Erasmus University Rotterdam

From Carnivores to Herbivores

A study on identifying the characteristics of people following meat-based and (partly) plant-based diets

MASTER THESIS

Erasmus School of Economics

DATA SCIENCE AND MARKETING ANALYTICS

Author :	C.F.J. Spit
Student ID :	506524
Supervisor :	C.S. Bellet
Second assessor :	T.B.A.

Rotterdam, The Netherlands, June 29, 2023

From Carnivores to Herbivores

A study on identifying the characteristics of people
following meat-based and (partly) plant-based diets

C.F.J. Spit

June 29, 2023

Abstract

Plant-based diets such as vegetarianism and veganism have gained popularity during the last couple of years. The same holds for replacing animal-based meat with plant-based alternatives. Who is most likely to shift their dietary pattern and start following a more plant-based diet? Are these people different from 10 years ago? With food survey data it is possible to determine what the important factors for identifying these people are and that these have changed a bit in the Netherlands over the last decade. Data on food surveys from the years 2007-2010, 2012-2016 and 2019-2021 are researched with machine learning algorithms. Over time, the most popular diet is still the so-called meat-lover. This segment follows someone who is male, higher educated, 28 years or older and has a household size of 1, or more than 3 people. The most important features for following (partly) plant-based diets are sex, education and BMI level. Women and the higher educated are most likely to follow these. The meat substitute segment follows someone who has a BMI level of 'seriously underweight' or 'normal weight', a household size of 1 or more than 7 people and is between 29-40 years old.

Contents

Chapter 1

Introduction

1.1 Problem statement

The production of meat had tripled due to increasing demand worldwide since 1960 (Cuffey, Chenarides, Li, & Zhao, 2023). In high-income countries such as the Netherlands, meat consumption is declining. Still, consuming a lot of meat consumption is normative in this country (Godfray et al., 2018; Verain, Dagevos, & Jaspers, 2022). Many people consume meat on a daily basis. Even though it is associated with a high environmental burden. Eating a lot is even associated with a negative impact on one's health (Verain et al., 2022). Therefore, it is recommended to limit our meat consumption. Transitioning to more plant-based diets might even be key to addressing climate change.

Consuming plant-based meat alternatives (PBMA), also referred to as meat substitutes, instead of animal-based meat is one way to transition to a (partly) plant-based diet. The sale of PBMA in the United States increased significantly since the beginning of 2019 (Zhao, Wang, Hu, & Zheng, 2023). From a market share of 0.1% in 2017, in the class of total fresh meat sales, to 0.4% in mid-2020. It is still not even slightly comparable with the sale of animal-based meat, but it does show that meat substitutes are gaining popularity with our neighbours overseas.

In the Netherlands, providing plant-based alternatives in restaurants and offering multiple choices of PBMA on supermarket shelves is becoming more standardized every day. However, it is still not always possible to get a plant-based alternative or have some to choose from, especially outside of large cities.

Who are the people following a (partly) plant-based diet and who are those consuming meat substitutes? By identifying their most important characteristics, the results can be used to target people better. The characteristics will be investigated over the last decade, as following a more plant-based diet gained much popularity over the last couple of years. Therefore, these might

have changed over time. The results can, for instance, be used in further research to determine regions where people who are most likely to buy meat substitutes live. With that information, meat substitute-producing companies might be better able to coordinate their supply based on this. This will be touched upon in more detail in the managerial subsection. The research question of this paper follows:

"Is it possible to accurately determine with food survey data what important factors for identifying consumers who follow a specific food rule are and whether these have changed in the Netherlands over the last decade?"

In this paper, there are four food groups identified, also referred to as 'food rules'. The meat-lovers, plant-centered, flexitarians and those consuming meat substitutes. The plant-centered rule represents the pescatarians, vegetarians and vegans combined. As the shares of these groups were too small on their own. Pescatarians are consumers who do not consume meat but, do consume fish, seafood products and animal-derived products such as dairy products and eggs (Wozniak et al., 2020). Vegetarians are consumers who do not consume meat and fish but, do consume animal-derived products (Melina, Craig, & Levin, 2016). Vegans are consumers who do not consume meat, fish or any animal-derived products (Wozniak et al., 2020). Thus, for the latter, it holds that this diet is completely plant-based.

Flexitarians are consumers who do not eat meat on a daily basis but occasionally without having strict guidelines for the number of days (Wozniak et al., 2020). In this paper, the threshold of consuming meat is set at a maximum of once in two non-consecutive days. The type of meat is not taken into account, it might be spread on loaves of bread or as the main ingredient of dinner. The umbrella term for following a '(partly) plant-based food rule' refers to the flexitarian and the plant-centered food rules.

The so-called 'meat-lovers' are consumers who do not follow one of these two food rules. Hence, they do consume meat more than once in the two non-consecutive days.

Furthermore, the last food rule presents those consuming meat substitutes. The substitutes are imitation meat that is completely made out of plant-based ingredients. Therefore, respondents from all the food rules might consume these.

1.2 Research questions

As already presented, the research question follows: *"Is it possible to accurately determine with food survey data what important factors for identifying consumers who follow a specific food rule are and whether these have changed in the Netherlands over the last decade?"*.

The following sub-questions will be studied:

- What are the most common food rules to follow in the Netherlands and have these changed over the last decade?
- What are the most important features for predicting whether someone is a meat-lover?
- What are the most important features for predicting whether someone follows a (partly) plant-based food rule?
- What are the most important features for predicting whether someone consumes meat substitutes?
- How does gender influence the probability of following a (partly) plant-based food rule and consuming meat substitutes?
- How does age influence the probability of following a (partly) plant-based food rule and consuming meat substitutes?
- How does educational attainment influence the probability of following a (partly) plant-based food rule and consuming meat substitutes?
- Have the predictors changed in the Netherlands over the last decade?

1.3 Motivation

1.3.1 Academical

This study is academically relevant in multiple ways. First of all, it will investigate whether the predictors have changed over the last decade by using a recently published dataset of the RIVM. The 2019-2021 dataset on national food consumption has just been made available when this research started (RIVM, 2021). Therefore, it might present new insights. The predictors might have changed over time. As following a more plant-based diet gained popularity over the last couple of years, it might be that different kinds of people are encouraged to change their

diet. For instance, by their surroundings, advertisements or the increased number of plant-based alternatives in supermarkets. This might make the group who follows such a diet larger and more diverse. A more detailed picture can be drawn from a larger number of observations. Therefore, the number of important predictors might have increased as well. In addition, the predictors might shift from more straightforward ones such as 'gender', with limited answer possibilities, to more detailed thresholds for 'age' or 'household size' for example.

Furthermore, this research is relevant as only very few earlier studies have been conducted on non-vegetarians for studying the potential transition to a food rule including more plant-based foods. Therefore, the prevalence and characteristics of meat-lovers are considered as well.

The same holds for researching people consuming meat substitutes. This will be taken into account as a specific food rule. Earlier literature about meat substitutes has only been done on consumer spending behavior and not on consumer identification.

Little research has been done on flexitarians as well. Earlier studies have mostly been conducted on a single-year dataset instead and have not looked into transitions. This research also investigates whether the characteristics of flexitarians changed over the last decade and therefore, adds value to existing literature. Additionally, investigating whether flexitarians differ from the more dedicated plant-centered food rule followers, might gain new insights.

With earlier Dutch research the characteristics, prevalence and consumers' attitudes of flexitarians in the Netherlands are identified and studied whether these have changed over the last decade. However, this paper differs in multiple ways. The identification will be first of all done for multiple food rules. As in other earlier literature, it was suggested to not only take meat-lovers and flexitarians into account. Also, it focuses on important features as determinants of whether someone follows a specific food rule instead of attitudes. Investigating whether the characteristics of consumers have changed over the last decade is an addition to already existing literature, specifically for the Netherlands.

Accurate identifications will provide information about socio-demographic factors that play an important role in following one of the food rules. Identifying these for all different stages of 'meat eaters' is new to earlier research. Based on the outcomes, further research can be done. Such as practical applications for specific neighborhoods or new campaigns and policies that can be designed to motivate people to eat less meat or shift their dietary patterns completely.

1.3.2 Managerial

The outcomes of the predictions of which consumers are most likely to be meat-lovers, consume meat substitutes or follow a (partly) plant-based food rule are interesting for businesses like the agricultural industry, producers of meat substitutes, supermarkets, policymakers and non-profit institutions. Policymakers of the agricultural industry can be helped to adjust and adapt to possible changing shares of consumers buying meat. Policymakers and industry leaders can prepare for changes in demand and supply. This could include adjusting agricultural production methods, investing in new technologies, changing their marketing strategies or opening new marketing channels.

The outcomes of this research might also be interesting for producers of meat substitutes such as Garden Gourmet and the Dutch brand the Vegetarian Butcher. Based on the important features predicting which consumers use meat substitutes, they might be able to respond to this in some regions where a large share of consumers with the characteristics of buying meat substitutes lives. The companies can do this by scaling up their production or releasing new variants of substitutes. Before they might be able to present new releases, they need to invest in research about consumers' tastes and preferences. Expanding the assortment of substitutes is also a way of differentiating themselves from their competitors.

Supermarkets could use the information about changes in the prevalence of consumers buying meat, meat substitutes or following a (partly) plant-based food rule with regard to their supply of meat, meat substitutions, greens and legumes. They can capitalize on the demand for these and possibly adjust their supply and (expand) their assortment. This can be done in specific neighborhoods where the share of consumers with characteristics as the most important ones live, is high. Therefore, consumers will be less likely to find empty shelves. In addition, this might be a way for supermarket branches to attract new customers who buy meat substitutes and enjoy a broad variety of choices. As new customers might buy additional groceries as well. Consequently, this might increase sales.

In addition, the production of meat and other animal products is a major contributor to greenhouse gas emissions, deforestation and other forms of environmental degradation. Reducing the production and intake of meat is therefore an important topic. Dutch policymakers who are concerned about this can use the results about which consumers are most likely to be meat-lovers to design food-specific interventions. For instance, populations or individuals who are dedicated to consuming meat can be identified. For those, interventions can be developed to encourage them to reduce their meat intake.

This is related to the relevance for non-profit institutions such as the Dutch organization

Wakker Dier. Whether it is identified that specific populations, neighborhoods or individuals are less likely to leave meat out of their meals, they can be targeted with a new campaign. In that way, they can be made more aware of the consequences of their choices and stimulated to lower their meat consumption. The organizations can develop targeted interventions to support consumers in making more sustainable meat consumption choices. This could include providing information on the health and environmental impacts of different types of meat or developing incentives for reducing meat consumption.

Chapter 2

Literature Review

2.1 Theoretical framework

This study focuses on the characteristics of meat-lovers, plant-centered (pescatarians, vegetarians and vegans), flexitarians and those who consume meat substitutes. Furthermore, it is investigated whether important predictors have changed in the Netherlands over the last decade. Related literature focuses on motives, enablers and barriers to following a specific food rule and who consumers following such a rule are. Literature related to this topic is researched to gather insights with regard to the research question. Earlier studies have been done on identifying who consumers buying meat substitutes, flexitarians and vegetarians are. Little research has been done on vegans and pescatarians. Nor on changes in the identifications over time. Research that has already been performed is more explanatory than predictive. This paper will fill the gap by identifying consumer characteristics, as well as researching changes over time, what can be used for further insights. More recent findings might be helpful in targeting the right consumers with a focus on the environment in the Netherlands.

The related literature is grouped into multiple themes related to the research question *"Is it possible to accurately determine with food survey data what important factors for identifying consumers who follow a specific food rule are and whether these have changed in the Netherlands over the last decade?"*. The first theme is about the attitudes and beliefs of food rules. Reasons for committing to a specific food rule. This is important for researching motivations to follow a (partly) plant-based food rule and outlining the topic. The studies of Mullee et al. (2017) and Reuzé et al. (2022) investigated different attitudes. Overall respondents gave reasons for following a (partly) plant-based food rule such as: their health, taste, followed by food rule and physical environment-related arguments. The prevalence of consumers following a specific food rule

and customer identification is captured in the second theme. The relation between who the consumers are and what kind of food rule they follow is researched in the studies of Wozniak et al. (2020) and Deliens, Mullie, and Clarys (2022). An increase in the prevalence of Swiss vegetarians between 2005-2017 was found (Wozniak et al., 2020). For the Belgians, a decrease in the prevalence of omnivores and an increase in the prevalence of flexitarians were reported (Deliens et al., 2022). Furthermore, Wozniak et al. (2020) identified characteristics of food rule followers and the relation with health effects. The third theme is meat substitute spending patterns, changes over time and its relation to meat. Cuffey et al. (2023) stated that, even though only one-fourth of the respondents bought PBMA, the spendings on meat did not change largely. In contrast, a decrease in spendings on meat would have been expected when meat is being substituted for PBMA. Zhao et al. (2023) also did not find a decrease in the interest in meat, even though in the US the market share of meat substitutes four-folded between 2017 and mid-2010. This implies that consumer interest in meat substitutes increased, but their interest in meat did not decrease. The fourth theme is about shifting behavior from meat-based to plant-based food rules. Verain et al. (2022) studied different groups of Dutch flexitarians and stated that the prevalence of the die-hards decreased, but of the less strict flexitarians increased during the last decade. Graça, Godinho, and Truninger (2019) incorporated the barriers and enablers consumers experience to shift their food behavior from meat-based to plant-based food rules while looking at capability, opportunity and motivation. Therefore, the third and fourth themes are related to changes over time. The fifth and last theme is about the methods for investigating who the different food rule followers are and finding potential segments. Lusk (2017) researched with decision trees the characteristics of US vegetarians. Furthermore, Lee (2014) applied different resampling techniques to machine learning models created for the medical field.

Mullee et al. (2017) found that overall women are more likely to agree with positive statements towards a vegetarian food rule than men. The research investigated motives and beliefs about following specific food rules. The most popular motives for eating meat among omnivores and semi-vegetarians are 'good taste', 'habit' and 'this is how I was brought up'. 50% of them believed that eating meat is not unhealthy. For omnivores, Mullee et al. (2017) found that reasons to consider adapting to a more vegetarian food rule are 'my health', 'to discover new tastes' and 'to reduce weight'. They also found reasons for omnivores to not follow a vegetarian food rule. These are, among others, 'no interest', 'the taste', 'I never thought about it' and 'limited personal cooking skills'. For semi-vegetarians, the reasons reported were 'no reason', 'insufficient

vegetarian options' and 'limited personal cooking skills'. According to the research done by Reuzé et al. (2022), the strongest change-inducing motives for participants to reduce their meat consumption were 'good to vary both food rule and protein sources', 'healthier' and 'better for the physical environment to limit meat'. These motives are considered to be the most effective for changing food behavior. They are followed by less strong motives, but also mentioned by meat reducers, namely 'dislike for the taste of meat', 'healthier to avoid meat' and 'doctor's advice'. All change-inducing motivations were investigated for their association with meat reduction. Participants who reported 'healthier' were more likely to be women, older, had a higher income and were highly educated. Participants who reported 'physical environment' were more likely to be younger, highly educated and less likely to live with a child. Participants who reported 'doctor's advice' to be a motive to reduce their meat intake were men, older and those with lower education. Therefore, they may be most motivated by the information and expertise of health professionals.

The prevalence of food rule trends to analyze socio-demographic characteristics and associates with health effects is researched by Wozniak et al. (2020). They found an increase in the prevalence of vegetarians from 0.5% to 1.2% over the 13-year study period. Vegetarians were more likely to be female, younger, higher educated and had a lower income in comparison to omnivores. Furthermore, with regard to the cardiovascular risk factor as a health effect, it is presented that vegetarians are, compared to omnivores, less likely to be overweight, obese, hypercholesterolemia and hypertensive. Also, for flexitarians a reduced risk of being overweight, obese and hypertensive was found. For pescatarians, a reduced risk of being obese and hypercholesterolemia is presented. Therefore, it is stated that those who reduced their meat intake or excluded meat at all had a lower BMI, total cholesterol and hypertension than omnivores. In conclusion, following a food rule with a lower meat intake is associated with having a better cardiovascular risk profile than following a meat-based food rule. Deliens et al. (2022) presented in their Belgium study on Flanders adults that most participants were omnivores. With regard to the baseline year (2011), the relative number of omnivores decreased from 89% to 84.6% in 2016 and 72.7% in 2020. In contrast, the relative number of flexitarians increased in comparison with the baseline. From 5.3% in 2011 to 10% in 2016 and to 9.2% in 2020. Presenting a plateaued level of the prevalence of flexitarians. For the vegetarian/vegan groups, no differences of time were found and thus, no trends were observed. The descriptive statistics presented that the vegetarian/vegan group consisted of more females, youngsters (age group 18-34 years), higher educated and living in urban areas compared to omnivores.

The studies of Cuffey et al. (2023) and Zhao et al. (2023) both used Nielsen datasets to

investigate US consumer interests in meat substitutes. Cuffey et al. (2023) investigated consumer spending on first-time bought PBMA in a non-laboratory-environment. They found that spending on food in general increases with USD 50 before and after the moment of this first purchase. On PBMA this is only USD 8. This suggests that consumers change their overall spending and buy multiple products. This increase pulls in during the subsequent month. In addition, spending on meat did not change extensively. If consumers would be substituting meat for PBMA a decrease in spending on meat would be expected, ceteris paribus. A decrease is also not found by Zhao et al. (2023). They stated that when consumers' budget for food is higher, they are less likely to purchase PBMA than animal-based meats. In addition, they presented that when prices of PBMA increase with 1%, the demand decreases by 1.5%, while taking time, state and promotion effects into account. This makes PBMA the most price-elastic product of all products in the fresh meat category. When researching cross-price elasticity between the different kinds of products in the fresh meat category, Zhao et al. (2023) concluded that PBMA are complementary to beef and pork and substitutional for chicken, turkey and fish. Even if the price of PBMA changes with 1%, it affects the demand for all other types of meat with only less than 0.01%. Thus, changes in PBMA prices have a relatively minimal impact on the demand for meat products. Regarding expenditure elasticity, they presented that an increase of 1% in fresh meat expenditure leads to an increase in demand for PBMA of 0.780. While for animal-based fresh meats, the demand increase is around 1. Indicating that, when consumers' budget for food is higher, they are less likely to purchase meat substitutes than animal-based meats.

