CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Using blood metabolomics to identify dietary protein intake with Machine Learning methods

Master's thesis in Computer science and engineering

KLEIO GKOUTZOMITROU

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2023

Master's thesis 2023

Using blood metabolomics to identify dietary protein intake with Machine Learning methods

KLEIO GKOUTZOMITROU

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2023

A Chalmers University of Technology Master's thesis template for LaTeX

KLEIO GKOUTZOMITROU

Supervisor: Annikka Polster, Department of Biology and Biological Engineering
Advisor: Helen Lindqvist, Biochemistry and Food Science (University of Gothenburg)
Examiner: Jean-Philippe Bernardy, Department of Computer Science and Engineering

Master's Thesis 2023
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Description of the picture on the cover page (if applicable)

Typeset in LaTeX
Gothenburg, Sweden 2023

A Chalmers University of Technology Master's thesis template for LaTeX

KLEIO GKOUTZOMITROU
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

This thesis examines how metabolomics data may be used to classify individuals based on the sources of protein in their diets. Developing accurate classification models that can distinguish between omnivores, vegans, vegetarians, and pescetarians is the aim of the study. Principal Component Analysis (PCA), Random Forest (RF), Support Vector Machines (SVM), and neural networks are used in this process as data analysis tools.

The dataset, which was given by the Gothenburg University Department of Internal Medicine and Clinical Nutrition, included 120 healthy participants who followed various eating patterns. The subjects were chosen based on certain criteria, and blood samples and body composition were taken and examined. The dataset has been scaled and contains unidentified metabolites.

The metabolic profile of the sample was shown using principal component analysis (PCA). The overall PCA analysis revealed that there was substantial individual variation in the metabolomic profiles and that the food groups could not be effectively differentiated. The metabolic profiles of meat eaters and non-meat eaters might be used to distinguish them.

Random Forest, SVM, and neural networks were the three machine learning techniques that were utilized for categorization. Neural Networks performed worse than Random Forest and SVM models in classifying each dietaryăgroup separately. Random Forest classified omnivores and non-omnivores with a high degree of accuracy.

To measure the consumption of dairy, eggs, and meat, several scoring techniques were applied. The second method, which increased meat intake ratings by a factor of 1.5, produced the results with the highest degree of accuracy.

This study sheds light on the metabolic effects of omnivorous diets and improves our understanding of the complex relationship between nutrition, metabolism, and health outcomes. It also highlights the potential of metabolomics and machine learning in predicting dietary patterns and categorizing people into different dietary categories.

Keywords: metabolomics, machine learning, Principal Component Analysis, Random Forest, Support Vector Machines, Neural Networks.

Acknowledgements

I would like to express my heartfelt gratitude to Annikka Polster, my supervisor, for her unwavering support, guidance, and valuable input throughout this project. Her expertise and mentorship have been instrumental in shaping the direction and execution of this thesis.

I am also deeply grateful to Helen Lindqvist for generously providing me with the training data that served as a crucial foundation for my analysis. Her contribution has been pivotal in the success of this study, and I am sincerely appreciative of her assistance.

Lastly, I would like to express my gratitude to my family, friends, and colleagues for their continuous support, encouragement, and understanding throughout this endeavor.

Kleio Gkoutzomitrou, Gothenburg, 2023-06-16

Contents

List of Figures

List of Tables

1

Introduction

In this chapter, we will provide a brief background on the use of blood metabolomics, the investigation of blood's small molecules known as metabolites, to identify dietary protein intake with machine learning methods. We will discuss why accurately assessing dietary protein intake is an important research problem and how traditional methods have limitations in addressing this problem. We will then introduce the potential of blood metabolomics and machine learning to provide a more accurate and personalized approach. Specifically, we will focus on the problem of identifying dietary protein intake and the approach we have implemented. We will define the scope of our work, including any assumptions and limitations. Next, we will present the contributions of this thesis and outline the chapters that follow.

1.1 Background

A healthy diet is essential for good health and nutrition. In recent years, there has been growing concern about the negative health effects of meat consumption, particularly red meat [1]. This concern has prompted interest in vegetarianism or "flexitarianism," which involves eating less meat or being vegetarian but consuming fish. As a result, the "omnivore dietary category" now includes those who consume a variety of meat and fish, as well as those who consume less meat or no meat at all.

Some vegetarians, known as lacto-ovo vegetarians, replace meat with full-fat dairy foods like cheese and eggs. Others follow an essentially vegan diet, replacing dairy foods with novel alternatives based on soy, rice, or oats. A vegan diet forgoes all foods of animal origin. These contemporary changes in food consumption may account for the conflicting findings on health effects in research contrasting vegetarian and omnivorous diets [2].

We research nutrition because dietary intake affects our health in significant ways [3]. However, measuring diet with objectivity is challenging. Metabolomics could be useful for this task. Metabolomic analyses offer an option to study a comprehensive set of small molecules present in biofluids, cells or tissue, and allows for the profiling of thousands of molecules [4]. Blood metabolomics, the comprehensive analysis of small molecules in the blood, is a promising tool for assessing dietary protein intake and other aspects of nutrition. Machine learning (ML) is a powerful tool for analyzing large and complex data sets, including blood metabolomics data. The

combination of blood metabolomics and ML methods has the potential to provide new insights into the complex relationship between dietary protein intake and health outcomes [5].

1.2 Methods

The research methodology involved collecting blood samples from participants with varying dietary protein sources, including meat-eaters, fish-eaters, and vegetarians. The blood samples will be analyzed using Nuclear magnetic resonance (NMR), a powerful analytical technique for identifying and quantifying small molecules in complex biological samples. The resulting data will be preprocessed and analyzed using statistical methods and ML methods, including feature selection, dimensionality reduction, and classification algorithms.

The significance of this research lies in its potential to advance our understanding of the relationship between dietary protein intake and health outcomes. Accurately assessing dietary protein intake using blood metabolomics and ML methods could improve our ability to develop personalized nutrition recommendations that are tailored to an individual's specific needs and goals.

1.3 Structure

The structure of this Master's thesis is as follows:

- **Chapter 1** provides an overview of the research background, objectives, and significance.
- **Chapter 2** reviews the relevant literature on blood metabolomics, dietary protein intake, and ML methods.
- **Chapter 3** describes the research methodology in detail, including study design, data description, and analysis methods.
- **Chapter 4** presents the study's results, including biomarker identification, model development, and model evaluation.
- **Chapter 5** summarizes the main findings and conclusions, as well as recommendations for future research.

2

Theory

The theory chapter of this thesis covers the fundamentals of diet and how it affects the human body, including the different types of macronutrients and their role in the body. Furthermore, the chapter explains the concept of metabolomics and its importance in the study of nutrition, including the techniques used in metabolomics analysis such as nuclear magnetic resonance (NMR) spectroscopy. The chapter then delves into the statistical method of principal component analysis (PCA) and its applications in metabolomics research. The section on machine learning provides an overview of the different types of machine learning algorithms used in metabolomics, including supervised and unsupervised learning. The focus will be on the use of machine learning techniques, such as support vector machines (SVMs), random forest, and deep learning, to analyze metabolomics data for identifying biomarkers of dietary protein intake.

2.1 Diet

Diet is defined as the sum of foods consumed by an individual or population and plays a crucial role in maintaining good health and preventing diseases [6]. The human diet consists of macronutrients and micronutrients. Macronutrients are nutrients that are required in large quantities by the body and include proteins, carbohydrates, and fats. Micronutrients are nutrients required in smaller quantities by the body. These include vitamins and minerals, which are essential for various physiological processes in the body.

The macronutrients known as proteins are made up of amino acids, which are the body's building blocks. Proteins are necessary for the production of enzymes, hormones, and other compounds as well as for the development and upkeep of human tissues. There are 20 different types of amino acids, and the body can synthesize some of them, whereas others are essential and must be obtained from the diet. Sources of dietary protein include animal products such as meat, fish, eggs, and dairy, as well as plant-based sources such as legumes, nuts, and seeds [7]. Proteins from animal sources contain all essential amino acids, which are the building blocks of proteins that the body cannot produce on its own. In contrast, vegetarian sources of protein may lack one or more essential amino acids. Therefore, different plant-based protein sources must be combined to ensure an adequate intake of all the essential amino acids. This is known as protein complementation and is often

necessary for vegetarians and vegans to meet their daily protein requirements [8]. Additionally, vegetarian protein sources often contain dietary fiber, which can have a positive impact on digestion and overall health.

Protein content and quality can vary widely among different foods [9]. While most foods contain some amount of protein and amino acids, high protein sources are typically found in animal muscle products such as meat and fish. However, protein sources also differ in other metabolites such as carbohydrate content, fatty acids, and other micronutrients. For example, while meat and fish are low in carbohydrates, they may contain varying amounts of saturated and unsaturated fatty acids, depending on the type of animal and its diet. In contrast, vegetarian protein sources such as legumes, nuts, and seeds, may contain higher levels of carbohydrates, fiber, and other micronutrients such as vitamins and minerals [9].

Carbohydrates are a subset of macronutrients that are the main sources of energy in the body [10]. They contain both complex carbohydrates, such as starch and fiber, and simple sugars, such as glucose, fructose, and galactose. Honey, fruits, and vegetables all have simple carbohydrates, but grains and legumes all include complex carbohydrates. Most carbohydrates provide energy by being broken down into glucose, which is used by the body as fuel. However, there are some exceptions to this. For example, some carbohydrates, such as dietary fiber, are not fully broken down by the human body and therefore do not provide energy in the same way as other carbohydrates. However, dietary fiber plays an essential role in maintaining good health by promoting the growth of beneficial bacteria in the colon, which can improve digestion and reduce the risk of certain diseases. Additionally, some carbohydrates, such as sugar alcohols, are only partially absorbed and utilized by the body for energy [11].

Fats are macronutrients that are important for the absorption of fat-soluble vitamins and play a role in hormone production [12]. While dietary fats contribute to the body's energy supply, it is important to note that adipose tissue, which stores fat, is not directly equivalent to dietary fat. Adipose tissue serves as a storage site for excess energy in the form of triglycerides, which can be derived from dietary fats as well as other sources. Dietary fats consist of fatty acids, which can be categorized as saturated or unsaturated. Saturated fats are commonly found in animal products such as meat and dairy, while unsaturated fats are predominantly present in plant-based sources like nuts, seeds, and vegetable oils. Research suggests that excessive consumption of saturated fats has been associated with an increased risk of heart disease [13], whereas unsaturated fats are generally considered to be healthier options.

2.2 Metabolomics

An increasing number of research studies has demonstrated that metabolomics appear to be a possible objective tool to identify habitual intake of meat and other animal products in healthy subjects adhering to a vegan, vegetarian, or omnivore diet [14], [15], [16]. By studying metabolomics, we can identify the source and the

amount of protein intake.

Metabolomics is the study of small molecules or metabolites that are present in a biological sample [17]. The study of metabolomics is important in understanding the complex interactions between diet and health. A metabolite profile in the blood can reflect the metabolic state of an individual and provide insight into how different dietary patterns impact their health.

Proteins are broken down into amino acids during digestion. The metabolites produced during protein metabolism can be measured in blood, providing a snapshot of an individual's dietary protein intake. Carbohydrates and fats also have specific metabolic pathways and produce metabolites that can be measured, providing a broader picture of an individual's overall dietary intake.

The ability to measure and analyze metabolites in blood will lead to advancements in the field of personalized nutrition. By identifying specific metabolites associated with certain dietary patterns, individuals can be provided with personalized nutrition recommendations tailored to their unique metabolic profile. This has the potential to improve overall health outcomes and prevent chronic diseases associated with poor dietary habits.

2.3 Nuclear Magnetic Resonance (NMR)

NMR spectroscopy, a powerful analytical technique, is employed to investigate the chemical and physical properties of molecules [18]. It is based on how specific atomic nuclei, such as those of hydrogen, carbon, and nitrogen, interact with a magnetic field. These nuclei have the ability to both absorb and release radiofrequency radiation when put in a high magnetic field. The intensity of the magnetic field and the atoms surroundings affect the frequency of radiation that is absorbed.

In NMR spectroscopy, a sample is placed in a strong magnetic field, typically generated by a superconducting magnet. After that, radiofrequency radiation is applied to the sample, usually in the form of a pulse. The samples nuclei take in the radiation, and as a result of this absorption, they align with the magnetic field. The nuclei relax back to their initial condition once the pulse is switched off and begin to release radiofrequency radiation. A sensitive antenna picks up this radiation, and the signal is processed to create an NMR spectrum [19].

The chemical and physical characteristics of the molecules in the sample are revealed by the NMR spectrum. The various types of nuclei in the sample and their surroundings are represented by the peaks in the spectrum. The frequency of radiation absorbed, which is correlated with the magnetic field and atomic environment, determines the position of the peak. The amount of nuclei in that environment has an impact on the peaks strength. We can ascertain the kind, number, and structure of the molecules in the sample by examining the NMR spectra [18].

