterias, and restaurants. The exercise environment captures the distance to the nearest areas suitable for physical activity, such as public green spaces, parks, day recreation areas, forests, open dry natural lands, semi-public green spaces, sports grounds, and swimming pools. In appendix A the distribution plots of the food environment and exercise environment variables are shown.

Category	Overall (N)		Using Diabetes Medication	
Very strong urbanity (≥ 2500 surrounding addresses/km²)	12.94%	(37587)	11.85%	(4455)
Strong urbanity (1500-2500)	21.91%	(63630)	10.06%	(6398)
Moderate urbanity (1000-1500)	19.27%	(55946)	8.90%	(4977)
Little urbanity (500-1000)	21.16%	(61436)	8.43%	(5178)
Not urban (<500)	24.72%	(71779)	7.76%	(5568)

Table 2.6: Living Environment summary statistics of the study sample with a comparison between the overall study sample and individuals using diabetes medication. The percentage of people that comply with that variable is shown. Besides the percentages, also the absolute number of people (N) is shown.

2.12. Excluded Variables

Several variables were excluded because the variables were either very highly correlated with other variables or because another overlapping variable was more informative.

In appendix in figure C.1 the heatmap with all the correlations between variables is shown. Additionally the Pearson correlation coefficient [89] between variables that have a correlation of 0.45 (as absolute number) or higher are shown in the appendix C. It can be seen that there are certain groups of variables highly correlated.

All variables about the food environment are very highly correlated with each other. To avoid multicollinearity problems, we chose to include only the variables about the number of supermarkets and grocery stores within 3 km about the food environment. Regarding the exercise environment, only the variable about semi-public green space is left out, as it correlates too much with public green space.

Regarding the variables about the exposure to people with diabetes medication it was chosen for the family, colleague and neighbor network layers to include only the variables for the same gender and different gender. The values for the overall exposure for those layers are thus left out. This decision was made to prevent multicollinearity issues, as these overall exposure variables overlap with the data on gender-specific and non-gender-specific exposure.

For the exposure to people using diabetes medication in the household-related, only the variable about the overall household network layer was included, without distinguishing by gender. This choice was made because, for exposure to diabetes medication of the same gender, the exposure value is zero for almost everyone (except for around 100 individuals, which is 0.04% of the dataset). As a result, differentiating between same-gender and different-gender exposure would provide little meaningful information.

An oversight of all the included variables (65 in total) can be found in table 3.6 in chapter 3.

2.13. Handling Missing Data

The dataset contains a total of 385,073 individuals, of whom 94,695 have missing values in one or more variables. Both the random forest model and the logistic regression model from Scikit-Learn [90] cannot handle missing values. In this section, we will discuss the types of missing values present in the dataset, the options we considered for handling them, and the final approach we selected, along with the rationale behind our choice.

Table 2.7 presents the number of missing values for each variable with missing data. Relatively few missing values are observed in the living environment variables, while a substantial number of missing values are present in variables related to loneliness, exercise behavior, smoking behavior, drinking behavior, BMI and education level.

The missing values in the living environment variables are due to data not being available for certain individuals in the dataset obtained from CBS. The exact cause of this absence is unclear, but it might be that these individuals do not have a (registered) address in the Netherlands. These missing values are categorized as structural missing values, meaning their absence is not random but inherent to the dataset. Since the reasons for these missing values are unknown, there is little that can be said about their potential impact on the dataset's representativeness and bias when excluding them.

The second group of missing values all (except for the one missing in the 'place in the household' variable) arise from respondents not answering one or more questions in the health monitor survey. These missing values are classified as 'Item Non-response'. Determining the type of these missing values within the categories of 'missing completely at random' (MCAR), 'missing at random' (MAR), or 'missing not at random' (MNAR) is important. This classification helps in selecting appropriate handling strategies and, in the case of exclusion, helps understand the potential impact on the dataset's representativeness and bias. Below, we briefly describe each type:

- **MCAR:** The probability of missing data is entirely random and independent of both observed and unobserved values. This could occur if respondents skipped questions accidentally.
- **MAR:** The probability of missing data depends on observed data but not on the values of the missing data itself. For example, if older respondents are less likely to answer questions about lifestyle, the missing data could be classified as MAR because it relates to observable characteristics such as age.
- **MNAR:** The probability of missing data depends on the values of the missing data itself. For instance, respondents might avoid answering questions about exercise behaviour, alcohol consumption or loneliness due to embarrassment.

Examining the summary statistics for individuals with missing values (see appendix B), it is evident that they differ significantly from those included in the analysis (refer to summary statistics in Tables 2.1, 2.3, 2.4, 2.5, and 2.6). Specifically, excluded individuals exhibit a higher prevalence of diabetes medication use (12.44% compared to 9.15%), are generally older (68.6 years versus 64.2 years), have a greater representation of different origins, and include a larger proportion of retirees. Additionally, they show slightly higher exposure to diabetes medication use and slightly lower exposure to high education levels in their networks. Furthermore, while it is more challenging to assess due to the high level of missing data for those variables, it appears that these individuals, on average, exercise less, smoke less, drink less alcohol, have poorer experienced health and a lower education. Given these observed differences, the missing values cannot be classified as entirely random (MCAR). The group deviates on both the observed data and the values of the missing variables themselves, suggesting that this second group of missing values likely represents a mix of MAR and MNAR patterns.

While imputation is the preferred method for addressing missing values that are classified as MAR [91], additional data collection for affected individuals—though not feasible for this research—or direct modeling of the missing data (which is highly complex, given that the missingness depends on the values of the missing data itself) is recommended for handling MNAR data [91].

In this study, we opted to exclude all individuals with missing values for several reasons. First, the missing data is likely a combination of MAR and MNAR, making it challenging to address appropriately. Second, we aimed to avoid imputation-induced bias, which can arise from the imputation process itself. Imputation can introduce uncertainty and potentially distort the data, leading to inaccurate associations. Furthermore, imputing missing values would have impacted approximately one in four individuals, creating substantial uncertainty within the dataset. Finally, the dataset was large enough to allow deletion without facing issues related to sample size. Initially, the dataset contained 385,073 individuals, and after deletion of individuals with missing values (94,695), it retained 290,378 individuals.

2.14. Pre-processing Variables

Two models, a random forest and a logistic regression, were trained, and each required a distinct approach to handle numerical and categorical variables. This pre-processing is done using the Scikit-

Variable	number of missings
Social Network Variables	
place in the household	1
loneliness	28948
Lifestyle Variables	
minutes of light intensity exercise per week	27850
minutes of middle intensity exercise per week	27850
minutes of high intensity exercise per week	27850
adherence to exercise guidelines	27850
(ex-)smoker	27184
(ex-)drinker	15716
number of alcoholic drinks per week	31836
BMI	18371
experienced health	4678
Socioeconomic Variables	
education level	25460
Living Environment Variables	
number of supermarkets and grocery stores within 3 km	885
distance in meters to the nearest public green space	5307
distance in meters to the nearest park	5307
distance in meters to the nearest day recreational area	5307
distance in meters to the nearest forest	5307
distance in meters to the nearest open dry land	5307
distance in meters to the nearest sports field	5307
distance in meters to the nearest swimming pool	885
Total Unique Individuals with Missings	94695

Table 2.7: The number of missings per variable and the total number of unique missings. Only variables that have at least one missing are included.

learn library [90].

For the numerical variables:

- **Random Forest Model**
  In the random forest model, numerical data can be fed directly into the algorithm without additional pre-processing, therefore no standardization was done.
- **Logistic Regression Model**
  For the logistic regression model, numerical data was first scaled using z-score normalization. While z-score normalization does not affect the fit of an unpenalized linear model due to its linear transformation, it is relevant for penalized regression (e.g., L2 regularization). Scaling ensures that features with varying scales do not disproportionately influence the regularization penalty and also facilitates more efficient optimization during the fitting process.

For the categorical variables:

- **Random Forest Model**
  For a random forest model, when a categorical variable has only two possible values (binary), one category is dropped because it is implicitly captured in the other, reducing redundancy and preventing perfect multicollinearity in the model. When the categorical variable has more than 2 categories, all categories are retained for easier interpretation of results.
- **Logistic Regression Model**

For logistic regression the first category of every categorical variable was dropped to avoid multicollinearity.

2.15. Models

2.15.1. Random Forest Algorithm Background Information

The Random Forest [92] classifier is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the mode (majority vote) of the classes as the final prediction. It is a powerful algorithm for both classification and regression tasks due to its ability to handle high-dimensional data, manage non-linear relationships, and resist overfitting through averaging or voting.

A Random Forest consists of a collection of individual decision trees. Each decision tree is a model that recursively splits the data into subsets based on feature values, aiming to maximize the separation of classes at each split. The separation of classes is measured using the Gini impurity.

The strength of a Random Forest, compared to a single decision tree, lies in its use of bagging (Bootstrap Aggregating). During training, the algorithm generates multiple subsets of the training data by sampling with replacement (bootstrap sampling). Each decision tree is trained on a different bootstrap sample, creating a diverse ensemble of trees. Since each tree sees a slightly different version of the data, the model becomes more robust to overfitting.

In addition to using different bootstrap samples, the Random Forest algorithm also randomly selects a subset of features at each split. This process further reduces the risk of overfitting by ensuring that no single tree relies too heavily on any one feature. It enhances the model's ability to generalize to unseen data.

Once all trees are trained, the final classification is determined through a majority voting process. Each tree independently predicts a class label, and the class with the most votes becomes the final prediction for that instance. This majority vote reduces the overall variance of the model and improves prediction accuracy. The probability of the classification is defined by the proportion of votes for the predicted class and can be used to produce a precision/recall graph. A decision threshold (the proportion of votes of trees needed to classify an individual as using diabetes medication) can be chosen based on the preferred trade-off between precision and recall.

2.15.2. Logistic Regression Algorithm Background Information

Logistic regression is a widely employed statistical method for binary classification tasks, where the objective is to distinguish between two possible outcomes, such as positive and negative cases. This model estimates the probability that a given input belongs to a specific class. The advantages of logistic regression include its simplicity and ease of interpretation, allowing clear insights into how input variables influence outcomes. It is computationally efficient, enabling quick training and predictions, even with large datasets. Additionally, it performs well with linearly separable data.

The logistic regression model utilizes the logistic function (or sigmoid function) to transform a linear combination of the input features into a probability value ranging from 0 to 1. The probability of the prediction can be used to create a precision/recall graph. A decision threshold can be chosen. For example if the decision threshold is 0.5, then when the estimated probability exceeds a threshold of 0.5, the model classifies the input as belonging to class 1; otherwise, it classifies it as class 0.

Penalized Logistic Regression

A significant aspect of logistic regression is the application of penalized techniques, such as Lasso (L1 regularization) or Ridge (L2 regularization) regression. These methods introduce a penalty term that constrains the magnitude of the coefficients. This penalization mitigates the influence of less important variables, leading to a more parsimonious model. Consequently, penalized logistic regression enables the attribution of predictions to a limited set of significant variables, improving both interpretability and model performance.

2.15.3. Training of the Random Forest and Logistic Regression Model

A random forest classifier and a logistic regression model are trained to predict on a person level the prevalence of T2D (operationalized by the use of diabetes medication, see section 2.4) using the earlier described independent variables (see sections 2.7, 2.8, 2.9, 2.10 and 2.11). For training of these models, Scikit-Learn [90] is used, which is a widely used open-source Python library that provides simple and efficient tools for machine learning. This section explains how the models were trained, selected and evaluated. For reference, the code used for both models is included in Appendix G.

The dataset was randomly divided into training and testing subsets using the train_test_split function from scikit-learn [90]. Specifically, 80% of the data was allocated to the training set, while the remaining 20% was reserved for testing. The training set is used for model training. The test set is not used in any of the training and is only used in the end for the final evaluation of the best model.

The classes in the dataset are unbalanced, with around 9.04% of people belonging to the class 'using diabetes medication' and the remaining 90.96% in the class 'no use of diabetes medication'. This imbalance can lead to problems during model training, as the model may favor predicting the majority class ('0'). To address this, balanced class weights are used. Balanced class weights assign higher importance (or weight) to the minority class, meaning that incorrectly classifying someone who uses diabetes medication as someone who doesn't will incur a much higher penalty compared to the reverse. This ensures the model is encouraged to treat both classes more equally, rather than just favoring the majority class.

Hyperparameter tuning with 5-fold cross-validation is performed to test different combinations of hyperparameters, which are predefined settings that control the model's learning process. The model is repeatedly trained and validated on different subsets of the data, and the hyperparameter combination that provides the best average performance across all folds is selected to ensure good generalization. Hyperparameter tuning is performed to find the optimal model by selecting the best combination of hyperparameters. However, there is a balance in the number of hyperparameters to test, as it is important to avoid underfitting (by having too little complexity) while explaining as much variance in the data as possible. Additionally, hyperparameter tuning can help prevent overfitting to avoid overly complex models that fit the training data too closely and fail to generalize well to new data. There is however a limit to computation time and power, which constrains the number of hyperparameters that can be tested and it also constrains the values of the hyperparameters itself.

The following hyperparameters are tuned for the random forest model:

- **The number of decision trees in the forest**. A larger number typically increases performance but also computational cost.
- **The maximum depth of each tree**. Greater depth allows the model to capture more complex patterns, potentially improving performance. However, it also increases the risk of overfitting, as deeper trees may learn noise instead of general trends.
- **The minimum number of samples required to split an internal node** A larger number ensures that each split has sufficient data, reducing the risk of overfitting; however, it may also prevent the model from capturing important patterns, leading to underfitting.
- **The number of features to consider when looking for the best split**. Random selection of features ensures diversity among the trees. This promotes diversity among the trees, which can enhance the overall robustness and accuracy of the model. However, if too few features are chosen, the model may miss critical predictors, potentially leading to suboptimal splits and reduced predictive power.
- **For all other tunable hyperparameters** the defaults of Scikit-Learn [90] are used.

The various values tested for these hyperparameters are shown in table 2.8.

Hyperparameter	**Values**
Number of trees	50, 100, 150
Maximum depth of each tree	5, 10, 20, 30
Minimum number of samples per split	5, 10, 20
Maximum number of features to consider for best split	`sqrt(*)`, log2(*), None **

Table 2.8: Hyperparameters for the random forest model. The * stand for: 'total number of variables'. **: None means all features are considered.

The following hyperparameters are tuned for the logistic regression model:

- **The inverse of regularization strength (`C`).** Lower values of `C` apply stronger regularization, which helps prevent overfitting by penalizing large coefficients. However, too strong regularization can lead to underfitting, where the model fails to capture important patterns in the data.
- **The solver used**. The solver determines the optimization algorithm for finding the best coefficients. Choosing the right solver can improve computational efficiency and accuracy, but an unsuitable solver may fail to converge or result in suboptimal performance for certain datasets or regularization types.
- **The regularization penalty**. Regularization helps control overfitting by shrinking coefficients, simplifying the model, and improving generalization. However, excessive regularization can lead to underfitting by removing meaningful predictors, reducing the model's ability to learn from the data.
- **For all other tunable hyperparameters**, the defaults of Scikit-Learn [90] are used.

The various values tested for those logistic regression hyperparameters are shown in table 2.9. All hyperparameter combinations are tested, except for the combinations of L1 and LBFGS and L1 and Newton-Cholesky as these combinations are not compatible [90].

Hyperparameter	**Values**
Inverse of regularization strength (`C`)	0.01, 0.1, 1, 10, 100
Solver	`liblinear`, `lbfgs`, `saga`, `newton-cholesky`
Penalty	`l2`, `l1`

Table 2.9: Hyperparameters for the model

After having trained all the models with those different hyperparameter combinations, the best model is selected based on the highest average precision on the validation set (of the cross-validation). The average precision is the metric that quantifies the area under the Precision/Recall curve (for the definition of precision and recall, see table 2.10). It represents the average precision achieved at varying levels of recall and provides a single-value summary of the trade-off between precision and recall across different decision thresholds. Average Precision is a suitable metric to use when there is class imbalance, which is the case for this classifciation task (only 9.15% is using diabetes medication).

Next, the decision threshold is selected based on precision/recall threshold curves generated from the training set. A precision-recall threshold curve is a graphical representation that illustrates how precision, recall, and the F1-score change as the decision threshold varies in a binary classification model. The decision threshold is chosen where the F1 score is maximized. Setting the threshold at a maximal F1 score means that recall and precision are considered to be equally important and thus false positives are considered to be at equal costs of false negatives. Setting this threshold is done for evaluation purpose, however other thresholds might be preferred depending on where the model is used for, see section 4.5.2 in the discussion (chapter 4) for further detail.

Then the 'real' model performance is checked by evaluating the average precision, the precision, recall, negative predictive value and specificity (using the threshold where F1 is optimal) on the unseen test set. The definition of the metrics is shown in table 2.10.

2.16. Shapley

2.16.1. Shapley Background Information

SHapley Additive exPlanations (SHAP) [93] is a widely used method for interpreting machine learning models, leveraging Shapley values from cooperative game theory [94]. This technique provides a fair attribution of a model's output to its input features, offering interpretability for both simple models like logistic regression and complex 'black-box' models such as Random Forests.

At its core, SHAP quantifies the contribution of each feature to a model's prediction by examining all possible subsets of input features. For each feature, the Shapley value is calculated as its average marginal contribution to the prediction across all possible subsets of input features. Positive Shapley values indicate that a feature increases the prediction, while negative values suggest it decreases the

Metric	Formula
Precision	$\frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$
Recall	$\frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$
Negative Predictive Value	$\frac{\text{True Negatives}}{\text{True Negatives} + \text{False Negatives}}$
Specificity	$\frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$
F1	$\text{F}_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

Table 2.10: Formulas for evaluation metrics

prediction. In figure 2.2 a SHAP plot is shown with on the x-axis the SHAP value for a certain feature (age in this case) and the color indicates the feature value. It can be seen that for this feature a high value results in a SHAP value above zero, which means a high value increases the prediction for a certain outcome (in this case the use of diabetes medication).

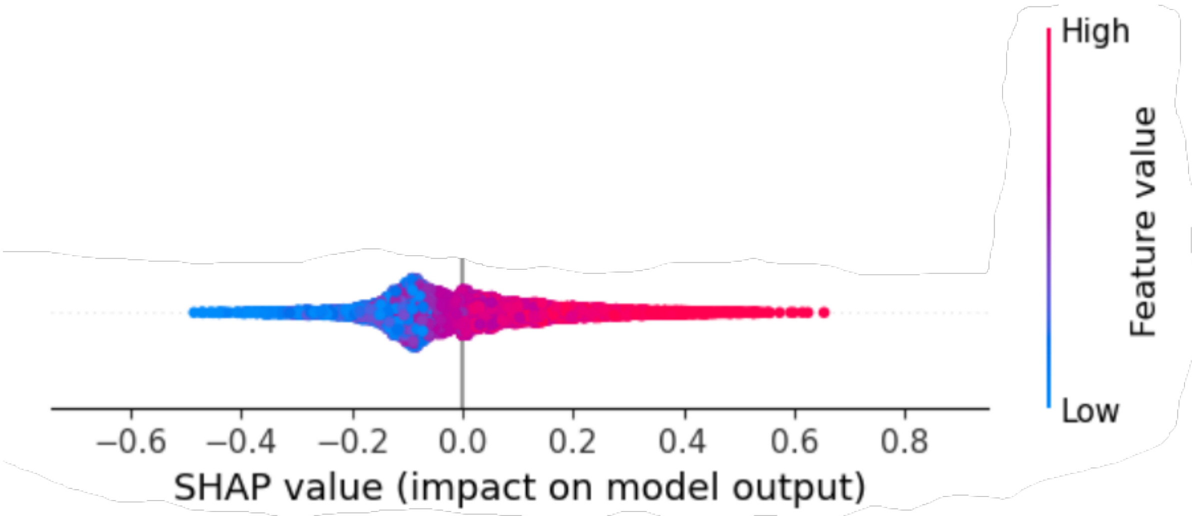


Figure 2.2: An example for the interpretation of SHAP values.

2.16.2. Calculation of Shapley Values

The SHAP library [78, 93] in Python is used for computing the Shapley values. SHAP analyzes feature importance, revealing how each feature impacts the model's predictions on average. To calculate SHAP values for a random forest, a specific method designed for tree-based models is used: the SHAP Tree Explainer [78]. To calculate SHAP values for a logistic regression model the (normal) SHAP explainer is used [93]. For the random forest model a random sample of 100,000 individuals of the dataset is used to calculate the SHAP values. For the logistic regression model a random sample of 10,000 individuals of the dataset is used to calculate the SHAP values. The SHAP Tree explainer (for the random forest model) was more efficient in calculating the SHAP values, therefore a sample of 100,000 could be included.

3 Results

In this chapter the results of training a random forest classifier and logistic regression model on social network, lifestyle, socioeconomic status, and living environment variables will be discussed. First, we will look into the performance of the models. Then we will look into the results of Shapley values, which are used to examine the influence (magnitude and polarity) of each variable on the model outputs. With this the associations of the variables with T2D prevalence can be evaluated. With these results it is possible to answer the research questions, which will be done in chapter 4.

3.1. Prediction Power of the Models

As described in section 2.15.3 random forest and logistic regression models are trained to predict on a person level the prevalence of T2D (operationalized by the use of diabetes medication).

3.1.1. Random Forest Model

There are in total 108 random forest models trained (see table 2.8 for all the hyperparameter combinations). In table 3.1 the hyperparameters and the average precision on the cross-validation sets of the train set are shown for the best, the second best, third best and worst random forest model of the gridsearch. The best model is used for evaluation and has an **average precision of 0.345 on the train set and an average precision of 0.291 on the test set**.