Verain et al. (2022) researched consumers' attitudes and motives regarding reducing their meat intake and shifts towards more flexitarian food rules in the 2010s. They found that the number of more 'dedicated' flexitarians (eating meat a maximum of two days a week) decreased significantly over the decade. Whereas the number of 'light' flexitarians (eating meat five or six days a week) increased significantly from 13.0% in 2011 to 42.9% in 2019. With regard to the attitudes and norms for eating meat, most beliefs remained the same over the decade. Only a decrease in the beliefs about positive health consequences due to a lower meat intake was found. Increases in positive attitudes towards reduction, the appreciation of eating less meat and giving more importance to animals and the environment were reported as well. The most dedicated flexitarians turned out to have higher personal norms regarding reducing their meat intake and appreciated meatless meals more than the meat-oriented. In addition, the environment and animal welfare were important for this group. However, no evidence was found that this group believed in the health effects of omitting meat. The characteris-

tics of the most dedicated flexitarians were most likely to be women, average aged around 50 years old, medium/high education, were single or living with a partner and were born in the Netherlands (Verain et al., 2022). Graça et al. (2019) investigated with a total of 110 articles the barriers and enablers for shifting from a meat-based to a plant-based food rule. 68.2% of them were made public in the last 6 years when the research was conducted in 2018. Therefore, this presented increasing interest in this topic for researchers. The focal topics were; meat reduction/curtailment, plant-based food rules and meals. The papers are investigated on focal topic, characteristics, design, sample and main theoretical framework. Resulting of investigating all papers, Graça et al. (2019) found that men were associated with unwillingness to adapt to a more plant-based food rule and increased meat consumption. For women, the contrary held. With regard to age, the articles presented quite different findings. Socio-economic status (SES) variables presented that following a plant-based food rule was in line with higher values for these variables. Lastly, living in urban areas was associated with plant-based food rules as well. With regard to the barriers and enablers of the COM-B model (capability, opportunity and motivation for understanding behaviour change), Graça et al. (2019) report that a barrier considering capability is learning new cooking skills. For opportunity, it is limited social support and prejudice. With regard to motivation, observational evidence presents barriers such as a lack of moral engagement, responsibility, concern and familiarity. In addition, respondents follow unhealthy lifestyles and have concerns about reduced meat food rules for their health. From experimental evidence, a barrier presented is that respondents hold strong beliefs that including meat in their food rules is healthy, climate-friendly and necessary. As enablers for opportunity based on observational evidence, supportiveness from close others and increased prices of meat are reported. Based on experimental evidence changes in service provision in collective meal contexts, for example in canteens, are presented. Enablers for motivation based on observational evidence are being interested in eating healthier, trying new foods, being environmentally conscious, having altruistic values and taking animal welfare into account. Based on experimental evidence, giving reminders, focusing on positive and appealing representations of plant-based food rules and highlighting the environmental impact of meat are enablers.

In one of the case studies conducted in the paper by Lusk (2017), the characteristics of consumers following a vegetarian or vegan food rule are researched. *"Who are the vegetarians/vegans?"* is investigated with a logit model and a decision tree. These methods are used to predict someone's vegetarian status by socioeconomics and demographic characteristics. A binary logit model was estimated. Furthermore, a decision tree was created to compare the

results. First, a large tree was created. Next, it was 'pruned' by cross-validation to make sure that only variables that add to the predictive power of the model are included. The tree had a better sensitivity score (true positives) than the logit model for both the test as well as the validation set. On the specificity score (true negatives), it performed a bit better. From the decision tree, it became clear that being liberal, participating in the Supplemental Nutrition Assistance Program (SNAP), having an income higher than $60,000, having children under 12 years old, being at least 35 years old or under 35 in combination with a household size of less than three, are most likely to be vegetarians/vegans. With resampling techniques to tackle the problem of a class-imbalanced dataset (meaning one that consists mostly of one class), Lee (2014) did research in the medical field. At first, sub-optimal results were attained due to the imbalanced classes. Thus, over- and undersampling are applied. The Classification and Regression Tree (CART) models were fitted on the training, oversampled training, weighted training and undersampled training datasets. With oversampling, a sample size of the small class comparable to the large class is made. With undersampling, a sample size of the large class comparable to the small class is made. Also, resampling case-to-control ratios of 1:1, 1:2 and 1:4 were investigated, and the performances of the models were researched. Demographic and socio-economic characteristics of respondents were included. It was found that CART performs better on the oversampled and undersampled training datasets than on the regular training dataset. This is based on the area under the receiver operating characteristic curve (AUC). Therefore, the performance of the decision trees could be improved by applying oversampling to the minority class. Afterwards, the performance should be investigated. An overview of the data used for each study and the key findings are presented in Table 2.1

Table 2.1: *Literature overview of data and key findings*

Paper	Data	Key findings
Mullee et al. (2017)	Case study. Cross-sectional study in Belgium among 2,436 adults. An online questionnaire to measure level of (dis)agreement was used to gather information in March 2011. Representative for general population for: sex, age, education and urbanization level. Prevalence of vegetarians with regards to population is low. High for semi-vegetarians and omnivores.	Only 10% of the respondents stated to not eat meat or fish on one or more days of the week. 25% believed that eating vegetarian meals often is unhealthy for you. Reasons for not following a vegetarian diet for omnivores were, among others, 'no interest', 'the taste', 'I never thought about it' and 'limited personal cooking skills'. For semi-vegetarians, reasons such as 'no reason', 'insufficient vegetarian options' and 'limited personal cooking skills' were reported.
Reuzé et al. (2022)	Cross-sectional study among 25,393 non-vegetarian adults from the French NutriNet-Santé cohort. Gathered data via an online questionnaire in 2018. As the participants might be biased against the topic, the cohort might not be representative of the French population.	This study presented that the most important motives for participants to induce a change in their meat or legume consumption were: health, nutrition, physical environment and taste. Followed by social influences, meat avoidance and dislike. Being a woman and highly educated for health motives were found to be associated with decrease in meat consumption and an increase in legume consumption.
Wozniak et al. (2020)	Yearly cross-sectional study in Geneva, Switzerland, conducted between 2005-2017. 10,797 individuals from population-based representative sample. Gathered from Bus Santé study. Questionnaires, anthropometric measurements and blood tests used for the study.	Study shows an increase in the prevalence of vegetarians from 0.5% to 1.2% over the study period of 13 years. Vegetarians were more likely to be females, younger, higher educated and had a lower income than omnivores. In addition, positive health associates were found for participants who reduced their meat intake/excluded it. The prevalence of pescatarians increased from 0.3% to 1.1%, for flexitarians it remained the same. In comparison with omnivores, both are more likely to be women and flexitarians are more likely to have a lower income as well.

Table 2.1: *Literature overview of data and key findings*

Paper	Data	Key findings
Deliens et al. (2022)	Cross-sectional study conducted in Flanders, Belgium. An online survey was held among 4,859 participants in five different representative cohorts in 2011, 2013, 2016, 2018 and 2020. Participants were sampled to be representative regarding sex, age, education level and urbanization level.	The share of omnivores decreased significantly in 2016 and 2020 with regards to the baseline in 2011. The relative number of participants being flexitarian nearly doubled in 2016 and remained stable till 2020. Only, no trends were observed for vegetarian/vegan group. Participants following a plant-based diet were associated with being women, younger, higher educated and living in urban areas.
Cuffey et al. (2023)	Nielsen Consumer Panel Data (CPD) from 52,022 US households between 2014-2019. Including recorded product-level purchase and socioeconomic information on panellists. Linked with USDA 2013 RuralâUrban Continuum Codes and the 2015 USDA Food Access Research Atlas.	When consumers first bought PBMA, overall spending on food increases with USD 50, on PBMA this is only USD 8. This increase pulls in during the subsequent month. The spending on meat does not change extensively. They stated that only a quarter of the sample used has ever purchased PBMA products during the timespan investigated.
Zhao et al. (2023)	Nielsen Scantrack scanner data of US households between 2017 and mid-2020. The data consists of 712 Universal Product Code (UPC) level meat alternative products. Merged with weekly new positive COVID-19 cases data from COVID Data Tracker at the Centers for Disease Control and Prevention.	When prices of PBMA increase with 1%, the demand decreases with 1.5%, ceteris paribus. This makes PBMA the most price elastic of all products in the fresh meat category. Cross-price elasticity between the different kinds of fresh meat showed that PBMA products are complementary to beef and pork and substitutional for chicken, turkey and fish. Even if the price of PBMA changes with 1%, it affects the demand for all other types of meat with only less than 0.01%.

Table 2.1: *Literature overview of data and key findings*

Paper	Data	Key findings
Verain et al. (2022)	Study on Dutch adults. Data is gathered via two online surveys. First on in 2011 and partly repeated in 2019. Participants were recruited by market research agency. 1,253 participants in 2011 and 2,383 in 2019. This study focuses on 2019 and the comparison.	The number of more 'dedicated' flexitarians decreased significantly over the decade. Whereas the number of 'light' flexitarians increased significantly. The most dedicated flexitarians had higher personal norms regarding reducing their meat intake, appreciated meatless meals more and found the environment and animal welfare important in comparison with the meat-oriented. The most dedicated flexitarians were most likely to be women, aged older, had a medium/high education, were single/with a partner and born in the Netherlands.
Graça et al. (2019)	Research from January 2018. 11 databases were studied which resulted in a total of 110 articles being included in qualitative synthesis which were in line with the inclusion criteria. The articles were published between 1989 and 2018 and were limited to English papers only. The focal topics were: meat reduction/curtailment, plant-based diets and meals.	The barriers and enablers of the topics were identified. In addition, the variables were classified into the components of the COM-B model, standing for capability, motivation and opportunity for understanding behavior change. Being female, having higher values for SES variables and living in urban areas were associated with being more likely to follow plant-based diets. Among barriers found are limited cooking skills, social support and prejudice, moral engagement and responsibility. Among enablers presented were supportiveness from close others, changes in service provision, being (environmentally) conscious and altruistic and taking animal welfare into account.

Table 2.1: *Literature overview of data and key findings*

Paper	Data	Key findings
Lusk (2017)	Three different studies performed. Data is of three years (June 2013-January 2016) collected from the online survey Food Demand Survey (FooDS) with data of over repeatedly delivered 1,000 US consumers measured monthly. Resulted in 32,683 survey responses. For third study also 5,175 observations between February-June 2016 are used to test predictive performance of models.	This study used a logit model and a decision tree to determine characteristics of vegetarians/vegans in the US. The tree performed better on sensitivity and slightly on specificity. The results show that people who are being liberal, on SNAP, an income of $60,000 or more, children under 12 years old, 35 years or older or under 35 with a maximum household size of 2, are the most likely to be vegetarian/vegan.
Lee (2014)	The Nutrition Examination Survey (NHANES) of 2009-2010 is used to determine resampling techniques in the medical field. It consists of 4,677 respondents over the age of 19 from the US.	It was found that CART perform better on the oversampled training and undersampled training datasets than on the regular training dataset, based on the AUC. The performances of the models were improved when applying undersampling. An AUC of 0.70 or higher could be achieved. Therefore, resampling methods can improve the classification power of CARTs.

2.2 Conceptual framework

Prevalence of following specific food rules

With regard to the most common food rules, expectations about the prevalence of people following a specific food rule in the Netherlands are drawn. The prevalence of Belgian omnivores decreased from almost 90% in 2016 to 72.7% in 2020 (Deliens et al., 2022). Therefore, the prevalence of Dutch meat-lovers is expected to be around 70% as well, as Belgium and the Netherlands are relatively comparable with regard to cultural differences. In addition, it is expected that the prevalence of consumers buying meat substitutes is at a maximum of 25%. The research of Cuffey et al. (2023) found that a quarter of the US sample they investigated had ever purchased a PBMA. The data used were from the last five years of the 2010s and thus with regard to time comparable to the dataset used for this report. In Belgium, the prevalence of flexitarians was found to increase from 5.3% in the baseline year 2011 to 10% in 2016 and to stay relatively constant over time with a share of 9.2% in 2020 (Deliens et al., 2022). Therefore, the prevalence of Dutch flexitarians is expected to be around 10%. The study of Wozniak et al. (2020) found that the prevalence for Swiss pescatarians and vegetarians increased between 2005-2017 from 0.3% to 1.1% and from 0.5% to 1.2% respectively. For vegans, it is more difficult to draw an expectation. As the vegan diet is even stricter, it is expected to be followed by fewer people. Therefore, the prevalence of vegans is not expected to exceed 1%. If for the pescatarians and vegetarians, there is an upward trend in committers, it is expected that the prevalence for the plant-centered will be, together with the vegans, around 3.3%. As the results are from already some years ago and more and more attention is being paid to meat-reducing food rules recently. Therefore, the first hypothesis states that:

H1: The prevalence of people following a specific food rule is expected to be in decreasing order; meat-lovers, consumers using meat substitutes, flexitarians and the plant-centered.

Identification of (partly) plant-based food rules, PBMA and meat-lovers

Based on the literature described, multiple socio-demographic characteristics might be important features for the identification of consumers that follow a specific food rule. An overview of the characteristics described in the papers for following a (partly) plant-based food rule is presented in Table 2.2. However, only the papers by Cuffey et al. (2023) and Mullee et al. (2017) present characteristics of consumers buying meat substitutes instead. These will be discussed along. Gender is expected to play an important role in the probability of whether someone follows a (partly) plant-based food rule or not. Women are more likely to follow a plant-based

food rule compared to men (Wozniak et al., 2020; Deliens et al., 2022; Verain et al., 2022; Graça et al., 2019). According to the research of Mullee et al. (2017), women are more likely to consume meat substitutes than men as well. Therefore, gender is expected to be of importance. In addition, age is also expected to be an important feature for determining whether a consumer follows a specific food rule. With regard to this variable, earlier literature is a bit divided. Younger aged are more likely to follow such a food rule, according to Wozniak et al. (2020) and Deliens et al. (2022). Whereas another study states that in combination with age, the composure of the household is also determined Lusk (2017). However, as those under 35 years can also be considered moderately young, it is expected that younger respondents are more likely to follow a specific food rule compared to older respondents. Also, the educational level attained is expected to be an important feature for the identification. Consumers who have attained a higher education are more likely to follow a (partly) plant-based food rule than consumers with a lower education (Wozniak et al., 2020; Deliens et al., 2022; Verain et al., 2022; Graça et al., 2019; Lusk, 2017). Based on the research of Cuffey et al. (2023) it is expected that education will also be an important determinant for predicting whether a respondent will buy meat substitutes. Therefore, being higher educated is expected to be of influence. Concluding, the variables gender, age and education are expected to have an influence on the identification of following a (partly) plant-based food rule. The hypotheses based on this follow:

H2: Women are more likely to follow a (partly) plant-based food rule and to buy meat substitutes than men.
H3: The younger aged are more likely to follow a (partly) plant-based food rule than the older aged.
H4: Higher educated are more likely to follow a (partly) plant-based food rule and to buy meat substitutes than lower educated.

Other important features that might have an influence on whether someone follows a (partly) plant-based food rule or not, are income and living area. The level of education might also be related to the income someone earns. The higher the educational attainment, the more likely someone's income will be high as well. Therefore, income is not surprisingly a candidate of importance. Consumers with a higher income or an income of over $60,000 are more likely to follow a specific food rule (Graça et al., 2019; Lusk, 2017). However, only the research of Wozniak et al. (2020) presented that a lower income would increase this likelihood. In contrast to the (partly) plant-based food rules, income is not expected to play an important role in

meat substitute purchasing behavior. Zhao et al. (2023) reported that in the US the meat substitute's market share four-folded in the late 2010s. Thus, increased interest among consumers was found. However, it is stated that an increased food budget leads to a lower likelihood of buying PBMA products than animal-based meats. This is in line with the finding that after the first time a consumer had bought PBMA, the increased spending on PBMA flattens in the subsequent month (Cuffey et al., 2023). Therefore, with what strength the influence of someone's income is on their PBMA purchasing behavior is questioned. Someone's living area can also be of importance whether people will follow a (partly) plant-based food rule. Consumers who live in urbanized areas are more likely to follow a specific food rule than consumers living in rural areas (Deliens et al., 2022; Graça et al., 2019). It might be that trends develop further and faster in urbanized areas as cities than in the rural countryside. Based on the research of Cuffey et al. (2023) it is expected that living area is also an important determinant for predicting whether a respondent will buy PBMA. Resulting from this, living in more urbanized areas is expected to be of influence. In addition, farms are located in, and practically the definition of rural areas. Therefore, it is not surprising that people who keep and raise animals on a farm are less likely to follow a (partly) plant-based food rule. In addition, based on these insights, it is expected that the variables income and living area also influence the chance that someone follows a (partly) plant-based food rule. Therefore, the hypotheses follow:

H5: People with a higher income are more likely to follow a (partly) plant-based food rule than people with a lower income.
H6: The variable income is not expected to be of influence for buying meat substitutes.
H7: Living in more urbanized areas will influence whether people will follow a (partly) plant-based food rule and buy meat substitutes.