The study of metabolism holds an interest in a number of NMR techniques we can see in Figure 2.1, including metabolomics analysis, metabolite identification

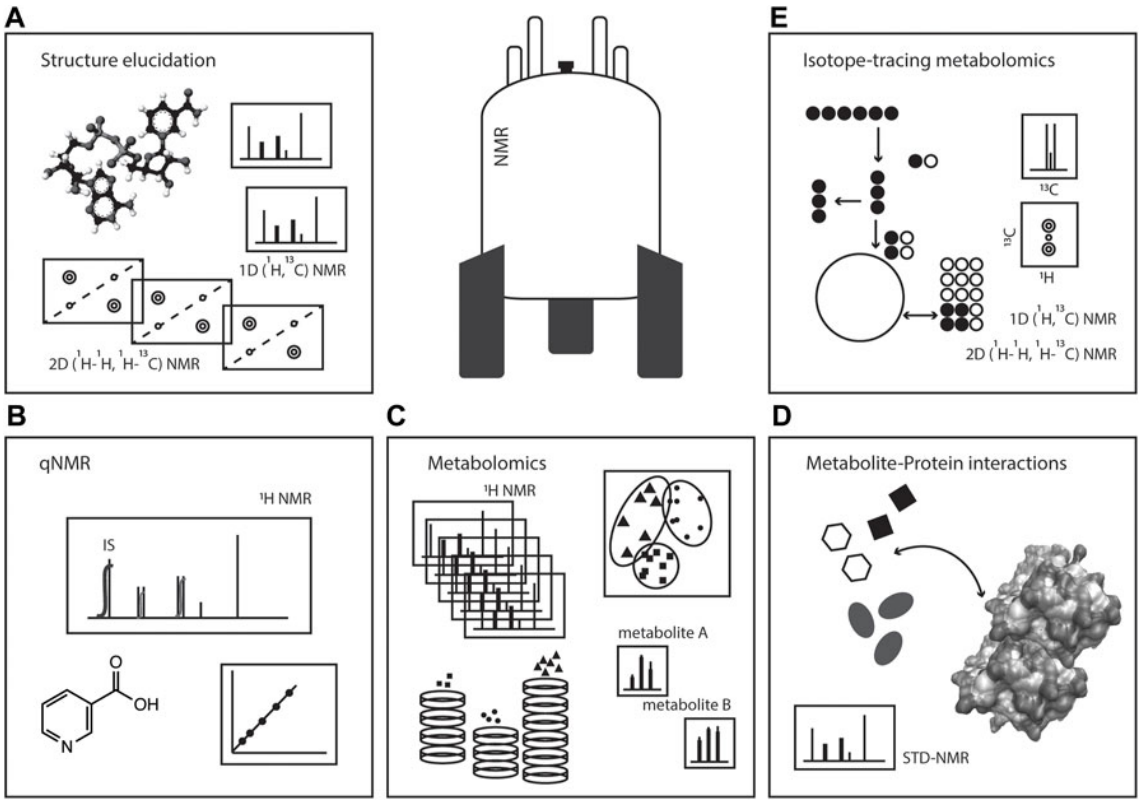

Figure 2.1: NMR spectroscopy: a toolset in metabolism studies. Pictorial representation of the various ways NMR spectroscopy can be used in metabolic studies, such as (A) structure elucidation, (B) quantitative NMR (qNMR), (C) metabolomics, (D) metabolite-protein interactions, and (E) isotope-tracing metabolomics or stable isotope resolved metabolomics (SIRM) [19].

and structure elucidation, metabolite quantification (qNMR), the use of stable isotopes in metabolism investigations, and metabolite-protein interactions, among others. NMR is a versatile spectroscopy that can be used to answer questions about metabolism in a variety of biological systems. This spectroscopy can also help to clarify fundamental biochemical concepts such as metabolite identification, quantification, and turnover, metabolic activities, organelle compartmentalization, and metabolite interactions with macromolecules for enzymology or regulatory events [19]. Although there are different methods for metabolomics analysis, such as LC-MS, NMR spectroscopy offers its own unique advantages [20]. The method used in this study is 1H-NMR.

To provide a visual representation of NMR spectroscopy, Figure 2.2 displays a spectrum along with identified metabolites. This figure illustrates the spectral peaks corresponding to different metabolites, allowing a clearer understanding of the analysis for readers who may be less familiar with spectroscopy.

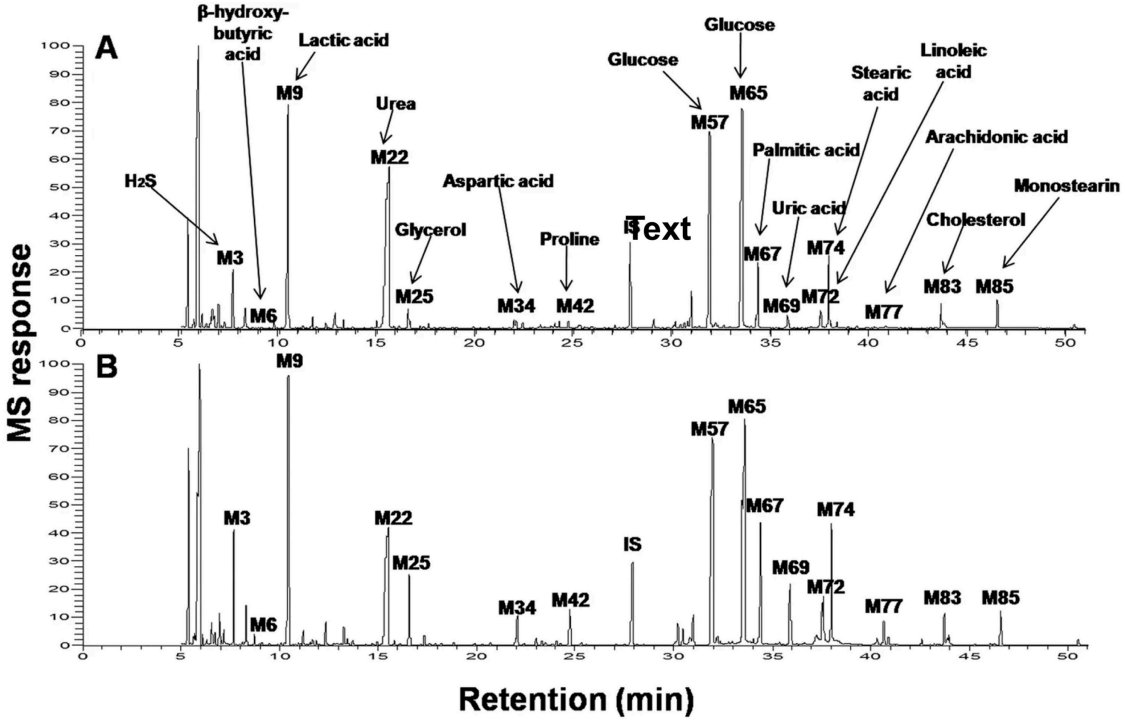

Figure 2.2: NMR spectrum with identified metabolites. A visual representation of an NMR spectrum showing the spectral peaks corresponding to different identified metabolites. This illustration helps to enhance understanding and interpretation of NMR spectroscopy [21].

2.4 Statistical Methods

2.4.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a widely used statistical technique for reducing the dimensionality of high-dimensional data by identifying and extracting the most important features that capture the majority of the variance in the data [22]. PCA is a linear transformation method that transforms the original data into a new set of variables, called principal components, which are linear combinations of the original variables. Each successive principal component effectively captures the largest possible part of the remaining variance, since they are orthogonal and uncorrelated [23]. The coefficients of this linear combination are called "loadings", and they represent the contribution of each variable to that particular principal component. PCA decreases the dimensionality of the data by projecting it onto a lower-dimensional space and identifying the directions in the data that hold the greatest information.

PCA (Principal Component Analysis) is a valuable tool in the analysis of metabolomics data, which typically involves numerous correlated variables (metabolites) [19]. It makes it possible to spot data patterns and trends that may not have been immediately obvious in the original dataset. Additionally, PCA makes feature selection easier by highlighting the metabolites that have the greatest impact on the data's overall variability. This is particularly helpful in metabolomics studies if there are more metabolites than samples. It is feasible to identify the main metabolites in charge of the variations across samples by using PCA to reduce the dimensionality of the data. This method can be useful in locating possible biomarkers linked to particular medical disorders or dietary changes. However, it is crucial to make sure that the findings of PCA are carefully interpreted, taking into consideration the constraints and probable causes of data variance. To use PCA's capabilities in metabolomics analysis successfully, other adjustments and considerations could be required.

2.5 Machine Learning

Machine learning is a field of computer science and artificial intelligence that focuses on the development of algorithms and statistical models that enable computers to automatically learn from and make predictions or decisions based on data. In essence, machine learning is the process of training computer programs to recognize patterns and make decisions based on input data.

Supervised learning and unsupervised learning are the two primary categories of machine learning. Other categories of machine learning are semi-supervised learning and reinforcement learning.

2.5.1 Supervised Learning

In supervised learning, a model is trained using data that has already been classified or labeled with the desired outcome. The model then applies these labeled data to fresh, unobserved data to generate predictions or categorize them [24].

2.5.2 Unsupervised Learning

Unsupervised learning, in contrast, entails developing a model using a dataset that has not been categorized or labeled. In order to uncover significant insights or groups, the model then independently searches the data for patterns or structure without using any pre-existing labels [25].

2.5.3 Semi-supervised Learning

Between supervised and unsupervised learning is semi- supervised learning. During the training phase, it blends a small quantity of labeled data with a big amount of unlabeled data and utilizes context to spot data trends [26]. This technique, for instance, can be applied to classification situations when the solution calls for a supervised learning algorithm but little labeling. Due to the fact that it uses a combination of labeled and unlabeled data, it is quicker than supervised learning. Generative models, low-density separation, Laplacian regularization, and heuristic methods are a few examples. There are not many reported applications for this method in the field of metabolomics.

2.5.4 Reinforcement Learning

Unsupervised ML was guided using the reinforcement learning approach, which rewards good behavior and penalizes undesirable behavior. The model's capacity to link desired inputs and outputs is strengthened by positive feedback [26]. In a number of fields, including game theory, operations research, and swarm intelligence, reinforcement learning has drawn a lot of interest. It provides a potential foundation for training models to make the best choices possible, depending on feedback from the outside world. We can speed up learning, improve the efficiency of our models in identifying complex patterns, and optimize results by using reinforcement learning.

2.5.5 Applications of Machine Learning

Machine learning is becoming increasingly important in many fields, including medicine, finance, and marketing, where large amounts of data are collected and analyzed to extract useful information and make predictions or decisions. Due to the enormous amount of data produced by high-throughput analytical methods, machine learning has also recently grown to be a crucial tool in metabolomics research [27]. This data may be analyzed and interpreted using machine learning algorithms, and predictions based on the found patterns can then be made. One of machine learning's main benefits is its capacity for handling massive, complicated data sets and learning from them, enabling more precise predictions and insights. In this thesis, we have used

machine learning techniques to predict the dietary source of protein intake from blood metabolomics data.

2.5.6 Machine Learning Algorithms

2.5.6.1 Random Forest

Random forest is a popular machine learning algorithm used for classification, regression, and other tasks. It leverages an ensemble approach, where multiple decision trees are created, and their predictions are combined to provide more accurate outcomes [28]. In this context, a decision tree is a tree-like model that recursively partitions the data based on selected features, aiming to achieve purity or a specific level of impurity. A random selection of characteristics and data samples from the original dataset is used to build each decision tree. The process of merging the predictions from each of these separate trees to get the final prediction is known as the ensemble approach. The random forest technique can capture a larger range of patterns and enhance overall performance by combining predictions from many trees.

Measures like Gini impurity or information gain are frequently used to evaluate the quality of the splits inside the decision trees. A decision tree node's impurity or homogeneity may be measured using the Gini impurity metric. It calculates the likelihood that a randomly selected element in the node will be classified incorrectly [29]. A purer node where the majority of the components fall within the same class or category is one with a lower Gini impurity score. Similarly, node impurity refers to the impurity or heterogeneity of a specific node in a decision tree.

In a random forest, a voting mechanism is used to integrate the forecasts of many decision trees. Each tree makes a forecast, and the prediction with the highest percentage is chosen as the outcome [28]. This method enhances the model's accuracy while reducing overfitting. With high-dimensional datasets, random forest can still maintain accuracy while handling missing data.

Random forest include knowledge regarding feature significance as one of its primary benefits. This is useful for locating a dataset's most crucial characteristics and learning more about the underlying data. Additionally, random forest is scalable, has high computational efficiency, and can handle both categorical and numerical data.

However, it might not work well when there are correlated features and can be sensitive to noisy data. Correlated features can cause problems including duplicated data, a decline in decision tree variety, and a higher chance of overfitting [30]. When characteristics are closely connected, the model could give them similar weights, which could distort judgment and prevent precise predictions. Additionally, linked features might restrict the variety of decision trees, making it more difficult for the model to fully capture the range of patterns in the data. On the other hand, noisy data might result in overfitting, cause inconsistencies, and affect evaluations of feature significance. Noisy data might result in inaccurate fluctuations that affect decision rules and reduce the model's overall reliability and accuracy. Correlated

characteristics and noisy data must be addressed and taken into consideration in the preprocessing step in order to guarantee the optimal performance of a random forest model.