Rank	Average Precision	Trees	Max Depth	Min Samples	Max Features
1	0.2751 ± 0.0077	150	10	20	All
2	0.2749 ± 0.0078	150	10	10	All
3	0.2748 ± 0.0078	100	10	20	All
108	0.2431 ± 0.0073	50	20	5	log2

Table 3.1: The best, second best, third best and worst poerforming model of the random forest gridsearch. The mean average precision (of all 5 cross-validation splits) is shown including the standard deviation. The accompanying hyperparameters (number of trees, maximum depth of eacht tree, minimum number of samples per split and the maximum number of features to consider for the best split) are also shown.

The precision/recall threshold curve for the tain set including the F1 curve is shown in figure 3.1. It can be seen that the F1 is optimal (0.40) for a threshold around 0.65. This threshold is used to evaluate the precision, recall, negative predictive value and the specificity (see table 2.10 for the definitions) for both the train and test set. The results are shown in table 3.2.

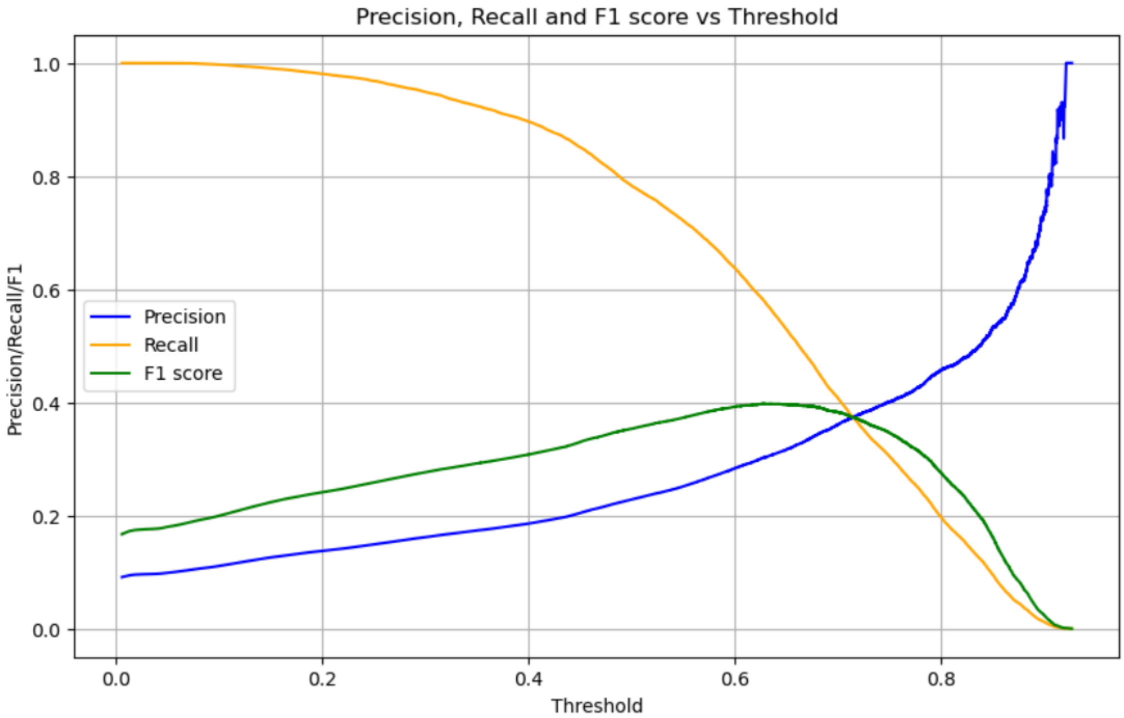


Figure 3.1: The precision/recall threshold curve for the train set of the random forest model.

Train Set				
group				support
	precision	recall	F1	
Diabetes medication	0.32	0.53	0.40	21,180
	negative predictive value	specificity		
No diabetes medication	0.95	0.89		211,122
Test Set				
group				support
	precision	recall	F1	
Diabetes medication	0.28	0.46	0.35	5,396
	negative predictive value	specificity		
No diabetes medication	0.94	0.88		52,680

Table 3.2: Classification report for a threshold of 0.65 for both the train and test sets for the random forest model for the 'No diabetes medication' group and the 'Diabetes medication' group.

The best random forest model of the gridsearch identified 46% of individuals using diabetes medication in the test set successfully for a threshold of 0.65. The precision for predicting diabetes medication use in the test set for that threshold was 32%, meaning that for every correct prediction, there were around two false positives.

When comparing the performance between the training and test sets, the average precision and the precision and recall in the training set were considerably higher than those in the test set (see table 3.2). This suggests that the random forest model is overfitted on the train set and generalizes less well to unseen data of the test set. In appendix D the precision recall curve and precision recall threshold curve for the test set are added for reference.

3.1.2. Logistic Regression Model

There are in total 30 logistic regression models trained (see table 2.9 for the hyperparameter combinations). In table 3.3 the hyperparameters and the average precision on the cross-validation sets of the train set are shown for the best,the second best, third best and worst logistic regression model of the gridsearch are shown. The best model is used for evaluation and has an **average precision of 0.276 on the train set and an average precision of 0.283 on the test set**.

Rank	Average Precision	C	solver	penalty
1	0.2753 ± 0.0095	1.00	saga	L2
2	0.2752 ± 0.0095	1.00	liblinear	L2
3	0.2752 ± 0.0095	1.00	Newton-Cholesky	L2
30	0.2738 ± 0.0096	0.01	saga	L1

Table 3.3: The best, second best, third best and worst poerforming model of the random forest gridsearch. The mean average precision (of all 5 cross-validation splits) is shown including the standard deviation. The accompanying hyperparameters (C: the inverse of the regularization strength, the solver and the penalty) are also shown.

The precision/recall threshold curve for the tain set including the F1 curve is shown in figure 3.2. It can be seen that the F1 is optimal (0.34) for a threshold around 0.65. This threshold is used to evaluate the precision, recall, negative predictive value and the specificity (see table 2.10 for the definitions) for both the train and test set. The results are shown in table 3.4.

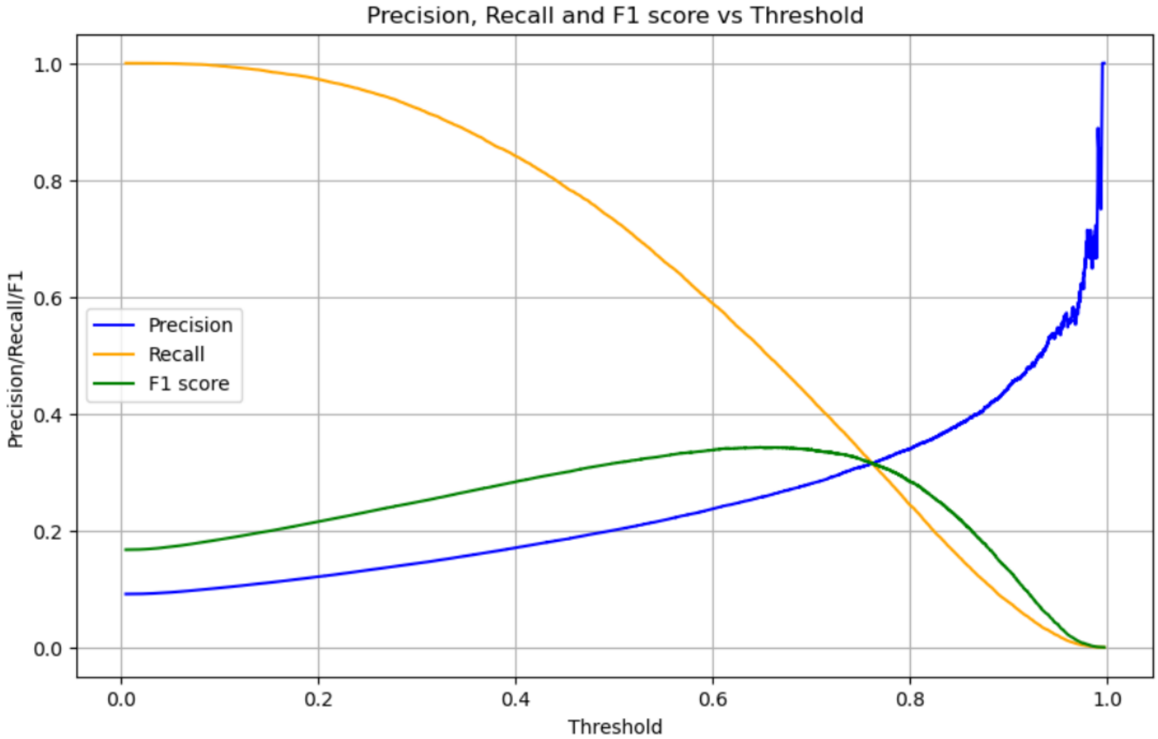


Figure 3.2: The precision/recall threshold curve for the train set of the logistic regression model.

Train Set				
group				support
	precision	recall	F1	
Diabetes medication	0.26	0.51	0.34	21,180
	negative predictive value	specificity		
No diabetes medication	0.95	0.85		211,122
Test Set				
group				support
	precision	recall	F1	
Diabetes medication	0.26	0.52	0.35	5,396
	negative predictive value	specificity		
No diabetes medication	0.95	0.85		52,680

Table 3.4: Classification report for a threshold of 0.65 for both the train and test sets for the logistic regression model for the 'No diabetes medication' group and the 'Diabetes medication' group.

The best logistic regression model of the gridsearch identified 52% of individuals using diabetes medication in the test set for a threshold of 0.65. The precision for predicting diabetes medication use in the test set was 26%, meaning that for every correct prediction, there were around three false positives.

When comparing the performance between the training and test sets (see table 3.4), the average precision and the precision and recall in the training set were very similar to that of the test set. The average precision and the recall are even slightly higher for the test set. This suggests that the logistic regression model is not overfitted on the train set and generalizes well to unseen data of the test set. In appendix D the precision recall curve and precision recall threshold curve for the test set are added for reference.

3.1.3. Post Hoc: Prediction power for using diabetes medication in the near future

Although it was not the primary focus of the research, it is expected that the models may also hold some predictive value for future outcomes. Specifically, the false positives (i.e., individuals incorrectly predicted to have T2D) could represent a particularly interesting group. These individuals may be at high risk of developing T2D in the near future, have undiagnosed T2D, or already have T2D but are not yet on medication, potentially starting diabetes medication in the future. To investigate this, the true negative and the false positive groups are compared and it is found that the model not only has predictive power for current diabetes medication use but also for future use (in 2022, six years later). The groups where no one used diabetes medication in 2016 are the true negatives (TN) and false positives (FP) groups. The FP group consists of individuals for whom the model predicted diabetes medication use, although they did not use it. However, by 2022, 7.91% of the FP group from the random forest model was using diabetes medication, compared to 2.65% of the TN group (see Table 3.5). This indicates that individuals labeled as false positives have, over a six-year period, a three times (2.99) higher chance of using diabetes medication than people that are true negatives. A similar pattern is observed for the logistic regression model, where by 2022, 7.20% of the FP group was using diabetes medication compared to 2.59% of the TN group (see Table 3.5).

	Random Forest		Logistic Regression	
	2016	2022	2016	2022
TN	0%	2.65%	0%	2.59%
FP	0%	7.91%	0%	7.20%
factor		2.99		2.78

Table 3.5: The percentage of diabetes medication use for the true negatives (TN) en false positives (FP) groups for the random forest and logistic regression models for the years 2016 and 2022. Additionally the factor, which is the ratio of the percentage of people using diabetes medication in 2022 in the FP group from 2016 using diabetes medication in 2022 to the percentage of people in the TN group from 2016 using diabetes medication in 2022, is shown.

3.2. The associations of variables with diabetes medication use

Table 3.6 presents the rankings of variables based on their predictive power for both the random forest and logistic regression models, as well as their associations (positive or negative) with diabetes medication use. For additional context, Shapley plots for each variable are provided in Appendix E and F.

The predictive power rankings in table 3.6 are calculated using the mean absolute Shapley values. These values represent a weighted average of the variable's contributions across all individuals in the dataset. As such, the rankings indicate the overall predictive power of a variable for the model as a whole, rather than its importance for individual predictions. This distinction is particularly relevant for nominal categorical variables with low representation in the dataset as those variables may rank low overall but have a significant impact on predictions for individuals that do comply with that category. Examples of variables where this may apply include those related to certain origins (see table 2.1 for representation details), socio-economic categories (see table 2.5), and household positions (see table 2.3).

Additionally, table 3.6 specifies the exact mean absolute Shapley values. Notably, there is a significant difference in predictive power between the highest and lowest-ranked variables, with differences spanning orders of magnitude up to 10^{-5}. This highlights that many variables contribute minimally to the overall model predictions. When comparing the predictive power rankings (based on Shapley values) between the random forest and logistic regression models, it is evident that the top-ranking variables are highly consistent, with eight out of the top ten variables appearing in both models. Notably, several social network factors demonstrate substantial predictive power in both models. Additionally, when focusing on the highest-ranking variables within each factor group and comparing their relative rankings, it becomes clear that lifestyle factors exhibit the strongest predictive power, followed by social network factors, socioeconomics, and finally, living environment variables. The following sections will provide a detailed discussion of the findings for each variable group.

Table 3.6: The direction, positive (+) or negative (-), of association with a variable and diabetes medication use. The variables are in the left column. The rankings of predictive power for the random forest model and the logistic regression model are in the ’RF’ and ’LR’ columns respectively. The SHAP column represents the mean absolute Shapley values. A X after a number means there is no association. A X without a number means that variable does not explicitly exist in the logistic regression model (because of dummy variables). A X! means there is a mixed association, so both positive as well as negative depending on the value of the variable.

Variable	RF	LR	SHAP RF	SHAP LR
self-reported experienced health	1-	1-	1.23-01	1.32-01
age	2+	3+	1.02-01	1.11-01
body mass index	3+	2+	7.99-02	1.18-01
woman	4-	4-	4.44-02	9.90-02
number of alcoholic drinks per week	5-	5-	4.15-02	4.61-02
exposure to family members of the same gender using diabetes medication	6+	7+	1.38-02	3.78-02
exposure to family members of a different gender using diabetes medication	7+	8+	1.22-02	3.61-02
number of minutes of middle intensity exercise per week	8-	22-	9.49-03	7.78-03
exposure to people with master education level	9-	9-	6.57-03	3.46-02
household income percentile	10-	29-	6.48-03	4.81-03
exposure to people with low education level	11+	20+	5.87-03	1.26-02
number of minutes of low intensity exercise per week	12-	30-	5.08-03	4.77-03
exposure to people with bachelor education levels	13-	26-	4.00-03	6.78-03
distance to nearest park	14X!	23-	2.45-03	7.57-03
distance to nearest recreation area	15X!	33+	2.44-03	4.08-03
number of minutes of high intensity exercise per week	16-	38-	2.43-03	3.67-03
adherence to exercise guidelines	17-	15-	2.27-03	2.28-02
distance to nearest forest	18X!	39+	2.02-03	3.58-03
distance to nearest swimming pool	19X!	27+	2.02-03	6.29-03
exposure to people with middle education level	20X	25+	1.97-03	7.25-03
distance to the nearest open dry land	21X!	36+	1.94-03	3.79-03
distance to nearest sports field	22X!	31+	1.85-03	4.24-03
exposure to neighbors of the same gender using diabetes medication	23+	24+	1.65-03	7.43-03
distance to nearest public green space	24X!	37-	1.56-03	3.67-03
alcohol drinker	25-	12-	1.54-03	2.85-02
exposure to colleagues of the same gender using diabetes medication	26+	18+	1.54-03	1.74-02
education level	27-	17-	1.34-03	1.84-02
number of supermarkets and grocery stores within 3 kilometers	28+	32+	1.16-03	4.09-03
never smoked	29-	X	1.16-03	X
exposure to neighbors of a different gender using diabetes medication	30X	35+	9.29-04	3.86-03
Dutch	31-	X	7.90-04	X
urbanity	32+	19+	6.49-04	1.30-02
ex-smoker	33+	11+	6.31-04	3.16-02

Continued on next page

Variable	RF	LR	SHAP RF	SHAP LR
working	34-	14-	5.28-04	2.68-02
Surinamese	35+	40+	5.27-04	2.99-03
retired	36+	13+	4.54-04	2.85-02
exposure to colleagues of a different gender using diabetes medication	37X	48+	4.31-04	7.45-04
partner in a married couple without children living at home	38+	6+	4.09-04	3.84-02
never drank alcohol	39+	X	2.97-04	X
exposure to household members using diabetes medication	40+	41+	2.48-04	2.52-03
loneliness	41-	21-	2.33-04	9.47-03
living alone	42+	10+	2.04-04	3.45-02
ex-alcohol drinker	43+	42-	1.94-04	1.57-03
receiving social benefits (Dutch: bijstand)	44+	55+	1.83-04	3.48-04
partner in married couple with children living at home	45-	28+	1.68-04	4.98-03
European (except Dutch)	46-	51-	1.64-04	5.29-04
smoker	47X	16+	1.42-04	2.16-02
parent in single-parent household	48X	44+	1.27-04	1.43-03
receiving unemployment benefits (Dutch: werkloosheidsuitkering)	49+	52+	9.95-05	4.99-04
Indonesian	50+	34+	8.87-05	4.05-03
unfit for work (Dutch: arbeidsongeschikt)	51X	X	8.15-05	X
partner in unmarried couple without children living at home	52-	43+	8.11-05	1.47-03
receiving social benefits (Dutch: sociale voorzieningen)	53+	50+	7.50-05	6.43-04
other Asian	54+	47+	7.04-05	8.80-04
no income	55-	45+	6.12-05	9.58-04
other African	56+	49+	5.68-05	6.97-04
Moroccan	57+	56+	4.35-05	3.01-04
other type of household member	58X	46+	3.65-05	9.55-04
Turkish	59+	57+	3.32-05	1.17-04
other American & Oceanian	60+	59-	2.69-05	4.67-05
living at parents home	61X	X	2.58-05	X
partner in unmarried couple with children living at home	62X	54-	2.21-05	3.64-04
Dutch Caribbean	63+	53+	2.15-05	4.58-04
reference person in other household	64X	60+	4.79-06	3.77-05
member of institutional household	65X	58-	1.20-06	1.13-04

3.2.1. Social Network

Exposure to Diabetes

Variables measuring exposure to people using diabetes medication demonstrate predictive power for one's own use of diabetes medication, which is in line with hypothesis H1 (see section 1.5). The predictive power of exposure to diabetes medication varies across different network layers. The family network layer shows very strong predictive power, followed by the work and neighborhood layers, and finally the household layer. That the predictive power varies between different network layers is in line with hypothesis H1 and H2 (see section 1.5). Exposure to family members with diabetes of the same and opposite gender ranks among the most predictive variables, placing sixth and seventh in the random forest model and seventh and eight in the logistic regression model. Exposure to people of the same gender who use diabetes medication within one's neighbor and colleague networks also ranks relatively high in predictive power, with positions of twenty-third and twenty-sixth in the random forest and twenty-fourth and eighteenth in the logistic regression models, respectively. All these exposure variables show a positive association with an individual's own use of diabetes medication. In contrast, exposures to individuals using diabetes medication in the neighbor and colleague networks who are of a different gender do not show clear associations in the random forest model, which means that the SHAP values (in addition to being very low) exhibit both positive and negative associations for different individuals. For the logistic regression model exposures to individuals using diabetes medication in the neighbor and colleague networks who are of a different gender show slightly positive associations, although having 23 (1.74×10^{-2} / 7.45×10^{-4}) times less (for the work network) and 1.92 (7.43×10^{-3} / 3.86×10^{-3}) times less (for the neigbor network) predictive power than the same gender exposure scores. This found same-gender effect (for neighbor and work networks) is in line with hypothesis H3 (see section 1.5). Lastly, exposure to people using diabetes medication within one's household (not categorized by gender) has the lowest predictive power but remains positively associated with diabetes medication use in both the random forest and logistic regression models.

Exposure to Education Level

Exposure to individuals with a master's, bachelor's, or low level of education shows clear associations with an individual's use of diabetes medication. Notably, exposure to those with a master's or bachelor's education is negatively associated in both models, while exposure to those with a low level education is positively associated, which is in line with hypothesis H4 (see section 1.5). In both models, the master's, bachelor's and low education level exposure scores rank relatively high (for the random forest model top 13 and in the logistic regression model top 26).

Position in the Household

In the random forest model, all position in the household variables rank at place thirty-eight or lower, while some have higher predictive power in the logistic regression model, such as living alone (rank ten), which is positively associated with diabetes medication use in both models, confirming hypothesis H5 (see section 1.5).

Loneliness

In both models, loneliness is negatively associated with the use of diabetes medication. In terms of predictive power, the loneliness variable ranks forty-first in the random forest model, whereas it ranks higher in the logistic regression model, at twenty-first. This negative association is the opposite of what was expected by hypothesis H6 (see section 1.5).

3.2.2. Living Environment

Food Environment & Urbanity

When analyzing the food environment, a higher number of supermarkets or grocery stores within 3 kilometer is positively associated with diabetes medication use. This applies to both the random forest and the logistic regression model. The variable ranks twenty-eighth for the random forest and thirty-second for the logistic regression model, respectively. Based on this finding, little can be said about hypothesis H7 (see section 1.5), for further explanation about this, see section 4.4.2 in chapter 4. Urbanity is in both models positively associated with diabetes medication use.

Exercise Environment

Looking into the exercise environment variables (distance to nearest public green space, park, recreation area, forest, open dry natural land, sports field and swimming pool) it can be seen that all variables show the same pattern for the random forest model. This pattern is that very short distances to those exercise environments are negatively associated with diabetes medication use, which is in line with hypothesis H8 (see section 1.5). After small distances there is either no association (a Shapley value around 0) with diabetes medication use or there is a slightly positive association with diabetes medication use and for relatively very far distances again a negative association. An example of this is shown in the dependence plot of the distance to nearest recreation area, see figure 3.3. This figure shows that distances of less than approximately 1 kilometer are associated with a lower likelihood of using diabetes medication. For the other exercise environment variables, this maximum distance that is negatively associated ranges from 250 meter to 1 kilometer.