Table 2.2: *Socio-demographic characteristics for (partly) plant-based/meat substitute food rule*

Paper	Gender	Age	Education	Income	Living area	Additional
Wozniak et al. (2020)	Women	Young	Higher	Low		
Deliens et al. (2022)	Women	Young	Higher		Urban	
Verain et al. (2022)	Women	Average 50	Medium and higher			Single or with partner, born in Netherlands
Graça et al. (2019)	Women	Different outcomes presented by literature	Higher	High	Urban	
Lusk (2017)		>35 OR <35 + household size = 2	Higher	Over $60,000		Liberal, kids <12 y/o, on SNAP
Mullee et al. (2017)	Women					
Cuffey et al. (2023)			Higher		Coast or metropolitan	

With regard to meat-lovers, the study by Reuzé et al. (2022) found that gender and education are also determinants of influence for this group. They namely stated that being a woman is associated with a decrease in meat consumption. Therefore, it is expected that meat-lovers will be associated with being a man. In addition, it was stated that being highly educated for health motives was also associated with reducing meat consumption (Reuzé et al., 2022). Based on this result, it is expected that being lower educated might be related to being a meat-lover. With regard to the meat-lovers, it is expected based on the described literature that the following hypotheses will hold:

H8: Men are more likely to be meat-lovers than women.

H9: Lower educated are more likely to be meat-lovers than higher educated.

Identification of changes over time

The study of Verain et al. (2022) is most in line with the study that will be conducted in this paper. They identified attitudes and norms of different levels of flexitarian consumers in the Netherlands and investigated whether and how these changed over the last decade. In this report, a comparable study will be performed. The identification will be done for flexitarians, as well as for meat-lovers, consumers who buy meat substitutes and the plant-centered. The research of Verain et al. (2022) compared the prevalence of different levels of flexitarians and consumers' attitudes over time. This report will focus, next to prevalence, on important features as determinants of whether someone follows a specific food rule. The outcomes of the most dedicated flexitarians might be comparable with the flexitarians in this paper. In addition, the time and place of both studies are relatively identical. With the research of Verain et al. (2022), it was found that women, with an average age of 50 years old, a medium to high education, single or living with a partner and born in the Netherlands are most likely to be dedicated flexitarians. Furthermore, at least no changes in the beliefs with regard to consumers' attitudes and norms towards eating meat were found (Verain et al., 2022). However, attitudes and norms are not the same as characteristics identifying consumers. These might have changed over time as plant-based diets gained much more popularity over the last couple of years. As little research is done about this, it is difficult to draw an expectation. Therefore, the hypothesis is still based on this outcome and follows:

H10: The important predictors for determining whether someone followed a (partly) plant-based food rule are not expected to have changed over the last 10 years.

Method related expectations

What the important factors are for identifying consumers who follow a specific food rule, decision trees will be built. On the one hand, based on the research of Lusk (2017), who researched *"Who are the vegetarians/vegans?"*, it is expected that the pruned trees will give clear insights into what the most important features are. On the other hand, based on the article of Lee (2014), it is expected that resampling (or weighting) will be needed for the classification tree to be a good prediction model. Thus, the last two hypotheses state:

H11: Post-pruned decision trees will provide clear insight into the most important features.
H12: Designing decision trees with resampling techniques to overcome the problem of the imbalanced dataset, will provide clear insights into the most important features.

Chapter 3

Data

This data section is divided into multiple paragraphs. First, the general information of the datasets used will be described. This includes the objective and recruitment process of how the data are gathered. Next, the dataset description follows which presents the content of the datasets. In addition, similarities and comparisons between the datasets are discussed. Furthermore, all variables included in the datasets are presented. After that, the dataset cleaning, variable creation and transformation paragraph captures how the datasets are cleaned, which new variables are created and which transformations are done. The section concludes with the descriptive statistics, per wave and per food rule.

3.1 General information

Objective

The datasets used in this study are the Dutch National Food Consumption Surveys (DNFCS) of 2007-2010 (RIVM, 2010), 2012-2016 (RIVM, 2016) and 2019-2021 (RIVM, 2021) from the Rijksinstituut voor Volksgezondheid en Milieu (RIVM). The aim of DNFCS is to gain insights into the diets of children and adults living in the Netherlands. The datasets are focused on the diet by subgroups of the population. For example, based on socio-demographic factors and on changes in food consumption in, among others, national meat consumption. It is made available for dietary environmental impact estimation. The DNFCS research is done in order of the Ministry of Health, Welfare and Sport. The data were collected by drawing a representative sample of a panel from the market research agency Kantar. Only for the wave 2007-2010, they were drawn from a panel of the GfK research agency instead. Therefore, some things with regard to the study population and recruitment process of this dataset differ from the other waves. Whenever that is the case, it is mentioned. For all datasets holds that the participants

were not selected based on the kind of diet they followed. The datasets match with the research that will be conducted about determining important features for identifying whether someone is a meat-lover, follows one of the (partly) plant-based food rules or consumes meat substitutes.

Recruitment process

The population targeted were males and females between 1 and 79 years old, living in the Netherlands and having a good command of Dutch. With the exception of pregnant and lactating women and institutionalized. Only for the 2007-2010 dataset, people between the ages of 7-69 years were targeted. People were invited to participate in the research by means of a written Dutch invitation letter. For the wave 2007-2010, this was either done by email. A different kind of letter was used for caregivers of the selected children (aged between 1-15 years old) and for adolescents (12-20 years old). If someone wanted to participate, a reply form could be filled out and sent back or an answer could be given online. When agreeing on participation, a general questionnaire was sent, preferably digitally. Different age groups got different ones. The age groups were: 1-3, 4-11, 12-18, 19-90 and 71-79 years old. For the 2007-2010 dataset, these were: 7-11, 12-18 and above 18 years old. The questions were about multiple topics, such as physical activity, educational level, family situation, smoking habits, alcohol consumption, consumption of specific foods, dietary supplements, as well as background and lifestyle factors. Within a time span of four weeks, the respondents were called twice on non-consecutive days by an interviewer for their 24-hour dietary recalls. Then, all food and drinks consumed by the respondent were reported to the interviewer. All days of the week were overall equally represented in the recalls. Children up to 8 years old and adults from the age of 70 years, were visited at home for their first interview and called for their second one. Children between 9-15 years old were visited at home for both interviews. Adults between 16-70 years old were called both times. The caregivers of children were interviewed about their child's drinking and eating habits. In addition, the child's height and weight were measured. For the 2007-2010 dataset, participants between 16-69 were called twice for the interviews. Children between 7-15 years old were visited at home, their caregivers were present during the interview. The computer-directed interview program GloboDiet was used to standardize the interviews and import the answers directly into the computer. Only for the 2007-2010 dataset, the program EPIC-Soft was used.

3.2 Dataset description

For the 2007-2010, 2012-2016 and 2019-2021 datasets (also referred to as waves) information of 5,502; 4,313 and 3,570 respondents respectively were collected. However, for the first dataset, only the information of 69% of the observations was complete. Thus, this resulted in 3,819 respondents. For all waves, there are two datasets provided. One with information about the participants' characteristics and one with the results of the 24-hour dietary recalls and their consumption behaviour. As over 10 years have gone by when the DNFCS 2007-2010 was conducted in comparison with the DNFCS 2019-2021, not all waves contain the same variables in the sets. The participant datasets provide information varying between 261 and 346 variables, which are way too many to be able to investigate. Many of them are about supplement intakes. There are only a few about characteristics and socio-demographics. These differed also between the three waves. Therefore, the variable list presented later will be relatively short. The consumption datasets contain between 94 and 193 variables, but only a couple on meat and meat substitute consumption are informative for this research. Thus, for both sets variable selection had to be done and the datasets had to be aligned. This resulted in the number of variables in the participant dataset being strongly reduced to 12. An overview of them is presented in Table 3.1. *Edu* for children between the age of 1-18, presents the highest education of the parents. *Migration_background* or any other ethnicity/origin-related variable was excluded from the wave 2007-2010. Therefore, this variable is only present for the 2012-2016 and 2019-2021 waves. *Hh_urb_5*, the level of urbanization presented with five categories, are not present for the 2012-2016 and 2019-2021 waves. However, for those a variable capturing this on a three-categoric scale is present. Therefore, *hh_urb_5* for wave 2007-2010 will be transformed later into three levels. In addition, a weighting factor on respond-level is included to generalize the results to the Dutch population. *W_demog_season_wk_wknd* contains weighting factors for demographic properties, the season and the day on which the recalls took place (weekday or weekend day). To derive the weighting factors for the wave 2007-2010, Dutch census data from 2008 is used as the reference population. For the 2012-2016 and 2019-2021 waves census data from 2014 and 2020 are used respectively. From the consumption dataset, only three variables are informative for further usage. These are presented in Table 3.2 and are mainly used to create the new variables.

Table 3.1: *Participation dataset*

Variable	Description	Type
p_id	Participant identification code	Nominal
age	Age of participant	Continuous
sex	Sex of participant: 1 - male, 2 - female	Nominal
BMI_cat	Evaluation of weight based on BMI: 1 - seriously underweight, 2 - underweight, 3 - normal weight, 4 - overweight, 5 - obesity, 999 - unknown (not for 1^{st} wave)	Nominal
edu	Highest completed education of participant: 0 - no education (not for 1^{st} wave), 1 - primary education, 2 - lower vocational education, 3 - advanced elementary education, 4 - intermediate vocational education, 5 - higher general secondary education, 6 - higher vocational education, 7 - university, 8 - other, 999 - unknown	Nominal
hh_size	Size of household: numerical + 999 - unknown (not for 3^{rd} wave)	Nominal
migration_background (not for 1^{st} wave)	Migration background of participant: 0 - Dutch, 1 - Western immigrant, 2 - Non-Western immigrant, 999 - unknown	Nominal
hh_urb_5 (not for 2^{nd} and 3^{rd} waves)	Level of urbanization of household location: 1 - very high, 2 - high, 3 - moderate, 4 - low, 5 - very low	Nominal
r_veg_meat	Vegetarian rule (no meat): 0 - false, 1 - true, 999 - unknown	Nominal
r_veg_meatfish	Vegetarian rule (no meat/fish): 0 - false, 1 - true, 999 - unknown	Nominal
r_vegan	Vegan rule: 0 - false, 1 - true, 999 - unknown	Nominal
w_demog_season_wk_wknd (1^{st} wave: w_demog_season_wk_wknd0)	Weighting factor for demographic properties, season (at 1^{st} recall day) and combination of both recall days (week or weekend)	Continuous

Table 3.2: *Consumption dataset*

Variable	Description	Type
p_id	Participant identification code	Nominal
group	Food or ingredient group: 7 - meat, meat products and substitutes	Nominal
subgroup	Food or ingredient sub-group. Subgroups of group 7 are: 1 - domestic mammals, 2 - poultry, 3 - game, 4 - processed meat, 5 - offals, 6 - meat substitutes (not for 1^{st} wave)	Nominal

3.3 Data cleaning, variable creation and transformation

Cleaning dataset

Furthermore, data cleaning, creating new variables and transformation are needed to be able to use the data properly. All datasets contained missing (NA's) and unknown ('999') values. As unknown values may be considered missing, respondents from whom no information is provided, are deleted from the datasets. The same holds for respondents who had missing values. In addition, all variables had to be investigated and checked for outliers and other notables. First, the 2007-2010 dataset is investigated. According to the study set-up women being pregnant and those lactating were not targeted for the questionnaires. However, the variables *pregnant* and *breastf* (breastfeeding) are included in this wave. It is checked whether respondents reported 'yes' for these variables. Only 'no' and 'unknown' for the variable *pregnant* were found. However, for *breastf* four respondents reported otherwise. These were deleted as something might have gone wrong with gathering the data. The variables *BMI_cat*, *edu* and *r_veg_meat* contained missing values for 2; 730 and 3 respondents respectively. *Hh_size* presented an unknown value for one respondent only. With regards to *edu*, two respondents had an unknown education level. As option 8 'other' was possible, it might not be the case that they have had no education at all. Therefore, these are excluded as well. *P_id* 2251, 2262, 2357, 2662 and 2816 reported both positive values for *r_veg_meat* (vegetarian rule, no meat) as well as for *r_veg_meatfish* (vegetarian rule, no meat/fish). This is seen as a reading mistake and therefore, these persons are considered 'dedicated' vegetarians (no meat/fish). As a result, the wave 2007-2010 contains data from 3077 respondents.

In the 2012-2016 dataset, the variables *BMI_cat* and *edu* contained missing values for 516 and 1472 respondents respectively. These were excluded. In addition, *BMI_cat* presented for

three respondents the value 'unknown', these were also excluded. For *migration_background* this was only the case of one respondent. *P_id* 3242 reported both positive values for *r_veg_meat* (vegetarian rule, no meat) and *r_veg_meatfish* (vegetarian rule, no meat/fish). This is again seen as a reading mistake and therefore, this person is considered a 'dedicated' vegetarian (no meat/fish). As 0 (no education) is not a possible answer option for the variable *edu* in the wave 2007-2010, but 8 (other) is, it might be assumed that those who have had no educational attainment are assigned to the latter. Therefore, respondents who had 0 in the 2012-2016 and 2019-2021 waves, are reassigned to category 8. This considered zero and three respondents respectively. As a result, the 2012-2016 wave contains data from 2321 respondents.

Furthermore, in the 2019-2021 dataset, the variables *edu* and *migration_background* contained missing values for 1251 and 23 respondents respectively. These are deleted. *P_id* 1748 reported both positive values for *r_veg_meat* (vegetarian rule, no meat) and *r_veg_meatfish* (vegetarian rule, no meat/fish). This is seen as a reading mistake and therefore, this person is considered a 'dedicated' vegetarian (no meat/fish). As a result, the 2019-2021 wave contains data from 2296 respondents.

Creating variables

Some new variables are created in all three datasets. The first newly created one is *wave*. It indicates to which of the three waves the respondents in the dataset belong, to the 2007-2010, 2012-2010 or 2019-2021 dataset. It is used for keeping the observations identifiable. In the 2007-2010 dataset only, the variable *migration_background* is added as this one was not present. A value of '888' is assigned to all respondents, indicating that information about this variable is not available. Furthermore, based on variables in the consumption dataset new variables are made. For all waves, the binary *flexitarian* is made and represents whether someone is a flexitarian (1) or not (0). A respondent is identified as a flexitarian if only once in those two non-consecutive days, an animal-based meat product was consumed. To determine whether someone consumes meat substitutes, the variables presented in Table 3.2 are used. Only respondents who consumed one product of group 7 (meat) in combination with one of the subgroups (except for the 6^{th} one), are given a positive value for *flexitarian*. Next, the variable *meatsub*, representing whether someone consumes meat substitutes (1) or not (0), is created. For the 2007-2010 wave, no information about meat substitutes is available. Therefore, it is not possible to create this variable for this wave. To determine whether someone used meat substitutes in the other waves, again the variables presented in Table 3.2 are used. Only respondents who consumed at least one product of group 7 (meat) in combination with one of the 6^{th} sub-

group, are given a positive value for *meatsub*. In addition, the variable *plantc* is created. This shows whether a respondent follows the vegetarian (*r_veg_meatfish*), pescatarian (*r_veg_meat*) or vegan (*r_vegan*) rule (1) or none of the three (0). The variables *flexitarian*, *meatsub* and *plantc* are added and assigned to the right respondent in the participation dataset based on *p_id*. Based on these new variables in the participation dataset, the binary variable *meatlover* is created. For all waves, it holds that when a respondent has a negative value for the variables *plantc* and *flexitarian*, he/she is assigned a positive value (1) for *meatlover*. Lastly, the binary variable *green* will present whether someone has a positive value for either *flexitarian* or *plantc* (1) or is a *meatlover* (0). It will show whether a (partly) plant-based food rule is followed. The flexitarian and plant-centered food rules are grouped together to create a larger group for performing one of the analyses only.

Transforming variables

Also, some variable transformation had to be performed. Firstly, the variable *sex* has been recoded to male (0) and female (1). Secondly, the variable *edu* has eight categories. A clearer picture can be drawn with fewer options. Therefore, it is recategorized into four categories, namely: 'low', 'middle', 'high' and 'other'. 'Low' captures primary school, lower vocational education and advanced elementary education. 'Middle' captures intermediate vocational education and higher general secondary education. 'High' captures higher vocational education and university. 'Other' is the same category as the already existing one. Thirdly, only in the wave 2007-2010 the degree of urbanization is presented by the variable *hh_urb_5* with five levels instead of three levels. Therefore, *hh_urb_5* is recategorized to be in line with the other waves. Level 1 indicates extremely/strongly urbanized (>=1500 addresses/km^2), level 2 moderately urbanized (1000-1500 addresses/km^2) and level 3 hardly/not urbanized (<1000 addresses/km^2). Hence, the old categories 'very high' and 'high' are assigned to level 1 and 'low' and 'very low' to level 3. Level 2 stays the same. Fourthly, the variable *w_demog_season_wk_wknd0* of the 2007-2010 wave is renamed to *w_demog_season_wk_wknd* as it captured the same information as the latter variable in the 2012-2016 and 2019-2021 waves. Lastly, all variables are transformed into factors, except for *age* and *w_demog_season_wk_wknd*, as those two contain continuous values.

3.4 Descriptive statistics

3.4.1 Per wave

The descriptive statistics of the continuous and categorical variables are presented in Table 3.3 and 3.4 respectively. It becomes clear that, in contrast to the minimal age, the maximum age in all three waves differs. It is the highest in wave 2019-2021. As well as the mean, laying more than 11 years above the means of the other waves. The sample weighted variable *w_demog_season_wk_wknd* presents values between 0.17 and 12.63, these are used to assign different weights to respondents who are underrepresented in the dataset compared to the population sample. The means of all three waves are comparable for this variable.