In general, random forest is an effective machine learning technique that may be used to solve many different types of issues. It has been used in many different industries, including biology, banking, and image recognition.

2.5.6.2 Support Vector Machine

Support Vector Machines (SVMs) are a type of supervised learning algorithm that can be used for both classification and regression tasks [31]. SVMs are based on the concept of finding the best hyperplane that separates different classes of data. The ideal boundary between classes is chosen to be the hyperplane with the largest distance from the closest data points. In other words, SVMs look for the decision boundary that maximizes the margin or distance between the data points and the decision border between classes of data [32].

SVMs work by converting the input data into a high-dimensional feature space, which makes it easier to distinguish between the data points. A kernel function is used to execute this transformation, mapping the input data from the original space to the higher-dimensional space [32]. The SVM algorithm then looks for the hyperplane that best separates the data points once the data has been processed.

In comparison to other classification algorithms, SVMs have a number of advantages, including the ability to handle high-dimensional data and the ability to handle non-linearly separable data using various kernel functions [33]. SVMs have also been demonstrated to work well in a wide range of applications and have a solid theoretical underpinning.

In the context of metabolomics, SVMs have been used to classify samples based on their metabolic profiles, such as distinguishing between different disease states or identifying different types of food intake [5]. SVMs have also been applied to feature selection to determine the most crucial metabolites for differentiating between various sample groups.

2.5.6.3 Neural Networks

Neural networks are a class of machine learning algorithms that are modeled after the structure and function of the human brain. A neural network is made up of a number of linked nodes, or neurons, that process and send information. A neural network's architecture generally consists of an input layer, one or more hidden layers, and an output layer, as shown in Figure 2.3. The connections between the neurons are weighted according to the strength of the relationships between them and connect neurons in each layer to neurons in the layer above them.

Neural networks are commonly used in supervised learning tasks, where the model is trained on labeled data to make predictions or classifications on new, unseen data. To reduce the difference between the projected output and the actual output, the

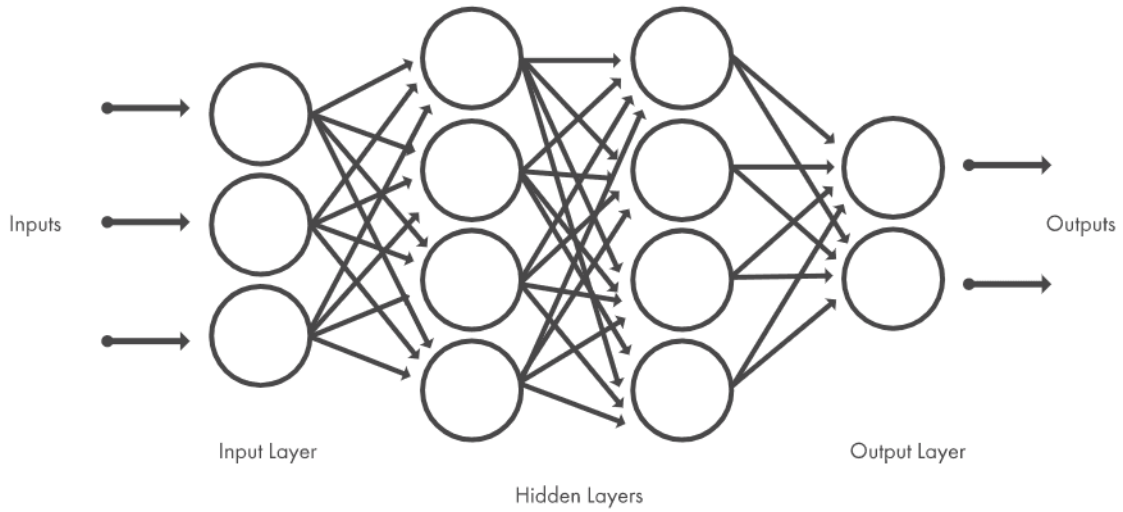

Figure 2.3: Neural networks, which are set up in layers and comprise a collection of linked nodes. Tens or even hundreds of hidden layers are common in networks. [34].

weights of the connections between neurons are altered during learning. The error between the projected output and the actual output is calculated in this procedure, which is also known as backpropagation, and it is then transmitted back through the network to change the weights [35].

Neural networks may be used for unsupervised learning tasks in addition to supervised learning, where the model is trained on unlabeled data to find patterns and correlations in the data. Unsupervised learning is frequently used for tasks like dimensionality reduction and clustering.

In a variety of applications, including computer vision, natural language processing, and speech recognition, neural networks have shown to be quite successful [36]. They may, however, be computationally expensive and need a lot of data to train well.

2.5.7 Cross Validation

In machine learning, the cross-validation approach is used to evaluate a model's performance. Cross-validation's fundamental idea is that the model should be trained and tested using several subsets of the data. A model is trained and tested k times using a distinct subset as the test set and the remaining subsets as the training set in k-fold cross-validation. The data is separated into k subsets [37].

The problem of overfitting, which can happen when a model is too complicated and catches noise in the training data, is helped by the use of cross-validation. We may obtain a more accurate approximation of the model's generalization performance by assessing the model's performance on various subsets of the data.

3

Methods

In this section, a description of the project's process is presented. The procedures employed are explained, along with an elucidation of the reasoning behind the decision-making process.

3.1 Data Collection

The dataset used in this project was provided by the Department of Internal Medicine and Clinical Nutrition at Gothenburg University. The data were collected as part of a study aimed at identifying techniques for measuring habitual dietary exposure [14]. Prior to this project, the dataset underwent extensive preprocessing steps to ensure its suitability for analysis. These steps are described in the following section.

The study enrolled 120 healthy volunteers, including 45 men and 75 women, who complied with habitual vegan, (lacto- ovo-)vegetarian, or omnivore diets [14]. Additionally, a fourth group of pescetarians was included.

Volunteers were recruited through advertisements and were considered suitable if they were between 18 and 65 years old, healthy, and had a BMI between 18 and 30. The screening process included clinical markers (to exclude participants with diseases), a short lifestyle questionnaire, a 4-day weighed food diary, and a food-frequency questionnaire (FFQ) developed specifically for the study. The FFQ included questions about food intake related to soy or soy products, legumes, vegetables, fruit and berries, milk products, eggs and egg-based foods, fish and shellfish, poultry, red meat, and cookies and confectionery [38].

Body composition was measured with bioimpedance analysis (ImpediMed Bioimp version 5.3.1.1), and volunteers who were pregnant, lactating, or who used nicotine products regularly were excluded. Participants provided written informed consent before entering the study. Serum samples were collected and analyzed for metabolites using 1H-NMR spectroscopy. The study was conducted in two periods, from April to May 2013 and from August to December 2015. Volunteers were not allowed to drink alcohol the night before sampling or consume food supplements 1 week before sampling [38], [14].

3.1.1 Unidentified Metabolites

It is significant to note that certain metabolites were either left undetermined or were only tentatively assigned during the examination of the blood samples using 1H-NMR spectroscopy. The presence of unidentified metabolites can stem from several factors inherent to metabolomics studies.

The complexity of the metabolomic profile and the limits of the available reference databases are two aspects that contribute to unidentified metabolites. There are still certain metabolites that have not been properly defined or included in the existing databases despite significant attempts to create comprehensive databases. As a result, it might not be possible to link particular NMR peak spectrum patterns to known metabolites.

Another factor is the lack of accurate mass spectral libraries for comprehensive identification. Mass spectral libraries are essential for metabolite identification because they make it possible to compare experimental and reference spectra. However, reliable identification may be limited by the lack of reference spectra for specific metabolites or by spectrum variability brought on by circumstances unique to a given sample.

3.2 Pre-processing

In the pre-processing section, the steps taken to prepare the dataset for analysis are described.

3.2.1 Missing Values

The first step in the data preprocessing was to check for missing values. As the dataset had been previously used for research purposes, we checked whether missing values had been addressed by the previous researchers [14], [38]. No missing values were found in the dataset we received for this study.

3.2.2 Scaling

The next step in the data preprocessing was to perform Unit Variable (UV) scaling, which is a widely used scaling method in spectroscopy, including NMR spectroscopy. It can be difficult to compare spectra between samples because to variances in signal intensities caused by variations in sample concentrations, which are addressed via UV scaling. Each data point is divided by the square root of the corresponding feature's standard deviation in order to address this problem. We adjust the data for variations in sample concentration using UV scaling, making it possible to compare and analyze NMR data in a more insightful manner.

One reason why UV scaling is a popular choice for NMR data is that it does not affect the shape of the spectra or the relative intensities of the signals [39]. This means that the scaling does not distort the data in any way, which can be important for subsequent analyses, such as feature selection and modeling.

3.3 Statistical Methods

We chose to use Principal Component Analysis (PCA) in this thesis project. PCA is a widely used method for reducing the dimensionality of data while retaining the most important information. It has been successfully applied in many different fields, including metabolomics [40]. In this study, we used PCA to select the most relevant features from the metabolomics dataset, which helped to reduce the dimensionality of the data and identify the most important variables. PCA is a powerful tool that can help to identify trends and patterns in large datasets.

In order to perform the PCA, the Python library scikitlearn was used, and the function PCA() was called with n_components = 4, which specifies that we want to keep the top 4 principal components [41]. The data were then fitted using the PCA function, which created a new dataset made up of the chosen principle components and was put in a new dataframe dubbed pca_metabolomics.

To visualize the results of the PCA, a 3D scatter plot was created using the matplotlib library [42], with each point in the plot representing a sample in the dataset. The x, y, and z coordinates of each point corresponded to the values of the first three principal components, respectively. Different colors and markers were used to distinguish between samples belonging to different classes.

After performing the PCA, the most important features in the dataset were identified by examining the absolute values of the components using the print statement at the end of the code. This step helped to identify which variables were contributing the most to the separation between the different sample classes in the PCA plot. The purpose of this step was to select and retain only the most influential features, thereby reducing the dimensionality of the dataset. This reduction can increase computational effectiveness and assist in focusing the emphasis on the significant variables that contribute to the observed patterns and changes in the data.

3.4 Machine Learning

3.4.1 Random Forest

Random forest is a popular machine learning algorithm used for classification tasks. It is an ensemble method that constructs multiple decision trees and combines their results to improve accuracy and prevent overfitting [28]. In this study, we utilized the scikitlearn library in Python to implement a random forest model.

The first step in building a random forest model is to split the dataset into training and testing sets. We used the train_test_split function from scikitlearn to randomly split the dataset into 80% training data and 20% testing data [43]. We set the random state to 42 to ensure reproducibility. Setting a specific random state value ensures that the same random sequence is generated each time the algorithm is run, making the results consistent and reproducible. The choice of the number 42 is arbitrary, and any other integer value could have been used instead.

Next, we defined the random forest classifier and created a parameter grid using the param_grid dictionary. The n_estimators parameter specifies the number of decision trees to be used in the random forest. The max_depth parameter controls the maximum depth of the decision trees. The min_samples_split parameter specifies the minimum number of samples required to split an internal node. The min_samples_leaf parameter specifies the minimum number of samples required to be at a leaf node. Finally, the max_features parameter specifies the maximum number of features to be considered when splitting a node.

We used the GridSearchCV function to perform a grid search over the parameter grid to find the best combination of hyperparameters [44]. The CV parameter specifies the number of cross-validation folds to be used. We used 5-fold cross-validation to obtain reliable estimates of the model's performance. In the specific case of using 5-fold cross-validation with the Random Forest model, it means that we divided the data into 5 subsets, and trained the model on 4 subsets while using the remaining subset as the test set. We repeated this process 5 times, each time using a different subset as the test set. This allowed us to obtain a more reliable estimate of the model's performance by averaging the performance metrics over the 5 test sets.

After finding the best hyperparameters, we created a new random forest classifier with the optimized hyperparameters and fit it to the training data using the fit function. We then used the fitted model to predict the classes of the testing data using the predict function.

To evaluate the performance of the random forest model, we calculated the confusion matrix and the classification report. The confusion matrix shows the number of true positive, true negative, false positive, and false negative predictions as shown in Table 3.1. The classification report provides metrics such as precision, recall, support and F1-score for each class. Accuracy, precision, recall, support and F1 score are commonly used metrics to evaluate the performance of a classification model. These metrics are calculated using the confusion matrix.

Table 3.1: Confusion Matrix

		Actual Class	
		Positive	Negative
Predicted Class	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Accuracy is the most commonly used metric, and it represents the proportion of correct predictions among all predictions. However, it can be misleading in cases where the classes are imbalanced, meaning that one class has many more observations than the other. For example, if a model is predicting whether a patient has a

rare disease or not, and 99% of the patients do not have the disease, a model that always predicts "no disease" will have an accuracy of 99%, but it is not useful in practice [45].