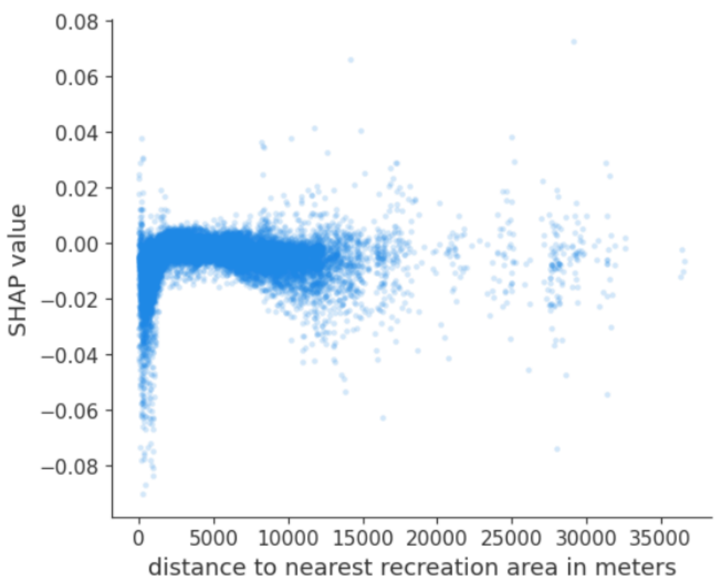


Figure 3.3: The Shapley values for the distance to the nearest recreation area in meters.

Looking at the predictive powers, it is seen for the random forest model that all seven variables about the exercise environment cluster in between ranks fourteen and twenty-four. For the logistic regression model, these variables also cluster , namely between twenty-third and thirty-nine. The associations with exercise environments in the logistic regression model differ with some showing a positive and some showing a negative association with diabetes medication use.

3.2.3. Demographics

The age variable is very predictive as it is for both models in the top three of the predictive power ranking. Age is strongly positively associated with diabetes medication use. The gender variable is also very predictive, it ranks fourth for both models in predictive power. Being female is associated with a lower likelihood of using diabetes medication. Origin does not rank among the most predictive factors, as all origin-related variables fall outside the top 30 in terms of predictive power and most are around the top 50 for both models. Having an origin from the Netherlands appears to have a slightly negative association with diabetes medication use. However, non-Dutch origins, except for 'other European' and 'other American & Oceanian', show a positive correlation with medication use in both models. However, origin is a nominal categorical variable, with many categories having low representation (see Table 2.7). This suggests that, although the overall predictive power may be low, being from a certain origin could still have a significant impact on predictions for individuals. For instance, the 'Surinamese' category demonstrates relatively high predictive power, despite representing only 0.78% of the sample population.

3.2.4. Lifestyle

(Self-reported) Experienced health is the most predictive variable in both models. Better experienced health is negatively associated with diabetes medication use.

Body Mass Index (BMI) is also very predictive for both models. It ranks third most predictive variable for the random forest model and second for the logistic regression model. BMI is positively associated with the use of diabetes medication. A BMI of category 0 (18.5-: underweight), 1 (18.5-20) and 2 (20-25) are negatively associated with diabetes medication use, while a BMI of category 3 (25-30: overweight) and category 4 (30+: obese) are positively associated with diabetes medication use.

Different levels of physical activity intensity (measured in minutes) and adherence to exercise guidelines show consistent and relatively high predictive power across both models regarding an individual's use of diabetes medication and all exercise variables are negatively associated with diabetes medication use. Specifically, time spent on moderate intensity exercise and adherence to exercise guidelines have relatively high predictive power in both models.

Higher alcohol consumption and current alcohol drinking are negatively associated with diabetes medication use and have strong predictive power in both models (ranking fifth for both models). In contrast, never drinking or being an ex-drinker show positive association with diabetes medication use in the random forest model, while in the logistic regression model being an ex-drinker is negatively associated with diabetes medication use.

Regarding smoking variables, the logistic regression model demonstrates stronger and clearer associations between smoking variables and diabetes medication use compared to the random forest model. Notably, current and former smoking are more predictive and positively associated with medication use in the logistic regression model, while these associations are weaker or absent in the random forest model.

3.2.5. Socioeconomic Status

One's own education level is negatively associated with diabetes medication use in both models and has higher overall prediction power in the logistic regression model than in the random forest model.

Employment ('working') is negatively associated with diabetes medication use and retirement positively. These variables are more predictive in the logistic regression model than in the random forest model. Other socioeconomic categories show weak predictive power (outside the top 40) for both models, however there is a notable observation, namely receiving social benefits (both Dutch: 'bijstand' and 'sociale voorzieningen') and receiving unemployment benefits (Dutch: 'werkloosheidsuitkering') are positively associated with diabetes medication use in both models. Although they exhibit low predictive power, the same principle applies as with origin as they are nominal categorical variables with low representation (ranging from 0.44% - 2.74%, see table 2.5) and thus they can still have a significant impact on predictions for individuals that comply with those categories.

Household income percentile is negatively associated with diabetes medication use and is quite predictive in both models.

3.3. Post Hoc: Training on only the people who work

Although the focus of the research was on the whole population of 40+ (see sections 2.2 and 2.5). We decided to do an additional analysis exclusively on working individuals. This decision was motivated by the fact that a significant portion (67.9%) of individuals lack a work network, which can be attributed to their retirement status (82%), voluntary or involuntary not employed (approximately 13%, including those who are unemployed, disabled, or not working), or being self-employed without colleagues (5%). Imputing zeros for these work exposure to diabetes medication scores may distort the predictive power of the work exposure variable. To better evaluate the true predictive power of exposure to colleagues who use diabetes medication a model was trained exclusively on working individuals. The hypothesis was that exposure to colleagues of the same gender using diabetes medication would demonstrate greater predictive power. Two models were trained, namely a random forest model and a logistic regression model, for both using the hyperparameter settings identified previously as optimal (see tables 3.1 and 3.3). For the random forest model, exposure to same gender colleagues using diabetes medication ranked eight in predictive power, while exposure to opposite-gender colleagues ranked twentie-eight. The random forest model achieved an average precision of 0.15 on the test set. For the logistic regression model, exposure to same-gender colleagues ranked fifth, and exposure to opposite-gender colleagues ranked thirtieth, with an average precision of 0.20 on the test set. These results show that focusing exclusively on the working population enhances the predictive power of exposure to colleagues, as it likely removes the noise introduced by imputing zeros for retired individuals. Additionally, the findings reaffirm that same-gender exposure has significantly higher predictive power than opposite gender exposure, thereby again confirming hypothesis H3 (see section 1.5).

4

Discussion & Conclusion

In this chapter the main research questions will be answered, the strengths and limitations of the research will be mentioned. A comparison with the outcomes of the research and the literature will be discussed. Additionally the suitability for policy making are addressed, policy recommendations will be given and the options for future research are mentioned.

4.1. Strengths of the Research

The strengths of this research lie in its use of a unique, comprehensive dataset that links individual-level data across an entire population network, allowing for an unprecedented analysis of the associations between T2D and social network, lifestyle, socioeconomic status and living environment at the population scale. In addition, it provides a comparison between the performance of both a random forest model, which can capture nonlinear interactions, with a logistic regression model, which can only capture linear patterns, providing a nuanced approach to understanding T2D risk factors. Lastly, the discussion section (see section 4.6) addresses how the key findings of this research can be translated into policy advice, aiming to enhance the relevance of this research for promoting social welfare.

4.2. Answers on the Research Questions

1. How do social networks, lifestyle, socioeconomic status, and living environment contribute to the prediction of Type 2 Diabetes prevalence among adults in the Netherlands?

This research highlights the significant association of social networks with the prevalence of T2D (operationalized through diabetes medication use). A strong association was found between diabetes medication use within one's social network and an individual's own use (confirming hypothesis H1). Diabetes medication use among family members emerged as particularly strong predictors, while medication use by same-gender colleagues also showed a significant predictive capacity. In contrast, the use of diabetes medication by colleagues of a different gender had little impact (confirming hypothesis H3).

A similar pattern was observed among neighbors: medication use by neighbors of the same gender was predictive, whereas use by neighbors of a different gender had less predictive power (confirming hypothesis H3). Interestingly, exposure to diabetes medication use within the household exhibited almost no predictive capacity. These findings highlight that the predictive influence of social networks on diabetes medication use varies significantly depending on the type of network (confirming hypothesis H2).

The education level within one's social network also played a key role in prediction. Networks with a higher average education level were negatively associated with one's own diabetes medication use (confirming hypothesis H4).

Living arrangement also played a role, with living alone being positively associated with diabetes medication use (confirming hypothesis H5), although the strength of this association was relatively weak. Contrary to expectations, loneliness was found to be negatively associated with diabetes medication use, albeit weakly (thereby disproving hypothesis H6).

Regarding the living environment, proximity to exercise facilities (within a maximum distance of approximately 1 km) was negatively associated with diabetes medication use (confirming hypothesis H8). However, due to data limitations, no conclusions could be established about the association between an unhealthy food environment and diabetes medication use (neither confirming nor disproving hypothesis H7).

As expected, lifestyle factors such as BMI and adherence to exercise guidelines were highly predictive of diabetes medication use. Specifically, a higher BMI and non-adherence to exercise guidelines were both positively associated with diabetes medication use.

With respect to socioeconomic status, household income and education level were negatively associated with diabetes medication use. Furthermore, receiving social benefits (in Dutch: 'bijstand' and 'sociale voorzieningen') and unemployment benefits (Dutch: 'werkloosheidsuitkering') were positively associated with diabetes medication use, although these associations were relatively weak.

When comparing the predictive power rankings of the variables between the models, consistent patterns, with eight of the top ten variables shared between the models and a similar hierarchy of predictive capacity observed, indicating that lifestyle is most predictive for diabetes medication use, followed by social network, socioeconomic status and, last, living environment.

2. How does the performance of a random forest model compare to that of a logistic regression model in the prediction of Type 2 Diabetes prevalence among adults in the Netherlands based on social network, lifestyle, socioeconomic status, and living environment?

Regarding the performance of the models, the random forest model outperforms the logistic regression model in capturing patterns within the training data, reaching an average precision of 0.345 on the train set against an average precision of 0.276 on the train set for the logistic regression model. Also, as anticipated, the random forest model captures non-linear relationships, in contrast to the logistic regression model, which is limited to identifying linear effects. This distinction was particularly evident for the exercise environment variables, where the random forest model detected non-linear relationships across all exercise environment variables, while the logistic regression model captured linear associations. However, when looking at the performance on the test set, the random forest model reaches a lower average precision, namely 0.291, indicating that it does not generalize well to unseen data. This while the logistic regression model does generalize well and reaches an average precision of 0.283 on the test set. This results in the fact that the random forest and logistic regression model have very similar performance on the unseen (test) data. Returning to hypothesis H9, it can therefore be concluded that the random forest model only very slightly outperforms the logistic regression model on the test dataset in terms of average precision, despite its ability to capture non-linear patterns. Therefore, hypothesis H9 cannot be confirmed.

4.3. Discussion of the Performance of the Models

The average precisions of the models is far from perfect, however it is plausible for two reasons. First, there is some noise in the data due to the presence of individuals with T1D, who make up approximately 10% of the diabetes population [7]. Second, it was anticipated that the set of variables used to train the model would not capture all factors influencing T2D prevalence, and thus, a perfect prediction was never expected. However, the performance being far from perfect calls for caution when interpreting aspects such as the predictive power ranking.

As the random forest model did not outperform the logistic regression model as initially expected, this suggest that there may not be a lot of non-linear relationships between T2D and the included factors. Additionally it is important to note that our study focuses on disease prevalence rather than onset and non-linear relationships might play a more significant role for disease onset.

Further, the random forest model exhibited overfitting on the training set, which is plausible as it is a characteristic of tree-based models, they tend to overfit due to their high flexibility in adapting to the data. However, a more extensive hyperparameter search using wider hyperparameter ranges and other optimization methods, such as Bayesian optimization, Hyperband, successive halving, and evolutionary algorithms could potentially have mitigated some of the overfitting. However for this study we were limited to computational constraints which limited the hyperparameter search.

4.4. Comparison with the Literature & Limitations of the Data

4.4.1. Social Network

Exposure to Diabetes Medication & Exposure to Education Level

Exposure to people using diabetes medication within one's family or workplace or neighborhood networks is predictive of an individual's own diabetes medication use. This indicates a clustering of T2D within personal networks. Very little prior research has explored this area [45], making this finding particularly innovative, especially given the large dataset used in this study. Regarding the association with family it is less surprising as genetics may also play a part [12]. Regarding the role of neighbors, previous research in the U.S. [49] found no clustering of obesity (which is related to T2D [46]) among neighbors. However, our study suggests that for the Netherlands and in the context of T2D, neighbors do exhibit clustering effects. As for colleagues, to the best of my knowledge, no prior research has investigated clustering of T2D within workplace networks. It is therefore a notable finding that colleagues appear to strongly cluster. The finding that gender does not significantly impact clustering within family networks contrasts with previous research on obesity, which showed stronger clustering among same-gender family connections than opposite-gender ones [49]. Additionally, this study offers an innovative perspective by examining the association between the education level of one's social network and the use of diabetes medication. The finding that a highly educated social network is associated with lower diabetes medication use, while a less educated social network is linked to higher diabetes medication use, represents a novel finding and also suggests that the existence of healthy and unhealthy social networks go hand in hand with high- and low-educated social networks.

Several potential explanations for the clustering of diabetes medication use within networks were discussed in the introduction (see Section 1.3.3). However, on the basis of this research we cannot determine whether the effect arises from social influence, shared underlying characteristics or other reasons. The observed gender effects however suggest the presence of social influence, as individuals are more likely to be influenced by those with whom they identify [49].

Due to several limitations of the social network data, we were unable to explore potential relevant associations between network characteristics and diabetes medication use. For instance, previous research has identified several associations between network characteristics, such as network size [45, 54, 84], network quality [45, 54, 84], network directionality [49], and degrees of influence within social networks [49], with T2D and obesity. First, network size was not a usable measure in this study, as the data collection was constrained by artificial caps on the number of colleagues (100) and neighbors (30) (see section 2.8). Second, network quality could not be operationalized, as the relationships were based on registry data, which does not confirm the actual existence or quality of these connections [52, 79, 83, 95]. Additionally, data on network directionality were unavailable. Lastly, the study was unable to assess the influence of degrees of separation for clustering of T2D as only the direct network contacts were considered (see section 2.8).

Moreover, potential relevant network type data was not avaialble. eFor instance, friendship networks could be particularly relevant, as clustering effects for obesity have been predominantly observed in such networks [49]. Furthermore, colleague networks are less relevant for retired individuals, who may instead participate actively in other types of networks, such as volunteer organizations. Similarly, parents who are not employed might be actively involved in school-related networks through their children. However, it is important to note that while these ”schoolyard parent” networks are not included in the (direct) exposure scores for diabetes medication use, they are accounted for in the (indirect) exposure scores for education level as the indirect exposure score incorporates school networks and considers both direct and indirect contacts across all age groups (see Section 2.8.2).

Furthermore, another limitation of the study is that when an individual lacks a specific network, the corresponding exposure score is filled with zero. This approach is particularly problematic for work exposure, as 67.9% of the individuals are not employed (see Table 2.2) and therefore lack a colleague network. Despite this, the exposure to diabetes medication within the work network was found to have relatively high predictive power. Additionally, a separate model trained exclusively on employed individuals demonstrated that this exposure variable had even greater predictive power in that context (see Section 3.3). It is also important to consider the implications of missing data and filling gaps with zeros in other network layers. For example, individuals without a family network may include minority groups, such as immigrants who have moved to the Netherlands without accompanying family members.

Household

The finding that living alone is consistently positively associated with diabetes medication use in both models, aligns with findings in the literature [45, 54].

Loneliness

The finding that loneliness is negatively associated with diabetes medication use is unexpected. Previous studies [55–57] suggested a positive association with the risk of developing T2D [42–44]. One possible explanation for this discrepancy is that high levels of loneliness may be more common among very old individuals, who, due to a survival effect, are often healthier overall.

4.4.2. Living environment

Food environment

No insights about an association between the food environment and diabetes medication use have emerged. This because of several reasons. First, as most variables about the food environment have been excluded from model trainings because they were very highly correlated and that would distort model training and outcomes (see section 2.12). Second, no ruling could be made if supermarkets are healthy as they offer a mix of both healthy and unhealthy food options [64].

Exercise Environment

The random forest model indicates that very close proximity to potential exercise environments is negatively associated with diabetes medication use. However, this relationship diminishes at greater distances, suggesting that beyond a certain threshold, all distances are perceived as 'far.' For these greater distances, other factors may influence the observed associations, such as levels of urbanization or alternative influences not accounted for in our dataset. The logistic regression model finds no clear associations for the exercise variables, likely because it is linear, whereas the findigns of the random forest model suggests that the relationships for these variables are non-linear.

4.4.3. Demographics, Lifestyle and Socioeconomics

Age and gender are highly predictive for diabetes medication use and associations were as expected from literature [10, 11]. The interpretation of the findings about origin is complex due to several factors. First, the representation of different origins in the health monitor data may be insufficiently comprehensive, with half of the origins comprising less than 0.5% of the sample. Second, the origin variable may capture a mix of influences, potentially reflecting genetic predispositions [12], cultural factors like dietary habits, or might even serve as a proxy for unrelated factors such as socioeconomic status within certain groups.

The findings about lifestyle are as expected by literature (see section 1.3.1), except for alcohol drinking. Alcohol use unexpectedly correlates with not using medication, this while it is well known that alcohol is a risk factor for T2D [33, 34]. This finding is likely reflecting medical advice to avoid alcohol when using diabetes medication rather than a protective effect. Further, it is noteworthy that self-reported health emerges as the strongest predictor, highlighting its utility as a simple yet powerful measure.

The findings about socioeconomics are in line with the findings in literature (see section 1.3.2). Further it is noteworthy that, although the association is relatively weak, receiving social or unemployment benefits is positively associated with usage of diabetes medication.

4.5. Suitability for Policy Making

To enhance the understanding of the applicability of the research findings for policy-making, this section examines the limitations and opportunities of the data and models, along with the associated ethical and legal considerations.

4.5.1. Data Limitations

Several limitations in the data impact its use for policy. First, there is no distinction between T1D and T2D in the data nor is there data available about people with T2D who do not use diabetes medication or who have pre-diabetes, the latter of which represents a significantly larger population than those with T2D [5–7]. Those limitations introduce noise into the data and limit the understanding of the relationship between T2D and its risk and protective factors. Second, the health monitor sample may not represent the entire Dutch population, as some groups might be underrepresented. Additionally, the removing of individuals with missings (see section 2.13) led to reduced representation of certain minority groups (for example certain origin groups) and can also introduce bias in the model (see section 2.13). Because of this extra caution is required when interpreting the results of the models. Furthermore, even if the study sample were a perfect representation of the study population, caution is still warranted as the model may underperform for minority groups, where 'minority' refers to any group that is underrepresented relative to the rest of the population. If the model is used to inform policy decisions, it is crucial to gain a deeper understanding of its biases and explore ways to mitigate them as studies have shown that machine learning models often exhibit performance disparities for minority groups [96]. Lastly, the data is from 2016, raising questions about its relevance to the current situation. Changes in work patterns post-COVID-19, such as the increase in remote work [97], could weaken associations like that between colleague networks and diabetes medication use. This is predicated on the assumption that social contagion contributes to the clustering of T2D within social networks, and that reduced in-person interactions affect social contagion. reduced in-person interactions affect social contagion.

4.5.2. Model Limitations & Possibilities

Model limitations also influence its policy applicability. First, the models predict the current presence of diabetes medication use rather than the onset of diabetes, which limits its predictive value for future cases. However, the models have shown some predictive power for the future (see section 3.1.3), indicating that it may be partially useful for this purpose. Additionally, if the models were to be adapted for future predictions (though it is not currently trained for this purpose), it is important to recognize that its decision threshold can be adjusted. For evaluation (see section 3.1), the models' thresholds are configured to balance the costs of false positives and false negatives equally, meaning recall and precision are given equal importance. However, by tuning this threshold, the costs of false positives and false negatives can be rebalanced to align with specific policy goals. For instance, precision could be prioritized (placing a higher cost on false positives) in scenarios where interventions involve substantial expenses, such as personalized lifestyle programs. Conversely, recall could be emphasized (placing a higher cost on false negatives) in cases where the goal is to reach as many at-risk individuals as possible, especially when subsequent actions, such as an informational campaign, involve minimal costs. It should also be noted that due to the models' insufficient performance at the individual level, it is not suitable for making predictions about specific individuals. However, it can still be effectively employed to compare groups and prioritize among them, helping to identify the highest-priority groups for targeted intervention programs.

4.5.3. Ethical and Legal Considerations

Ethical and legal considerations further shape the model's role in policy making. It is crucial to view the model as an informative, not a decisive, tool for policy decisions, given that the model is a simplified version of reality and many factors outside the model contribute to policy choices. Additionally, the use of personal data without explicit consent, linked solely for research, raises ethical and legal concerns about its application for policy guidance. Although, it is good to know that CBS data can never be used for research that can be traced back to individual persons; conclusions can only be drawn about (sufficiently large) groups. Finally, access to data and models is restricted, stored in a secure CBS environment, making it currently impossible for external use for policy development.

4.6. Policy Recommendations

The goal of this research was to identify key leverage points for effective interventions aimed at reducing the incidence of T2D (see section 1.4). Several findings from this study offer valuable insights for the development of preventive policies.