In Table 3.4, it is presented that the descriptive statistics of the categorical variables for all three waves are comparable. Namely, for all waves it holds that the variable sex is evenly divided. In addition, *BMI_cat* presents that most respondents belong to the category of 'normal weight', followed by 'overweight' and 'obesity'. Furthermore, with a value above 40%, most respondents have had a middle education, shown by *edu*. Household sizes of two and four people are the most popular, as presented with *hh_size*. *Migration_background* shows that most respondents have a Dutch background. The level of urbanization, presented with *hh_urb_3*, indicates that around half of the respondents live in strongly to extremely urbanized areas.

Table 3.3: *Descriptive statistics continuous variables per wave*

Variable	2007-2010				2012-2016				2019-2021			
Continuous	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**
age	33.51	17.7	12	69	33.47	18.21	12	71	45.01	22.02	12	79
w_demog_season_wk_wknd	1.00	0.34	0.30	2.98	1.49	0.99	0.21	5.56	1.35	1.15	0.17	12.63

Table 3.4: *Descriptive statistics categorical variables per wave*

Variable	2007-2010	2012-2016	2019-2021
Categorical	**Proportion**	**Proportion**	**Proportion**
sex			
male	50.0%	50.8%	49.6%
female	50.0%	49.2%	50.4%
BMI_cat			
Seriously underweight	0.9%	0.6%	1.2%
Underweight	3.5%	4.1%	4.0%
Normal weight	55.1%	53.6%	48.3%
Overweight	27.7%	27.8%	30.3%
Obesity	12.9%	14.0%	16.2%
edu			
Low	35.9%	27.0%	28.1%
Middle	46.7%	46.8%	42.0%
High	17.4%	25.9%	29.2%
Other	0.0%	0.3%	0.7%
hh_size			
1	15.1%	12.3%	17.8%
2	27.7%	28.3%	35.5%
3	17.1%	18.5%	12.1%
4	25.9%	27.7%	22.1%
5	10.3%	9.9%	9.7%
6	2.9%	2.5%	2.0%
7	0.8%	0.4%	0.6%
8	0.2%	0.3%	0.1%
9	0.03%	0.04%	0.0%
10	0.0%	0.04%	0.0%
11	0.0%	0.04%	0.0%
migration_background			
Dutch	N.A.	91.9%	88.0%
Western immigrant		2.6%	5.8%
Non-Western immigrant		5.5%	6.2%
hh_urb_3			
Extremely/strongly urbanized	46.2%	48.1%	52.8%
Moderately urbanized	22.2%	19.6%	17.2%
Hardly/not urbanized	31.6%	32.3%	30.0%
meatlover			
True	89.0%	89.5%	84.5%
False	11.0%	10.5%	15.5%
plantc			
True	1.9%	3.0%	5.1%
False	98.9%	97.0%	94.9%
flexitarian			
True	9.1%	7.5%	10.4%
False	90.9%	92.5%	89.6%
meatsub			
True	N.A.	3.7%	9.1%
False		96.3%	90.9%
Number of observations N	3077	2321	2296

With regard to the food groups, Table 3.4 presents, that around 85 to 90% of the respondents are categorized as *meatlover*. The prevalence of respondents following the plant-centered food rule, presented with *plantc*, increased over time. Namely, 1.9% of the respondents reported a positive value in the 2007-2010 wave, in comparison to 3.0% and 5.1% in the 2012-2016 and 2019-2021 waves respectively. A slightly different result is found for flexitarians. From wave 2007-2010 to

wave 2012-2016 the prevalence of flexitarians decreased from 9.1% to 7.5%. It increased in wave 2019-2021 to more than 10%. The prevalence of consumers using meat substitutes increased between wave 2012-2016 and wave 2019-2021 from 3.7% to 9.1%. The same conclusions can be drawn from the distributions of all food groups in Figure 3.1 till 3.4. Figure 3.1 shows that the number of meat-lovers decreased over time but still dominates the other food groups in all waves. Figure 3.2 and 3.4 both present slight increases in the number of participants following the plant-centered rule and consuming meat substitutes respectively. The dip and subsequent increase in respondents following a flexitarian food rule also become clear from Figure 3.3. From this, it can be stated that based on the descriptive statistics, the interest in (partly) plant-based food rules increased over time. As well as the interest in the consumption of meat substitutes.

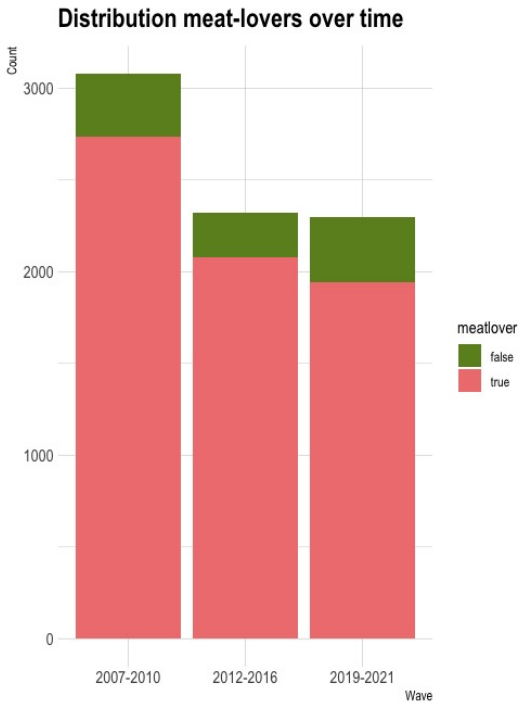

Figure 3.1: *Distribution of meat-lovers*

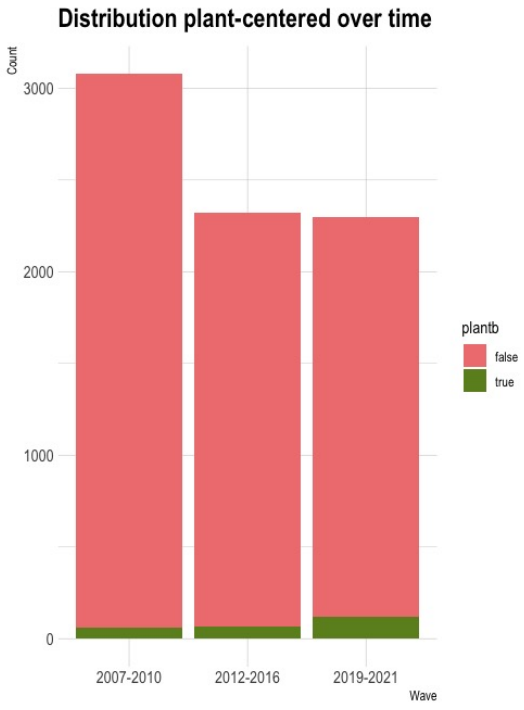

Figure 3.2: *Distribution of plant-centered*

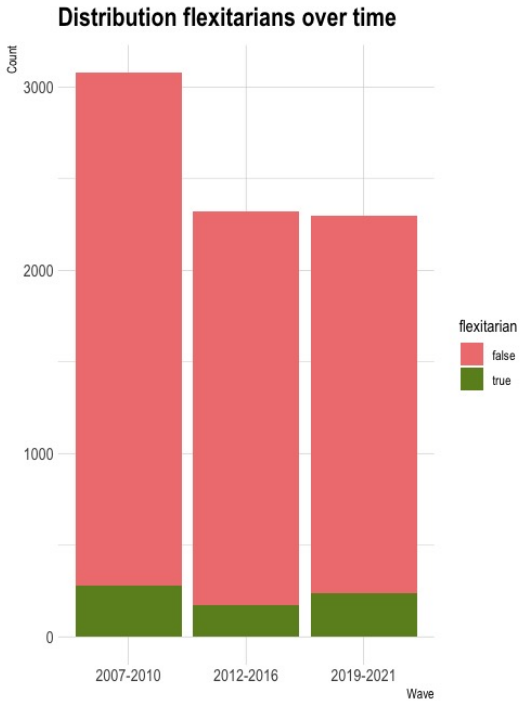

Figure 3.3: *Distribution of flexitarians*

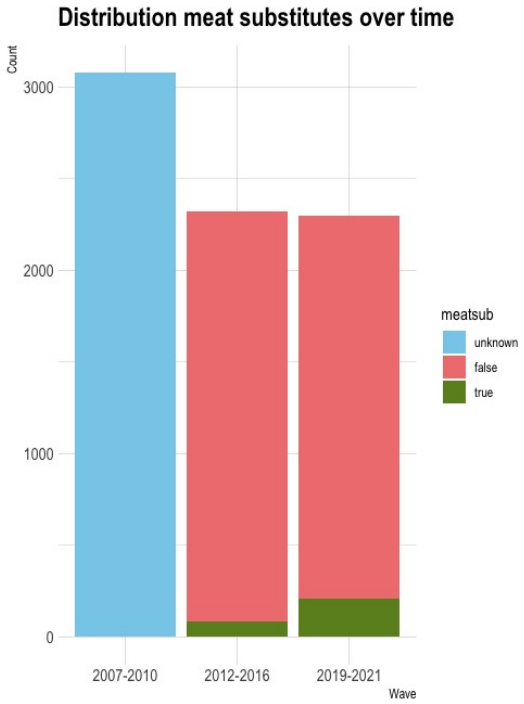

Figure 3.4: *Distribution of meat substitutes*

3.4.2 Per food rule

In addition, the descriptive statistics of the respondents are studied by the different food rules. The datasets are merged to do this. Thus, each food rule is now represented by its followers of all waves. The results of the continuous and categorical variables are presented in Table 3.6 and 3.7 respectively. From looking at Table 3.6, it becomes clear that the mean of age for the meat substitute rule is the highest. The means of the other three food rules are relatively comparable. *W_demog_season_wk_wknd* shows a comparable result. The mean of the meat substitute food rule is just above those of the other food rules.

With regard to *sex* in Table 3.7, most meat-lover respondents are male, in contrast to the plant-centered, flexitarians and those consuming meat substitutes. *BMI_cat* shows that the latter three overall have a lower BMI level. They are less obese and overweighted than the meat-lovers. In addition, *edu* presents that the plant-centered, flexitarians and those consuming meat substitutes are often higher educated than the meat-lovers. Also, in contrast to the meat-lovers, for the three other food rules, a household size (*hh_size*) of three people is less common than of one person. The other household sizes are comparable. *Migration_background* presents for all food rules that a Dutch background is represented the most. The meat-lovers and flexitarians have comparable proportions, both around 50% Dutch and 40% unknown. The level of

urbanization, shown with *hh_urb_3*, presents that the proportion of respondents living in extremely/strongly urbanized areas is slightly higher for the plant-centered, flexitarian and meat substitute rules than for the meat-lover rule. This exploratory analysis already gives some first insights. The characteristics will be further researched in the result section.

Table 3.6: *Descriptive statistics continuous variables per food rule*

Variable	meat-lovers				plant-centered				flexitarians				meat substitutes			
Continuous	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**
age	36.95	20.01	12	79	36.97	19.26	12	73	36.99	19.82	12	79	41.01	20.43	12	78
w_demog_season_wk_wknd	1.25	0.88	0.17	12.63	1.32	0.94	0.22	7.03	1.25	0.86	0.18	6.97	1.60	1.37	0.22	12.63

Table 3.7: *Descriptive statistics categorical variables per food rule*

Dietary pattern	Meatlover	Plantc	Flexitarian	Meatsub
Variable				
sex				
Male	52.4%	27.4%	36.2%	37.8%
Female	47.6%	72.6%	63.8%	62.2%
BMI_cat				
Seriously underweight	0.8%	1.9%	0.9%	1.4%
Underweight	3.5%	9.3%	5.5%	4.1%
Normal weight	51.9%	61.1%	56.1%	62.5%
Overweight	29.2%	20.0%	24.9%	22.6%
Obesity	14.6%	7.8%	12.7%	9.5%
edu				
Low	31.6%	15.9%	29.5%	15.5%
Middle	45.9%	44.4%	40.3%	38.5%
High	22.2%	39.6%	30.0%	45.9%
Other	0.3%	0.0%	0.1%	0.0%
hh_size				
1	14.1%	27.4%	20.2%	27.4%
2	30.3%	27.0%	30.7%	27.4%
3	16.4%	11.9%	14.7%	14.2%
4	25.7%	21.1%	23.3%	18.6%
5	10.2%	8.5%	8.1%	10.1%
6	2.5%	3.3%	2.4%	2.0%
7	0.7%	0.4%	0.4%	0.3%
8	0.2%	0.4%	0.1%	0.0%
9	0.03%	0.0%	0.0%	0.0%
10	0.01%	0.0%	0.0%	0.0%
11	0.01%	0.0%	0.0%	0.0%
migration_background				
Dutch	53.9%	63.3%	51.7%	83.8%
Western immigrant	2.4%	4.8%	2.7%	6.4%
Non-Western immigrant	3.2%	8.1%	5.0%	9.8%
Unknown	40.5%	23.7%	40.5%	0.0%
hh_urb_3				
Extremely/strongly urbanized	47.9%	58.1%	54.0%	63.5%
Moderately urbanized	20.2%	14.4%	18.9%	12.5%
Hardly/not urbanized	31.9%	27.4%	27.1%	24.0%
Num. obs. N	6755	245	694	296

Notes: $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Chapter 4

Methodology

The methods that will be used in this research are logistic regressions and decision trees. The first one will function as the benchmark model. Both will investigate the most important features for predicting what kind of food rule someone follows and whether these have changed over the last decade. For each food rule, a logistic regression model will be developed. To make a comparison over time, models for two waves are built. Resulting in a total of eight logistic regression models. This method is suitable for the study as it will present the probability that someone follows a specific food rule. The probability that someone follows, for instance, the flexitarian food rule (outcome = true). Decision trees will be built to gain even more insights into what these features are. In addition, potential segments can be determined. A decision tree can be visualized, which will improve interpretability and is, therefore, a good match. It is a supervised learning technique as the data are labelled. As in this case, it is priorly known to which food rule, class, a respondent belongs. Furthermore, a classification task and not a regression task is performed, as the respondents are classified based on their characteristics. It is priory known to which class a respondent belongs. Thus, as the answer to the question *"Do you follow food rule X?"* is yes or no, a classification task is suitable to perform. In addition, no interaction terms need to be researched as a machine learning model investigates these themselves. The first decision tree will be on being a meat-lover vs. following a (partly) plant-based food rule. One with that dataset of wave 2007-2010 and one with the dataset of wave 2019-2021 will be made to determine whether the results have changed over the last decade. The *meatsub* food rule is excluded from this analysis, as meat substitutes can be consumed by either meat-lovers, as well as flexitarians and the plant-centered. It is not exclusive and therefore, this food rule gets its own decision tree. The second decision tree will therefore present whether someone consumes meat substitutes vs. does not consume them. As the wave 2007-2010 has no information available on meat substitutes, for this a comparison of 5 years will be made

(comparing wave 2012-2016 with wave 2019-2021). An overview of which dataset will be used for which analysis is presented visually in the flowchart in Figure 4.1. Furthermore, the (predictive) performance of the models will be investigated. All analyses will be performed in the RStudio version 2023.03.0+386, named 'Cherry Blossom'. The logistic regression method will be described first, followed by the decision tree description and different measurements for the models' performances.

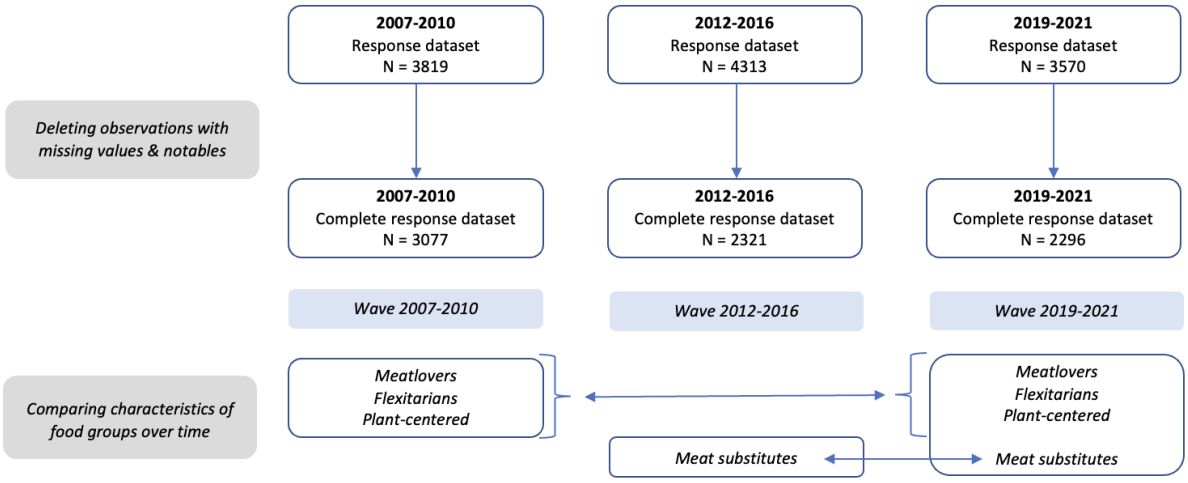

Figure 4.1: *Flowchart of dataset usage*

4.1 Logistic Regression

A logistic regression is a machine learning algorithm that is able to map an input to a prediction of an event occurring. The main use is, in classification tasks, where the data needs to be classified to a binary outcome. The output provides a value between 0 and 1 indicating the probability of an event occurring given the data. During classification, the continuous data is converted into a binary output by checking if the probability is higher or lower than a given threshold (Pant, 2019). The basis on which the regression is built is the sigmoid function or also called the logistic curve. This curve has an S-like shape and is described by the following formula:

$$f(x) = \frac{1}{1 + e^{-x}}$$

In the context of logistic regression, this function is reinterpreted and rewritten to include a term for x that allows us to reshape the sigmoid curve. This ability provides the possibility to fit the curve to the data. The adjusted sigmoid curve that can be fitted, follows the function:

$$f(x) = \frac{1}{1 = e^{-(B_0 + B_1 X + B_2 X^2 ...) + B_n}}$$

The goal is to find the optimum values for B_n that fit the data best. Before the data can be fit to the sigmoid curve, a cost function is needed to provide the goodness-of-fit for a given set of parameters B_n and the data. This cost function can be optimized to find the optimum parameters B_n that describe the data best. Finding this optimum can be done by, for instance, using the gradient descent optimization algorithm. When the optimal values for B_n are found, the sigmoid curve will describe the data as closely as possible. It can then be used to probe the likelihood of existing data occurring or to make predictions on hypothetical data.