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{3.1}$$

Precision represents the proportion of true positives among all predicted positives. It is a useful metric when the cost of false positives is high [45]. For example, consider a predictive model designed to identify the presence of a particular medical condition in a patient, based on a set of observable symptoms. In this case, a false positive (a patient who is predicted to have the condition but actually does not) could lead to unnecessary medical procedures, such as surgery or medication, that can have negative side effects on the patient's health. On the other side, a false negative (a patient who is expected not to have the illness but really has) might cause a delay in treatment, which may result in the disease progressing and potentially even death. In order to guarantee the accuracy of the model's predictions, precision is a critical parameter in medical diagnosis.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{3.2}$$

Recall represents the proportion of true positives among all actual positives. It is a useful metric when the cost of false negatives is high. For example, in the case of a model that predicts whether a patient has a disease or not, a false negative (a patient that has the disease but is classified as healthy) can be life-threatening [45].

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.3}$$

The support metric, commonly referred to as the sample size or the number of instances, offers important insights into how classes are distributed across datasets. It displays the proportion of observations in the dataset that correspond to each type. The support indicator is crucial for assessing classification performance since it determines how reliable the outcomes are. A larger number of cases for a given class are indicated by a higher support value, indicating a more accurate and trustworthy evaluation. A lower support value, on the other hand, denotes a fewer number of instances, which might result in less accurate estimations of performance indicators. The samples included in the test set are represented by the support numbers in the evaluation metrics, not the dataset's overall sample size. This distinction results from the division of the data into training and testing subsets for the purpose of evaluating the effectiveness of the classification model. As a result, the evaluation metrics' support numbers may look lower than the dataset's overall sample sizes.

F1 score is the harmonic mean of precision and recall, and it provides a balance between the two metrics. It has a scale from 0 to 1, with 1 denoting optimal performance, which minimizes both false positives and false negatives. When both

types of errors are significant and must be taken into account when assessing the performance of the model, the F1 score can be particularly helpful [45].

$$\text{F1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.4}$$

Precision, recall, and F1-score may still be calculated and used to assess the model's performance while making predictions for many classes. However, as compared to binary categorization settings, their interpretation becomes more complex.

When using multiple classes, accuracy, which measures the percentage of true positives among all occurrences predicted to fit in a given class, is determined separately for each class. Recall, often referred to as sensitivity, is determined for each class and denotes the percentage of cases that really fit in a given category out of all occurrences. The F1-score can also be calculated for each class using the harmonic mean of precision and recall.

Overall, we chose to use random forest because it is a powerful and flexible algorithm that can handle high- dimensional data and nonlinear relationships between features and target variables. The use of grid search and cross- validation allowed us to find the optimal hyperparameters and obtain reliable estimates of the model's performance.

3.4.2 Support Vector Machine

To implement SVM, we first split the data into training and testing sets using an 80-20 ratio [43]. SVM is a binary classification algorithm by default, meaning it separates data into two classes. However, it can be extended to handle multiclass classification tasks through techniques like one-vs-rest or one-vs-one, where multiple binary classifiers are trained to distinguish each class from the rest [46]. We then normalized the data using the StandardScaler function, which standardizes the data by removing the mean and scaling to unit variance. Next, we defined the SVM model using the Support Vector Classification (SVC) function from the scikitlearn library [41].

To find the best hyperparameters for the SVM model, we used a grid search with cross-validation. We defined a hyperparameter grid with different values for the regularization parameter C, the kernel function, and the gamma parameter. The regularization parameter (C) determines the trade-off between achieving a low training error and a low complexity model. To investigate various levels of regularization, we took into account a range of C values, namely [0.001, 0.01, 0.1, 1, 10]. In addition, we investigated the linear, RBF, and polynomial kernel functions. The gamma parameter regulates how much each training sample has an impact on the decision boundary. For gamma, we took into account the values [0.001, 0.01, 0.1, 1, 10] to evaluate the effects of various degrees of influence. It is important to note that the selection of hyperparameters is influenced by the particular dataset and issue at hand. The grid search performed a five-fold cross-validation to evaluate the model

with each combination of hyperparameters and returned the set of hyperparameters that yielded the best performance on the training data.

We then trained an SVM model with the best hyperparameters on the training data and evaluated its performance on the testing data using the classification_report function from scikit learn. The classification report provides a summary of the precision, recall, and F1 score for each class, as well as the the overall accuracy of the model.

3.4.3 Neural Network

Neural networks are a type of machine learning model that can learn to recognize patterns in data. Specifically, we used a feedforward neural network with three layers: an input layer with 237 nodes, a hidden layer with 64 nodes, and another hidden layer with 32 nodes. The output layer had four nodes, each corresponding to one of the four diet types in the dataset. The neural network is built using the Keras API with a sequential model structure.

The first layer has 64 neurons and uses the rectified linear unit (ReLU) activation function, which is commonly used in neural networks for its ability to handle non-linearity. The input dimension of this layer is set to 237, which is the number of features in the input data.

The second layer has 32 neurons and also uses the ReLU activation function. This layer is followed by a dropout layer, which helps prevent overfitting by randomly dropping out some of the neurons during training [47].

The final layer is a dense layer with 4 neurons, which corresponds to the number of classes in the output data. This layer uses the softmax activation function, which is commonly used in multi-class classification problems.

The model is compiled using the categorical cross-entropy loss function, which is commonly used for multi-class classification problems. The optimizer used is Adam, which is an adaptive learning rate optimization algorithm that is commonly used in deep learning [48].

During training, the model is fed with mini-batches of 32 data points and trained over 100 epochs. The performance of the model is evaluated using the accuracy metric, which measures the proportion of correctly classified samples.

Overall, the neural network architecture we used is a relatively simple feedforward network with three dense layers. The ReLU activation function is used in the hidden layers to introduce non-linearity, and the softmax activation function is used in the output layer to predict the probabilities of the different classes [49].

3.5 Challenges and Limitations

The present study acknowledges certain limitations inherent in the methodologies employed for the execution of this thesis project.

One limitation of PCA is that it assumes a linear relationship between the variables, which may not be true for metabolomics data. Complex interactions between metabolites can provide non- linear connections between variables. When using PCA, this may lead to the loss of crucial data. In addition, PCA may be sensitive to outliers in the data, which could affect the findings and provide false conclusions. In metabolomics, PCA has been used to identify biomarkers for various diseases and conditions, including cancer, diabetes, and obesity. However, the use of PCA in metabolomics has been criticized for its limitations, and alternative methods such as Partial Least Squares (PLS) have been proposed as more suitable alternatives for certain types of data [50].

Random Forest is a popular machine-learning algorithm that is used in metabolomics to predict the metabolite concentrations of unknown samples. However, one of the main limitations of random forest is that it is prone to overfitting, especially when the number of features is large [51]. This can be a problem in metabolomics, where there are often many thousands of metabolites that can be measured. Additionally, random forest can be computationally expensive, especially when there are many trees in the forest.

Support Vector Machines (SVMs) are another popular machine learning algorithm that can be used for metabolomics data analysis. One of the limitations of SVMs is that they can be sensitive to the choice of kernel function. Additionally, SVMs can be sensitive to outliers, which can be a problem in metabolomics where there may be systematic errors in the measurement of metabolite concentrations. Finally, SVMs can be computationally expensive, especially when there are many samples and/or features [51].

Neural Networks is a powerful machine learning algorithm that can be used for various applications, including metabolomics. However, one of the main limitations of neural networks is that they are prone to overfitting, especially when there are many parameters to be learned. In metabolomics, neural networks can also be limited by the availability of large datasets, as they require large amounts of data to be trained effectively. Additionally, neural networks can be computationally expensive, especially when there are many layers and/or neurons in the network.

In conclusion, the use of algorithms such as PCA, random forest, SVMs and neural networks in metabolomics has both advantages and limitations. While these algorithms can provide valuable insights into complex biological systems, they can also be limited by their assumptions, computational requirements and sensitivity to different types of data. As such, it is important for researchers to carefully consider the strengths and limitations of these algorithms when analyzing metabolomics data, and using a range of approaches in order to obtain the most comprehensive insights possible.

3.6 Objective

The primary objective of this Master's thesis is to develop and evaluate a machine learning model that can accurately predict the type of dietary protein consumed

by an individual based on their blood metabolomics data. Specifically, the study aims to investigate whether the consumption of meat, fish, or vegetarian sources of protein can be determined through blood metabolomics data analysis. The ability to accurately assess the type of dietary protein intake is essential for research in various fields, including nutrition, metabolism, and public health. However, traditional methods for assessing dietary protein intake have limitations, such as reliance on self-reported dietary intake or incomplete nutrient databases. Blood metabolomics, on the other hand, offers a promising approach for assessing dietary protein intake, as it allows for the identification of specific metabolites that can serve as biomarkers of protein consumption. Our goal is to identify these biomarkers and develop a machine learning model that can use them to predict an individual's type of dietary protein intake with a high degree of precision. Overall, this thesis aims to advance the field of dietary assessment by providing a novel approach for accurately predicting an individual's dietary protein intake using blood metabolomics and machine learning.

4

Results

In this chapter, we present the results of our research, which aimed to investigate the metabolic differences between different dietary groups.

4.0.1 PCA Results

PCA was employed to visualize the metabolic profile of the dataset.

4.0.1.1 Overall PCA analysis results

The 3D plot (Figure 4.1) does not show a clear separation between the four dietary groups, indicating that the metabolomic profiles of the groups are not distinct enough to be separated by the chosen number of principal components.

4.0.1.2 Gender-specific PCA analysis results

Furthermore, two separate 3D PCA plots were created for men and women, respectively. The plots (Figure 4.2 and 4.3) showed that there is not a clear separation between the two groups in both cases, indicating that there are no clear sex-specific metabolic differences.

4.0.1.3 Meat Consumption-specific PCA analysis

In addition, a 3D PCA plot was generated to compare the metabolic profiles of meat eaters and non-meat eaters. Figure: 4.4 showed a clear separation between the two groups, indicating that meat consumption has an impact on the metabolic profile.

4.0.1.4 Outliers

From the 3D plots of the PCA analysis (Figures 4.1, 4.2, 4.3), it was clear that there were some outliers present in the dataset. To identify and remove these outliers, we used the Z-score method which computes the deviation of each data point from the mean in terms of standard deviation. We then set a threshold of 2.45 standard deviations, which is a commonly used threshold to identify outliers. In total, we identified and removed 10 outliers, 5 men and 5 women. Notably, none of the outliers belonged to the pescetarian group. The removal of outliers improved the quality and accuracy of our analysis, allowing us to draw more reliable conclusions from the data.

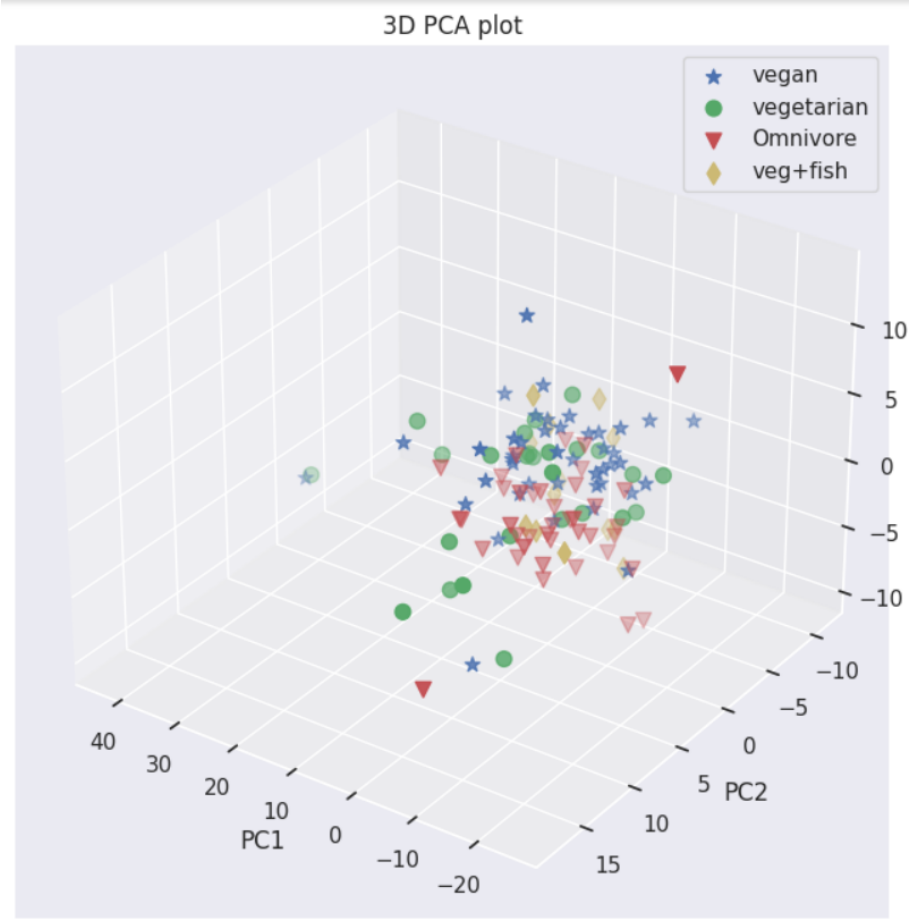

Figure 4.1: 3D plot of the PCA method.