Firstly, lifestyle interventions remain a critical pillar for policy focus. Encouraging and facilitating a healthy BMI and ensuring that individuals meet minimum weekly physical activity guidelines could significantly mitigate the risk of T2D. Targeting social networks as a basis for policy interventions appears particularly promising due to the observed clustering of T2D cases within these social networks.

Within family networks, it is important to also inform and support family members of individuals diagnosed with T2D about T2D and preventive practices as they seem to have a heightened risk.

In terms of workplace networks, targeting low-SES colleague groups with a focus on gender-based differences could be a promising area for intervention. While workplace interventions have been identified as potential policy avenues in prior research [98], gender-based approaches are unexplored. Policies could place legal responsibility on employers to promote physical activity during work hours, enabling employees to meet recommended physical activity guidelines during the workday. Such regulatory measures would align with existing recommendations that advocate for more comprehensive workplace health policies, particularly in low-SES environments [98]. Current policy advice emphasizes the need to improve the food environment in workplaces [98], but mandating physical activity as a legal requirement for employers is an innovative approach. Additionally, this also seems highly necessary, as the Netherlands is holding the title for the most sitting in Europe [99].

Policies must also be inclusive, targeting all societal groups, including those who are unemployed. Unemployed individuals, particularly those receiving social benefits, seem to have a heightened risk of developing T2D. Therefore, policy initiatives could extend beyond financial assistance for people receiving social benefits to also include lifestyle support programs.

Neighborhood-level policies should also be strengthened, focusing on social networks within communities and ensuring the availability of spaces conducive to walking and exercise within a radius 1 km. This approach should include targeted interventions in low-SES neighborhoods and incorporate gender-specific strategies. Current policy advice already recommends this as well [98], however, here again the gender aspect is innovative advice. An added advantage of focusing on community-level policies is their inclusivity, potentially benefiting retirees who do not have access to workplace networks and who also face increased T2D risk.

Finally, greater emphasis on educational diversity within work, and neighborhood social networks could be beneficial. The research highlights that the educational level of an individual's network has a strong correlation with T2D risk. This suggests that educational segregation in society may contribute to health disparities. On a positive note, educational segregation has decreased in recent years in the Netherlands [52, 53].

4.7. Future Research

For future research three main focus points are interesting. The first is to focus on being more informing for policy making by improving and changing data and models. The second thing is to apply the research now done for T2D in other areas. The third point is to investigate the feasibility of the policy recommendations (described in chapter 4.6).

Concerning the first focus, the models could be more useful and informing in several ways:

Prognostic Model

First, it is valuable for policymakers to have a model specifically designed to forecast future T2D. This as a prognostic model can help identify people at high risk for T2D earlier and provide them with personalized lifestyle guidance, which is in line with the objective of the Diabetes Fund [19]. Additionally, a prognostic model could be used to simulate policy effects. For instance, if a policy is estimated to increase physical activity by 10%, the model could then project the resulting reduction in T2D cases. While the current model does offer some predictive insights, this has been more of an incidental benefit. Developing a model using longitudinal data that is intentionally focused on future prediction would be highly beneficial.

Optimize for a Top-Performing Model

The focus for our research was on understanding the associations between variables and T2D and not on creating a top-performing model. However, for a prognostic model, optimizing model performance even further is beneficial. This could involve feature selection (based on the mean average SHAP values (see table 3.6) or by doing additional ablation analyses), leveraging insights from the random forest and logistic regression models to create hard-coded features that capture interaction effects and discrete transitions, as well as more comprehensive hyperparameter tuning. Additionally, exploring other high-performing machine learning models, like XGBoost [100], might enhance the model's predictive accuracy.

A deeper exploration is required into how various factors are associated with T2D for different demographic and minority groups, with the ultimate goal of mitigating bias.

Running separate models for different demographic groups

First, it can be helpful to run separate models for different demographic groups, as for example the context for older adults often differs significantly from that of younger, working individuals, which may hinder the model's ability to identify relevant patterns. Creating separate models for different age groups, distinguishing between working individuals and retirees might provide deeper insights into their unique relationships and dynamics, ultimately enhancing the understanding of how different factors influence each group.

Examine model performance for minority groups

Additionally, gaining a better understanding on the model's performance with respect to minority groups is needed. This includes investigating whether the model consistently underperforms for certain minority groups, resulting in higher rates of false positives or false negatives for these groups. If this is the case, potential avenues for addressing these issues include the use of imputation methods, training separate models specifically for minority populations, or incorporating weighting mechanisms to adjust for the penalties associated with misclassifications.

Investigate the cause and impact of 'Item Non Response' missings

Furthermore, focusing on a deeper investigation of item non-response of the health monitor survey would be helpful to understand and mitigate bias. Understanding why individuals choose not to answer certain questions, whether due to embarrassment (e.g., related to lack of physical activity), survey fatigue or other reasons, could provide valuable insights. Methods such as follow-up phone interviews or additional qualitative research could help identify underlying reasons, which can help to understand the impact of these missings on the model outcomes.

Create T2D Personas

When a clearer understanding is obtained of how various factors are associated with different demographic and minority groups, it becomes possible to develop T2D personas. These personas are fictional, data-driven representations of at-risk groups, characterized by attributes such as age, lifestyle, and socioeconomic background. The creation of T2D personas aligns with the Diabetes Fund's objective of identifying individuals at high risk for T2D and enabling personalized prevention strategies [19]. By leveraging these personas, policymakers can design and implement more targeted and effective interventions to address specific needs.

Understand Underlying Mechanisms

It would also be valuable to gain a deeper understanding of the mechanisms underlying the relationship between social network factors and their associations with T2D. As mentioned, this relationship could potentially be explained by social contagion, shared underlying characteristics (see Section 1.3.3), or other contributing factors. To explore these dynamics more comprehensively, longitudinal data could be utilized. Such data would enable the investigation of causality, as well as the identification of mediation and moderation within these associations.

For the second focus—applying those research methods to other areas to inform policy—several avenues for future research would be valuable:

Transfer learning about Depression
It is interesting to look at other health outcomes. Specifically, depression, since research already indicates that psychological symptoms can spread in social networks and work environments [101–105].

Investigate the effect of educational segregation on health outcomes
This research suggests that there is existence of healthy social networks and unhealthy social networks and that those overlap with educational social networks suggesting that educational segregation in society may contribute to health inequities, providing an intriguing opportunity for further research if this is indeed the case and if so what the mechanisms are that control how educational segregation permeates health outcomes.

Concerning the third focus, to investigate the feasibility of the policy recommendations, the following would be valuable to further investigate:

Investigate possibility to legally mandate employers
Regarding the policy advice to place legal responsibility on employers to promote physical activity during work hours (see section 4.6) research is needed to explore the feasibility of this. This includes examining legal frameworks, considering the needs of people with disabilities, and evaluating potential negative side effects.

4.8. Take Home Message

While existing research identifies lifestyle behavior as a key determinant of T2D, our findings underscore the significant role of an individual's social network in shaping T2D risk. As a result, prevention and intervention strategies in the Netherlands should extend beyond individual-level approaches to incorporate group-based interventions tailored to specific at-risk populations. An innovative policy recommendation emerging from this research is to place legal responsibility on employers to encourage physical activity during work hours. This approach could help employees meet recommended exercise guidelines. The most interesting avenue for future research seems to develop a robust prognostic model to be able to identify high-risk individuals early and support them with personalized lifestyle interventions to prevent or delay the onset of T2D.

References

1. Abdul Basith Khan, M. *et al.* Epidemiology of type 2 diabetes—global burden of disease and forecasted trends. *Journal of epidemiology and global health* **10,** 107–111 (2020).
2. Who, W. H. O. Diabetes. *World Health Organization: WHO.* `https://www.who.int/news-room/fact-sheets/detail/diabetes` (Apr. 2023).
3. Lovic, D. *et al.* The growing epidemic of diabetes mellitus. *Current vascular pharmacology* **18,** 104–109 (2020).
4. Prasad, R. B. & Groop, L. Genetics of type 2 diabetes—pitfalls and possibilities. *Genes* **6,** 87–123 (2015).
5. Nielen, M., Poos, R. & Korevaar, J. Diabetes mellitus in Nederland. *Prevalentie en incidentie: heden, verleden en toekomst. Utrecht: Nivel* (2020).
6. *Diabetes mellitus | Leeftijd en geslacht | Volksgezondheid en Zorg* [Online; accessed 28. Mar. 2024]. Mar. 2024. `https://www.vzinfo.nl/diabetes-mellitus/leeftijd-en-geslacht`.
7. *Diabetes in cijfers* [Online; accessed 28. Mar. 2024]. Mar. 2024. `https://www.diabetesfonds.nl/over-diabetes/diabetes-in-het-algemeen/diabetes-in-cijfers`.
8. *Ranglijsten | Aandoeningen op basis van zorguitgaven | Volksgezondheid en Zorg* [Online; accessed 29. Mar. 2024]. Mar. 2024. `https://www.vzinfo.nl/ranglijsten/aandoeningen-op-basis-van-zorguitgaven`.
9. *Diabetes mellitus | Zorguitgaven | Volksgezondheid en Zorg* [Online; accessed 29. Mar. 2024]. Mar. 2024. `https://www.vzinfo.nl/diabetes-mellitus/zorguitgaven`.
10. *Diabetes mellitus* **|** *Volksgezondheid en Zorg* `https://www.vzinfo.nl/diabetes-mellitus`. (Accessed on 10/14/2024).
11. Wild, S., Roglic, G., Green, A., Sicree, R. & King, H. Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. *Diabetes care* **27,** 1047–1053 (2004).
12. Galicia-Garcia, U. *et al.* Pathophysiology of type 2 diabetes mellitus. *International journal of molecular sciences* **21,** 6275 (2020).
13. *Levensverwachting mensen met diabetes aanzienlijk lager | RIVM* [Online; accessed 28. Mar. 2024]. Mar. 2024. `https://www.rivm.nl/nieuws/levensverwachting-mensen-met-diabetes-aanzienlijk-lager`.
14. *Diabetes - long-term effects* [Online; accessed 29. Mar. 2024]. Mar. 2024. `https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/diabetes-long-term-effects`.
15. Collaboration, E. R. F. *et al.* Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. *The lancet* **375,** 2215–2222 (2010).
16. Geurten, R. J. *et al.* Identifying and delineating the type 2 diabetes population in the Netherlands using an all-payer claims database: characteristics, healthcare utilisation and expenditures. *BMJ open* **11,** e049487 (2021).

17. Cholerton, B., Baker, L. D., Montine, T. J. & Craft, S. Type 2 diabetes, cognition, and dementia in older adults: toward a precision health approach. *Diabetes Spectrum* **29,** 210–219 (2016).
18. *Diabetes tot 2025. Preventie en zorg in samenhang | RIVM* `https://www.rivm.nl/publicaties/diabetes-tot-2025-preventie-en-zorg-in-samenhang`. (Accessed on 09/29/2024).
19. *'Diabetescrisis': veel meer mensen met voorstadium diabetes* `https://nos.nl/artikel/2520281-diabetescrisis-veel-meer-mensen-met-voorstadium-diabetes`. (Accessed on 09/29/2024).
20. *Nationaal Preventieakkoord | Convenant | Rijksoverheid.nl* `https://www.rijksoverheid.nl/documenten/convenanten/2018/11/23/nationaal-preventieakkoord`. (Accessed on 09/29/2024).
21. Hu, F. B. *et al.* Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. *New England journal of medicine* **345,** 790–797 (2001).
22. Ojo, O. *Dietary intake and type 2 diabetes* 2019.
23. Petroni, M. L. *et al.* Nutrition in patients with type 2 diabetes: present knowledge and remaining challenges. *Nutrients* **13,** 2748 (2021).
24. Del Carmen Fernández-Fígares Jiménez, M. Plant foods, healthy plant-based diets, and type 2 diabetes: a review of the evidence. *Nutrition Reviews,* nuad099 (2023).
25. Schellenberg, E. S., Dryden, D. M., Vandermeer, B., Ha, C. & Korownyk, C. Lifestyle interventions for patients with and at risk for type 2 diabetes: a systematic review and meta-analysis. *Annals of internal medicine* **159,** 543–551 (2013).
26. Uusitupa, M. *et al.* Prevention of type 2 diabetes by lifestyle changes: a systematic review and meta-analysis. *Nutrients* **11,** 2611 (2019).
27. Tuomilehto, J. *et al.* Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. *New England journal of medicine* **344,** 1343–1350 (2001).
28. Mendenhall, E., Kohrt, B. A., Norris, S. A., Ndetei, D. & Prabhakaran, D. Non-communicable disease syndemics: poverty, depression, and diabetes among low-income populations. *The Lancet* **389,** 951–963 (2017).
29. Mendenhall, E., Newfield, T. & Tsai, A. C. Syndemic theory, methods, and data. *Social Science & Medicine (1982)* **295,** 114656 (2022).
30. Chater, N. & Loewenstein, G. The i-frame and the s-frame: How focusing on individual-level solutions has led behavioral public policy astray. *Behavioral and Brain Sciences* **46,** e147 (2023).
31. Campagna, D. *et al.* Smoking and diabetes: dangerous liaisons and confusing relationships. *Diabetology & metabolic syndrome* **11,** 1–12 (2019).
32. *Quitting smoking cuts your risk of developing type 2 diabetes by 30–40%* `https://www.who.int/news/item/14-11-2023-quitting-smoking-cuts-your-risk-of-developing-type-2-diabetes-by-30-40`. (Accessed on 04/29/2024). Nov. 2023.
33. Kim, S.-J. & Kim, D.-J. Alcoholism and diabetes mellitus. *Diabetes & metabolism journal* **36,** 108–115 (2012).
34. Wannamethee, S., Shaper, A., Perry, I. & Alberti, K. Alcohol consumption and the incidence of type II diabetes. *Journal of Epidemiology & Community Health* **56,** 542–548 (2002).

35. Knutson, K. L., Ryden, A. M., Mander, B. A. & Van Cauter, E. Role of sleep duration and quality in the risk and severity of type 2 diabetes mellitus. *Archives of internal medicine* **166,** 1768–1774 (2006).

36. Shan, Z. *et al.* Sleep duration and risk of type 2 diabetes: a meta-analysis of prospective studies. *Diabetes care* **38,** 529–537 (2015).

37. Maddatu, J., Anderson-Baucum, E. & Evans-Molina, C. Smoking and the risk of type 2 diabetes. *Translational Research* **184,** 101–107 (2017).

38. Willi, C., Bodenmann, P., Ghali, W. A., Faris, P. D. & Cornuz, J. Active smoking and the risk of type 2 diabetes: a systematic review and meta-analysis. *Jama* **298,** 2654–2664 (2007).

39. *StatLine - Gezondheid en zorggebruik; persoonskenmerken, 2014-2021* `https://opendata.cbs.nl/statline/#/CBS/nl/dataset/83005NED/table?dl=F25F`. (Accessed on 05/21/2024).

40. *Sociaal economische Gezondheidsverschillen (SEGV)* `https://www.pharos.nl/factsheets/sociaaleconomische-gezondheidsverschillen-segv/`. (Accessed on 05/21/2024).

41. Wang, J. & Geng, L. Effects of socioeconomic status on physical and psychological health: lifestyle as a mediator. *International journal of environmental research and public health* **16,** 281 (2019).

42. Kelly, S. J. & Ismail, M. Stress and type 2 diabetes: a review of how stress contributes to the development of type 2 diabetes. *Annual review of public health* **36,** 441–462 (2015).

43. Mishra, D. N. *et al.* Stress etiology of type 2 diabetes. *Current diabetes reviews* **18,** 50–56 (2022).

44. Sharma, K., Akre, S., Chakole, S. & Wanjari, M. B. Stress-induced diabetes: a review. *Cureus* **14** (2022).

45. Schram, M. T., Assendelft, W. J., van Tilburg, T. G. & Dukers-Muijrers, N. H. Social networks and type 2 diabetes: a narrative review. *Diabetologia* **64,** 1905–1916 (2021).

46. Yashi, K. & Daley, S. F. Obesity and Type 2 Diabetes (2023).

47. Leitner, D. R. *et al.* Obesity and type 2 diabetes: two diseases with a need for combined treatment strategies-EASO can lead the way. *Obesity facts* **10,** 483–492 (2017).

48. Zhang, S., De La Haye, K., Ji, M. & An, R. Applications of social network analysis to obesity: a systematic review. *Obesity reviews* **19,** 976–988 (2018).

49. Christakis, N. A. & Fowler, J. H. The spread of obesity in a large social network over 32 years. *New England journal of medicine* **357,** 370–379 (2007).

50. Smith, N. R., Zivich, P. N. & Frerichs, L. Social influences on obesity: Current knowledge, emerging methods, and directions for future research and practice. *Current nutrition reports* **9,** 31–41 (2020).

51. Berkman, L. F., Kawachi, I. & Glymour, M. M. *Social epidemiology* (Oxford University Press, 2014).

52. *Opleidingssegregatie in Nederland gedaald | CBS* `https://www.cbs.nl/nl-nl/nieuws/2023/15/opleidingssegregatie-in-nederland-gedaald`. (Accessed on 04/17/2024).

53. *Opleidingssegregatie* `https://dashboards.cbs.nl/v4/opl_segregatie/`. (Accessed on 04/17/2024).

54. Brinkhues, S. *et al.* Social network characteristics are associated with type 2 diabetes complications: the Maastricht study. *Diabetes care* **41,** 1654–1662 (2018).

55. Henriksen, R. E., Nilsen, R. M. & Strandberg, R. B. Loneliness increases the risk of type 2 diabetes: a 20 year follow-up–results from the HUNT study. *Diabetologia* **66,** 82–92 (2023).

56. Rosenkilde, S. *et al.* Loneliness and the risk of type 2 diabetes. *BMJ Open Diabetes Research and Care* **12,** e003934 (2024).

57. Song, Y. *et al.* Social isolation, loneliness, and incident type 2 diabetes mellitus: results from two large prospective cohorts in Europe and East Asia and Mendelian randomization. *EClinicalMedicine* **64** (2023).

58. Cohen, S. Social relationships and health. *American psychologist* **59,** 676 (2004).

59. Swinburn, B. A. *et al.* The global syndemic of obesity, undernutrition, and climate change: the Lancet Commission report. *The lancet* **393,** 791–846 (2019).

60. De Ruijter, A. *et al.* Tussen Mens En Ruimte. De (On) gezonde Voedselomgeving Als Omgevingswaarde (Between People and Space. The (Un) healthy Food Environment as an Environmental Value). *De (On) gezonde Voedselomgeving Als Omgevingswaarde (Between People and Space. The (Un) healthy Food Environment as an Environmental Value)(December 6, 2023). Amsterdam Law School Research Paper* (2023).

61. Downs, S. M., Ahmed, S., Fanzo, J. & Herforth, A. Food environment typology: advancing an expanded definition, framework, and methodological approach for improved characterization of wild, cultivated, and built food environments toward sustainable diets. *Foods* **9,** 532 (2020).

62. Wansink, B. & Sobal, J. Mindless eating: The 200 daily food decisions we overlook. *Environment and Behavior* **39,** 106–123 (2007).

63. De Krom, M., Vonk, M. & Muilwijk, H. *Voedselconsumptie veranderen: bouwstenen voor beleid om verduurzaming van eetpatronen te stimuleren* (PBL, Planbureau voor de Leefomgeving, 2020).

64. Poelman, M. *et al. Monitoring van de mate van gezondheid van het aanbod en de promoties van supermarkten en out-of-home-ketens: Inzicht in de huidige stand van zaken en aanbevelingen voor het opzetten van een landelijke monitor* (Wageningen University & Research, 2021).

65. Hendriksen, A. *et al.* How healthy and processed are foods and drinks promoted in supermarket sales flyers? A cross-sectional study in the Netherlands. *Public Health Nutrition* **24,** 3000–3008 (2021).

66. Van Erpecum, C.-P. L., van Zon, S. K., Bültmann, U. & Smidt, N. The association between fast-food outlet proximity and density and Body Mass Index: Findings from 147,027 Lifelines Cohort Study participants. *Preventive Medicine* **155,** 106915 (2022).

67. Ntarladima, A.-M. *et al.* Associations between the fast-food environment and diabetes prevalence in the Netherlands: a cross-sectional study. *The Lancet Planetary Health* **6,** e29–e39 (2022).

68. Poelman, M. *et al.* Relations between the residential fast-food environment and the individual risk of cardiovascular diseases in The Netherlands: A nationwide follow-up study. *European journal of preventive cardiology* **25,** 1397–1405 (2018).

69. Harbers, M. C. *et al.* Residential exposure to fast-food restaurants and its association with diet quality, overweight and obesity in the Netherlands: a cross-sectional analysis in the EPIC-NL cohort. *Nutrition Journal* **20,** 56 (2021).

70. Poelman, M. P. *et al.* Does the neighbourhood food environment contribute to ethnic differences in diet quality? Results from the HELIUS study in Amsterdam, the Netherlands. *Public health nutrition* **24,** 5101–5112 (2021).

71. Hoenink, J. C., Eisink, M., Adams, J., Pinho, M. G. & Mackenbach, J. D. Who uses what food retailers? A cluster analysis of food retail usage in the Netherlands. *Health & place* **81,** 103009 (2023).

72. Smith, D. M. & Cummins, S. Obese cities: how our environment shapes overweight. *Geography Compass* **3,** 518–535 (2009).

73. Frank, L. D., Saelens, B. E., Powell, K. E. & Chapman, J. E. Stepping towards causation: do built environments or neighborhood and travel preferences explain physical activity, driving, and obesity? *Social science & medicine* **65,** 1898–1914 (2007).

74. Poortinga, W. Perceptions of the environment, physical activity, and obesity. *Social science & medicine* **63,** 2835–2846 (2006).