The logistic regression is different from the linear regression model as it makes use of other assumptions. First, the Bernoulli distribution presents the conditional distribution as the outcome variable is binary. Second, the outcome is a predicted probability between 0 and 1 and not a predicted outcome itself. Beforehand, all numerical variables are scaled as they are measured in different units, thus the coefficients are standardized. Otherwise, they would present a distorted picture.

4.2 Decision Tree

A decision tree is used to calculate the probability that a respondent belongs to a specific class (Quinlan, 1986). Potential segments of respondents can be identified. A decision tree is a visual representation of possible solutions to a decision based on certain conditions. The conditions are presented by nodes in the tree. It starts with the root node and splits based on a certain condition about a single variable into possible outcomes. For instance, the first variable could be age and split on the condition whether someone is 30 years or older. The first node presents the best predictor of all independent variables included in the analysis. A top-down approach is used. The outline format is like a tree with leaves. The branches present the independent predictor variables and the leaves are the outcome variable. After the first split other, internal, nodes follow with corresponding conditions. The lower a variable is placed in the tree, the less important it is. This branches off into the end nodes where the outcome is shown. A categorical outcome with a probability that someone belongs to a specific class.

Before the data are split into the training and test sets for the decision trees, the Random Over-Sampling Examples (ROSE) package (Lunardon, Menardi, & Torelli, 2014), can be used to tackle the presence of the imbalanced classes. As respondents belonging to some classes are more 'rare' events than those belonging to another class, the data are imbalanced. This binary classification problem can be dealt with, with the use of resampling techniques. With ROSE, balanced samples can be created. By means of oversampling, the number of observations from the minority class is replicated to generate more observations. The dataset is randomly split

into a training and test set with an 80/20 ratio. Furthermore, the sets will be checked whether they are shuffled well enough. This is to double-check the resampling procedure. The training set is used to build the tree and consequently trained to make the predictions for the test set. The Gini Index is used to determine the place of a variable in the tree, it follows the formula:

$$G = 1 - \sum_{i=1}^{C} (p_i)^2.$$

Indicating to subtract the summons of all squared possibilities from 1, the independent variable with the lowest score is searched for. The larger the drop in the Gini Index, in the node impurity, the more important a variable is and the higher up in the tree it will be used to split the data on. Therefore, the variable with the lowest score, the most optimal one, will be the root node. This procedure is repeated with all remaining independent variables to determine the sequence of the nodes in the tree. It is done until no more splits can be made and consequently, the tree is created.

However, it can become very large very quickly. In addition, there is a risk of overfitting. A smaller, less complex, tree has less variance but more bias. Thus, a trade-off between the size and the error rate must be made. To prevent building an immense tree, the 'maxdept' argument can be pre-pruned (Therneau & Atkinson, 2022). This set the maximum depth of the tree without counting the root node. In that way, an optimal decision tree can be created. This tree will be used to make the predictions for the test set. In conclusion, the most important variables to predict whether someone follows a specific food rule, are the ones on which the first couple of splits are made. As a tree can be visualized, these are found easily.

4.3 Model performance measures

The model performance measures that will be used for the benchmark model are the Log Likelihood and the Akaike Information Criterion (AIC). The Log Likelihood measures a model's goodness-of-fit. It presents the likelihood that the model determined the outcome variable correctly. The highest Log Likelihood score is preferred. It is used in the AIC and investigates how well the model fits the data. The AIC adds a penalization for every extra added variable to the model to avoid the model from overfitting. The highest AIC score with the lowest number of variables involved is preferred. The AIC formula follows (Sakamoto, Ishiguro, & Kitagawa, 1986):

$$AIC = -2log(\mathcal{L}) + 2K$$

When $log(\mathcal{L})$ is the Log Likelihood and K is the number of parameters of the model (variables). Also stated as adding 2 times the number of parameters in the model to minus 2 times the maximum log-likelihood of the model.

One of the performance measures that will be used for the decision trees is the balanced accuracy. By using the balanced accuracy, it is checked how well the models perform on the test sets. The balanced accuracy is used as the food rules have imbalanced classes. Therefore, an equal weight is given to both classes. The formula for the balanced accuracy follows (Kuhn et al., 2020):

$$Balanced\ Accuracy = \frac{\frac{TP}{(TP+FN)} + \frac{TN}{(TN+FP)}}{2}$$

TP and TN are the true positives and true negatives respectively. FN and FP are the false negatives and false positives respectively. Also, indicated as the Specificity added to the Sensitivity divided by two. The outcome is between 0 and 1. The higher the value the better the model performs, with 1 indicating a perfect classification. The False Positive Rate (FPR) gives the proportion of the incorrectly classified negatives divided by the total negatives. Presented with the formula (Fawcett, 2006):

$$False\ Positive\ Rate\ = \frac{FP}{(TN + FP)}\ = 1\ - Specificity$$

Furthermore, the Receiver Operator Characteristics (ROC) curve plots the Sensitivity against the FPR for different values of the threshold. It tries to separate the signal and the noise to show how well the model performs at different classification thresholds. The closer the curve is to the upper left corner the more precise. It means a high true positive rate and a low false positive rate.

The Area Under the Curve (AUC) measures the model's performance for binary classifiers by investigating how well the model can distinguish between the positive and negative classes. It presents the probability that the model will estimate a random observation belonging to the positive class higher than a random observation belonging to the negative class (Fawcett, 2006). The AUC has a score between 0 and 1, with 0.5 being completely random. The higher the value, the better the model is in distinguishing the positive and the negative class. A score of a minimum of 0.70 is considered acceptable (Mandrekar, 2010).

Chapter 5

Results

In this section, the results will be presented and the hypotheses discussed. First, the model will be identified. The binary outcome variable y presents whether, for a specific food rule, the participant follows this rule (true) or not (false). This can be stated for all four food rules separately, namely the meat-lovers, plant-centered, flexitarians and those consuming meat substitutes. Such that:

$$y = \begin{cases} 0 & \text{if } \textit{food rule} = \text{false} \\ 1 & \text{if } \textit{food rule} = \text{true} \end{cases}$$

The *meatlover*, *plantc* and *flexitarian* food rule models are created with the following set of parameters $\{age, sex, BMI\_cat, hh\_size, hh\_urb\_3, edu\}$. As the variable *migration_background* is only present in wave 2012-2016 and wave 2019-2021, it could only be included for the food rule *meatsub* models. Therefore, this model is created with the following set of parameters $\{age, sex, BMI\_cat, hh\_size, hh\_urb\_3, edu, migration\_background\}$. First, statistical analyses will be done to determine whether the variables of the food rules differ significantly from each other. Second, the benchmark model will be discussed. This is a logistic regression model, investigating the probability that someone follows a specific food rule. Next, the outcomes of the decision trees and the models' performances will be presented and interpreted. Decision trees on being a meat-lover vs. following a (partly) plant-based food rule are presented first. The *meatsub* food rule is excluded from this analysis, as meat substitutes can be consumed by either meat-lovers, as well as flexitarians and those following the plant-centered food rule. Thus, the *meatsub* rule gets its own decision trees. Consequently, they will be on whether someone consumes meat substitutes vs. does not consume them. Lastly, the hypotheses will be discussed.

5.1 Statistical analysis

The four food rules are studied with statistical analyses per wave whether they are comparable or different based on their characteristics. T-tests and Pearson's chi-squared tests are used for testing the continuous and categorical variables respectively. All food rules are researched whether the characteristics differ for those who do follow a specific rule ('engaged') in contrast to those who do not follow that specific rule ('non-engaged'). The engaged group is considered to have a positive outcome (true) for y. The non-engaged group is considered to have a negative outcome (false) for y. Only the coefficients which are statistically significant at α level of 1% are investigated. Differences are expected to be found, as described in the literature review section.

First, the only continuous variable *age* is investigated. It is determined whether the age of the respondents differs for the engaged and non-engaged groups of a specific food rule. Therefore, the null hypothesis follows:

H0: There is no difference between the mean of age when comparing the engaged and the non-engaged group.

The results are presented in Table 5.1 with the t-statistic. All coefficients are statistically insignificant at an α level of 1%. For none of the food rules in none of the waves does the mean age significantly differ between the engaged and non-engaged groups. Concluding, there is not enough evidence for one of the food rules to reject the null hypothesis.

Second, the categorical variables are investigated with Pearson's chi-squared test. It is researched whether they differ between the engaged and non-engaged groups of a specific food rule. The null hypothesis follows:

H0: There is no association between the categorical variables in the sample; they are independent.

The results are presented in Table 5.2 with the X^2-statistic. A couple of the results are statistically significant at an α level of 1%, especially for wave 2019-2021. For those variables, the null hypothesis can be rejected and thus, an association is found. It is expected that, in the models that will be developed in the following subsections, those variables will be among the most important ones.

For *meatlover* wave 2007-2010, only *sex* is statistically significant at an α level of 1%. For wave 2019-2021, this is the case for all variables except for *hh_urb_3*. Based on this, the *meatlover*

wave 2019-2021 model built later is expected to present significant results.

The outcome of the *plantc* group is comparable. Only for wave 2007-2010, in addition to *meatlover*, the null hypothesis can also be rejected for the *BMI_cat* variable.

The *flexitarian* waves present less statistically significant coefficients at an α level of 1%. In wave 2007-2010 *sex* and in wave 2019-2021 *sex* and *edu* are the only significant results.

For *meatsub* wave 2007-2010, only *edu* is statistically significant at an α level of 1%. For wave 2019-2021, this is the case for all variables except for *migration_background*. Based on this, the later built *meatsub* wave 2019-2021 model is expected to present significant results. The characteristics of all food rules will be further researched with the benchmark model and the decision trees.

Table 5.1: *Results of the t-tests*

	Meatlover vs not		Plantc vs not		Flexitarian vs not		Meatsub vs not	
	wave A	wave C	wave A	wave C	wave A	wave C	wave B	wave C
age	1.18	1.87	0.32	-2.08^{*}	-1.44	-0.67	-0.53	-0.16

Notes: $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$. Wave A refers to wave 2007-2010. Wave B to wave 2012-2016. Wave C to wave 2019-2021.

Table 5.2: *Results of the chi-squared tests*

	Meatlover vs not		Plantc vs not		Flexitarian vs not		Meatsub vs not	
	wave A	wave C	wave A	wave C	wave A	wave C	wave B	wave C
sex	**28.53*****	**51.99*****	**13.40*****	**26.57*****	**15.01*****	**23.58*****	3.58	**14.33*****
BMI_cat	8.97	**15.91*****	**17.19****	**15.39****	2.93	5.21	8.51	**17.18****
edu	3.90	**26.98*****	6.55*	**13.54****	1.1	**15.19*****	**19.42*****	**37.82*****
hh_size	18.38*	**24.42*****	11.28	**25.79****	11.44	8.79	12.63	**26.33*****
hh_urb_3	2.20	6.98*	0.55	4.67	1.94	3.45	8.41*	**10.99*****
migration_background							2.72	7.70*

Notes: $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$. Wave A refers to wave 2007-2010. Wave B to wave 2012-2016. Wave C to wave 2019-2021.

5.2 Benchmark model

The results of the standardized logistic regression are presented in Table 5.3. Only the coefficients which are statistically significant at α level of 1% are investigated. For the meat-lovers in wave 2007-2010, it is shown that the variables *sex* and *hh_size* are the most important. When someone is a female the log-odds of being a meat-lover decrease by 0.32, ceteris paribus. When the size of someone's household increases by one person, the log-odds of being a meat-lover increase by 0.26, ceteris paribus. For the meat-lovers in wave 2019-2021, it is shown that the variables *sex*, *BMI_cat* and *edu* are the most important. When someone is a female the log-odds of being a meat-lover decrease by 0.46, ceteris paribus. This demonstrates a larger magnitude of increase compared to the wave A model. Thus, *sex* has nowadays a more determining role in identifying meat-lovers than it used to have. In addition, when someone has a BMI categorized one level higher, the log-odds of being a meat-lover increase by 0.20, ceteris paribus. When someone has an educational attainment of one level higher, the log-odds of being a meat-lover decrease by 0.31, ceteris paribus. Thus, over time, *hh_size* is no longer an important predictor and is replaced by *BMI_cat* and *edu*. In conclusion, being a male has a positive influence on the log-odds of being a meat-lover, ceteris paribus. As well as, having a high BMI level, ceteris paribus, and a low education level, ceteris paribus.

With regard to the plant-centered rule in wave 2007-2010, it is presented that the variables *sex* and *BMI_cat* are the most important. Being a female increases the log-odds of following a plant-centered food rule by 0.52, ceteris paribus. When someone has a BMI categorized one level higher, the log-odds of following a plant-centered food rule decrease by 0.48, ceteris paribus. With regard to the plant-centered rule in wave 2019-2021, it is presented that *edu* is one of the most important variables, besides the variables *sex* and *BMI_cat*. When someone has an education obtained of one level higher, the log-odds of following a plant-centered food rule increase by 0.35, ceteris paribus. Someone's education became an important predictor over the decade. It might be that people were used to following this food rule due to their taste or because they found it sad for the animals. Over time, it became more popular to eat plant-based. It might be that people were informed better about the health consequences of consuming meat and about the effects it has on climate change. For instance, also children at schools. That higher educated people might care more about their health, the environmental impact and also have the means to. In addition, being a female increases the log-odds of following a plant-centered food rule by 0.53, ceteris paribus. This demonstrates a comparable magnitude to the wave A model. It might be that eating meat is still associated with masculinity. Consequently, males might be less likely than females to shift their consumption patterns. When someone has

a BMI categorized one level higher, the log-odds of following a plant-centered food rule decrease by 0.33, ceteris paribus. This demonstrates a smaller magnitude of decrease compared to the wave A model. Thus, *BMI_cat* has nowadays a less determining role in identifying plant-centered food rule followers than it used to have. In conclusion, being a female has a positive influence on the log-odds of following a plant-centered food rule, ceteris paribus. As well as, having a low BMI level, ceteris paribus, and a high education level, ceteris paribus.

For the flexitarians in wave 2007-2010, it is shown that the variables *sex* and *hh_size* are the most important. When someone is a female the log-odds of being a flexitarian increase by 0.25, ceteris paribus. When the size of someone's household increases by one person, the log-odds of being a flexitarian decrease by 0.25, ceteris paribus. For the flexitarians in wave 2019-2021, it is shown that the variables *sex* and *edu* are the most important. When someone is a female the log-odds of being a flexitarian increase by 0.36, ceteris paribus. This demonstrates a larger magnitude of increase compared to the wave A model. Thus, *sex* has nowadays a more determining role in identifying flexitarians than it used to have. It might be that males are even less likely to lower their meat consumption due to the role of social expectations about masculinity. When someone has an education obtained of one level higher, the log-odds of being a flexitarian increase by 0.25, ceteris paribus. It might be for the same reasons as for the plant-centered food rule that education became an important predictor over time. Reducing one's meat intake and being able to replace it with more greens and legumes, might be assigned to the higher educated. They might have been made more aware over time of the environmental impact of meat consumption and have the means to buy more (expensive) healthier foods. Thus, over time, *hh_size* is replaced by *edu*. In conclusion, being a female has a positive influence on the log-odds of being a flexitarian, ceteris paribus. As well as, having a high education level, ceteris paribus.

With regard to consuming meat substitutes in wave 2012-2016, it is presented that the variable *edu* is important. When someone has an educational attainment of one level higher, the log-odds of consuming meat substitutes increase by 0.36, ceteris paribus. With regard to consuming meat substitutes in wave 2019-2021, it is presented that besides the variable *edu*, also *sex* and *BMI_cat* are important. When someone has an education obtained of one level higher, the log-odds of consuming meat substitutes increase by 0.40, ceteris paribus. This demonstrates a slightly larger magnitude of increase compared to the wave B model. Thus, *edu* has nowadays a slightly more determining role in identifying meat substitute consumers than it used to have. It might be straightforward that higher educated are more likely to buy meat substitutes. Meat substitutes are not that much cheaper than animal-based meats. Some kinds, such as burgers,

are sometimes even more expensive. As overall the higher educated might earn more than the lower educated, they can afford it more easily. When someone is a female the log-odds of consuming meat substitutes increase by 0.32, ceteris paribus. When someone has a BMI categorized one level higher, the log-odds of consuming meat substitutes decrease by 0.28, ceteris paribus. In conclusion, being a female has a positive influence on the log-odds of consuming meat substitutes, ceteris paribus. As well as, having a low BMI level, ceteris paribus, and a high education level, ceteris paribus.