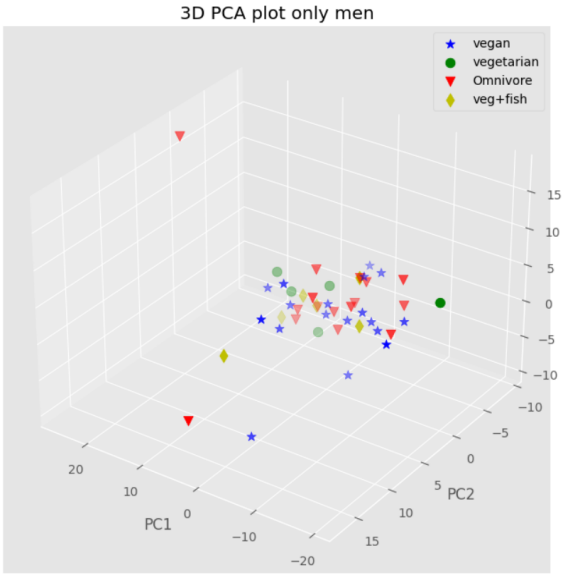

Figure 4.2: 3D plot of the PCA method only for the men samples.

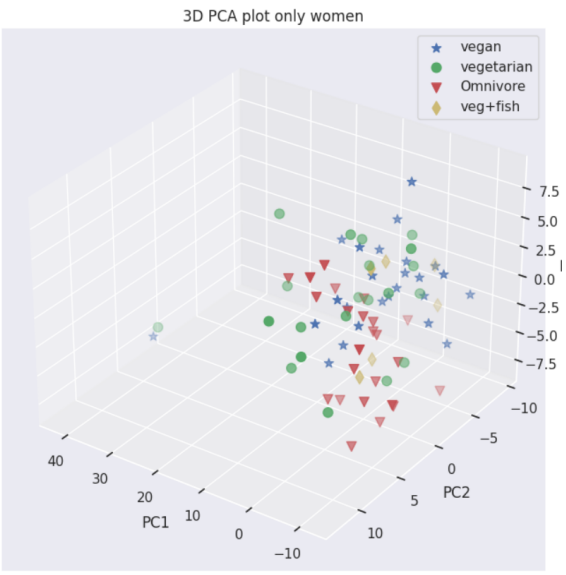

Figure 4.3: 3D plot of the PCA method only for the women samples.

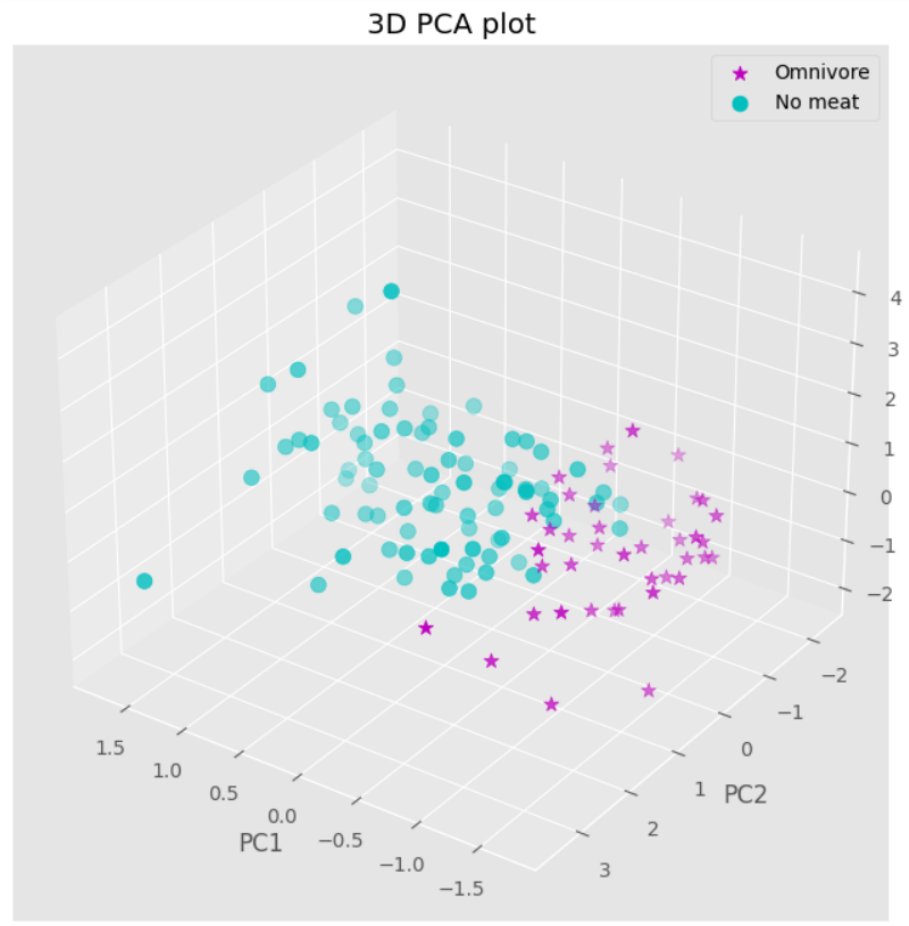

Figure 4.4: 3D plot of the PCA method for meat eaters and non-meat eaters.

4.0.1.5 Most Important Features

To identify the most important features in our dataset, we used principal component analysis (PCA) and extracted the first principal component (PC1). The absolute values of the loadings for each feature on PC1 were then calculated using the 'pca.components_' attribute. We selected the top 100 features with the highest loadings on PC1 as the most important features for our analysis. This allowed us to focus on the most informative variables in our dataset and improve the accuracy and efficiency of our analysis.

In addition to the full dataset, we also performed the same procedure for the datasets containing only men and only women. However, after comparing the results, we decided to keep the most important features found from the full dataset for further analysis. We believe that this approach will provide a more generic and accurate representation of the data, rather than focusing solely on gender-specific patterns.

4.0.2 Classification Models Results

Based on the research question of classifying individuals into four groups (vegan, vegetarian, pescetarian, and omnivore) based on their dietary habits, we applied three popular machine learning algorithms: Random Forest, Support Vector Machines (SVM), and Neural Networks. We attempted multiple tasks with each algorithm. The first task involved classifying each group separately. The second task separated the samples into omnivores and all the other groups together. The third task separated the samples into vegans and all the other groups together. Finally, the fourth task separated the samples into omnivores and pescetarians together and vegans and vegetarians together.

4.0.2.1 First Task: classifying each group separately

After running the Random Forest algorithm to classify each group separately, the results showed an overall accuracy of 57%. Table 4.1 shows the precision, recall, f1-score, and support for each of the four groups in the first task of using random forest for classification. The numbers for the support metric refers to the test samples. Due to that we observe smaller values of support compared with the total samples of the dataset.

Table 4.1: Performance metrics for random forest model (Task 1)

Group	Precision	Recall	F1-score	Support
Vegans	0.58	0.64	0.61	11
Vegetarians	0.00	0.00	0.00	3
Omnivores	0.86	0.75	0.80	8
Pescetarians	0.00	0.00	0.00	1

After analyzing the results, we proceeded to extract the top 20 most important features for this model. Figure 4.5 displays these features in descending order of importance.

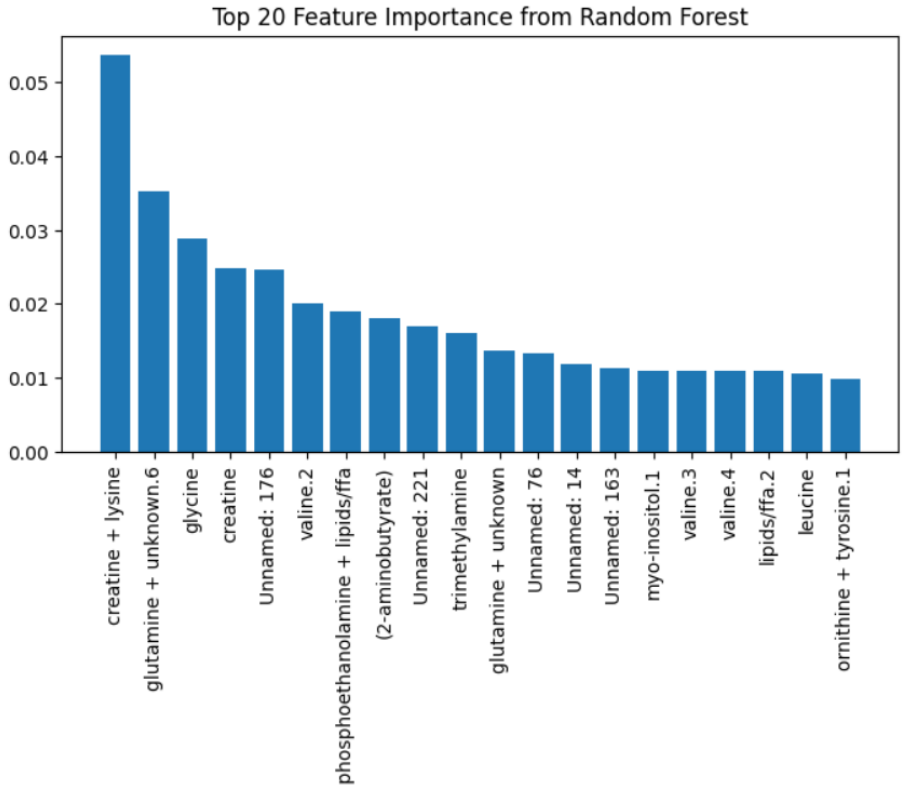

Figure 4.5: 20 most important features from RF model (Task 1). The seventh most important feature is a combination of phosphocholine, acetylcholine, phosphoethanolamine and lipids/ffa, but it is not visible in the figure.

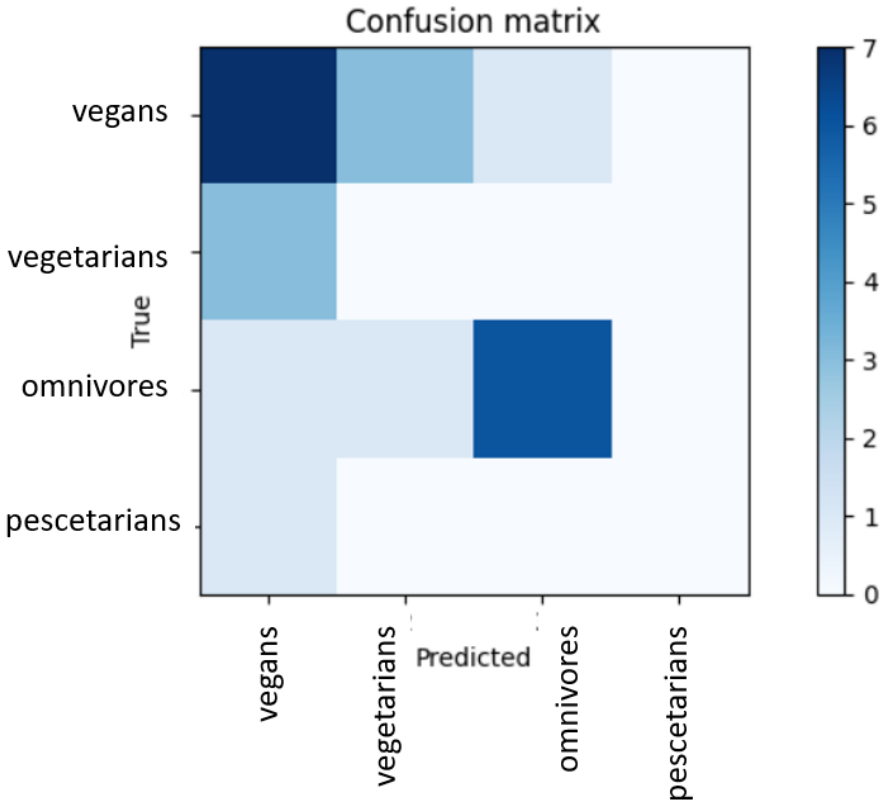

Figure 4.6: Confusion matrix for RF in the task 1.

In addition to the classification results and the feature importance analysis, we also calculated and visualized the confusion matrix (Figure 4.6) for the best-performing random forest classifier in task 1.

Following the random forest classifier, we employed a support vector machine (SVM) model for the same dataset. The obtained accuracy of 74% was recorded, and the performance metrics, namely precision, recall, F1-score, and support, are presented in Table 4.2.

Table 4.2: Performance metrics for SVM model (Task 1)

Group	Precision	Recall	F1-score	Support
Vegans	0.69	1.00	0.81	11
Vegetarians	0.00	0.00	0.00	3
Omnivores	0.86	0.75	0.80	8
Pescetarians	0.00	0.00	0.00	1

In addition to the classification results and the feature importance analysis, we also calculated and visualized the confusion matrix (Figure 4.7) for the best-performing SVM classifier in task 1.