75. Giles-Corti, B. & Donovan, R. J. Socioeconomic status differences in recreational physical activity levels and real and perceived access to a supportive physical environment. *Preventive medicine* **35,** 601–611 (2002).

76. Li, F., Fisher, K. J., Brownson, R. C. & Bosworth, M. Multilevel modelling of built environment characteristics related to neighbourhood walking activity in older adults. *Journal of Epidemiology & Community Health* **59,** 558–564 (2005).

77. Vreke, J., Donders, J., Langers, F., Salverda, I. & Veeneklaas, F. *Potenties van groen!: de invloed van groen in en om de stad op overgewicht bij kinderen en op het binden van huishoudens met midden-en hoge inkomens aan de stad* tech. rep. (Alterra, 2006).

78. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence* **2,** 56–67 (2020).

79. Van der Laan, J., de Jonge, E., Das, M., Te Riele, S. & Emery, T. A whole population network and its application for the social sciences. *European sociological review* **39,** 145–160 (2023).

80. *Nation-scale social networks – SODAS - University of Copenhagen* `https://sodas.ku.dk//projects/nation-scale-social-networks/`. (Accessed on 11/28/2024).

81. De Zoete, B. *Measuring Social Capital in a Population-scale Social Network* MA thesis (Sept. 1, 2022). `https://theses.liacs.nl/2319` (2023). published.

82. Bos, B. *et al.* Persoonsnetwerken en criminaliteit van Neder-landse jongeren. *Tijdschrift voor Criminologie* **64,** 170 (2022).

83. *Herkomstsegregatie in Nederland: een netwerkanalyse | CBS* `https://www.cbs.nl/nl-nl/longread/statistische-trends/2024/herkomstsegregatie-in-nederland-een-netwerkanalyse`. (Accessed on 04/17/2024).

84. Hempler, N. F., Joensen, L. E. & Willaing, I. Relationship between social network, social support and health behaviour in people with type 1 and type 2 diabetes: cross-sectional studies. *BMC public health* **16,** 1–7 (2016).

85. *De Gezondheidsmonitors | Gezondheidsmonitor* `https://www.monitorgezondheid.nl/`. (Accessed on 10/01/2024).

86. Centraal Bureau voor de Statistiek. A Person Network of the Netherlands. *Centraal Bureau voor de Statistiek.* https://www.cbs.nl/nl-nl/achtergrond/2022/20/a-person-network-of-the-netherlands (May 2022).

87. Voor de Statistiek, C. B. *Tijdreeks persoonsnetwerkbestanden: overzicht van de verschillen met de eerste versie van het persoonsnetwerk* https://www.cbs.nl/-/media/cbs-op-maat/microdatabestanden/documents/2023/36/overzicht_verschillen_oude_nieuwe_persoonsnetwerk.pdf. [Accessed 23-12-2024]. 2023.

88. *Beweegrichtlijnen 2017* [Online; accessed 9. Dec. 2024]. Dec. 2024. https://www.gezondheidsraad.nl/documenten/adviezen/2017/08/22/beweegrichtlijnen-2017.

89. Pearson, K. VII. Note on regression and inheritance in the case of two parents. *proceedings of the royal society of London* **58,** 240–242 (1895).

90. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12,** 2825–2830 (2011).

91. Hariharan, S. Statistical test for MCAR in python... - Towards Data Science. *Medium.* https://towardsdatascience.com/statistical-test-for-mcar-in-python-9fb617a76eac (Dec. 2021).

92. Breiman, L. Random forests. *Machine learning* **45,** 5–32 (2001).

93. Lundberg, S. & Lee, S. *A unified approach to interpreting model predictions. Part of Advances in Neural Information Processing Systems 30* in *31st Conference on Neural Information Processing Systems (NIPS December 2017), Long Beach, CA. NeuroIPS Proceedings* (2017).

94. Shapley, L. S. A value for n-person games. *Contribution to the Theory of Games* **2** (1953).

95. *Mensen met Nederlandse herkomst hebben meest gesegregeerde netwerk | CBS* https://www.cbs.nl/nl-nl/nieuws/2024/08/mensen-met-nederlandse-herkomst-hebben-meest-gesegregeerde-netwerk. (Accessed on 04/17/2024).

96. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* **54,** 1–35 (2021).

97. TNO. *Aantal thuiswerkuren sinds coronapandemie fors gestegen* Accessed: 2024-11-05. https://www.tno.nl/nl/newsroom/2023/10/corona-thuiswerkuren-gestegen/.

98. Boer, J. *et al.* Preventief gezondheidsbeleid 2006-2018. Wat zijn de effecten en lessen? (2021).

99. *Nederland Europees kampioen zitten | TNO* [Online; accessed 7. Jan. 2025]. Dec. 2024. https://www.tno.nl/nl/newsroom/2024/02/nederland-europees-kampioen-zitten.

100. Chen, T. & Guestrin, C. *XGBoost: A Scalable Tree Boosting System* in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, San Francisco, California, USA, 2016), 785–794. ISBN: 978-1-4503-4232-2. http://doi.acm.org/10.1145/2939672.2939785.

101. Lee, D. & Lee, B. The role of multilayered peer groups in adolescent depression: A distributional approach. *American Journal of Sociology* **125,** 1513–1558 (2020).

102. Bakker, A. B., Le Blanc, P. M. & Schaufeli, W. B. Burnout contagion among intensive care nurses. *Journal of advanced nursing* **51,** 276–287 (2005).

103. Alho, J. *et al.* Transmission of Mental Disorders in Adolescent Peer Networks. *JAMA psychiatry* (2024).

104. Kensbock, J. M., Alkærsig, L. & Lomberg, C. The epidemic of mental disorders in business—How depression, anxiety, and stress spread across organizations through employee mobility. *Administrative Science Quarterly* **67,** 1–48 (2022).

105. Rosenquist, J. N., Fowler, J. H. & Christakis, N. A. Social network determinants of depression. *Molecular psychiatry* **16,** 273–281 (2011).

A
Living Environment Graphs

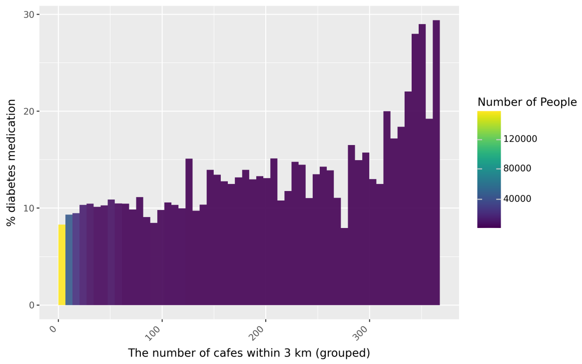


Figure A.1: Cafe

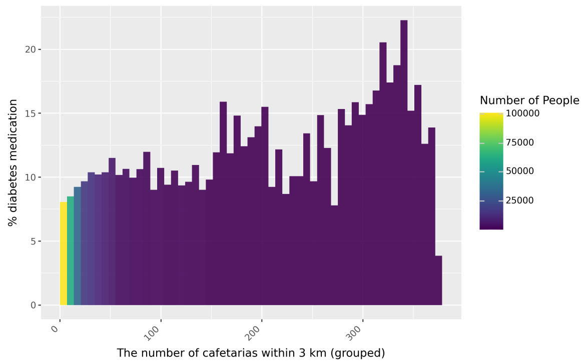


Figure A.2: Cafetaria

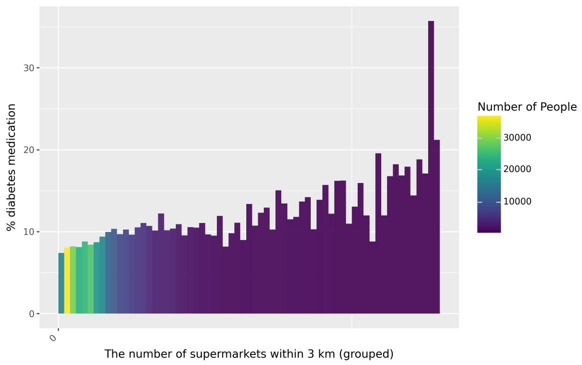


Figure A.3: Supermarkets & Grocery stores

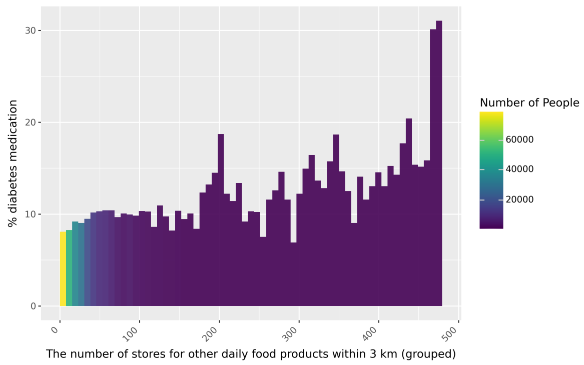


Figure A.4: other daily food stores

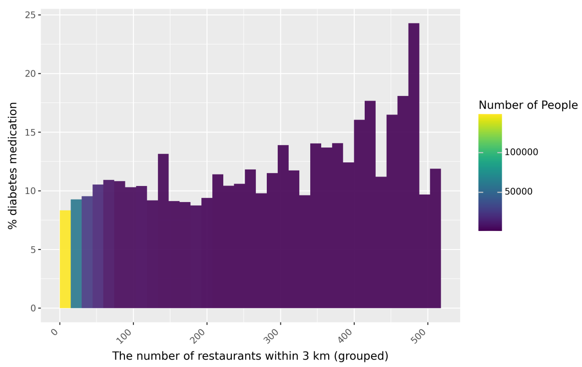


Figure A.5: Restaurant

Figure A.6: The distribution plots of the food environment variables. The x-axis shows the distance in meters to the nearest one (variable of the exercise environment). The y-axis shows the percentage of diabetes medication use. For privacy reasons and readability only bins that include at least 50 people are included.

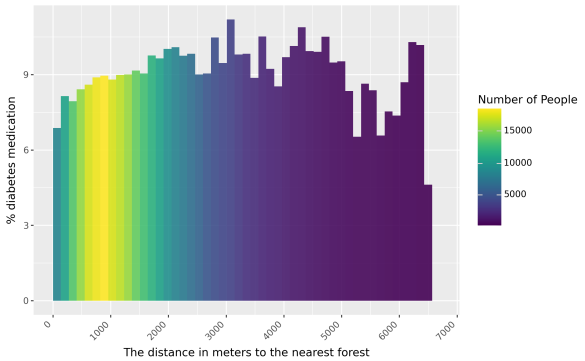


Figure A.7: Forest

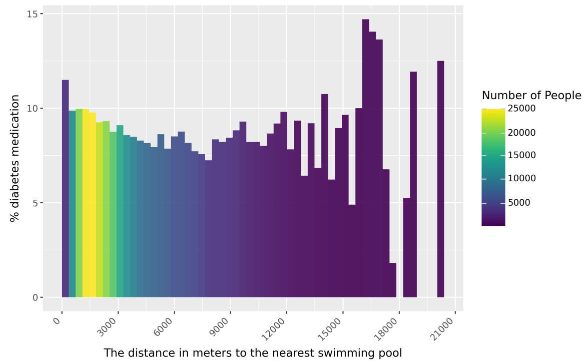


Figure A.8: Swimming Pool

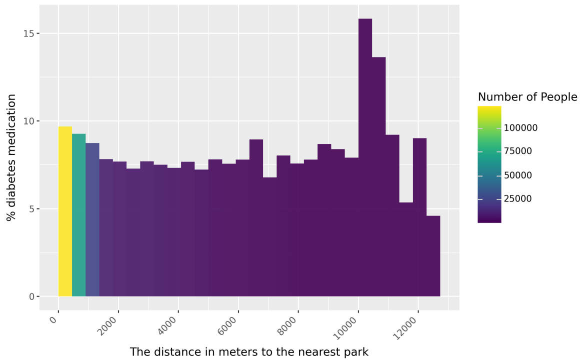


Figure A.9: Park

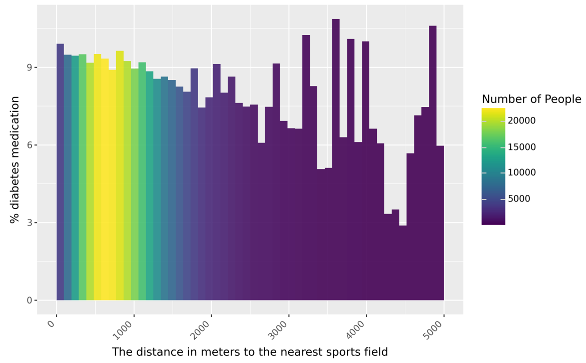


Figure A.10: Sports Field

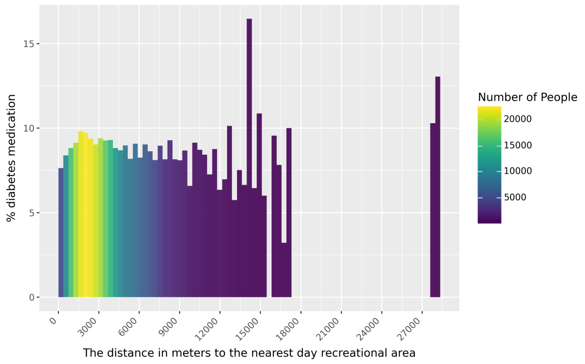


Figure A.11: Recreational Area

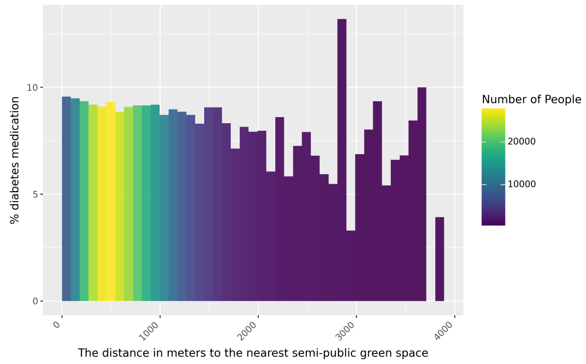


Figure A.12: Semi-public green space

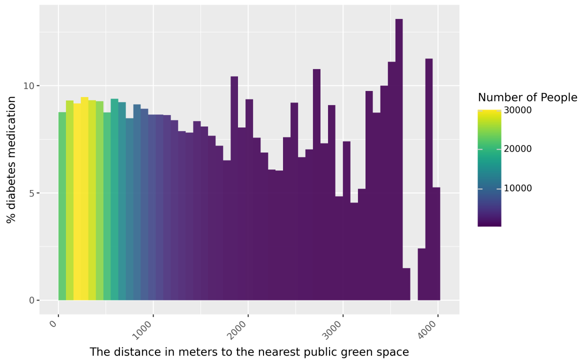


Figure A.13: Public Green Space

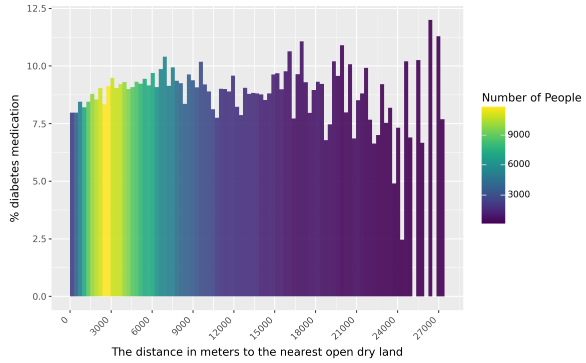


Figure A.14: Open Dry Land

Figure A.15: The distribution plots of the exercise environment variables. The x-axis shows the distance in meters to the nearest one (variable of the exercise environment). The y-axis shows the percentage of diabetes medication use. For privacy reasons and readability only bins that include at least 50 people are included.

B

Left Out Data

In total 94,695 individiuals had missings for one or more variables. Those individuals were left out (see section 2.13). Of those left out individuals, 12.44% (11,777 individuals) is using diabetes medication. The summary statistics of those individuals is shown in tables B.1, B.2, B.3, B.4 and B.5.

B.1. Demographic Summary Statistics Left Out Individuals

Category	**Overall (N)**		**Using Diabetes Medication**	
Average age	68.55 ± 12.31 years		72.96 ± 9.33 years	
Men	41.41%	(39216)	14.04%	(5505)
Women	58.59%	(55479)	11.31%	(6272)
Dutch	85.81%	(81256)	11.94%	(9704)
Other European	6.97%	(6601)	12.83%	(847)
Turkish	0.53%	(502)	19.72%	(99)
Moroccan	0.50%	(473)	23.26%	(110)
Surinamese	1.26%	(1193)	29.84%	(356)
Dutch Caribbean	0.41%	(391)	16.62%	(65)
Indonesian	2.67%	(2526)	14.45%	(365)
Other African	0.48%	(454)	16.30%	(74)
Other Asian	1.01%	(953)	13.12%	(125)
Other American & Oceanian	0.37%	(346)	9.25%	(32)

Table B.1: Demographic summary statistics of the left out individuals with a comparison between the overall left out individuals and the individuals within this left out group that are using diabetes medication. Either the average of the variable value or the percentage of people that comply with that variable is shown. Besides the percentages, also the absolute number of people (N) is shown.

B.2. Social Network Summary Statistics Left Out Individuals

Category	Overall (N)		Using Diabetes Medication	
Living at parents home	0.32%	(305)	10.82%	(33)
Living alone	29.82%	(28237)	14.91%	(4210)
Partner in unmarried couple without children living at home	3.66%	(3469)	10.15%	(352)
Partner in married couple without children living at home	48.08%	(45527)	13.08%	(5957)
Partner in unmarried couple with children living at home	1.78%	(1688)	3.73%	(63)
Partner in married couple with children living at home	12.34%	(11683)	6.25%	(730)
Parent in single-parent household	2.56%	(2427)	10.55%	(256)
Reference person in other household	0.17%	(163)	9.82%	(16)
Other household member	1.02%	(969)	13.62%	(132)
Member of institutional household	0.24%	(226)	12.39%	(28)
Not lonely	35.68%	(33791)	10.95%	(3699)
Slightly lonely	26.55%	(25141)	14.19%	(3567)
Lonely	4.72%	(4473)	16.23%	(726)
Very lonely	2.47%	(2342)	16.82%	(394)
Average loneliness **	0.63 ± 0.64		0.71 ± 0.68	
Indirect exposure to people * across all network layers	0.05 ± 0.06		0.07 ± 0.08	
Exposure to family members *	0.07 ± 0.12		0.10 ± 0.15	
Exposure to family members of the same gender *	0.05 ± 0.13		0.08 ± 0.17	
Exposure to family members of a different gender *	0.07 ± 0.15		0.11 ± 0.19	
Exposure to household members *	0.08 ± 0.26		0.12 ± 0.33	
Exposure to household members of the same gender *	0.00 ± 0.05		0.00 ± 0.06	
Exposure to household members of a different gender *	0.08 ± 0.26		0.12 ± 0.33	
Exposure to neighbors *	0.09 ± 0.07		0.11 ± 0.08	
Exposure to neighbors of the same gender *	0.09 ± 0.10		0.11 ± 0.10	
Exposure to neighbors of a different gender *	0.09 ± 0.10		0.11 ± 0.11	
Exposure to colleagues *	0.01 ± 0.03		0.01 ± 0.04	
Exposure to colleagues of the same gender *	0.01 ± 0.04		0.01 ± 0.05	
Exposure to colleagues of a different gender *	0.01 ± 0.04		0.00 ± 0.03	
Exposure to colleagues * (***)	0.03 ± 0.05		0.05 ± 0.08	
Exposure to colleagues * of the same gender (***)	0.03 ± 0.06		0.06 ± 0.13	
Exposure to colleagues * of a different gender (***)	0.03 ± 0.08		0.04 ± 0.08	
Exposure to people with master education	0.12 ± 0.12		0.10 ± 0.10	
Exposure to people with bachelor education	0.23 ± 0.13		0.21 ± 0.12	
Exposure to people with middle education	0.46 ± 0.16		0.47 ± 0.16	
Exposure to people with low education	0.19 ± 0.14		0.22 ± 0.15	

Table B.2: Social Network summary statistics of the left out individuals with a comparison between the overall left out individuals and the individuals within this left out group that are using diabetes medication. Either the average of the variable value or the percentage of people that comply with that variable is shown. Besides the percentages, also the absolute number of people (N) is shown. The '*' in the table stands for: using diabetes medication. **: The average loneliness ranges from 0 (not lonely) to 3 (very lonely). ***: Only the working population is included for those variables as only that group can have colleagues.

B.3. Lifestyle Summary Statistics Left Out Individuals

Category	Overall (N)		Using Diabetes Medication	
Minutes of light intensity exercise *	1135.13 ± 865.99		921.69 ± 772.26	
Minutes of middle intensity exercise *	682.63 ± 702.49		546.85 ± 620.34	
Minutes of high intensity exercise *	22.78 ± 82.09		14.09 ± 66.81	
Adherence to exercise guidelines	22.97%	(21754)	7.57%	(1647)
Never smoked	28.00%	(26517)	11.68%	(3098)
Ex-smoker	32.58%	(30849)	14.21%	(4383)
Smoker	10.71%	(10145)	11.68%	(1185)
Never drank alcohol	12.11%	(11463)	18.12%	(2077)
Alcohol drinker	65.14%	(61688)	10.40%	(6418)
Ex-alcohol drinker	6.15%	(5828)	21.00%	(1224)
Number of alcoholic drinks *	5.87 ± 7.37		4.72 ± 6.48	
Under weight (BMI: 18.5-)	0.95%	(901)	4.77%	(43)
Normal weight (BMI: 18.5-20)	2.08%	(1967)	3.51%	(69)
Normal weight (20-25)	30.27%	(28660)	7.12%	(2041)
Overweight (BMI: 25-30)	33.14%	(31382)	12.63%	(3965)
Obese (BMI: 30+)	14.17%	(13414)	23.84%	(3198)
Very good experienced health	9.67%	(9153)	2.29%	(210)
Good experienced health	49.22%	(46610)	8.83%	(4117)
Moderate experienced health	29.72%	(28148)	19.31%	(5434)
Bad experienced health	5.60%	(5301)	24.52%	(1300)
Very bad experienced health	0.85%	(805)	25.34%	(204)

Table B.3: Lifestyle summary statistics of the left out individuals with a comparison between the overall left out individuals and the individuals within this left out group that are using diabetes medication. Either the average of the variable value or the percentage of people that comply with that variable is shown. Besides the percentages, also the absolute number of people (N) is shown. The * stands for: 'per week'.