Table 5.3: *Standardized Logistic Regression Coefficients*

	Meatlover		Plantc		Flexitarian		Meatsub	
	wave A	wave C	wave A	wave C	wave A	wave C	wave B	wave C
(Intercept)	**2.16*****	**1.83*****	**−4.14*****	**−3.07*****	**−2.35*****	**−2.25*****	**−3.47*****	**−2.46*****
	(0.06)	(0.06)	(0.16)	(0.11)	(0.07)	(0.07)	(0.13)	(0.08)
age	0.14*	0.17*	0.06	−0.22	−0.19*	−0.12	−0.06	−0.07
	(0.07)	(0.08)	(0.15)	(0.12)	(0.08)	(0.09)	(0.14)	(0.10)
sex	**−0.32*****	**−0.46*****	**0.52*****	**0.53*****	**0.25*****	**0.36*****	0.26*	**0.32*****
	(0.06)	(0.06)	(0.14)	(0.10)	(0.06)	(0.07)	(0.11)	(0.08)
BMI_cat	0.14*	**0.20****	**−0.48*****	**−0.33****	−0.06	−0.12	−0.31*	**−0.28*****
	(0.06)	(0.07)	(0.14)	(0.10)	(0.07)	(0.08)	(0.13)	(0.08)
hh_urb_3	0.04	0.13*	−0.00	−0.16	−0.05	−0.10	−0.18	−0.17*
	(0.06)	(0.06)	(0.13)	(0.10)	(0.06)	(0.07)	(0.12)	(0.08)
hh_size	**0.26*****	0.17*	−0.30*	−0.19	**−0.25*****	−0.14	−0.34*	−0.15
	(0.07)	(0.08)	(0.15)	(0.12)	(0.07)	(0.09)	(0.13)	(0.09)
edu	−0.05	**−0.31*****	0.24	**0.35*****	−0.01	**0.25*****	**0.36****	**0.40*****
	(0.06)	(0.06)	(0.13)	(0.10)	(0.06)	(0.07)	(0.12)	(0.08)
migration_background							0.09	0.13*
							(0.09)	(0.06)
AIC	2098.57	1884.10	598.96	943.59	1863.16	1496.18	717.05	1342.24
Log Likelihood	−1042.29	−935.05	−292.48	−464.79	−924.58	−741.09	−350.52	−663.12
Num. obs. N	3077	2296	3077	2296	3077	2296	2321	2296

Notes: $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$. Wave A refers to wave 2007-2010. Wave B to wave 2012-2016. Wave C to wave 2019-2021. A *sex* of 0 is male and of 1 is female. A *BMI_cat* value is between 1 (underweight) - 5 (obesity). The variable *edu* presents: 1 = low, 2 = middle, 3 = high. *Hh_urb_3* has a value of 1 (extremely/strongly), 2 (moderately) or 3 (hardly/not) urbanized. For *migration_background* holds: 0 = Dutch, 1 = Western immigrant and 2 = nonwestern immigrant.

The models' performance measures are presented at the bottom of Table 5.3. Following from the AIC, the *plantc* model wave 2007-2010 and *meatsub* model wave 2012-2016 are performing relatively the best. As the *meatlover*, *plantc* and *flexitarian* use the same number of variables, but the AIC score for *plantc* 2007-2010 is the lowest. In comparison, both models of the *meatsub*

food rule use one extra variable, therefore *meatsub* 2012-2016 performs best out of these two. From investigating the Log Likelihood, both models are also relatively best at predicting the observed data. The most important features for predicting to which food rule a person belongs will be further researched and discussed with decision trees in the following subsection.

5.3 Decision Trees

There are imbalanced classes in the datasets as described in the Methodology section and presented with the descriptive statistics in Table 3.4. This issue must be tackled, otherwise, the machine learning models will be very bad at predicting the minority class. What might result in not applicable outcomes ('NA') for the true positives. Hence, resampling is applied. With oversampling, observations from the minority classes are copied and added to the dataset, also referred to as sampling with replacement. However, one should be careful. The first tree designed to make balanced classes used a relatively high ratio (around 1:20). The predictive performance on the test set was so good, that resulted in balanced accuracy scores of around 99.0%. Additionally, the decision tree was still huge and unreadable even after post-pruning. These were signals of overfitting. The model was too good that it became bad. It might be because of something like, for instance, there are only a few flexitarians in the dataset. If these happen to be almost all women of a younger age, then due to resampling with replacement, a lot of other younger aged women are copied and added to the dataset. This might result in a very good predictive performance model on the test set. However, it gives a distorted picture and thus, a different approach had to be studied. First, the decision trees indicating whether someone is a meat-lover or not will be described. Second, decision trees presenting whether someone consumes meat substitutes or not are shown and discussed. A minimum of 5% of the respondents is used as a requirement for describing potential segments and most important features. All trees are built thrice, fully identical. Only another value for the function 'set.seed' is used, indicating that different observations are randomly put in the training and test set. This is done to create more robust results and determine whether the models' performances are not just due to one-time luck. Therefore, the performance measures are presented with an interval and hence, the best and worst scores are shown.

5.3.1 Meat-lover vs. a (partly) plant-based food rule

First of all, the variable *green* comes at hand here for representing the (partly) plant-based rules. As *meatlover*, *plantc* and *flexitarian* are exclusive, it is possible to research whether someone is a meat-lover (*meatlover*) or follows a (partly) plant-based food rule (*green*). Based on the

presented descriptives in Table 3.4, the proportion of respondents belonging to *green* is 11.0% in wave 2007-2010 (1.9% *plantc* + 9.1% *flexitarian*) and 15.5% in wave 2019-2021 (5.1% *plantc* + 10.4% *flexitarian*). Thus, it makes the group already a bit larger by combining them. In addition, it becomes clear that the prevalence of those belonging to the *meatlover* food rule is 89.0% and 84.5% in wave 2007-2010 and 2019-2021 respectively. As the proportion of meat-lovers is still larger than the proportion of those belonging to *green*, oversampling is applied to the minority group. For determining the correct ratio, different values are researched with a grid search. The results of both waves are presented with balanced accuracy (BA) and AUC scores in Table 7.1 in the Appendix. Based on the BA score, the optimal ratio found for both waves is 1:3.

Next, to prevent the trees from becoming immense and unreadable again, the 'maxdepth' argument of the decision tree is decided on. The optimal value for 'maxdepth' is found via a grid search as well, with 10-fold cross-validation. The results of both waves are presented with BA and AUC scores in Table 7.2 in the Appendix. The optimal values are 6 and 5 for wave 2007-2010 and wave 2019-2021 respectively, based on the BA score. The BA score of wave 2019-2021 for 'maxdepth' ratio 1 is 'NA'. This model could not determine those having a positive value for *green* with a ratio that low.

Furthermore, the dataset is divided into a training and test set with an 80/20 ratio. In addition, the variable *w_demog_season_wk_wknd* is added as function weight to the model to give all observations a weight in line with the reference population. Together with the optimal values for the oversampling ratio and the 'maxdepth' argument, it resulted in small and readable decision trees for both wave 2007-2010 as well as for wave 2019-2021. The decision tree for wave 2007-2010 will be described first, followed by the decision tree for wave 2019-2021. Next, a comparison over time is made. The trained models are used to make a prediction on the test set. Lastly, the models' performances will be discussed.

The decision tree for wave 2007-2010 is shown in Figure 5.1. It represents whether someone is a meat-lover (true) or he/she follows a (partly) plant-based food rule (false). The most important variables are *sex*, *hh_size* and *age*. The segment found presents: someone who is male, lives in a household size of 1, 6 or more than 9 people and is younger than 49 years, has a 58% chance of being a meat-lover.

The second decision tree is built for wave 2019-2021 and shown in Figure 5.2. The most important ones are *sex*, *edu*, *age* and *hh_size*. The segment found presents: someone who is male, higher educated, 28 years or older and has a household size of 1, or more than 3 people, has a 60% chance of being a meat-lover.

Over time, only *edu* is added to the important variables. Males are still most likely to be a

meat-lover. Besides living on your own, the household size shifted from 6 or more than 9 to more than 3 people. The threshold of age shifted downwards, from younger than 49 to younger than 28. Education showed that highly educated are most likely to be meat-lovers.

Furthermore, the results can be compared with those of the logistic regressions presented in Table 5.3. For wave 2007-2010, the results of the variables *sex* and *hh_size* were similar. Being male and living in a household size of one extra person are positively associated with the log-odds of being a meat-lover, ceteris paribus. For wave 2019-2021, only the result of *sex* was similar. *Edu* presented a negative association with the log-odds of being a meat-lover, ceteris paribus. This will be further elaborated on in the hypotheses discussion.

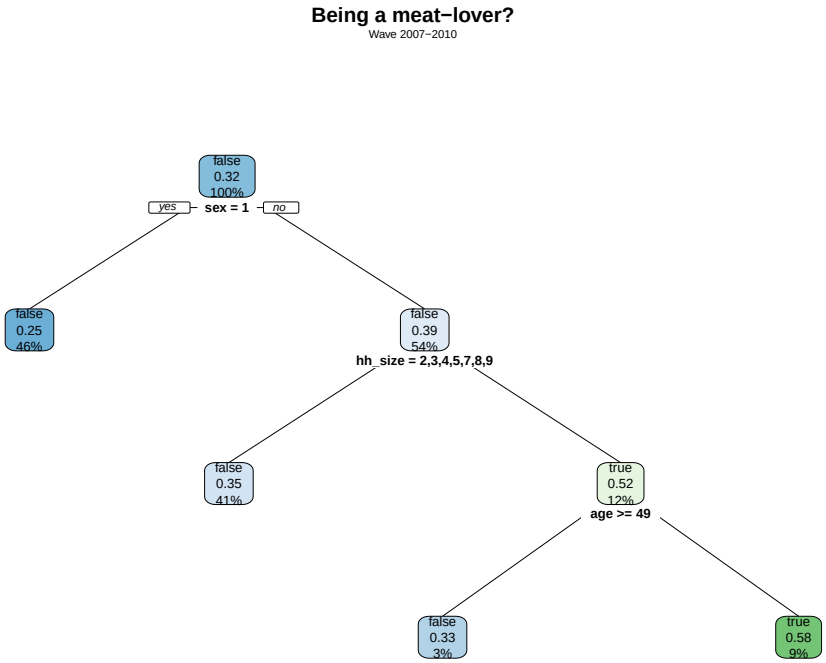

Figure 5.1: *Decision tree meat-lover wave 2007-2010*

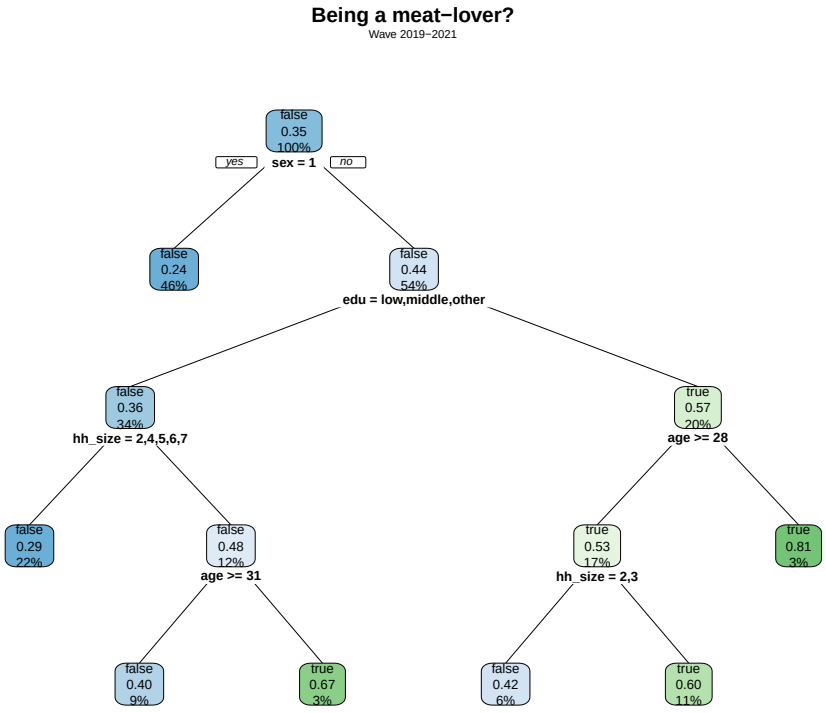

Figure 5.2: *Decision tree meat-lover wave 2019-2021*

The predictive performances of the decision trees are presented in Table 5.4 by different measures. Overall, the models are relatively fine for the fact that the minority class was relatively small. The worst balanced accuracy of both is 61.4% and can be considered fine. However, from investigating the AUC scores, it can be stated that the models are performing not very well. The worst score is 54.4% as a minimum of 70% is considered acceptable (Mandrekar, 2010). The models' performances can also be visually presented with the ROC-curve. They are shown for waves 2007-2010 and 2019-2021 in Figure 7.1 and Figure 7.2 in the Appendix respectively. As the curves are not that close to the upper left corner, they present not to be that precise.

Table 5.4: *Predictive performance of meat-lover trees*

	Wave 2007-2010	**Wave 2019-2021**
Balanced Accuracy - interval	(0.614, 0.664)	(0.650, 0.673)
AUC - interval	(0.544, 0.561)	(0.564, 0.587)

5.3.2 Consuming meat substitutes vs. not

The same procedure is applied for building the decision trees presenting whether someone consumes meat substitutes (true) or not (false). These trees are built for the 2012-2016 and 2019-2021 waves. Based on the presented statistics in Table 3.4, the proportion of respondents belonging to *meatsub* is 3.7% in wave 2012-2016 and 9.1% in wave 2019-2021. As the percentages are relatively low, resampling is applied for this minority group. For determining the correct ratio, different values for the ratio are again researched with a grid search. The results of both waves are presented with BA and AUC scores in Table 7.3 in the Appendix. Based on the BA score, the optimal ratio found is 1:3.5 for both waves.

Furthermore, the 'maxdepth' argument is also decided on. To find the optimal value, a grid search with 10-fold cross-validation is used. The results of both waves are presented with BA and AUC scores in Table 7.4 in the Appendix. The optimal value for the 2012-2016 and 2019-2021 waves are 8 and 7 respectively. For the latter, the value is based on the fourth decimal. For wave 2012-2016 'maxdepth' values 1 and 2 show 'NA' values for the BA scores. These models were not able to identify respondents consuming meat substitutes.

Next, the dataset is divided into a training and test set with an 80/20 ratio. The variable *w_demog_season_wk_wknd* is added as function weight to the model to give all observations a weight in line with the reference population. The decision trees are created with the optimal values for the ratio and the 'maxdepth' argument. They resulted in readable trees. First, the

decision tree for wave 2012-2016 will be described, followed by the decision tree for wave 2019-2021. Also, a comparison over time is made. The trained models are used to make a prediction on the test set. Lastly, the models' performances will be discussed.

The first decision tree is built for wave 2012-2016 and shown in Figure 5.3. The most important variables are *edu*, *age*, *hh_urb_3*, *hh_size* and *BMI_cat*. With *age* accounting for three splits. The segment found presents: someone with a middle or high education, living in urbanization level 1 ('extremely/strongly' urbanized), living in a household size of 1, 2 or over 8 people, having BMI level 1 or 3 ('seriously underweight' and 'normal weight' respectively) and is between the age of 42 and 62 years old, has an 81% chance of consuming meat substitutes.

The second decision tree built for wave 2019-2021 is shown in Figure 5.4. When taking the 5% minimum of respondents into account, three segments can be identified. For the first one, the most important variables are *BMI_cat*, *age*, *migration_background* and *hh_size*. The segment found presents: someone with a BMI level of 2, 4 or 5 ('underweight', 'overweight' and 'obesity' respectively), under the age of 43, a migration background other than Dutch and living in a household size of 1, 2, 4 or more than 5 people, has a 69% chance of consuming meat substitutes.

For the second one, the most important variables are *BMI_cat*, *hh_size*, *edu* and *age*. With *age* accounting for two splits. The segment found presents: someone with a BMI level of 1 or 3 ('seriously underweight' and 'normal weight' respectively), a household size between 2 and 7 people, higher educated and between the age of 39 and 47, has a 75% chance of consuming meat substitutes.

For the third one, the most important variables are *BMI_cat*, *hh_size* and *age*. With *age* accounting for three splits. The segment found presents: someone with a BMI level of 1 or 3 ('seriously underweight' and 'normal weight' respectively), a household size of 1 or more than 7 people and between the age of 29 and 40, has an 84% chance of consuming meat substitutes. As this segment presents the highest chance of consuming meat substitutes, the conclusions will be based on this one. Thus, it can be stated that the variables *BMI_cat*, *hh_size* and *age* are among the most important for segmenting consumers of meat substitutes.

Over time, the segment of wave 2012-2016 will be compared with the third one of wave 2019-2021. As the third segment presents the highest chance of consuming meat substitutes. The variables *edu* and *hh_urb_3* were not included in the list of the 2019-2021 wave. BMI level presented the same results. Household size shifted a bit from 1, 2 or more than 8 to 1 or more than 7 people. The threshold of age shifted downwards, from 42-62 to 29-40 years old. The tree of the wave 2012-2016 does not provide as many potential segments as the 2019-2021 wave.

This is not very surprising, as the prevalence of respondents consuming meat substitutes was relatively low. What might have made it more difficult.