The third classifier that was used for Task 1 is neural networks. The model was trained with 100 epochs and a batch size of 32. After training, we evaluated the model using the test set and obtained an accuracy of 30%. We also calculated

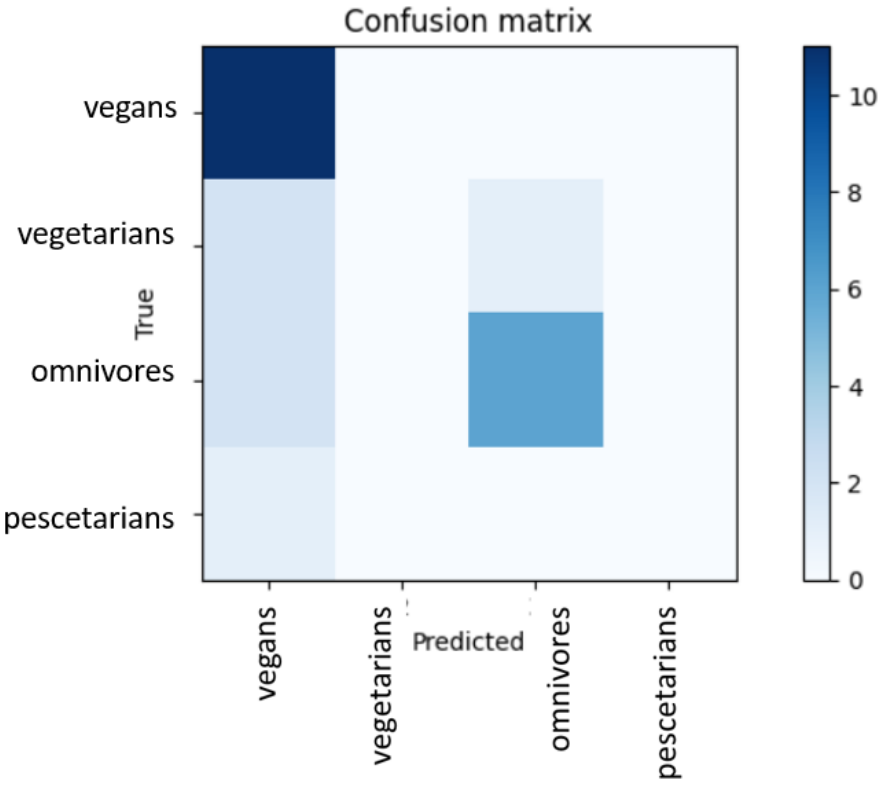

Figure 4.7: Confusion matrix for SVM in the task 1.

precision, recall, F1-score and support for each class. These metrics are shown in Table 4.3

Table 4.3: Performance metrics for neural network model (Task 1)

Group	Precision	Recall	F1-score	Support
Vegans	0.00	0.00	0.00	11
Vegetarians	0.00	0.00	0.00	3
Omnivores	0.32	0.88	0.47	8
Pescetarians	0.00	0.00	0.00	1

4.0.2.2 Second Task: classifying omnivores and a combined group of vegetarians, vegans, and pescetarians

The Second Task involved grouping all non-omnivorous samples together and treating them as a single class while considering omnivorous samples as the other class. This task allowed us to investigate whether the metabolomics data can distinguish between omnivorous and non-omnivorous diets, which could be useful for developing biomarkers of dietary intake and assessing the health effects of different dietary patterns. For this task, we utilized the Random Forest and SVM models to classify the samples.

The accuracy of the best random forest classifier for the second task was 87%. The

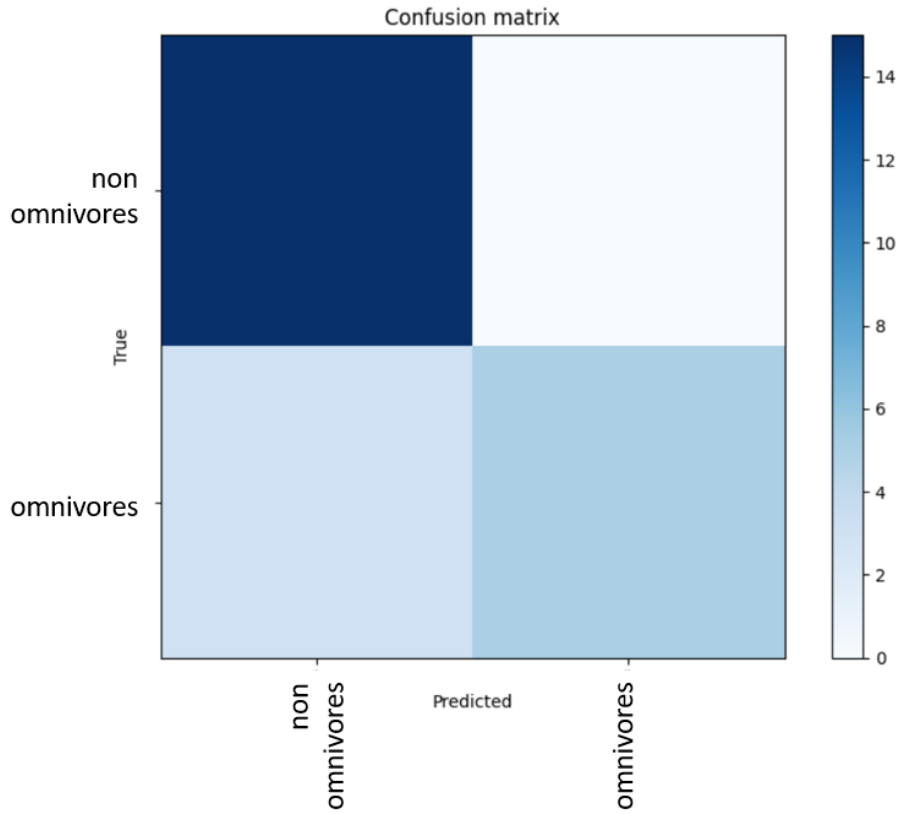

Figure 4.8: Confusion matrix for the RF in the task 2.

other metrics can be seen in Table 4.4 and the confusion matrix is shown in Figure 4.8.

Table 4.4: Performance metrics for random forest model (Task 2)

Group	Precision	Recall	F1-score	Support
Non-omnivores	0.83	1.00	0.91	15
Omnivores	1.00	0.62	0.77	8

In addition to classifying omnivores and non-omnivores, we also wanted to understand which features are the most important for this classification. To achieve this, we used the Random Forest classifier and obtained the feature importance. We then kept the top 50 features with the highest importance scores and used them as input for the SVM model. This allowed us to not only achieve a high accuracy of 87% but also to identify the most relevant features that contribute to the classification. These features and their importance scores can be seen in the Figure 4.9.

Along with the prior classification tasks, we further explored the relationship between dietary patterns and metabolomics data by attempting to predict the amount of meat consumed by individuals. Leveraging the information available in the dataset, which indicated the frequency of meat consumption in the four days preceding the sample collection, we categorized the individuals into three groups: 0 (no meat consumption), 1 (meat consumed 1-4 times), and 2 (meat consumed 5-6 times).

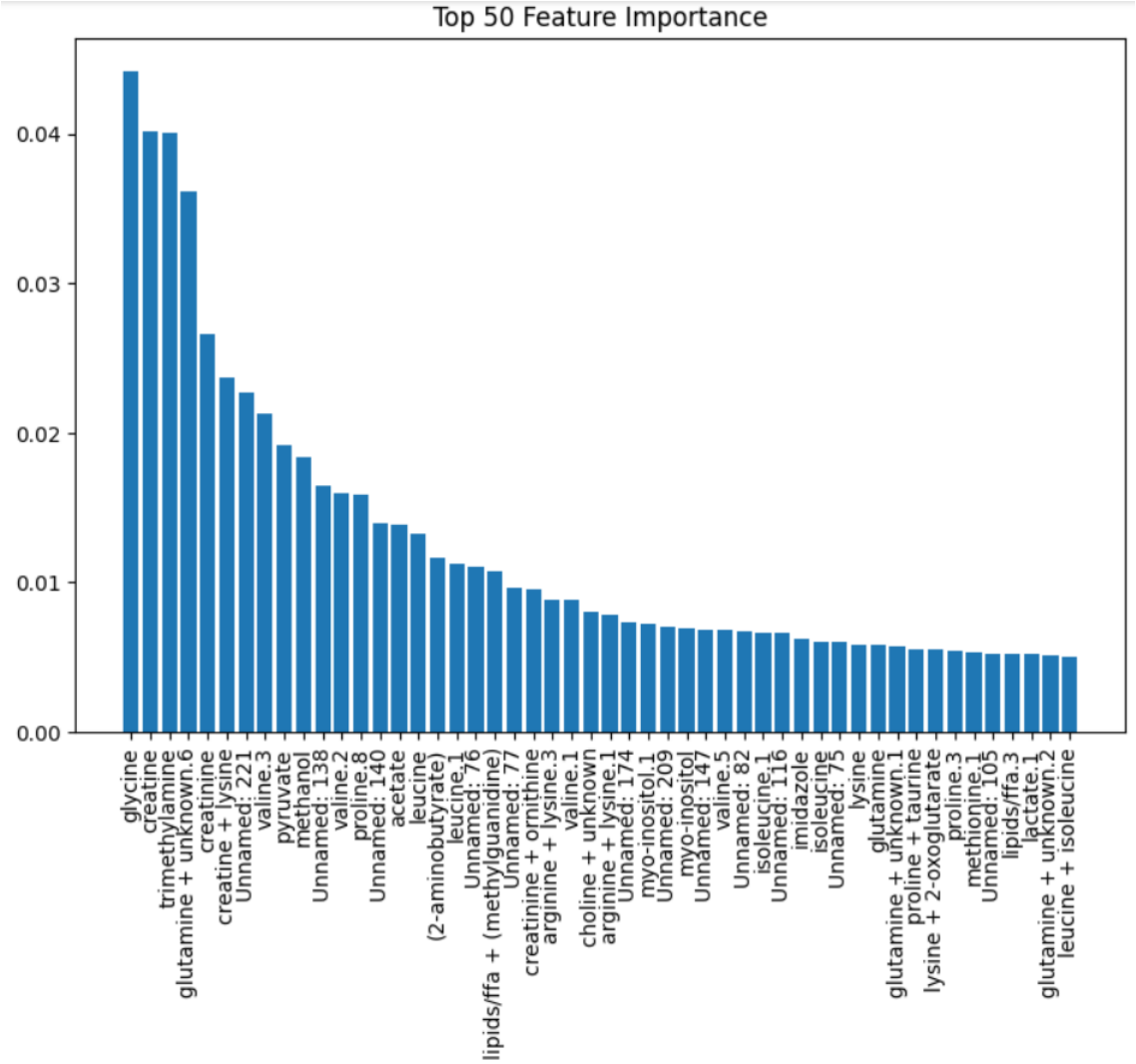

Figure 4.9: The 50 most important features from RF model (Task 2).

While our predictive models were unable to accurately estimate the exact amount of meat consumed, they demonstrated moderate success in distinguishing individuals who abstained from meat consumption (group 0) from those who consumed meat to varying degrees.

For this specific task of predicting the amount of meat consumption, we employed the Random Forest algorithm. The accuracy score achieved by the model was 72%, indicating moderate success in the prediction task. However, it is important to note that the model struggled to accurately predict the amount of meat consumed, particularly for the categories with lower sample sizes. The precision, recall, and F1-score varied across the three groups and can be seen in Table 4.5.

Table 4.5: Performance metrics for RF, meat consumption score.

Group	**Precision**	**Recall**	**F1-score**	**Support**
Group 0	0.68	1.00	0.81	15
Group 1	0.00	0.00	0.00	3
Group 2	1.00	0.20	0.33	5

Group 0 (no meat consumption) achieved the highest precision and recall, while Group 1 (meat consumed 1-4 times) and Group 2 (meat consumed 5-6 times) exhibited lower scores. This indicates that the model was more successful in identifying individuals who abstained from meat consumption, compared to differentiating between different levels of meat consumption. Overall, the results highlight the challenges in precisely estimating the amount of meat consumed based on metabolomics data and emphasize the need for further research.

Continuing our exploration of dietary patterns and their metabolic associations, we proceeded to investigate the consumption of meat, dairy, and eggs among individuals in the dataset. For this particular task, we opted to exclude individuals following a vegan diet from our analysis. Since vegans abstain from consuming meat, dairy, and eggs altogether, their exclusion allowed us to focus specifically on individuals with varying levels of meat, dairy, and egg consumption. By removing vegans from the dataset, we aimed to develop a predictive model that could estimate the amount of consumption among non-vegan individuals.

In our analysis of meat, dairy, and egg consumption, we employed three different scoring methods to quantify the dietary patterns of individuals. The first scoring task involved calculating the cumulative score based on the total number of times an individual reported consuming meat, dairy, and eggs. In the second method, we introduced a scoring adjustment by multiplying the number of times individuals consumed meat by a factor of 1.5. This modification aims to assign a higher score to individuals who consume protein from animal sources, reflecting the potentially greater impact of animal-based protein consumption on metabolic profiles. Lastly, the third task involved utilizing a scoring system derived from previous research.