B.4. Socioeconomic Summary Statistics Left Out Individuals

Category	Overall (N)		Using Diabetes Medication	
Low education (primary education)	11.15%	(10560)	19.83%	(2094)
Middle 1 education (Dutch: MAVO, LBO)	33.37%	(31598)	13.41%	(4236)
Middle 2 education (Dutch: HAVO, VWO, MBO)	16.98%	(16078)	9.63%	(1548)
High education (HBO, WO)	11.62%	(10999)	7.82%	(860)
Unfit for work (Dutch: arbeidsongeschikt)	2.56%	(2424)	13.08%	(317)
Social benefits (Dutch: bijstand)	1.38%	(1308)	14.14%	(185)
No income	3.68%	(3487)	7.14%	(249)
Retired	68.70%	(65052)	15.26%	(9924)
Social benefits (Dutch: sociale voorzieningen)	0.47%	(445)	10.56%	(47)
Working	22.12%	(20945)	4.74%	(993)
Using unemployment benefits *	1.09%	(1034)	6.00%	(62)
Average household income percentile	49.42 ± 25.57		41.41 ± 22.87	

Table B.4: Socioeconomic summary statistics of the left out individuals with a comparison between the overall left out individuals and the individuals within this left out group that are using diabetes medication. Either the average of the variable value or the percentage of people that comply with that variable is shown. Besides the percentages, also the absolute number of people (N) is shown. *: Dutch: werkloosheidsuitkering.

B.5. Living Environment Summary Statistics Left Out Individuals

Category	Overall (N)		Using Diabetes Medication	
Very strong urbanity (≥ 2500 surrounding addresses/km²)	13.81%	(13079)	15.72%	(2056)
Strong urbanity (1500-2500)	22.45%	(21261)	13.30%	(2827)
Moderate urbanity (1000-1500)	18.61%	(17625)	11.99%	(2113)
Little urbanity (500-1000)	20.28%	(19204)	11.61%	(2229)
Not urban (<500)	24.84%	(23525)	10.85%	(2552)

Table B.5: Living Environment summary statistics of the left out individuals with a comparison between the overall left out individuals and the individuals within this left out group that are using diabetes medication. The percentage of people that comply with that variable is shown. Besides the percentages, also the absolute number of people (N) is shown.

C
Correlation between Variables

C.1. Heatmap

In figure C.1 the heatmap for the correlations between all variables is shown. For the meaning of the variable names, see table C.1.

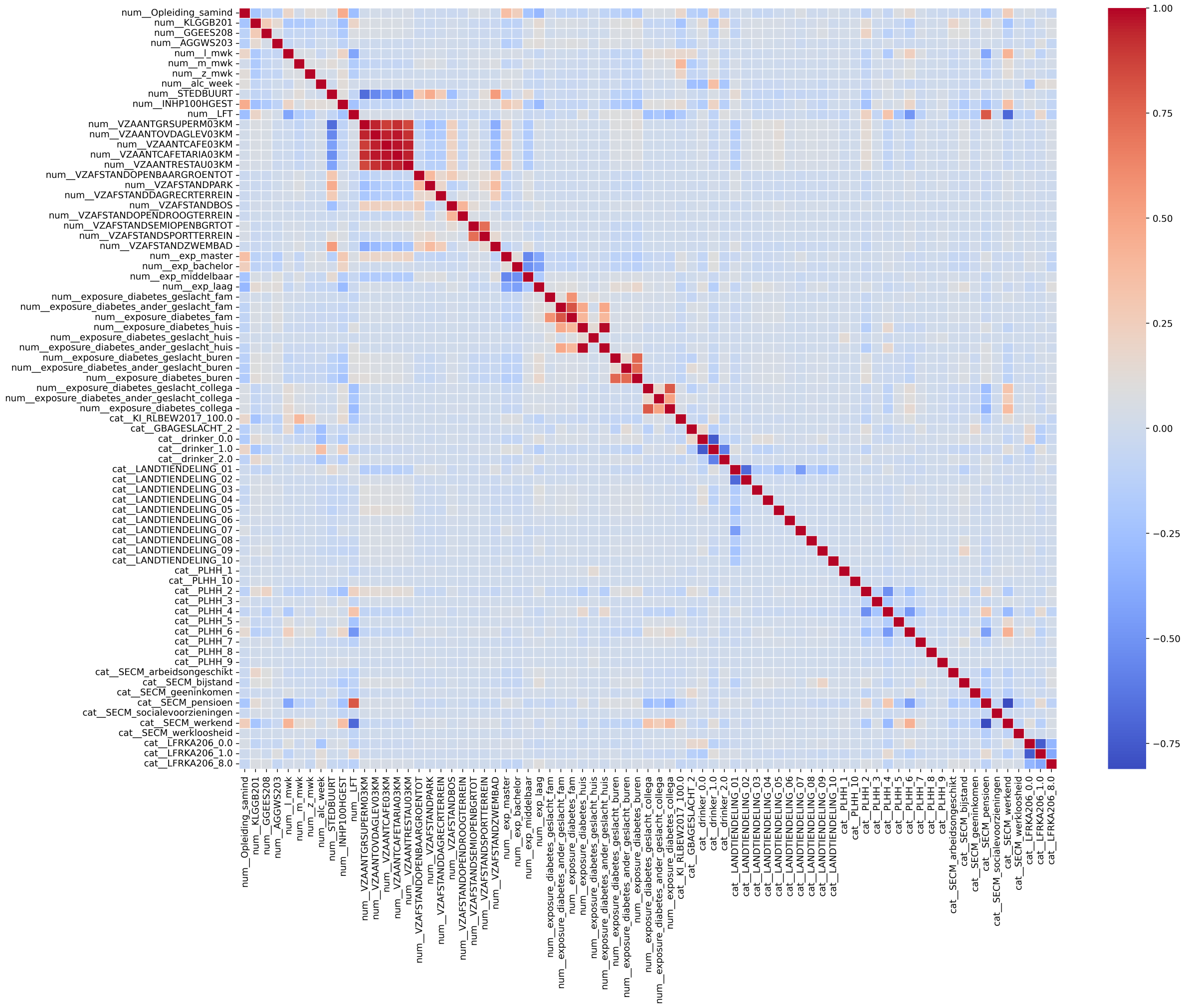


Figure C.1: The heatmap of all the variables showing if and how much variables are positively or negatively correlated.

Table C.1: Legend for variable names and meanings.
With * standing for: using diabetes medication

Variable Name	Variable meaning
num__Opleiding_samind	Education level
num__KLGGB201	Experienced health
num__GGEES208	Loneliness
num__AGGWS203	Body mass index
num__l_mwk	Minutes of light intensity exercise
num__m_mwk	Minutes of moderate intensity exercise
num__z_mwk	Minutes of high intensity exercise
num__alc_week	Number of alcoholic drinks per week
num__STEDBUURT	Urbanity
num__INHP100HGEST	Average household income percentile
num__LFT	Age
num__VZAANTGRSUPERM03KM	Number of supermarkets and grocery stores within 3 km
num__VZAANTOVDAGLEV03KM	Number of other daily food stores within 3 km
num__VZAANTCAFE03KM	Number of cafes within 3 km
num__VZAANTCAFETARIA03KM	Number of cafeterias within 3 km
num__VZAANTRESTAU03KM	Number of restaurants within 3 km
num__VZAFSTANDOPENBAARGROENTOT	Distance in meters to nearest public green space
num__VZAFSTANDPARK	Distance in meters to nearest park
num__VZAFSTANDDAGRECRTERREIN	Distance in meters to nearest day recreational area
num__VZAFSTANDBOS	Distance in meters to nearest forest
num__VZAFSTANDOPENDROOGTERREIN	Distance in meters to nearest open dry land
num__VZAFSTANDSEMIOPENBGRTOT	Distance in meters to nearest semi-public green space
num__VZAFSTANDSPORTTERREIN	Distance in meters to nearest sports field
num__VZAFSTANDZWEMBAD	Distance in meters to nearest swimming pool
num__exp_master	Exposure to people with master education
num__exp_bachelor	Exposure to people with bachelor education
num__exp_middelbaar	Exposure to people with middle education
num__exp_laag	Exposure to people with low education
num__exposure_diabetes_geslacht_fam	Exposure to family members of the same gender *
num__exposure_diabetes_ander_geslacht_fam	Exposure to family members of a different gender *
num__exposure_diabetes_fam	Exposure to family members *
num__exposure_diabetes_geslacht_huis	Exposure to household members of the same gender *
num__exposure_diabetes_ander_geslacht_huis	Exposure to household members of a different gender *
num__exposure_diabetes_huis	Exposure to household members *
num__exposure_diabetes_geslacht_buren	Exposure to neighbors of the same gender *
num__exposure_diabetes_ander_geslacht_buren	Exposure to neighbors of a different gender *

Variable Name	Variable meaning
num__exposure_diabetes_buren	Exposure to neighbors *
num__exposure_diabetes_geslacht_collega	Exposure to colleagues of the same gender *
num__exposure_diabetes_ander_geslacht_collega	Exposure to colleagues of a different gender *
num__exposure_diabetes_collega	Exposure to colleagues *
cat__KI_RLBEW2017_100.0	Adherence to exercise guidelines
cat__GBAGESLACHT_2	Women
cat__drinker_0.0	Never drank alcohol
cat__drinker_1.0	Alcohol drinker
cat__drinker_2.0	Ex-alcohol drinker
cat__LANDTIENDELING_01	Dutch
cat__LANDTIENDELING_02	Other European
cat__LANDTIENDELING_03	Turkish
cat__LANDTIENDELING_04	Moroccan
cat__LANDTIENDELING_05	Surinamese
cat__LANDTIENDELING_06	Dutch Caribbean
cat__LANDTIENDELING_07	Indonesian
cat__LANDTIENDELING_08	Other African
cat__LANDTIENDELING_09	Other Asian
cat__LANDTIENDELING_10	Other American and Oceanian
cat_PLHH_1	Living at parents home
cat_PLHH_2	Living alone
cat_PLHH_3	Partner in unmarried couple without children living at home
cat_PLHH_4	Partner in married couple without children living at home
cat_PLHH_5	Partner in unmarried couple with children living at home
cat_PLHH_6	Partner in married couple with children living at home
cat_PLHH_7	Parent in single-parent household
cat_PLHH_8	Reference person in other household
cat_PLHH_9	Other household member
cat_PLHH_10	Member of institutional household
cat__SECM_arbeidsongeschikt	Unfit for work (Dutch: arbeidsongeschikt)
cat__SECM_bijstand	Receiving social benefit (Dutch: bijstand)
cat__SECM_geeninkomen	No income
cat__SECM_pensioen	Retired
cat__SECM_socialevoorzieningen	Receiving social benefits (Dutch: sociale voorzieningen)
cat__SECM_werkend	Working
cat__SECM_werkloosheid	Receiving unemployment benefits (Dutch: werkloosheidsuitkering)
cat__LFRKA206_0.0	Never smoked
cat__LFRKA206_1.0	Smoker

Variable Name	Variable meaning
cat__LFRKA206_8.0	Ex-smoker

C.2. Correlations between Variables

The Pearson correlations between variables if the absolute correlation is equal or higher than 0.45. For the meaning of the variable names, see table C.1.

Variable 1	Variable 2	Correlation
num__exposure_diabetes_huis	num__exposure_diabetes_ander_geslacht_huis	0.99
num__VZAANTCAFE03KM	num__VZAANTCAFETARIA03KM	0.97
num__VZAANTOVDAGLEV03KM	num__VZAANTCAFETARIA03KM	0.97
num__VZAANTCAFETARIA03KM	num__VZAANTRESTAU03KM	0.96
num__VZAANTOVDAGLEV03KM	num__VZAANTCAFE03KM	0.95
num__VZAANTCAFE03KM	num__VZAANTRESTAU03KM	0.95
num__VZAANTGRSUPERM03KM	num__VZAANTOVDAGLEV03KM	0.95
num__VZAANTGRSUPERM03KM	num__VZAANTCAFETARIA03KM	0.93
num__VZAANTOVDAGLEV03KM	num__VZAANTRESTAU03KM	0.92
num__VZAANTGRSUPERM03KM	num__VZAANTCAFE03KM	0.89
cat__KI_RLBEW2017_0.0	cat__KI_RLBEW2017_100.0	-0.86
num__VZAANTGRSUPERM03KM	num__VZAANTRESTAU03KM	0.85
num__exposure_diabetes_ander_geslacht_fam	num__exposure_diabetes_fam	0.82
cat__SECM_pensioen	cat__SECM_werkend	-0.81
num__LFT	cat__SECM_pensioen	0.80
num__exposure_diabetes_geslacht_collega	num__exposure_diabetes_collega	0.79
num__exposure_diabetes_geslacht_buren	num__exposure_diabetes_buren	0.75
num__exposure_diabetes_ander_geslacht_buren	num__exposure_diabetes_buren	0.75
num__VZAFSTANDSEMIOPENBGRTOT	num__VZAFSTANDSPORTTERREIN	0.72
num__LFT	cat__SECM_werkend	-0.69
cat__LANDTIENDELING_01	cat__LANDTIENDELING_02	-0.69
num__STEDBUURT	num__VZAANTGRSUPERM03KM	-0.67
cat__drinker_0.0	cat__drinker_1.0	-0.67
cat__LFRKA206_0.0	cat__LFRKA206_1.0	-0.65
num__exposure_diabetes_geslacht_fam	num__exposure_diabetes_fam	0.56
num__STEDBUURT	num__VZAANTOVDAGLEV03KM	-0.55
cat__PLHH_2	cat__PLHH_4	-0.55
num__exp_master	num__exp_middelbaar	-0.54
num__STEDBUURT	num__VZAFSTANDZWEMBAD	0.54
num__STEDBUURT	num__VZAANTCAFETARIA03KM	-0.52
num__exposure_diabetes_ander_geslacht_collega	num__exposure_diabetes_collega	0.52
cat__drinker_nan	cat__LFRKA206_nan	0.51
num__exp_bachelor	num__exp_middelbaar	-0.51
cat__drinker_1.0	cat__drinker_2.0	-0.51
num__exposure_diabetes_ander_geslacht_fam	num__exposure_diabetes_ander_geslacht_huis	0.49
num__LFT	cat__PLHH_6	-0.49
num__totspier	cat__KI_RLBEW2017_100.0	0.48
num__exposure_diabetes_ander_geslacht_fam	num__exposure_diabetes_huis	0.48
num__STEDBUURT	num__VZAFSTANDPARK	0.46
num__Opleiding_samind	num__INHP100HGEST	0.45

Table C.2: Pearson Correlations that are higher than (absolute of) 0.45 between different variables.

D

Precision Recall Curves Test Set

D.1. Random Forest Model

```
average precision: 0.29077661013788064
```

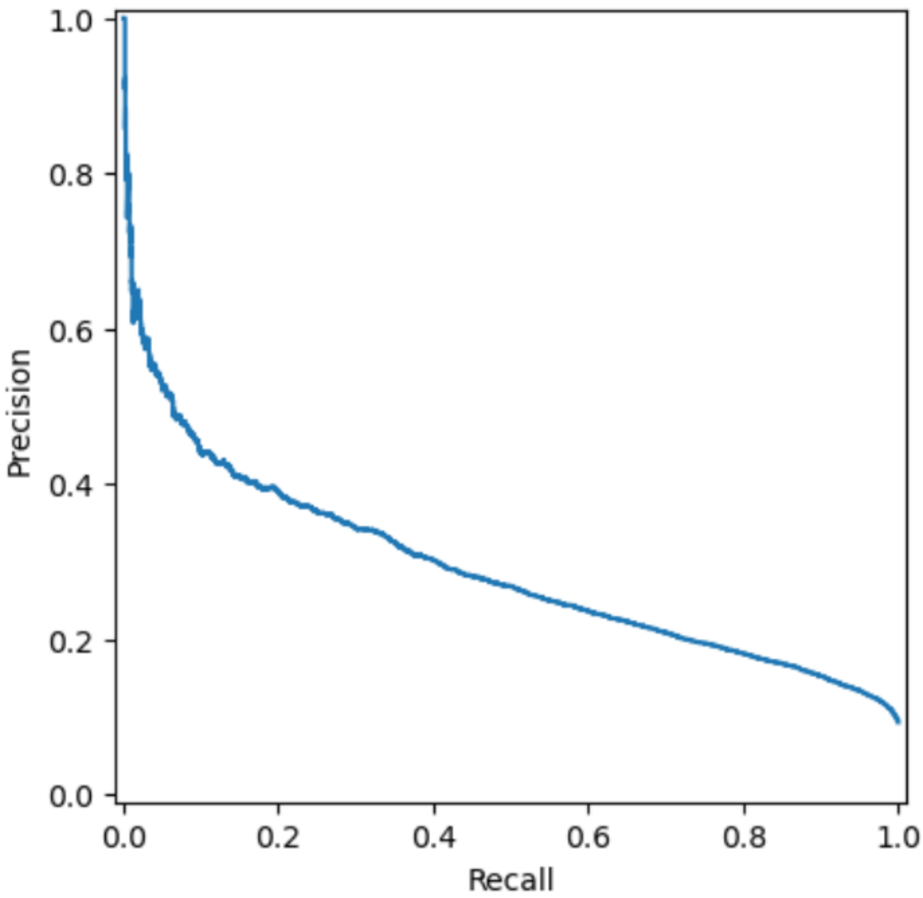


Figure D.1: The precision/recall curve for the test set of the random forest model.

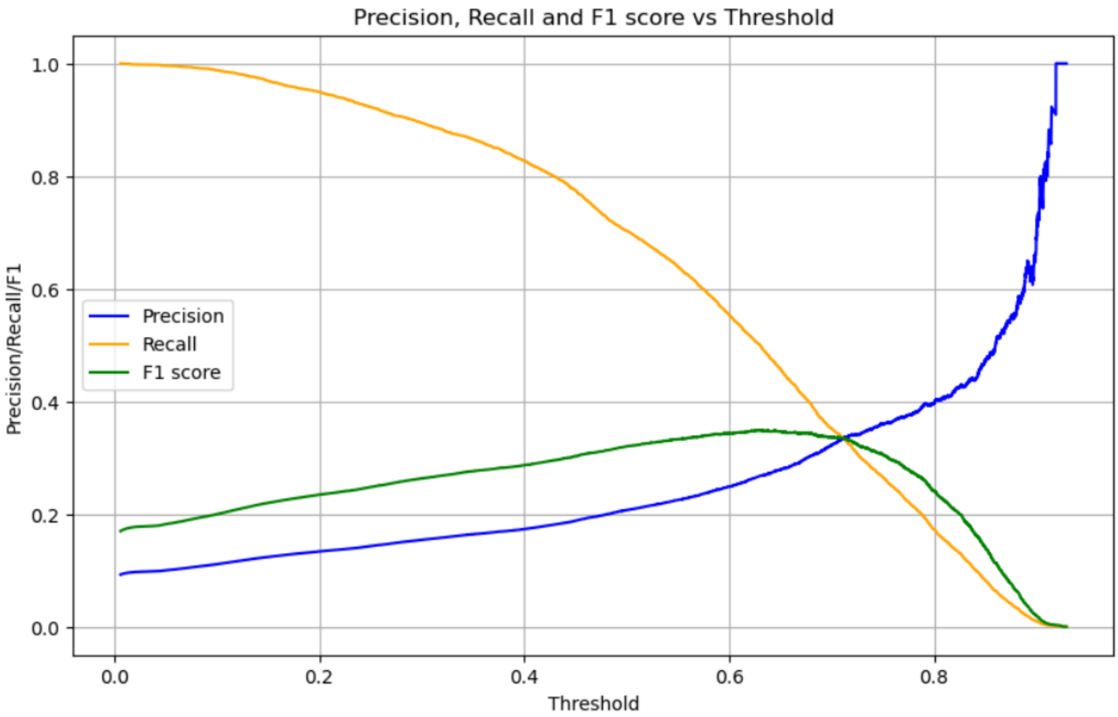


Figure D.2: The precision/recall and F1 threshold curve for the test set of the random forest model.