Furthermore, the results can be compared with those of the logistic regressions presented in Table 5.3. For wave 2012-2016, the result of the variable *edu* is similar. Having an education of one level higher is positively associated with the log-odds of consuming meat substitutes, ceteris paribus. For wave 2019-2021, only the result of *BMI_cat* and *edu* could be compared. Having a BMI of one level lower is namely positively associated with the log-odds of consuming meat substitutes, ceteris paribus. This is in line with the second and third segments. Having an education of one level higher is positively associated with the log-odds of consuming meat substitutes, ceteris paribus. This is in line with the second segment.

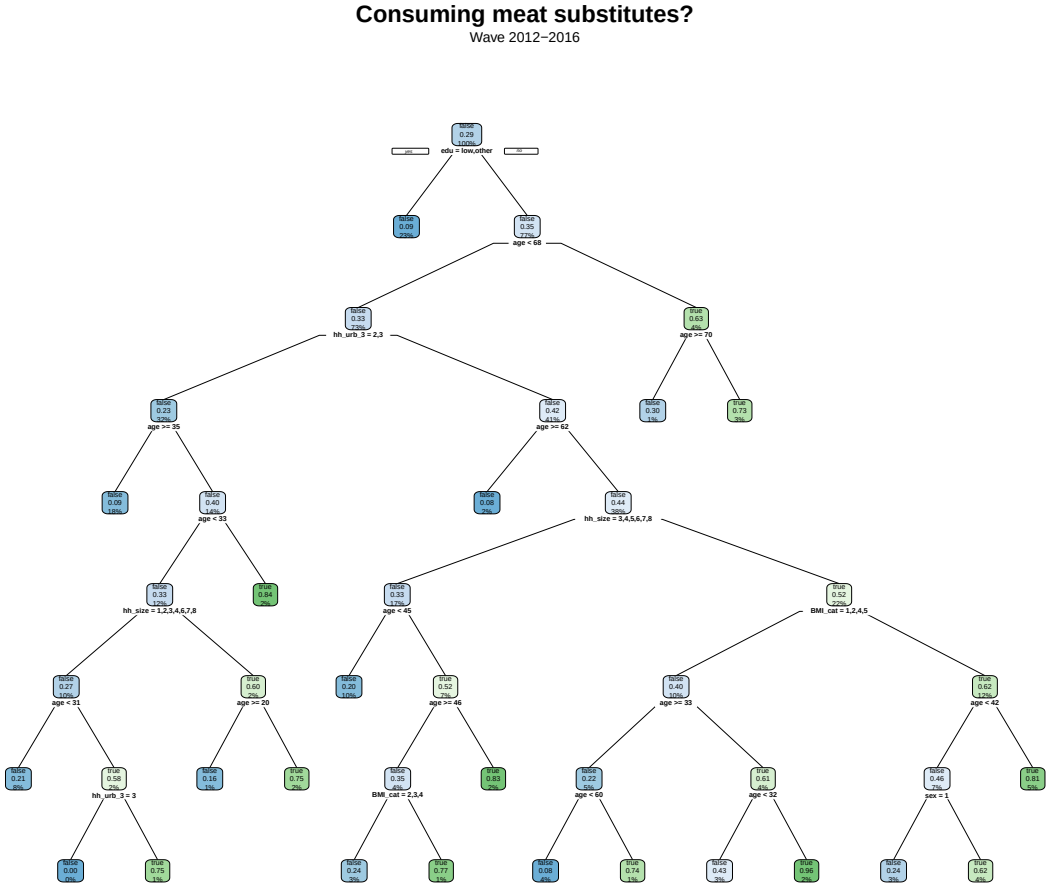

Figure 5.3: *Decision tree meat substitutes wave 2012-2016*

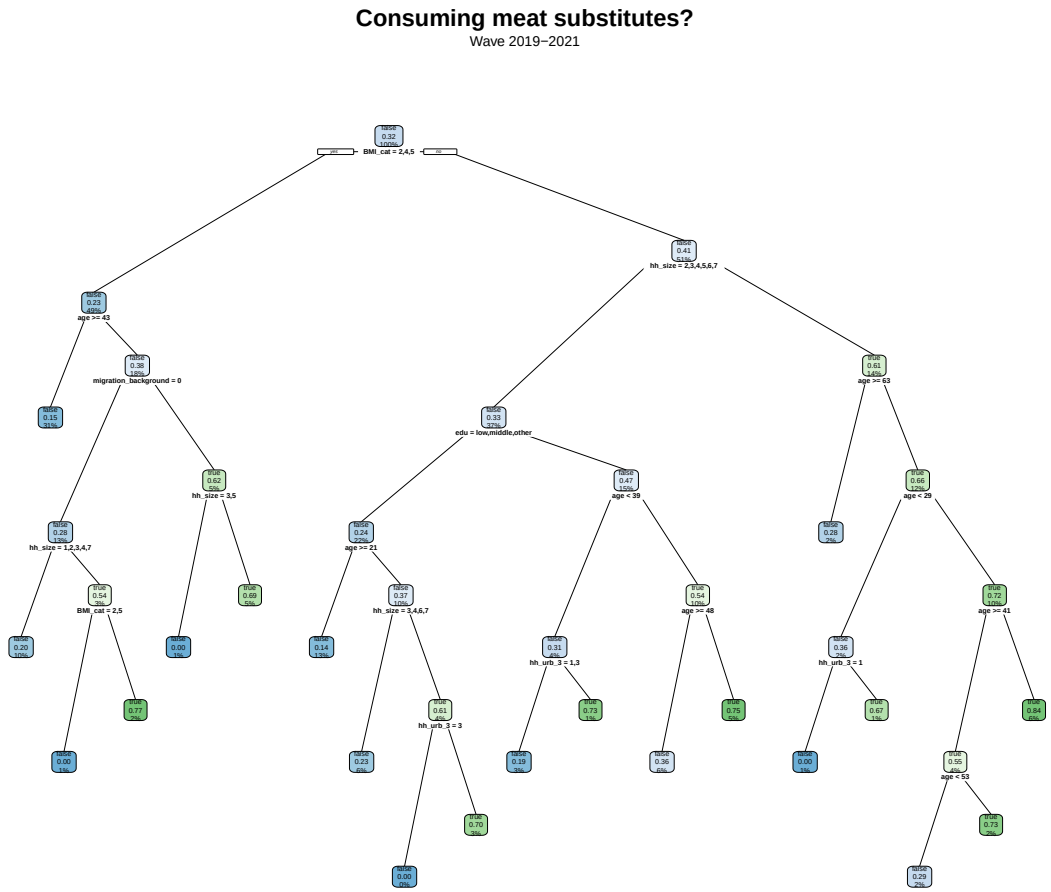

Figure 5.4: *Decision tree meat substitutes wave 2019-2021*

The predictive performance of the decision trees is presented in Table 5.5 by different measures. Overall, the models are relatively fine considering that the minority class was also for this food rule relatively small. The worst BA of both waves is 67.5% and is therefore good. However, from investigating the AUC scores, it can be stated that the models are performing not very well. The worst score is, with 57.3%, just better than complete randomness. This is from the wave 2019-2021 model. The worst AUC score of the wave 2012-2016 model is higher than that of the 2019-2021 model. The same holds for the best score. First of all, an interval is provided such that no conclusions are drawn based on one-time luck. The same holds for the three presented here. The data were randomly divided into a training and test set. Thus, it could just be due to coincidence that this occurs. However, it could be that, as the prevalence of meat substitute consumers is lower in the first model, more observations had to be resampled with replacement. This might be in line with problems that occurred with the earlier developed 99.0% accurate models. It might still be overfitting a bit. Thus, this limits the results. The models' performances can also be visually presented with the ROC-curve. They are presented

for wave 2012-2016 and wave 2019-2021 in Figure 7.3 and Figure 7.4 in the Appendix respectively. As also these curves are not that close to the upper left corner, they present not to be that precise.

Table 5.5: *Predictive performance of meat substitute trees*

	Wave 2012-2016	**Wave 2019-2021**
Balanced Accuracy - interval	(0.684, 0.732)	(0.675, 0.677)
AUC - interval	(0.608, 0.647)	(0.573, 0.636)

5.4 Hypotheses discussion

Based on the outcomes of the descriptive statistics, statistical analyses, logistic regressions and decision trees, it can be determined whether the hypotheses hold, do not hold or whether it was not possible to draw conclusions. The results of the logistic regressions are mostly used for drawing conclusions about those following a (partly) plant-based food rule (food rules *plantc* and *flexitarian*). The outcomes of the decision trees are mainly used for the hypotheses about the meat-lovers and those consuming meat substitutes. As these provide more detailed identifications in comparison with the logistic models. If a hypothesis does not make a statement about a transition over time, the most recent results are used to draw conclusions (of wave 2019-2021). In this subsection, all hypotheses are discussed. Table 5.6 presents an overview of them and their corresponding result.

The first hypothesis is rejected, based on the descriptive statistics presented in Table 3.4. The prevalence in decreasing order for the 2019-2021 wave are: meat-lovers (84.5%), flexitarians (10.4%), meat substitutes (9.1%) and plant-centered (5.6%). Thus, the prevalence of flexitarians is larger than the prevalence of consumers using meat substitutes. The prevalence of people consuming meat substitutes was expected to be higher, around the 25% based on the research of Cuffey et al. (2023).

The second hypothesis is accepted, for both the (partly) plant-based food rules as well as the *meatsub* rule. The logistic regression results in Table 5.3 present that being female has a positive influence on the log-odds of following a (partly) plant-based food rule and on consuming meat substitutes. This is in line with the expectation based on earlier literature that women are more likely to follow a plant-based food rule compared to men (Wozniak et al., 2020; Deliens et al., 2022; Verain et al., 2022; Graça et al., 2019) and to consume meat substitutes (Mullee et al., 2017).

The third hypothesis is rejected. The coefficients of the variable *age* in the logistic regressions

of the wave 2019-2021 models *plantc* and *flexitarian* in Table 5.3 are not statistically significant at an α level of 1%. This is not surprising as the null hypotheses of the t-tests were not rejected based on the results in Table 5.1. However, the negative associations of the coefficients do show that the younger someone is, the more likely they are to belong to one of these food rules. Earlier research was already a bit divided on the importance of someone's age, but it was expected that the younger aged (possibly in combination with their household size) are more likely to follow a (partly) plant-based food rule (Wozniak et al., 2020; Deliens et al., 2022; Lusk, 2017).

The fourth hypothesis is accepted, for both the (partly) plant-based food rules as well as the *meatsub* rule. The logistic regression results in Table 5.3 present that being higher educated has a positive influence on the log-odds of following a (partly) plant-based food rule and on consuming meat substitutes. This is in line with earlier research, as it was expected that consumers with a higher education level are more likely to follow a (partly) plant-based food rule than consumers with a lower one (Wozniak et al., 2020; Deliens et al., 2022; Verain et al., 2022; Graça et al., 2019; Lusk, 2017). The same was expected for those consuming meat substitutes Cuffey et al. (2023).

Unfortunately, the fifth and sixth hypotheses were not possible to test with the data present as the variable *income* was not available. Therefore, these are inconclusive.

The seventh hypothesis is rejected. The coefficients of the variable *hh_urb_3* in the logistic regressions of the wave 2019-2021 models *plantc*, *flexitarian* and *meatsub* in Table 5.3 are not statistically significant at an α level of 1%. However, the negative associations of the coefficients do show that living in more urbanized areas (a lower value for *hh_urb_3*) increases the chance of belonging to one of these food rules as the coefficients presented are negative as well. The earlier literature of Deliens et al. (2022); Graça et al. (2019); Cuffey et al. (2023) expected that living in more urbanized areas has an influence on following a (partly) plant-based food rule and on consuming meat substitutes.

The eighth hypothesis is accepted, based on the decision tree about meat-lovers. Both trees, but focusing on the one of wave 2019-2021 presented in Figure 5.2, showed that being male is an important characteristic of being a meat-lover. This is in line with the results of the logistic regression *meatlover* wave 2019-2021 model in Table 5.3. The variable *sex* presented, statistically significant at an α level of 1%, that being female has a negative effect on the log-odds of being a meat-lover. This aligns with the earlier study done by Reuzé et al. (2022).

The ninth hypothesis is rejected, based on the decision tree about meat-lovers. The tree of wave 2019-2021 shown in Figure 5.2, presented that being higher educated is an important

characteristic of being a meat-lover. This is in contrast with the results of the logistic regression *meatlover* wave 2019-2021 model in Table 5.3 and earlier literature. The variable *edu* presented, statistically significant at an α level of 1%, that being higher educated has a negative effect on the log-odds of being a meat-lover. Based on the research of Reuzé et al. (2022), it was expected that the lower educated are more likely to be meat-lovers than the higher educated. However, the segment presented by the decision tree showed something else. As education might be a proxy for income, it might for instance be that the people in this segment are higher educated, have higher incomes and can afford to buy expensive kinds of meat such as steaks. Therefore, the result might differ.

The tenth hypothesis is rejected. Based on the decision trees as well as on the logistic regression models in Table 5.3, it can be stated that not all but some important predictors have changed over the last decade. Or for the meat substitutes, during the last five years. It was difficult to draw an expectation based on earlier research as it has not yet been studied. Verain et al. (2022) researched differences in the beliefs with regard to consumers' attitudes and norms towards eating meat. They did not find a difference. However, as characteristics are something different, another result is not surprising.

The eleventh hypothesis is rejected. The post-pruned decision trees were expected to provide insights into the most important features for determining which food rule someone follows (Lusk, 2017). However, as they became immense and unreadable, they could not provide clear insights. In addition, there were signs of overfitting.

The twelfth hypothesis is accepted. The decision trees with applied oversampling have provided insights into the most important features for determining which food rule someone follows. It was expected beforehand that resampled decision trees might provide clear insights as there were imbalanced classes in the dataset (Lee, 2014).

Table 5.6: *Hypotheses overview*

Identification	Hypothesis	Result
H1	The prevalence of people following a specific food rule is expected to be in decreasing order; meat-lovers, consumers using meat substitutes, flexitarians and plant-centered.	Rejected
H2	Women are more likely to follow a (partly) plant-based diet and to buy meat substitutes than men.	Accepted
H3	The younger aged are more likely to follow a (partly) plant-based diet than the older aged.	Rejected
H4	Higher educated are more likely to follow a (partly) plant-based diet and to buy meat substitutes than lower educated.	Accepted
H5	People with a higher income are more likely to follow a (partly) plant-based diet than people with a lower income.	Inconclusive
H6	The variable income is not expected to be of influence for buying meat substitutes.	Inconclusive
H7	Living in more urbanized areas will influence whether people will follow a (partly) plant-based diet and buy meat substitutes.	Rejected
H8	Men are more likely to be meat-lovers than women.	Accepted
H9	Lower educated are more likely to be meat-lovers than higher educated.	Rejected
H10	The important predictors for determining whether someone followed a (partly) plant-based diet are not expected to have changed over the last 10 years.	Rejected
H11	Post-pruned decision trees will provide clear insight into the most important features.	Rejected
H12	Designing decision trees with resampling techniques to overcome the problem of the imbalanced dataset, will provide clear insights into the most important features.	Accepted

Chapter 6

Conclusion

In this section, a summary of the study is given. The sub-questions and finally the research question will be answered. This is followed by a discussion of the limitations of the research. To conclude with recommendations for further research.

6.1 Summary findings

Summarizing, data from the Dutch National Food Consumption Surveys of the RIVM are used to investigate different diets of people in the Netherlands while focusing on their meat consumption. Therefore, the following four food rules are researched: being a meat-lover, following a plant-centered food rule (pescatarian, vegetarian or vegan), being a flexitarian (consumed meat at maximum once in two non-consecutive days) and consuming meat substitutes. It is studied what the most important characteristics of respondents following the food rules are and potential segments are designed. By identifying them, they can be targeted more easily in the future. Survey data from three waves are available, namely 2007-2010, 2012-2016 and 2019-2021. This made it possible to study whether the most important characteristics changed over the last decade (or the last five years for consuming meat substitutes). First, a benchmark logistic regression model is made. One for each food rule and for each wave. Next, decision trees are built to further identify the meat-lovers vs. the (partly) plant-based food rules and those consuming meat substitutes vs. those who do not. Based on these results, the hypotheses were discussed and linked to previous literature. Next, all sub-questions will be answered and finally the research question.

The first sub-question follows *"What are the most common food rules to follow in the Netherlands and have these changed over the last decade?"*. The order of the most important food rules based on the prevalence in decreasing order has only changed due to the meat substitute food rule that

became available in wave 2012-2016. Namely, the most common food rules in decreasing order are: meat-lovers, flexitarians and plant-centered for wave 2007-2010 and: meat-lovers, flexitarians, meat substitutes and plant-centered for wave 2019-2021. Concluding, being a meat-lover is still the most popular food rule.

The second one follows *"What are the most important features for predicting whether someone is a meat-lover?"*. The most important ones are *sex*, *edu*, *age* and *hh_size*. Someone who is male, higher educated, 28 years or older and has a household size of 1, or more than 3 people, has a 60% chance of being a meat-lover. Over time, *edu* was added to the list of important ones.

The third one follows *"What are the most important features for predicting whether someone follows a (partly) plant-based food rule?"*. For the plant-centered rule, *sex*, *BMI_cat* and *edu* are the most important features. Sex and education have a positive influence on the log-odds of following a plant-centered food rule, ceteris paribus, BMI level a negative, ceteris paribus. Over time, *edu* became important as well. For the flexitarian rule, *sex* and *edu* are the most important ones. Both have a positive influence on the log-odds of being a flexitarian, ceteris paribus, just like for the plant-centered rule. *Hh_size* was used to be important but is replaced by *edu*. Concluding, overall sex, education and in some cases, BMI level are the most important features for predicting whether someone follows a (partly) plant-based food rule.