After calculating the cumulative consumption of these products, we classified individuals into three groups based on their consumption frequency: group 1 represented

individuals who consumed meat, dairy, and eggs 0-4 times, group 2 included individuals who consumed them 5-9 times, and group 3 consisted of individuals who consumed them 10-14 times within a specified period. Our goal was to develop predictive models capable of estimating an individual's group based on their metabolic profile.

Using the Random Forest classifier with the best hyperparameters found during the model optimization process, we obtained an accuracy score of 50.36%. The confusion matrix revealed that the model struggled to accurately classify the individuals into the respective consumption groups. There were misclassifications across all three groups, resulting in low precision, recall, and f1-scores for each group (Table 4.6).

Table 4.6: Performance metrics for RF, meat/dairy/eggs consumption score without vegans.

Group	**Precision**	**Recall**	**F1-score**	**Support**
Group 1	0.00	0.00	0.00	4
Group 2	0.33	0.75	0.46	4
Group 3	0.40	0.33	0.36	6

In the method where we multiplied the score of meat consumption by 1.5 and separated individuals into two groups, we aimed to capture the differential impact of animal-based protein intake on metabolic profiles. The two groups were defined as follows: Group 1 included individuals with a score ranging from 0 to 8, and Group 2 consisted of individuals with a score ranging from 9 to 16.5, with an average score of 8.25. Using the Random Forest classifier with the best hyperparameters found during the model optimization process, we obtained an accuracy score of 75.76%. The precision, recall, and F1-score for each group varied as we can see in Table 4.7

Table 4.7: Performance metrics for RF, meat/dairy/eggs consumption score 1.5 x for omnivores, without vegans.

Group	**Precision**	**Recall**	**F1-score**	**Support**
Group 1	0.36	0.67	0.47	6
Group 2	0.50	0.22	0.31	9

Group 1 exhibited a precision of 36%, recall of 67%, and F1-score of 47%. Group 2 showed a precision of 50%, recall of 22%, and F1-score of 31%. The overall performance indicates that the model had limited success in accurately predicting the groupings based on the multiplied meat consumption score.

In the third method, we employed the Omnivore Index, which was previously described in the paper titled "Identification of Single and Combined Serum Metabolites Associated with Food Intake." [38] This index serves as a metric for assessing an individual's dietary pattern and quantifying their level of omnivorous consumption. The scores ranged from 2 to 14, and our objective was to classify individuals into either a high or low group. The two groups were defined as follows: Group 1 consisted of individuals with scores ranging from 2 to 9, while Group 2 included individuals with scores ranging from 10 to 14.

Using the Random Forest classifier with the best hyperparameters obtained during the model optimization process, we achieved an accuracy score of 73% correctly classifying 9 out of 15 individuals. The precision, recall, and F1-score for each group indicate the performance of the classifier in distinguishing between high and low omnivore index groups can be seen in Table 4.8

Table 4.8: Performance metrics for RF, meat/dairy/eggs consumption score from previous research, without vegans.

Group	Precision	Recall	F1-score	Support
Group 1	0.50	1.00	0.67	6
Group 2	1.00	0.33	0.50	9

4.0.2.3 Third Task: classifying vegans and a combined group of vegetarians, omnivores, and pescetarians

In this task, our objective was to classify vegans separately from all the other groups. We utilized the Random Forest classifier to train and evaluate the model's performance. By using the GridSearchCV function with cross-validation, we performed hyperparameter tuning to find the best combination of hyperparameters that maximizes the model's performance. The corresponding best accuracy score achieved was 73.07%.

Both classes' precision, recall, and F1-score can be observed in Table 4.9. For class 1 (vegan), the precision was 0.60, recall was 0.27, and F1-score was 0.37. For class 3 (non-vegan), the precision was 0.56, recall was 0.83, and F1-score was 0.67.

Table 4.9: Performance metrics for the RF model (Task 3)

Group	Precision	Recall	F1-score	Support
Vegans	0.60	0.27	0.37	11
Other groups	0.56	0.83	0.67	12

When evaluating the model's performance on the test set, the confusion matrix, Figure 4.10, showed that out of the 23 samples, 3 were correctly classified as class 1 (vegan), while 10 were correctly classified as class 3 (non-vegan). However, there were 8 misclassifications for class 1 and 2 misclassifications for class 3.

In order to gain a deeper understanding of the factors influencing the classification of vegans and non-vegans, we performed an analysis to determine the most important features for this task. From this analysis, we identified the top 50 most important features that can be seen in Figure 4.11

Similarly, we applied the SVM (Support Vector Machine) algorithm to classify vegans and non-vegans. Utilizing the best hyperparameters obtained from the grid search, we trained the SVM model and examined its performance. The SVM model trained with these hyperparameters achieved an accuracy of 78% in predicting the dietary groups. Table 4.10 demonstrates the metrics precision, recall, F1-score and

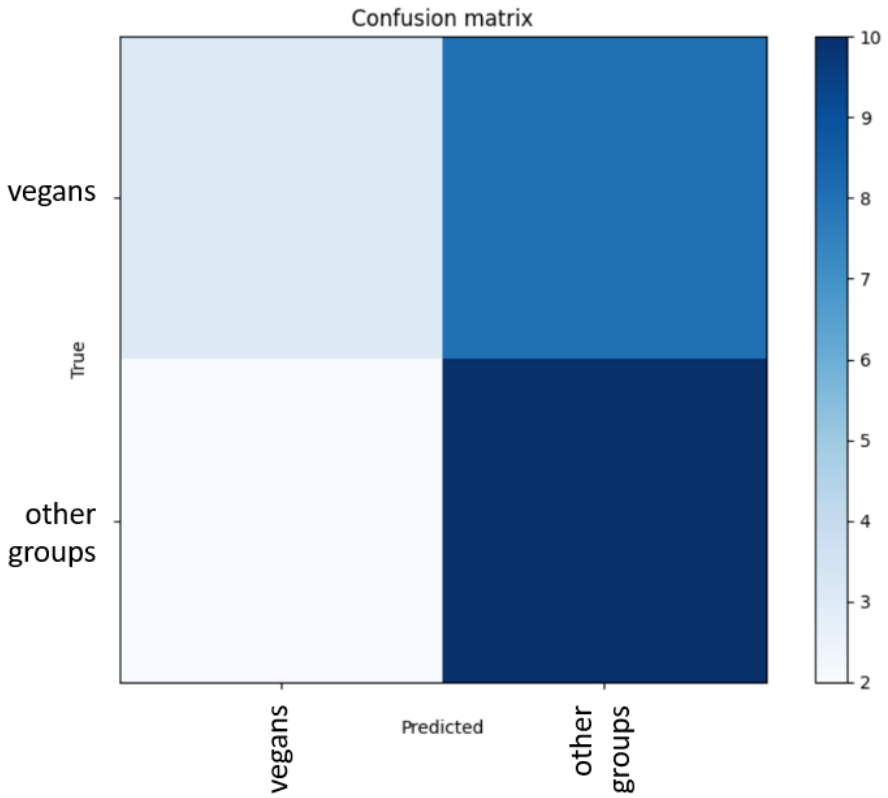

Figure 4.10: Confusion matrix for Random Forest in the task 3.

support. The precision, recall, and F1-score were also calculated for each class, demonstrating satisfactory performance with precision values of 0.80 for class 1 (vegans) and 0.77 for class 3 (non-vegans). The recall values were 0.73 for class 1 and 0.83 for class 3, indicating good performance in correctly identifying instances from each class. The F1-score, which considers both precision and recall, was 0.76 for class 1 and 0.80 for class 3.

Table 4.10: Performance metrics for SVM model (Task 3)

Group	Precision	Recall	F1-score	Support
Vegans	0.80	0.73	0.76	11
Other groups	0.77	0.83	0.80	12

For the SVM model, since it used a non-linear kernel, feature importance analysis using coefficients was not applicable. Therefore, we focused solely on the Random Forest model to identify the most influential features.

4.0.2.4 Fourth Task: classifying vegans and vegetarians together and omnivores and pescetarians together

In the last task, we focused on classifying vegans and vegetarians together against omnivores and pescetarians. We employed both Random Forest and Support Vector Machine (SVM) models to perform the classification task.

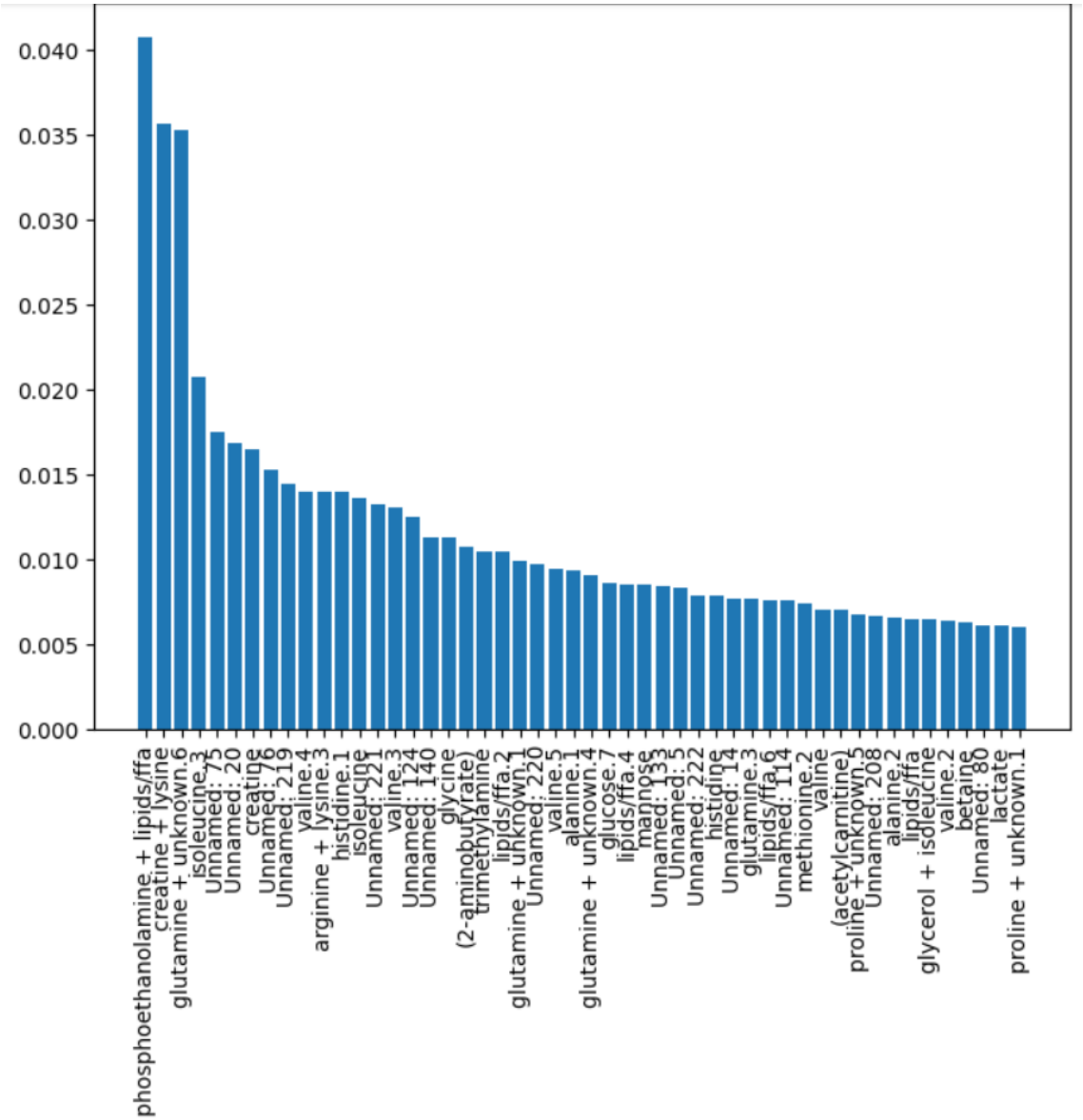

Figure 4.11: The 50 most important features from RF model (Task 3). The first most important feature is a combination of phosphocholine, acetylcholine, phosphoethanolamine, and lipids/ffa, but it is not visible in the figure.

Upon applying the Random Forest (RF) model to classify the combined group of vegans and vegetarians and the combined group of omnivores and pescetarians, we obtained the following outcomes. The best hyperparameters for the RF model were determined as follows: max_depth = 10, max_features = 'sqrt', min_samples_leaf = 4, min_samples_split = 2, and n_estimators = 500. With these parameters, the RF model achieved an accuracy score of 82.03%.

When evaluating the performance of the Random Forest model, we obtained the classification metrics on the test set, shown in Table 4.11.