D.2. Logistic Regression Model

```
average precision: 0.2834576226969896
```

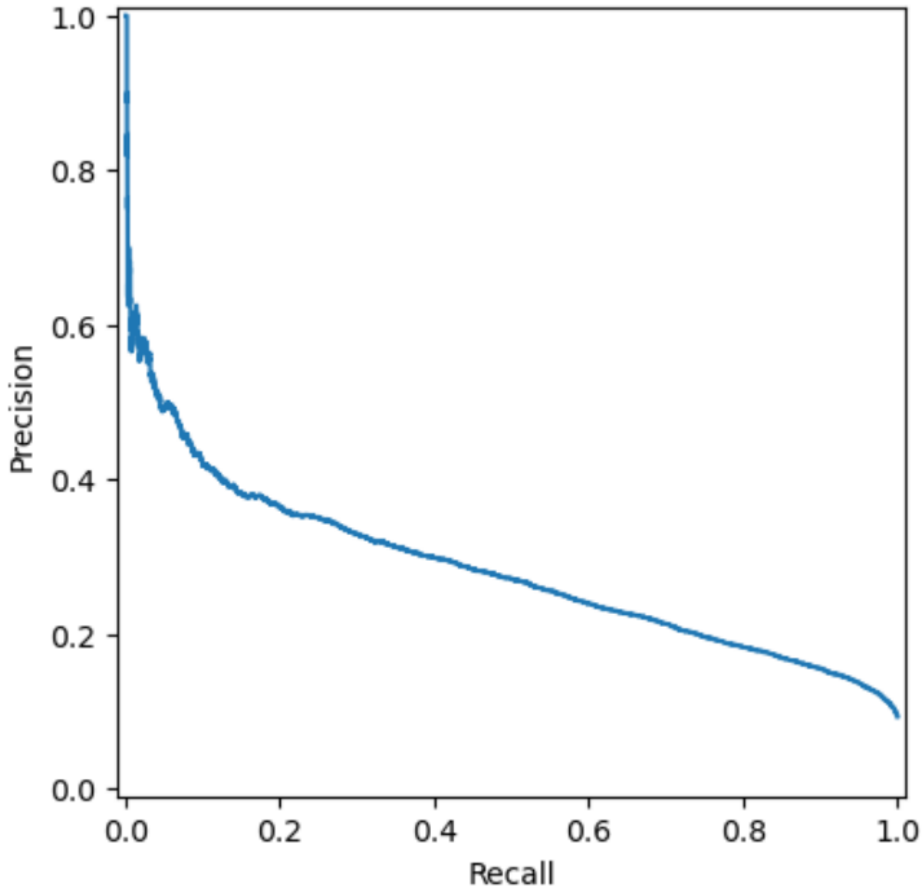


Figure D.3: The precision/recall curve for the test set of the logistic regression model.

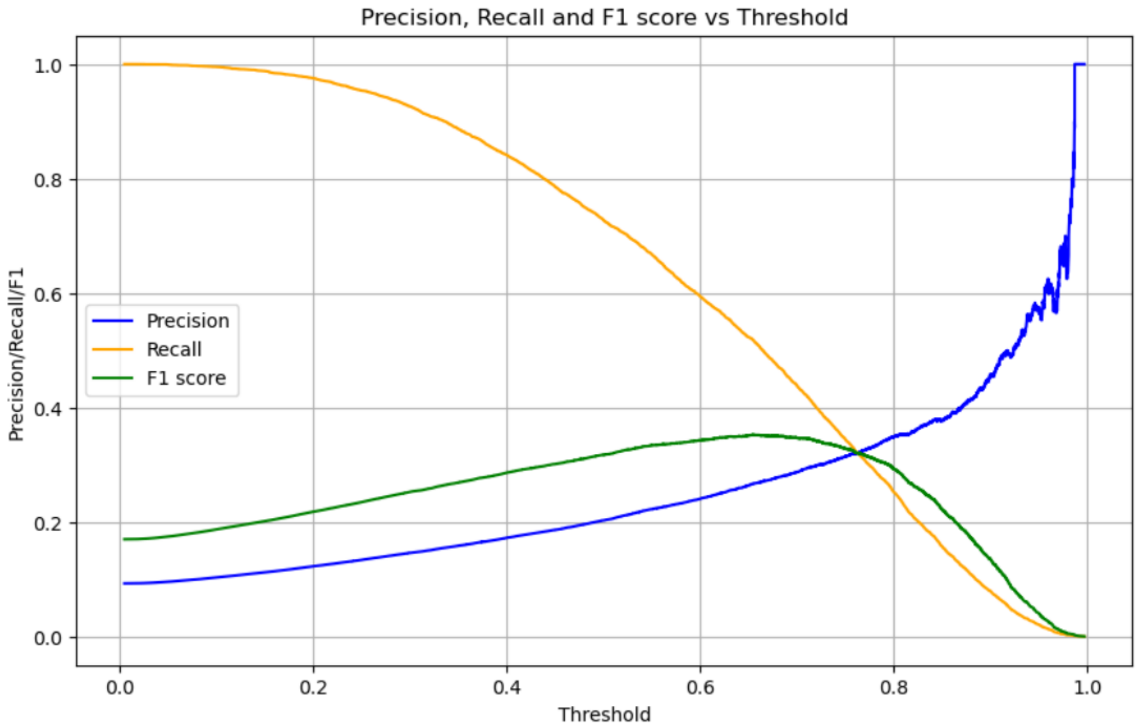
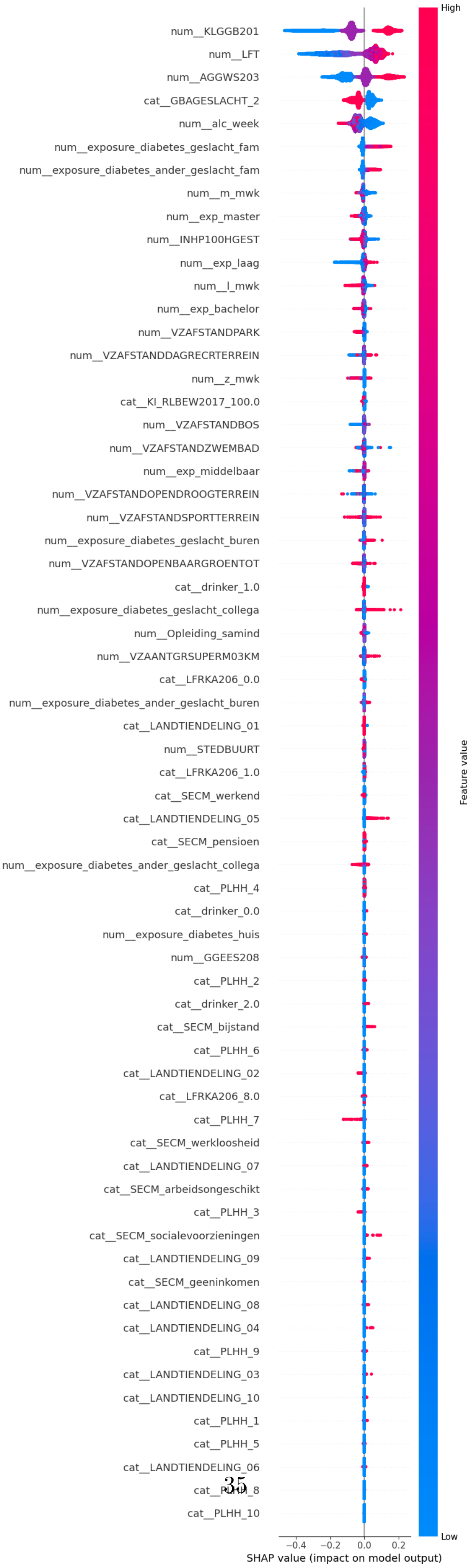


Figure D.4: The precision/recall and F1 threshold curve for the test set of the logistic regression model.

E

Shapley Graph Random Forest

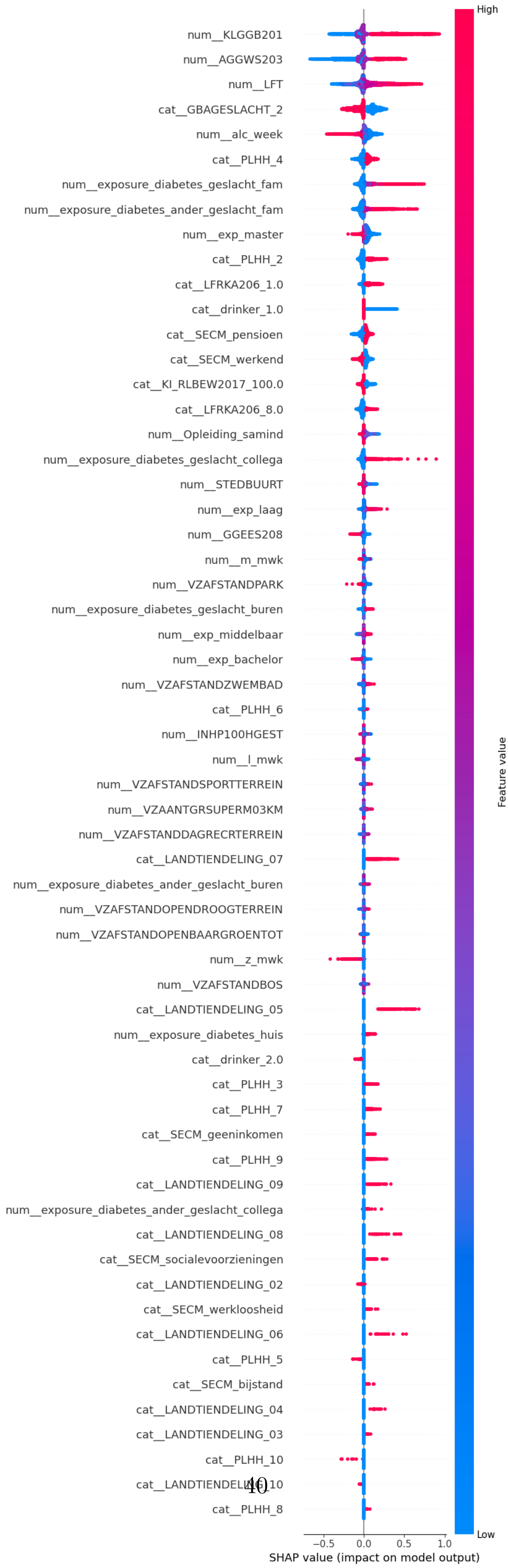
For the interpretation of this graph, see section 2.16.2. For the meaning of the variable names, see table C.1.


num__KLGGB201
num__LFT
num__AGGWS203
cat__GBAGESLACHT_2
num__alc_week
num__exposure_diabetes_geslacht_fam
num__exposure_diabetes_ander_geslacht_fam
num__m_mwk
num__exp_master
num__INHP100HGEST
num__exp_laag
num__l_mwk
num__exp_bachelor
num__VZAFSTANDPARK
num__VZAFSTANDDAGRECRTERREIN
num__z_mwk
cat__KI_RLBEW2017_100.0
num__VZAFSTANDBOS
num__VZAFSTANDZWEMBAD
num__exp_middelbaar
num__VZAFSTANDOPENDROOGTERREIN
num__VZAFSTANDSPORTTERREIN
num__exposure_diabetes_geslacht_buren
num__VZAFSTANDOPENBAARGROENTOT
cat__drinker_1.0
num__exposure_diabetes_geslacht_collega
num__Opleiding_samind
num__VZAANTGRSUPERM03KM
cat__LFRKA206_0.0
num__exposure_diabetes_ander_geslacht_buren
cat__LANDTIENDELING_01
num__STEDBUURT
cat__LFRKA206_1.0
cat__SECM_werkend
cat__LANDTIENDELING_05
cat__SECM_pensioen
num__exposure_diabetes_ander_geslacht_collega
cat__PLHH_4
cat__drinker_0.0
num__exposure_diabetes_huis
num__GGEES208
cat__PLHH_2
cat__drinker_2.0
cat__SECM_bijstand
cat__PLHH_6
cat__LANDTIENDELING_02
cat__LFRKA206_8.0
cat__PLHH_7
cat__SECM_werkloosheid
cat__LANDTIENDELING_07
cat__SECM_arbeidsongeschikt
cat__PLHH_3
cat__SECM_socialevoorzieningen
cat__LANDTIENDELING_09
cat__SECM_geeninkomen
cat__LANDTIENDELING_08
cat__LANDTIENDELING_04
cat__PLHH_9
cat__LANDTIENDELING_03
cat__LANDTIENDELING_10
cat__PLHH_1
cat__PLHH_5
cat__LANDTIENDELING_06
cat__PLHH_8
cat__PLHH_10
−0.4
−0.2
0.0
0.2
SHAP value (impact on model output)
High
Low
Feature value

F

Shapley Graph Logistic Regression

For the interpretation of this graph, see section 2.16.2. For the meaning of the variable names, see table C.1.



num__KLGGB201
num__AGGWS203
num__LFT
cat__GBAGESLACHT_2
num__alc_week
cat__PLHH_4
num__exposure_diabetes_geslacht_fam
num__exposure_diabetes_ander_geslacht_fam
num__exp_master
cat__PLHH_2
cat__LFRKA206_1.0
cat__drinker_1.0
cat__SECM_pensioen
cat__SECM_werkend
cat__KI_RLBEW2017_100.0
cat__LFRKA206_8.0
num__Opleiding_samind
num__exposure_diabetes_geslacht_collega
num__STEDBUURT
num__exp_laag
num__GGEES208
num__m_mwk
num__VZAFSTANDPARK
num__exposure_diabetes_geslacht_buren
num__exp_middelbaar
num__exp_bachelor
num__VZAFSTANDZWEMBAD
cat__PLHH_6
num__INHP100HGEST
num__l_mwk
num__VZAFSTANDSPORTTERREIN
num__VZAANTGRSUPERM03KM
num__VZAFSTANDDAGRECRTERREIN
cat__LANDTIENDELING_07
num__exposure_diabetes_ander_geslacht_buren
num__VZAFSTANDOPENDROOGTERREIN
num__VZAFSTANDOPENBAARGROENTOT
num__z_mwk
num__VZAFSTANDBOS
cat__LANDTIENDELING_05
num__exposure_diabetes_huis
cat__drinker_2.0
cat__PLHH_3
cat__PLHH_7
cat__SECM_geeninkomen
cat__PLHH_9
cat__LANDTIENDELING_09
num__exposure_diabetes_ander_geslacht_collega
cat__LANDTIENDELING_08
cat__SECM_socialevoorzieningen
cat__LANDTIENDELING_02
cat__SECM_werkloosheid
cat__LANDTIENDELING_06
cat__PLHH_5
cat__SECM_bijstand
cat__LANDTIENDELING_04
cat__LANDTIENDELING_03
cat__PLHH_10
cat__LANDTIENDELING_10
cat__PLHH_8
−0.5
0.0
0.5
1.0
SHAP value (impact on model output)
High
Low
Feature value

G
Code

G.1. Random Forest Code

250109_Code_RF

January 9, 2025

1 Code Random Forest

[ ]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import random
import time
import warnings
import joblib
import seaborn as sns
import shap
import xgboost as xgb
from joblib import Parallel, delayed
from sklearn.decomposition import PCA

from scipy.stats import pearsonr, pointbiserialr, chi2_contingency
from sklearn.linear_model import LinearRegression
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import OneHotEncoder, StandardScaler
from sklearn.impute import SimpleImputer
from sklearn.pipeline import Pipeline
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import (
    train_test_split, cross_val_score, StratifiedKFold, GridSearchCV
)
from sklearn.metrics import (
    accuracy_score, confusion_matrix, classification_report, recall_score,
    make_scorer, precision_score, f1_score, precision_recall_curve,
    PrecisionRecallDisplay, average_precision_score
)
from sklearn.inspection import PartialDependenceDisplay, permutation_importance
from sklearn.tree import plot_tree
from sklearn.preprocessing import FunctionTransformer

from tabulate import tabulate
from fancyimpute import KNN, IterativeImputer
from sklearn.linear_model import BayesianRidge
```

```
import multiprocessing
multiprocessing.cpu_count()

from sklearn.experimental import enable_iterative_imputer
from sklearn.impute import IterativeImputer
```

1.0.1 Document imports

[ ]:
```
df = pd.read_parquet('CoupledData.parquet')
```

1.0.2 Select only group who is 40+

[ ]:
```
df = df[df['LFT']>= 40]
print(f' The shape of df is: {df.shape}')
```

1.0.3 Fill missing values familie, huis, buren, colleague

[ ]:
```
columns_to_fill =['exposure_diabetes_fam','exposure_diabetes_geslacht_fam',␣
 ↪'exposure_diabetes_ander_geslacht_fam',
                  'exposure_diabetes_huis','exposure_diabetes_geslacht_huis',␣
 ↪'exposure_diabetes_ander_geslacht_huis',
                  'exposure_diabetes_buren','exposure_diabetes_geslacht_buren',␣
 ↪'exposure_diabetes_ander_geslacht_buren',
                  'exposure_diabetes_collega',␣
 ↪'exposure_diabetes_geslacht_collega',␣
 ↪'exposure_diabetes_ander_geslacht_collega']
df[columns_to_fill] = df[columns_to_fill].fillna(0)
```

1.0.4 De input for training are only rows that have no missings

[ ]:
```
df_nomis = df.dropna()
len(df_nomis)
print(f'The difference is: {len(df)-len(df_nomis)}')
print(len(df_nomis))
```

[ ]:
```
len(df)
```

1.0.5 Check missings per column

[ ]:
```
missing_per_column = df.isnull().sum()
print(missing_per_column)
```

1.0.6 Check exposure missings

[ ]:
```
df_check_exp_missings = df_nomis[['RINPERSOON']].merge(df_famexp,␣
↪on='RINPERSOON', how='left')\
   .merge(df_huisexp, on='RINPERSOON', how='left')\
   .merge(df_burenexp, on='RINPERSOON', how='left')\
   .merge(df_collegaexp, on='RINPERSOON', how='left')
```

[ ]:
```
df_check_exp_missings.isna().sum()
```

2 Train machine learning model

[ ]:
```
df = df_nomis
```

[ ]:
```
# Separate features and target
X = df.drop('DIABETESMED_2016', axis=1)
y = df['DIABETESMED_2016']
```

[ ]:
```
# Identify numerical and categorical columns
categorical_cols = ['KI_RLBEW2017', 'GBAGESLACHT', 'drinker',
                    'LANDTIENDELING','PLHH','SECM', 'LFRKA206']

numerical_cols = ['Opleiding_samind','KLGGB201', 'GGEES208', 'AGGWS203',␣
↪'l_mwk',
                  'm_mwk', 'z_mwk', 'alc_week', 'STEDBUURT','INHP100HGEST',
                  'LFT',
                 ␣
↪'VZAANTGRSUPERM03KM','VZAFSTANDOPENBAARGROENTOT','VZAFSTANDPARK',
                 ␣
↪'VZAFSTANDDAGRECRTERREIN','VZAFSTANDBOS','VZAFSTANDOPENDROOGTERREIN',
                  'VZAFSTANDSPORTTERREIN','VZAFSTANDZWEMBAD',
                  'exp_master','exp_bachelor','exp_middelbaar','exp_laag',
                  'exposure_diabetes_geslacht_fam',␣
↪'exposure_diabetes_ander_geslacht_fam',
                  'exposure_diabetes_huis',
                  'exposure_diabetes_geslacht_buren',␣
↪'exposure_diabetes_ander_geslacht_buren',
                  'exposure_diabetes_geslacht_collega',␣
↪'exposure_diabetes_ander_geslacht_collega']
```

[ ]:
```
numerical_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='mean')) # there are NO missing values␣
↪(are already deleted/filled)
])

categorical_transformer = Pipeline(steps=[
```

```
    ('onehot', OneHotEncoder(drop='if_binary', handle_unknown='ignore')) ␣
 ↪#there are no unkowns
])

# Combine preprocessing steps
preprocessor = ColumnTransformer(
    transformers=[
        ('num', numerical_transformer, numerical_cols),
        ('cat', categorical_transformer, categorical_cols)
    ])
```

2.0.1 Change the parameters for the GridSearch

[ ]:
```
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,␣
 ↪random_state=42)

# Define hyperparameter grid for RandomForest
param_grid = {
    'model__n_estimators': [50, 100, 150],
    'model__max_depth': [5, 10, 20, 30],
    'model__min_samples_split': [5, 10, 20],
    'model__max_features':['sqrt', 'log2', None], #[0.1, 0.4, 0.8, None]
    'model__random_state':[42]
}

# Create the pipeline
pipeline = Pipeline(steps=[
    ('preprocessor', preprocessor),
    ('model', RandomForestClassifier(random_state=42, class_weight='balanced'))
])
```

[ ]:
```
# Initialize GridSearchCV
grid_search = GridSearchCV(pipeline, param_grid, cv=5,␣
 ↪scoring='average_precision', n_jobs=-10)
```

[ ]:
```
# Train the model with GridSearchCV
start_time = time.time()
grid_search.fit(X_train, y_train)
end_time = time.time()

training_time = (end_time - start_time) / 60
print(f'Training time: {training_time} minutes')
```

[ ]:
```
# check results of the grid search
gridsearch_results = pd.DataFrame(grid_search.cv_results_)
```

```
gridsearch_results = gridsearch_results.sort_values(by=['rank_test_score'])

gridsearch_results
```

Save the gridsearch data

```
[ ]: # save the data
gridsearch_results.to_parquet('241228_gridsearchRF_results.parquet',␣
 ↪index=False)
# read the data
gridsearch_results = pd.read_parquet('241228_gridsearchRF_results.parquet')
```

Save the best model

```
[ ]: best_model = grid_search.best_estimator_

# save the best model to a file
joblib.dump(best_model, '241228_best_model_RF_NEW.pkl')
best_model = joblib.load('241228_best_model_RF_NEW.pkl')
```

2.1 Calculate predicted probabilities

```
[ ]: # Predict probabilities on the TRAIN set with the best model
y_pred_proba_train = best_model.predict_proba(X_train)[:, 1]
```

```
[ ]: # Predict probabilities on the TEST set with the best model
y_pred_proba = best_model.predict_proba(X_test)[:, 1]
```

3 Precision Recall Curve

3.0.1 Train set

```
[ ]: # Train Set Raw Numbers
precision, recall, thresholds = precision_recall_curve(y_train,␣
 ↪y_pred_proba_train)
disp = PrecisionRecallDisplay(precision=precision, recall=recall)
disp.plot()
average_precision = average_precision_score(y_train, y_pred_proba_train)
print(f'average precision: {average_precision}')