The fourth one follows *"What are the most important features for predicting whether someone consumes meat substitutes?"*. For the meat substitute rule, the variables *BMI_cat*, *hh_size* and *age* are the most important. Someone with a BMI level of 1 or 3 ('seriously underweight' and 'normal weight' respectively), a household size of 1 or more than 7 people and between the age of 29 and 40, has an 84% chance of consuming meat substitutes. Over time, *edu* and *hh_urb_3* were not included anymore.

The fifth sub-question follows *"How does gender influence the probability of following a (partly) plant-based food rule and consuming meat substitutes?"*. The variable *sex* presents for the plant-centered, flexitarian and meat substitute rules a positive influence on the log-odds of following one of these specific rules, ceteris paribus. Meaning that being female positively influences the log-odds, ceteris paribus.

The sixth one follows *"How does age influence the probability of following a (partly) plant-based food rule and consuming meat substitutes?"*. For the (partly) plant-based food rules, *age* presented negative associations with the log-odds of following a plant-centered food rule or being a flexitarian. The coefficients were not statistically significant at an α level of 1%. For consuming meat substitutes, being between 29 and 40 years old positively influences the probability of consuming them.

The seventh sub-question follows *"How does educational attainment influence the probability of following a (partly) plant-based food rule and consuming meat substitutes?"*. *Edu* presents a positive influence on the log-odds of following a plant-centered food rule, being a flexitarian and consuming meat substitutes, ceteris paribus. Meaning that a higher educational attainment level positively influences the log-odds, ceteris paribus.

The eighth and final sub-question follows *"Have the predictors changed in the Netherlands over the last decade?"*. It can be stated that they have changed a bit but not very much. Some became less important, others more. Overall, the results became more detailed as more coefficients of the benchmark model were statistically significant at an α level of 1% in wave 2019-2021 than in wave 2007-2010 or wave 2012-2016. Also, the decision trees presented more thresholds.
In conclusion, the research question follows:

"Is it possible to accurately determine with food survey data what important factors for identifying consumers who follow a specific food rule are and whether these have changed in the Netherlands over the last decade?"

It can be stated that it is indeed possible to determine with food survey data what the important factors for identifying consumers who follow a specific food rule are and that these have changed a bit in the Netherlands over the last decade. However, as presented earlier with the performance measures of the models, the results are fine but the models are not performing that well. The accuracy of the models does not present an excellent result. This is due to the imbalanced dataset and the fact of overfitting that had to be tackled.

6.2 Discussion

Before discussing the research limitations, there are two points that are worth touching upon. First, a strength of this study is that by using the weighted factor for all respondents, a representative sample of the Dutch population is accounted for. This causes representative outcomes. However, one should be careful when comparing the results with other European countries as the prevalence of people following a specific food rule might differ (Deliens et al., 2022).

Second, one should take into account that the food intake is the actual and not a predicted intake. It is first checked whether the respondents who characterized themselves as vegetarian no meat (*r_veg_meat*), vegetarian no meat/fish (*r_veg_meatfish*) and vegan (*r_vegan*) indeed also did not eat meat/fish. Next, these respondents are categorized into the plant-centered food rule. In addition, flexitarians are grouped based on their actual consumption and not because a respondent thought to be one himself. The same holds for the meat-lovers. As the requirements

of belonging to a specific food rule might differ from person to person or people simply do not know the difference. Deliens et al. (2022) had found that respondents who ate no meat for one day a week categorized themselves as omnivores if the options to choose from were limited. In this study, that would not present a clear and transparent view. Therefore, to make sure that the respondents are grouped based on the same requirements, everyone is categorized based on their actual consumption. Therefore, this study presents the actual food rule a respondent belongs to.

There are some limitations of this study. Regarding the data gathering, first of all, there is a non-response bias. People get an invitation to participate in the research and are not obliged to (would be ethically questionable). Those who do respond are therefore different from those who do not. This causes a non-response bias. In the 2019-2021 dataset, for instance, there was a response call of 37% (RIVM, 2021). Non-response bias might have an influence on the results, as the sample is potentially biased. However, this is with survey data impossible to overcome.

Second, the respondents self-reported their consumption during the 24-hour dietary recalls. Respondents could forget things they ate or drank, unintentionally or on purpose. As the respondents are not monitored, it cannot be stated with 100% certainty that the reported consumption was the actual consumption.

Third, there might be cases of misclassification bias. It is not considered for how long a respondent has been following a specific food rule. During the four weeks of data gathering, respondents were asked about their diet on two non-consecutive days. Since someone's consumption might differ between months, it could lead to misclassification bias. If one of these three biases with regard to the data is present, it is not expected to be a huge limitation for the study as they may be systematic. They apply to all respondents across all waves and therefore are equally present across time. For instance, the misclassification bias presents a snapshot, but this is the case for everyone.

Fourth, another limitation of the study is that only four food rules are taken into account. The pescatarians, vegetarians and vegans are combined in the plant-centered rule as the proportions on their own were too small. Furthermore, besides the exclusive plant-centered, flexitarian and meat-lover food rules, there are no other options (as the meat substitutes can be consumed by anyone). The food rules are limited. Further research could take multiple levels of flexitarians (little, middle, very) into account as well, for instance. Or investigating the pescatarians, vegetarians and vegans separately when these groups are large enough.

Fifth, the fact that some food rules presented relatively low prevalence rates, oversampling

with replacement was applied. This is a limitation as it might not fully represent the actual population as the sample is adjusted. One should keep this in mind when interpreting the results of these classes. In addition, this might also be the reason for the bit better performing decision tree on consuming meat substitutes for wave 2012-2016 than wave 2019-2021. It might be that due to the lower prevalence and consequently the more resampling with replacement in the first model than in the latter, the results of the first might seem to perform better. The worst AUC score of the wave 2012-2016 model is higher than that of the 2019-2021 model. The same holds for the best score. This might be in line with problems that occurred with the earlier developed 99.0% accurate models. It might still be overfitting a bit.

Sixth, not all variables are independent of the outcome determining if someone belongs to a specific food rule, a case of endogeneity. *BMI_cat* is included, but the kind of food rule that is being followed might influence someone's BMI level. Thus, it is influenced by the dependent variable. In addition, it might also hold a bit for *edu*. As the food rule someone follows might, via things such as nutrition, energy, focus and behavioural well-being, influence one's education somehow. However, as not many identical variables were available in all three datasets, only six or seven were included in the models. There were not a lot of variables as options to include. Therefore, this study especially focused more on describing the most important ones and on potential segments.

Seventh, unfortunately, there was no variable present in all datasets regarding respondents' income. Education might serve as a proxy for income. Based on earlier literature by Graça et al. (2019), Lusk (2017) and Wozniak et al. (2020), income might be one of the important variables for belonging to a (partly) plant-based food rule. It is a limitation that the influence of income could not be fully studied and thus a recommendation for further research.

Another recommendation is, now the important characteristics of people are identified, to map where people following a specific food rule live on a neighborhood scale in the Netherlands. With a practical application, useful information can be gathered for sellers of meat, meat substitutes and supermarkets. For instance, in an urbanized area like Rotterdam, there are still lots of differences between neighborhoods and the people living in them. This provides insights into where to sell what. In addition, it can be used for non-profit organisations. It is easier to determine where local campaigns would be useful. By encouraging people they create attention and can make them aware of the health consequences of eating meat.

Chapter 7

Appendix

Table 7.1: *Results of ratio grid search meat-lover*

Ratio	**1:2**	**1:2.5**	**1:3**	**1:3.5**
Wave 2007-2010 BA	(0.603)	(0.608)	**(0.667)**	(0.630)
Wave 2007-2010 AUC	(0.602)	(0.544)	(0.558)	(0.553)
Wave 2019-2021 BA	(0.588)	(0.637)	**(0.666)**	(0.600)
Wave 2019-2021 AUC	(0.587)	(0.629)	(0.584)	(0.549)

Table 7.2: *Results of 'maxdepth' grid search meat-lover*

Maxdepth	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Wave 2007-2010 BA	(0.581)	(0.581)	(0.596)	(0.596)	(0.609)	**(0.610)**	(0.610)	(0.610)	(0.610)	(0.610)
Wave 2007-2010 AUC	(0.581)	(0.581)	(0.591)	(0.596)	(0.608)	(0.608)	(0.608)	(0.608)	(0.608)	(0.608)
Wave 2019-2021 BA	(NA)	(0.616)	(0.616)	(0.614)	**(0.634)**	(0.634)	(0.634)	(0.634)	(0.634)	(0.634)
Wave 2019-2021 AUC	(0.500)	(0.569)	(0.569)	(0.591)	(0.592)	(0.592)	(0.592)	(0.592)	(0.592)	(0.592)

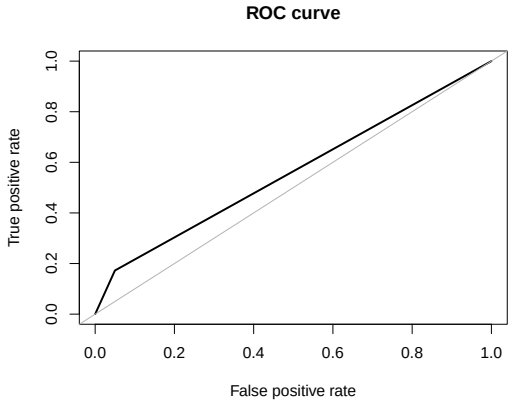

Figure 7.1: *ROC-curve of decision tree meat-lover wave 2007-2010*

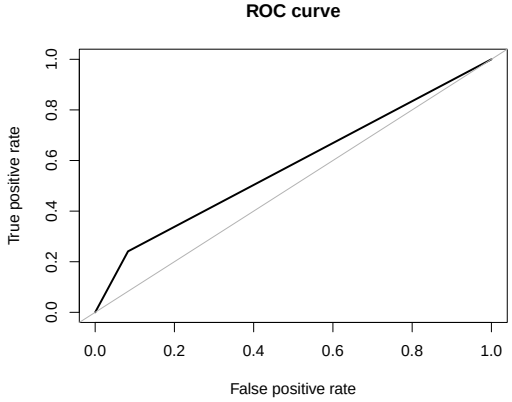

Figure 7.2: *ROC-curve of decision tree meat-lover wave 2019-2021*

Table 7.3: *Results of ratio grid search meat substitutes*

Ratio	**1:2**	**1:2.5**	**1:3**	**1:3.5**
Wave 2012-2016 BA	(0.707)	(0.734)	(0.733)	**(0.796)**
Wave 2012-2016 AUC	(0.704)	(0.729)	(0.691)	(0.738)
Wave 2019-2021 BA	(0.666)	(0.648)	(0.634)	**(0.698)**
Wave 2019-2021 AUC	(0.663)	(0.637)	(0.582)	(0.570)

Table 7.4: *Results of 'maxdepth' grid search meat substitutes*

Maxdepth	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Wave 2012-2016 BA	(NA)	(NA)	(0.628)	(0.700)	(0.711)	(0.704)	(0.722)	**(0.727)**	(0.727)	(0.727)
Wave 2012-2016 AUC	(0.500)	(0.500)	(0.604)	(0.609)	(0.643)	(0.694)	(0.704)	(0.703)	(0.703)	(0.703)
Wave 2019-2021 BA	(0.568)	(0.585)	(0.630)	(0.644)	(0.660)	(0.696)	**(0.696)**	(0.696)	(0.696)	(0.696)
Wave 2019-2021 AUC	(0.555)	(0.530)	(0.533)	(0.554)	(0.554)	(0.605)	(0.584)	(0.584)	(0.584)	(0.584)

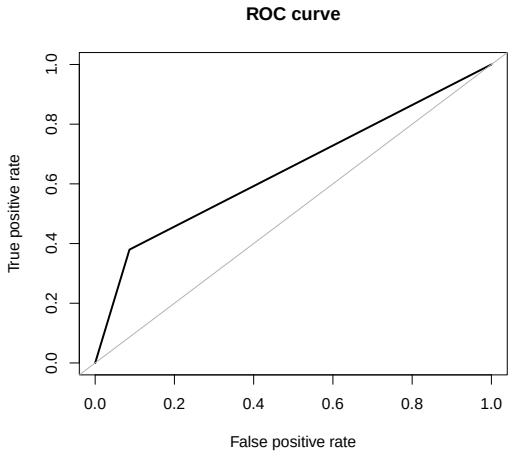

Figure 7.3: *ROC-curve of decision tree meat substitutes wave 2012-2016*

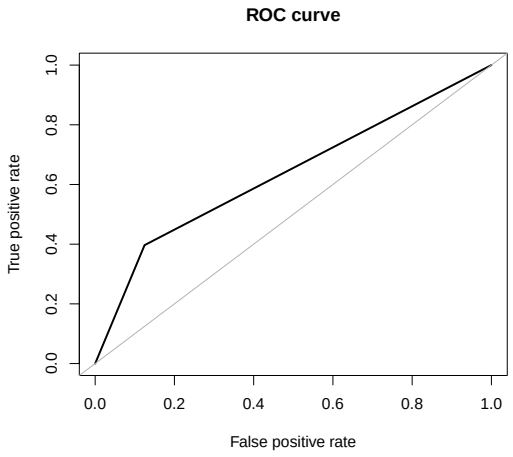

Figure 7.4: *ROC-curve of decision tree meat substitutes wave 2019-2021*

References

Cuffey, J., Chenarides, L., Li, W., & Zhao, S. (2023). Consumer spending patterns for plant-based meat alternatives. *Applied Economic Perspectives and Policy*, *45*(1), 63–85.

Deliens, T., Mullie, P., & Clarys, P. (2022). Plant-based dietary patterns in flemish adults: a 10-year trend analysis. *European Journal of Nutrition*, *61*(1), 561–565.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, *27*(8), 861–874.

Godfray, H. C. J., Aveyard, P., Garnett, T., Hall, J. W., Key, T. J., Lorimer, J., ... Jebb, S. A. (2018). Meat consumption, health, and the environment. *Science*, *361*(6399), eaam5324.

Graça, J., Godinho, C. A., & Truninger, M. (2019). Reducing meat consumption and following plant-based diets: Current evidence and future directions to inform integrated transitions. *Trends in Food Science & Technology*, *91*, 380–390.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., ... others (2020). Package âcaretâ. *The R Journal*, *223*(7).

Lee, P. H. (2014). Resampling methods improve the predictive power of modeling in class-imbalanced datasets. *International journal of environmental research and public health*, *11*(9), 9776–9789.

Lunardon, N., Menardi, G., & Torelli, N. (2014). Rose: a package for binary imbalanced learning. *R journal*, *6*(1).

Lusk, J. L. (2017). Consumer research with big data: applications from the food demand survey (foods). *American Journal of Agricultural Economics*, *99*(2), 303–320.

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, *5*(9), 1315–1316.

Melina, V., Craig, W., & Levin, S. (2016). Position of the academy of nutrition and dietetics: vegetarian diets. *Journal of the Academy of Nutrition and Dietetics*, *116*(12), 1970–1980.

Mullee, A., Vermeire, L., Vanaelst, B., Mullie, P., Deriemaeker, P., Leenaert, T., ... others (2017). Vegetarianism and meat consumption: A comparison of attitudes and beliefs between vegetarian, semi-vegetarian, and omnivorous subjects in belgium. *Appetite*, *114*, 299–305.

Pant, A. (2019). *Introduction to logistic regression.* Retrieved from `https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148`. (Accessed: June 14, 2023)

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*, 81–106.

Reuzé, A., Méjean, C., Carrère, M., Sirieix, L., Druesne-Pecollo, N., Péneau, S., ... Allès, B. (2022). Rebalancing meat and legume consumption: change-inducing food choice motives and associated individual characteristics in non-vegetarian adults. *International Jour-*

nal of Behavioral Nutrition and Physical Activity, *19*(1), 112.

RIVM. (2010). DNFCS *2007-2010, 7-69 years.* Retrieved from https://www.rivm.nl/en/dutch-national-food-consumption-survey/overview-surveys/dnfcs-2007-2010. (Accessed: April 19, 2023)

RIVM. (2016). DNFCS *2012-2016, 1-79 years.* Retrieved from https://www.rivm.nl/en/dutch-national-food-consumption-survey/overview-surveys/dnfcs-2012-2016. (Accessed: April 19, 2023)

RIVM. (2021). DNFCS *2019-2021, 1-79 years.* Retrieved from https://www.rivm.nl/en/dutch-national-food-consumption-survey/overview-surveys/dnfcs-2019-2021-1-79-years. (Accessed: April 19, 2023)

Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986). Akaike information criterion statistics. *Dordrecht, The Netherlands: D. Reidel*, *81*(10.5555), 26853.

Therneau, T., & Atkinson, B. (2022). *Rpart: Recursive partitioning and regression trees version 4.1.19 from* CRAN. Retrieved from https://rdrr.io/cran/rpart/. (Accessed: May 4, 2023)

Verain, M. C., Dagevos, H., & Jaspers, P. (2022). Flexitarianism in the netherlands in the 2010 decade: Shifts, consumer segments and motives. *Food Quality and Preference*, *96*, 104445.

Wozniak, H., Larpin, C., de Mestral, C., Guessous, I., Reny, J.-L., & Stringhini, S. (2020). Vegetarian, pescatarian and flexitarian diets: sociodemographic determinants and association with cardiovascular risk factors in a swiss urban population. *British journal of nutrition*, *124*(8), 844–852.

Zhao, S., Wang, L., Hu, W., & Zheng, Y. (2023). Meet the meatless: Demand for new generation plant-based meat alternatives. *Applied Economic Perspectives and Policy*, *45*(1), 4–21.