# calculate F1 scores
f1_scores = 2*(precision*recall)/(precision+recall)
f1_scores = np.nan_to_num(f1_scores) # handle any NAN from division by zero
# plot precisie en recall als functie van de thresholds
plt.figure(figsize=(10,6))
plt.plot(thresholds, precision[:-1], label='Precision', color='blue')
plt.plot(thresholds, recall[:-1], label='Recall', color='orange')
```

```
plt.plot(thresholds, f1_scores[:-1], label='F1 score', color='green')
plt.xlabel('Threshold')
plt.ylabel('Precision/Recall/F1')
plt.title('Precision, Recall and F1 score vs Threshold')
plt.legend()
plt.grid()
plt.show()
```

3.0.2 Test set

[ ]:
```
# Test Set Raw Numbers
precision, recall, thresholds = precision_recall_curve(y_test, y_pred_proba)
disp = PrecisionRecallDisplay(precision=precision, recall=recall)
disp.plot()
average_precision = average_precision_score(y_test, y_pred_proba)
print(f'average precision: {average_precision}')

# calculate F1 scores
f1_scores = 2*(precision*recall)/(precision+recall)
f1_scores = np.nan_to_num(f1_scores) # handle any NAN from division by zero
# plot precisie en recall als functie van de thresholds
plt.figure(figsize=(10,6))
plt.plot(thresholds, precision[:-1], label='Precision', color='blue')
plt.plot(thresholds, recall[:-1], label='Recall', color='orange')
plt.plot(thresholds, f1_scores[:-1], label='F1 score', color='green')
plt.xlabel('Threshold')
plt.ylabel('Precision/Recall/F1')
plt.title('Precision, Recall and F1 score vs Threshold')
plt.legend()
plt.grid()
plt.show()
```

3.0.3 Set decision threshold

[ ]:
```
optimal_threshold = 0.65
```

3.0.4 Confusion matrix Train set

[ ]:
```
# Apply the optimal threshold to the TRAIN set
# when the predicted probability is bigger than the optimal_threshold, a 1 will␣
↪be put (so has diabetes), otherwise a 0 will be put (no diabetes)
optimal_threshold = optimal_threshold
y_pred_adjusted_train = (y_pred_proba_train >= optimal_threshold).astype(int)
```

[ ]:
```
# Classification report TRAIN set
class_report = classification_report(y_train, y_pred_adjusted_train)
```

```
print('Classification Report TRAIN set:')
print(class_report)
```

3.0.5 Confusion matrix Test set

[ ]:
```
# Apply the optimal threshold to the test set
# when the predicted probability is bigger than the optimal_threshold, a 1 will␣
↪be put (so has diabetes), otherwise a 0 will be put (no diabetes)
optimal_threshold = optimal_threshold
y_pred_adjusted = (y_pred_proba >= optimal_threshold).astype(int)
```

[ ]:
```
# Classification report
class_report = classification_report(y_test, y_pred_adjusted)
print('Classification Report:')
print(class_report)
```

3.0.6 Analyse TP, FP, TN, FN (of the TEST set) in further detail

[ ]:
```
results_df = pd.DataFrame({
    'Actual': y_test, #actual outcomes
    'Predicted': y_pred_adjusted #predicted outcomes
})

results_df = pd.concat([results_df, X_test], axis=1)

false_positives = results_df[(results_df['Predicted']==1) &␣
↪(results_df['Actual']==0)]
false_negatives = results_df[(results_df['Predicted']==0) &␣
↪(results_df['Actual']==1)]
true_positives =  results_df[(results_df['Predicted']==1) &␣
↪(results_df['Actual']==1)]
true_negatives =  results_df[(results_df['Predicted']==0) &␣
↪(results_df['Actual']==0)]
```

[ ]:
```
# check percentage of people who are in the true negative group and do use␣
↪diabetesmedicatie in 2022
print(true_negatives['DIABETESMED_2022'].value_counts())
TN1_2022 = (true_negatives['DIABETESMED_2022'].value_counts()[1]/
↪true_negatives['DIABETESMED_2022'].value_counts()[0])*100
print(f'percentage: {TN1_2022}')
```

[ ]:
```
# check percentage of people who are in the false positive group and do use␣
↪diabetesmedicatie in 2022
print(false_positives['DIABETESMED_2022'].value_counts())
#vergelijk dat met de True negatives hoeveel die omhoog gaan
```

```
FP1_2022 = (false_positives['DIABETESMED_2022'].value_counts()[1]/
↪false_positives['DIABETESMED_2022'].value_counts()[0])*100
print(f'percentage: {FP1_2022}')
```

[ ]:
```
# factor
FP1_2022/TN1_2022
```

4 Shapley

[ ]:
```
# Transform the data sing the preprocessor
sample_size = 100000 #00
X_sample = X_train.sample(sample_size, random_state=42)
X_transformed = preprocessor.fit_transform(X_sample)

# convert transformed data to dataframe
X_transformed_df = pd.DataFrame(X_transformed, columns = preprocessor.
↪get_feature_names_out())
```

[ ]:
```
X_transformed.shape
```

[ ]:
```
explainer = shap.TreeExplainer(best_model.named_steps['model'], feature_names =␣
↪preprocessor.get_feature_names_out()) #, X_transformed)
```

[ ]:
```
# For all plots except the dependence plot
start_time = time.time()
shap_values = explainer(X_transformed)
end_time = time.time()
shapley_time = (end_time - start_time) / 60
print(f'Shapley time: {shapley_time} minutes')
```

[ ]:
```
joblib.dump(shap_values, '241228_shap_values_RF_NEW.pkl')
shap_values = joblib.load('241228_shap_values_RF_NEW.pkl')
```

[ ]:
```
# For the dependence plot
start_time = time.time()
shap_values_2 = explainer.shap_values(X_transformed)
end_time = time.time()
shapley_time = (end_time - start_time) / 60
print(f'Shapley 2 time: {shapley_time} minutes')
```

[ ]:
```
joblib.dump(shap_values_2, '241228_shap_values_2_RF_NEW.pkl')
shap_values_2 = joblib.load('241228_shap_values_2_RF_NEW.pkl')
```

4.0.1 Global interpretability

4.0.2 Get the Ranking

[ ]:
```
mean_abs_shap_values = np.mean(np.abs(shap_values[:,:,1].values), axis=0)
sorted_features = np.array(shap_values[:,:,1].feature_names)[np.
↪argsort(-mean_abs_shap_values)]
print(sorted_features)
```

[ ]:
```
np.sort(-mean_abs_shap_values)
# t/m variable 29 gaat het tot de macht -3, daarna tot de macht -4
```

[ ]:
```
shap.summary_plot(shap_values[:,:,1], max_display=150)#, X_transformed)
```

4.0.3 Afstand dag recreatie terrein

[ ]:
```
dot_size = 10
jitter = 0.3
alpha = 0.2
interaction_index = 'auto'
```

[ ]:
```
feature = 'num__VZAFSTANDDAGRECRTERREIN'
shap.dependence_plot(feature, shap_values_2[:,:,1], X_transformed_df,␣
↪dot_size=dot_size, x_jitter=jitter, alpha=alpha, interaction_index=None,␣
↪show=False)
# Set custom x and y axis labels
plt.xlabel('distance to nearest recreation area in meters')
plt.ylabel('SHAP value')
#plt.title('Custom Title')
plt.savefig('dependence_plot_RF_VZAFSTANDDAGRECRTERREIN.png', dpi=300,␣
↪bbox_inches='tight')

plt.show()
```

G.2. Logistic Regression Code

250109_Code_LR

January 9, 2025

1 Code Logistic Regression

[ ]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import random
import time
import warnings
import joblib
import seaborn as sns
import shap
import xgboost as xgb
from joblib import Parallel, delayed
from sklearn.decomposition import PCA

from scipy.stats import pearsonr, pointbiserialr, chi2_contingency
from sklearn.linear_model import LinearRegression
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import OneHotEncoder, StandardScaler
from sklearn.impute import SimpleImputer
from sklearn.pipeline import Pipeline
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import (
    train_test_split, cross_val_score, StratifiedKFold, GridSearchCV
)
from sklearn.metrics import (
    accuracy_score, confusion_matrix, classification_report, recall_score,
    make_scorer, precision_score, f1_score, precision_recall_curve,
    PrecisionRecallDisplay, average_precision_score
)
from sklearn.inspection import PartialDependenceDisplay, permutation_importance
from sklearn.tree import plot_tree

from tabulate import tabulate
from fancyimpute import KNN, IterativeImputer
from sklearn.linear_model import BayesianRidge
import multiprocessing
```

```
multiprocessing.cpu_count()

from sklearn.linear_model import LogisticRegression
```

1.0.1 Document imports

[ ]:
```
df = pd.read_parquet('CoupledData.parquet')
```

1.0.2 Select only group who is 40+

[ ]:
```
df = df[df['LFT']>= 40]
print(f' The shape of df is: {df.shape}')
```

1.0.3 Fill missing values familie, huis, buren, (collega, maar al eerder gedaan) netwerken

[ ]:
```
columns_to_fill =['exposure_diabetes_fam','exposure_diabetes_geslacht_fam',␣
 ↪'exposure_diabetes_ander_geslacht_fam',
                  'exposure_diabetes_huis','exposure_diabetes_geslacht_huis',␣
 ↪'exposure_diabetes_ander_geslacht_huis',
                  'exposure_diabetes_buren','exposure_diabetes_geslacht_buren',␣
 ↪'exposure_diabetes_ander_geslacht_buren',
                  'exposure_diabetes_collega',␣
 ↪'exposure_diabetes_geslacht_collega',␣
 ↪'exposure_diabetes_ander_geslacht_collega']
df[columns_to_fill] = df[columns_to_fill].fillna(0)
```

1.0.4 De input for training are only rows that have no missings

[ ]:
```
df_nomis = df.dropna()
len(df_nomis)
print(f'The difference is: {len(df)-len(df_nomis)}')
```

2 Train machine learning model

[ ]:
```
# Separate features and target
df = df_nomis.drop('RINPERSOON', axis=1)
X = df.drop('DIABETESMED_2016', axis=1)
y = df['DIABETESMED_2016']
```

[ ]:
```
# Identify numerical and categorical columns
categorical_cols = ['KI_RLBEW2017', 'GBAGESLACHT', 'drinker',
                    'LANDTIENDELING','PLHH','SECM', 'LFRKA206']
```

```
numerical_cols = ['Opleiding_samind','KLGGB201', 'GGEES208', 'AGGWS203', 'l_mwk',
                  'm_mwk', 'z_mwk', 'alc_week', 'STEDBUURT','INHP100HGEST',
                  'LFT',
                  'VZAANTGRSUPERM03KM', 'VZAFSTANDOPENBAARGROENTOT','VZAFSTANDPARK',
                  'VZAFSTANDDAGRECRTERREIN','VZAFSTANDBOS','VZAFSTANDOPENDROOGTERREIN',
                  'VZAFSTANDSPORTTERREIN','VZAFSTANDZWEMBAD',
                  'exp_master','exp_bachelor','exp_middelbaar','exp_laag',
                  'exposure_diabetes_geslacht_fam', 'exposure_diabetes_ander_geslacht_fam',
                  'exposure_diabetes_huis',
                  'exposure_diabetes_geslacht_buren', 'exposure_diabetes_ander_geslacht_buren',
                  'exposure_diabetes_geslacht_collega', 'exposure_diabetes_ander_geslacht_collega']
```

[ ]:
```
# Preprocessing for numerical data
numerical_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='mean')), # there are NO missing values (are already deleted/filled)
    ('scaler', StandardScaler())
])

# Preprocessing for categorical data (one hot encoding)
categorical_transformer = Pipeline(steps=[
    ('onehot', OneHotEncoder(drop='first', handle_unknown='ignore')) #there are no unkowns
])

# Combine preprocessing steps
preprocessor = ColumnTransformer(
    transformers=[
        ('num', numerical_transformer, numerical_cols),
        ('cat', categorical_transformer, categorical_cols)
    ])
```

[ ]:
```
X_prepp = preprocessor.fit_transform(X)
```

[ ]:
```
X_prepp.shape
```

3 Logistic Regression

[ ]:
```
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,␣
↪random_state=42)

param_grid = {
    'model__C':[0.01, 0.1, 1, 10, 100],
    'model__solver':['liblinear', 'lbfgs', 'saga', 'newton-cholesky'],
    'model__penalty':['l2', 'l1'],
    'model__random_state':[42]
}

# Create the pipeline
pipeline = Pipeline(steps=[
    ('preprocessor', preprocessor),
    ('model', LogisticRegression(class_weight='balanced', max_iter=200,␣
↪random_state=42))
])
```

[ ]:
```
# Initialize GridSearchCV
grid_search = GridSearchCV(pipeline, param_grid, cv=5,␣
↪scoring='average_precision', n_jobs=-10, verbose = 1,)

# Train the model with GridSearchCV
start_time = time.time()
grid_search.fit(X_train, y_train)
end_time = time.time()

training_time = (end_time - start_time) / 60
print(f'Training time: {training_time} minutes')
```

[ ]:
```
# check results of the grid search
gridsearch_results = pd.DataFrame(grid_search.cv_results_)
gridsearch_results = gridsearch_results.sort_values(by=['rank_test_score'])
gridsearch_results
```

Save the gridsearch data

[ ]:
```
# save the data
gridsearch_results = pd.read_parquet('gridsearch_LR_250102.parquet')
gridsearch_results
```

Save the best model

[ ]:
```
best_model = grid_search.best_estimator_
```

```
# save the best model to a file
joblib.dump(best_model, 'best_model_LR_250102.pkl')
best_model = joblib.load('best_model_LR_250102.pkl')
```

Use the best model

3.1 Calculate predicted probabilities

```
[ ]: # Predict probabilities on the TRAIN set with the best model
y_pred_proba_train = best_model.predict_proba(X_train)[:, 1]
```

```
[ ]: # Predict probabilities on the test set with the best model
y_pred_proba = best_model.predict_proba(X_test)[:, 1]
```

4 Precision Recall Curve

4.0.1 Train set

```
[ ]: # Test Set Raw Numbers LR
precision, recall, thresholds = precision_recall_curve(y_train,␣
 ↪y_pred_proba_train)
disp = PrecisionRecallDisplay(precision=precision, recall=recall)
disp.plot()
average_precision = average_precision_score(y_train, y_pred_proba_train)
print(f'average precision: {average_precision}')

# calculate F1 scores
f1_scores = 2*(precision*recall)/(precision+recall)
f1_scores = np.nan_to_num(f1_scores) # handle any NAN from division by zero
# plot precisie en recall als functie van de thresholds
plt.figure(figsize=(10,6))
plt.plot(thresholds, precision[:-1], label='Precision', color='blue')
plt.plot(thresholds, recall[:-1], label='Recall', color='orange')
plt.plot(thresholds, f1_scores[:-1], label='F1 score', color='green')
plt.xlabel('Threshold')
plt.ylabel('Precision/Recall/F1')
plt.title('Precision, Recall and F1 score vs Threshold')
plt.legend()
plt.grid()
plt.show()
```

4.0.2 Test set

```
[ ]: # Test Set Raw Numbers LR
precision, recall, thresholds = precision_recall_curve(y_test, y_pred_proba)
disp = PrecisionRecallDisplay(precision=precision, recall=recall)
disp.plot()
average_precision = average_precision_score(y_test, y_pred_proba)
print(f'average precision: {average_precision}')

# calculate F1 scores
f1_scores = 2*(precision*recall)/(precision+recall)
f1_scores = np.nan_to_num(f1_scores) # handle any NAN from division by zero
# plot precisie en recall als functie van de thresholds
plt.figure(figsize=(10,6))
plt.plot(thresholds, precision[:-1], label='Precision', color='blue')
plt.plot(thresholds, recall[:-1], label='Recall', color='orange')
plt.plot(thresholds, f1_scores[:-1], label='F1 score', color='green')
plt.xlabel('Threshold')
plt.ylabel('Precision/Recall/F1')
plt.title('Precision, Recall and F1 score vs Threshold')
plt.legend()
plt.grid()
plt.show()
```

4.0.3 Set decision threshold

```
[ ]: optimal_threshold = 0.65
```

4.0.4 Confusion matrix Train set

```
[ ]: # Apply the optimal threshold to the TRAIN set
# when the predicted probability is bigger than the optimal_threshold, a 1 will␣
↪be put (so has diabetes), otherwise a 0 will be put (no diabetes)
optimal_threshold = optimal_threshold
y_pred_adjusted_train = (y_pred_proba_train >= optimal_threshold).astype(int)
```

```
[ ]: # Classification report TRAIN set
class_report = classification_report(y_train, y_pred_adjusted_train)
print('Classification Report TRAIN set:')
print(class_report)
```

4.0.5 Confusion matrix Test Set

```
[ ]: # Apply the optimal threshold to the test set
# when the predicted probability is bigger than the optimal_threshold, a 1 will␣
↪be put (so has diabetes), otherwise a 0 will be put (no diabetes)
optimal_threshold = optimal_threshold
```

```
y_pred_adjusted = (y_pred_proba >= optimal_threshold).astype(int)
```

[ ]:
```
# Classification report
class_report = classification_report(y_test, y_pred_adjusted)
print('Classification Report:')
print(class_report)
```

4.0.6 Analyse TP, FP, TN, FN (of the TEST set) in further detail

[ ]:
```
results_df = pd.DataFrame({
    'Actual': y_test, #actual outcomes
    'Predicted': y_pred_adjusted #predicted outcomes
})

results_df = pd.concat([results_df, X_test], axis=1)

false_positives = results_df[(results_df['Predicted']==1) &␣
 ↪(results_df['Actual']==0)]
false_negatives = results_df[(results_df['Predicted']==0) &␣
 ↪(results_df['Actual']==1)]
true_positives =  results_df[(results_df['Predicted']==1) &␣
 ↪(results_df['Actual']==1)]
true_negatives =  results_df[(results_df['Predicted']==0) &␣
 ↪(results_df['Actual']==0)]
```

[ ]:
```
# check percentage of people who are in the true negative group and do use␣
 ↪diabetesmedicatie in 2022
print(true_negatives['DIABETESMED_2022'].value_counts())
MedUse2022_TN = (true_negatives['DIABETESMED_2022'].value_counts()[1]/
 ↪true_negatives['DIABETESMED_2022'].value_counts()[0])*100
print(MedUse2022_TN)
```

[ ]:
```
# check percentage of people who are in the false positive group and do use␣
 ↪diabetesmedicatie in 2022
print(false_positives['DIABETESMED_2022'].value_counts())
MedUse2022_FP = (false_positives['DIABETESMED_2022'].value_counts()[1]/
 ↪false_positives['DIABETESMED_2022'].value_counts()[0])*100
print(MedUse2022_FP)
```

[ ]:
```
# factor
MedUse2022_FP/MedUse2022_TN
```

4.0.7 Coefficients

[ ]:
```
coefficients = best_model.named_steps['model'].coef_
feature_names = preprocessor.get_feature_names_out()
coef_df = pd.DataFrame(coefficients, columns=feature_names)
coef_df.T.sort_values(by=0, ascending=False).head(60)
```

4.0.8 Shapley

[ ]:
```
# Transform the data sing the preprocessor
sample_size = 10000
X_sample = X_test.sample(sample_size, random_state=42)
X_transformed = best_model.named_steps['preprocessor'].transform(X_sample)
##X_transformed_train = best_model.named_steps['preprocessor'].
↪transform(X_train)

# convert transformed data to dataframe
X_transformed_df = pd.DataFrame(X_transformed, columns = preprocessor.
↪get_feature_names_out())
```

[ ]:
```
# define the SHAP explainer
explainer = shap.KernelExplainer(best_model.named_steps['model'].predict, shap.
↪sample(X_transformed_df, 5), feature_names = preprocessor.
↪get_feature_names_out())
```

[ ]:
```
# For all plots except the dependence plot
start_time = time.time()
shap_values = explainer(X_transformed)
end_time = time.time()
shapley_time = (end_time - start_time) / 60
print(f'Shapley time: {shapley_time} minutes')
```

[ ]:
```
joblib.dump(shap_values, 'shap_values_LR_250102.pkl')
shap_values = joblib.load('shap_values_LR_250102.pkl')
```

[ ]:
```
# For the dependence plot
start_time = time.time()
shap_values_2 = explainer.shap_values(X_transformed)
end_time = time.time()
shapley_time = (end_time - start_time) / 60
print(f'Shapley 2 time: {shapley_time} minutes')
```

[ ]:
```
joblib.dump(shap_values_2, 'shap_values_2_LR_250102.pkl')
shap_values_2 = joblib.load('shap_values_2_LR_250102.pkl')
```

4.0.9 Global interpretability

[ ]:
```
shap_values.shape
```

4.0.10 Get the Ranking

[ ]:
```
mean_abs_shap_values = np.mean(np.abs(shap_values.values), axis=0)
sorted_features = np.array(shap_values.feature_names)[np.
 ↪argsort(-mean_abs_shap_values)]
print(sorted_features)
```

[ ]:
```
np.sort(-mean_abs_shap_values)
```

[ ]:
```
shap.summary_plot(shap_values, max_display=150)#, X_transformed)
```

4.0.11 Afstand dag recreatie terrein

[ ]:
```
dot_size = 10
jitter = 0.3
alpha = 0.2
interaction_index = 'auto'
```

[ ]:
```
feature = 'num__VZAFSTANDDAGRECRTERREIN'
shap.dependence_plot(feature, shap_values_2, X_transformed_df,␣
 ↪dot_size=dot_size, x_jitter=jitter, alpha=alpha, interaction_index=None,␣
 ↪show=False)
# Set custom x and y axis labels
plt.xlabel('distance to nearest recreation area in meters')
plt.ylabel('SHAP value')
#plt.title('Custom Title')
plt.savefig('dependence_plot_LR_VZAFSTANDDAGRECRTERREIN.png', dpi=300,␣
 ↪bbox_inches='tight')

plt.show()
```