Table 4.11: Performance metrics for random forest model (Task 4)

Group	**Precision**	**Recall**	**F1-score**	**Support**
Vegans	0.81	0.93	0.87	14
Other groups	0.86	0.67	0.75	9

The confusion matrix revealed that out of the 23 instances, 13 were correctly classified as belonging to the combined group of vegans and vegetarians, while 6 were accurately identified as belonging to the combined group of omnivores and pescetarians. However, there was one misclassification in classifying the combined group of vegans and vegetarians and three misclassifications in classifying the combined group of omnivores and pescetarians.

The 50 most important features are shown in Figure 4.12.

Moving on to the SVM model, we also performed a grid search with cross-validation to determine the best hyperparameters. The best hyperparameters for the SVM model were determined as C = 0.01, gamma = 0.001, and kernel = 'linear'. The model achieved an accuracy of 74%, indicating a moderately accurate classification.

The classification report for the SVM model on the test set showed the following results (Table 4.12):

Table 4.12: Performance metrics for SVM model (Task 4)

Group	**Precision**	**Recall**	**F1-score**	**Support**
Vegans	0.75	0.86	0.80	11
Other groups	0.71	0.56	0.63	12

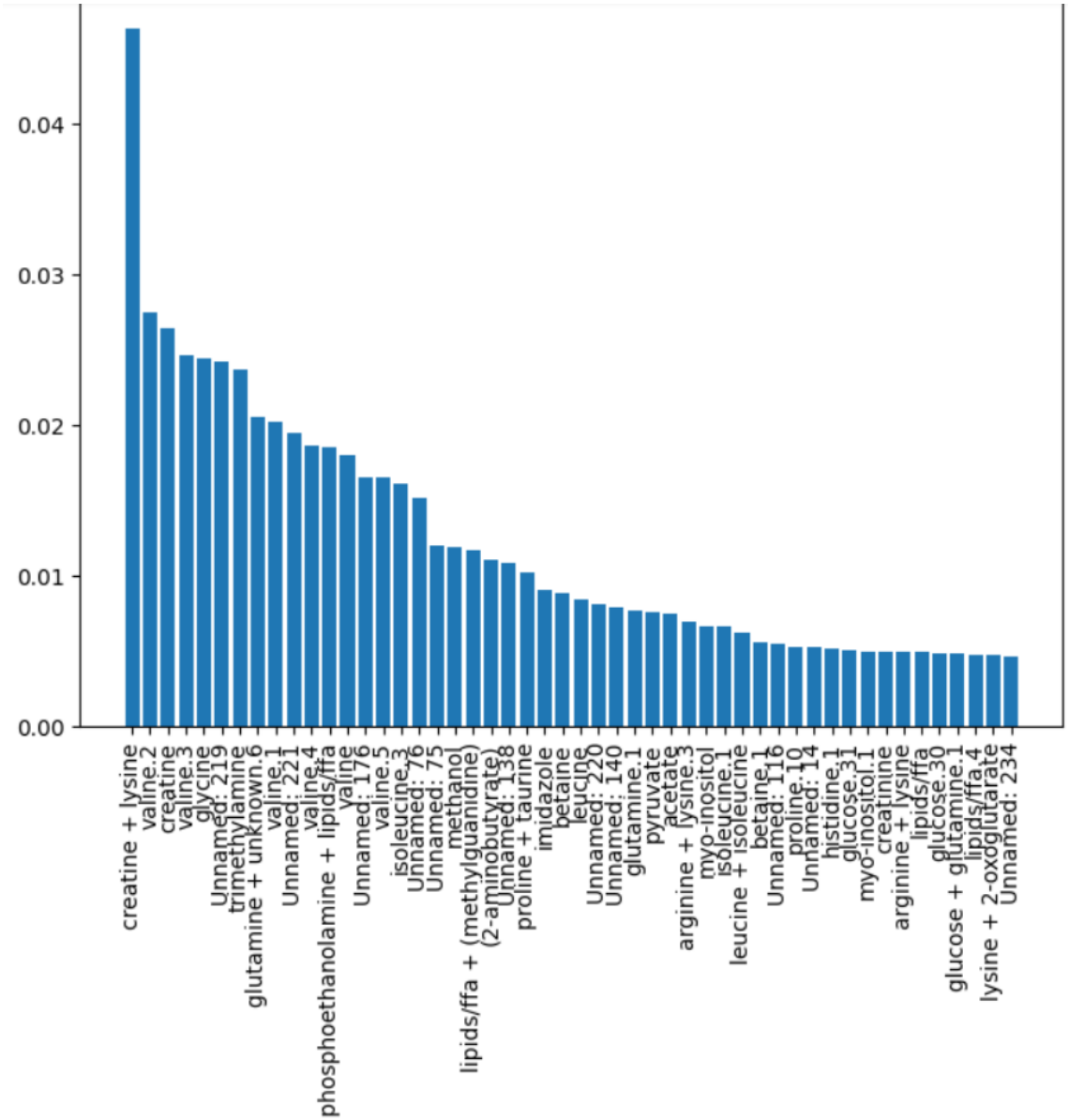

Figure 4.12: The 50 most important features from the RF model (Task 3). The 12th most important feature is a combination of phosphocholine, acetylcholine, phosphoethanolamine, and lipids/ffa, but it is not visible in the figure.

5

Discussion and Conclusion

The Results chapter presented the findings of our research, which aimed to investigate the metabolic differences between different dietary groups. In this discussion section, we will interpret and analyze the results, provide insights into the implications of the findings, discuss the limitations of the study, and suggest potential directions for future research.

5.1 Discussion

The principal component analysis (PCA) was employed to visualize the metabolic profile of the dataset. The four dietary groups were not clearly distinguished by the overall PCA analysis, indicating that there may be significant individual variation in the metabolomic profiles of each group. This variation may result from causes other than diet, suggesting that individual variations or other factors may have an impact on metabolic profiles. Another explanation is that although sharing the same dietary group, the diets within a category may really demonstrate significant variances.

The principal component analysis (PCA) was employed to visualize the metabolic profile of the dataset. The overall PCA analysis revealed that the four dietary groups exhibited some degree of overlap in their metabolic profiles. However, it is worth noting that omnivores were more clearly distinguished from the other three groups in Figure 4.4. This indicates that there are discernible metabolic differences between omnivores and the remaining dietary groups. Nonetheless, there is still notable individual variation within each group, suggesting that factors other than diet, such as genetic variations or lifestyle factors, may contribute to the observed metabolic variations. Another explanation is that although sharing the same dietary group, the diets within a category may really demonstrate significant variances.

Additionally, gender-specific PCA analyses were carried out separately for men and women. But in all instances, the 3D PCA plots failed to clearly separate the two groups. This suggests that within the population under study, there are no obvious sex-specific metabolic differences.

Interestingly, when comparing the metabolic profiles of meat eaters and non-meat eaters, a clear separation was observed between the two groups. This suggests that meat consumption has an impact on the metabolic profile and can be a distinguishing

factor in the analysis. However, it is important to note that this separation does not necessarily imply causality.

The Z-score approach allowed outliers to be located and eliminated from the dataset, which enhanced the analysis's accuracy and quality. The removal of outliers allowed for more reliable conclusions to be drawn from the data.

The identification of the most important features in the dataset was performed using PCA, focusing on the first principal component (PC1). The top 100 features on PC1 with the highest loadings were chosen as the most crucial features. This method enabled a more concentrated study on the dataset's most important variables, potentially increasing the analysis's accuracy and effectiveness. Both men and women were analyzed separately to investigate potential gender-specific differences in metabolism and dietary patterns. The results from the entire dataset, however, were assessed to be more representative and were chosen for further analysis in light of the goal of collecting a thorough and accurate picture of the data. The differences in metabolization and diet between men and women, as well as the impact of disparities in muscle mass on the patterns found, were also of interest to investigate.

Moving on to the classification models, three popular machine learning algorithms, namely Random Forest, Support Vector Machines (SVM), and Neural Networks were applied to classify individuals into four dietary groups based on their habits.

In the first approach of classifying each group separately, the Random Forest algorithm achieved an overall accuracy of 57%. The precision, recall, and F1-score for each group varied, indicating different levels of classification performance. The SVM model performed better, achieving an accuracy of 74%. However, the Neural Network model yielded lower accuracy of 30%. These results suggest that the classification of individuals into specific dietary groups based on metabolomic data is challenging and may require more sophisticated approaches or additional data sources to achieve higher accuracy.

In the second approach, where omnivores and non-omnivores were classified together, the Random Forest model achieved an accuracy of 87%. The results indicate that the metabolomics data can distinguish between omnivorous and non-omnivorous diets, highlighting the potential of metabolomics in developing biomarkers for dietary intake assessment and exploring the health effects of different dietary patterns.

Dietary studies can benefit from the ability to predict the absence of meat intake, since it provides information on the metabolic signatures connected to non-meat-eating habits. Our models demonstrated the ability of metabolomics data to capture variations in dietary patterns and give information on the metabolic impacts of meat consumption, despite the difficulty of correctly measuring meat consumption.

Certain metabolites found in Approach 2 (Table 4.9), including glycine, creatine, trimethylamine, glutamine, and valine, indicate a likely link between protein intake and meat eating [14]. Glycine, a crucial amino acid essential in protein synthesis, is frequently included in meals high in protein, such as meat [52]. Creatine, which is mostly present in meat and other animal products, plays a role in energy metabolism [53]. Trimethylamine (TMA), which is produced by gut bacteria when they break

down nutrients like the choline in meat, may signify increasing consumption of nutrients obtained from animals [54]. A vital amino acid called glutamine is a component of proteins, suggesting that eating meat may result in a larger protein intake. Animal protein sources are a good supply of valine, another important amino acid [55]. These metabolites likely reflect the metabolic processes related to digestion, metabolism, and utilization of proteins obtained from meat.

In our analysis of meat, dairy, and egg consumption, we utilized three different scoring methods to quantify individuals' dietary patterns and explore their relationship with metabolic profiles. With an accuracy score of just 50.36%, the first scoring method, which was based on cumulative consumption, had poor classification performance. Low accuracy, recall, and F1-scores for each group show that the model had difficulty correctly classifying individuals into the corresponding consumption categories. This suggests that simply counting the number of times individuals reported consuming these products the last 4 days before the sample taking, may not be sufficient to capture the nuances of their dietary patterns.

The model's accuracy was greater while using the second score (75.76%), which multiplied the meat consumption score by a factor of 1.5. However, there were differences in the precision, recall, and F1-scores between the two groups, demonstrating that the adjusted score couldn't properly predict consumption habits. The performance metrics for Group 1 were superior to those for Group 2, indicating that those with lower meat intake levels were more accurately classified.

The Omnivore Index, drawn from earlier research, was used as the third score. The model performed quite well in identifying groups with high and low omnivore indexes, with an accuracy score of 73%. However, there were some misclassifications between the two groups based on the precision, recall, and F1-score differences.

Comparing the three scoring methods, the second approach, involving the multiplication of meat consumption scores by 1.5, yielded the highest accuracy. This indicates that better classification performance may be achieved by taking into consideration the various effects that consuming animal-based protein may have on metabolic profiles. The addition of a weighting factor accounts for the increased protein content and possible metabolic effects of sources of protein produced from animals.

The first scoring approach, based on cumulative consumption, may have limited accuracy due to its simplistic nature. It does not take into account potential changes in dietary patterns, such as frequency or various types of meat, dairy, and egg products.

The third approach, using the Omnivore Index showed differences in performance indicators for the two groups while reaching a reasonably high accuracy score. The complex nature of people's eating habits and the inherent difficulties in representing their variation within a single scoring system may be to blame for this.

Overall, the second strategy performed the best, but it is crucial to keep in mind that dietary habits are complex and affected by a variety of factors in addition to consuming meat, dairy, and eggs. The accuracy and predictive value of models designed to evaluate dietary patterns based on metabolic profiles may be further

improved by the inclusion of additional dietary and lifestyle factors in subsequent research.

In the third approach, our objective was to classify vegans separately from all other dietary groups. To complete this challenge, we used the Support Vector Machine (SVM) method and the Random Forest (RF) classifier. The accuracy score for the RF model was 73.07%, while the accuracy score for the SVM model was 78%.

Numerous metabolites that were relevant in the categorization of vegans and non-vegans were discovered through the examination of critical features. According to the RF model, phosphocholine, acetylcholine, phosphoethanolamine, and lipids/ffa were the most significant features, followed by creatine and lysine, glutamine and an unidentified metabolite, and separately, creatine and valine.

These significant features, as depicted in Figure 4.11, provide information on how the metabolic profiles of vegans and non-vegans differ, based on an analysis of serum samples from individuals. The interaction of lipids/ffa, phosphoethanolamine, acetylcholine, and phosphocholine appears to be a particularly important element in distinguishing between the two groups. The discriminating ability of the model is also enhanced by creatine, lysine, glutamine, and the unidentified metabolite.

The classification performance of both models suggests that distinguishing between vegans and non-vegans based on metabolic profiles is challenging. The low recall for class 1 (vegans) in both models suggests that it may be challenging to accurately identify every vegan, maybe as a result of the diversity of the vegan population. Vegans may adhere to various dietary subtypes, such as processed vegan diets or whole-food plant-based diets, which might alter their metabolic profiles. Furthermore, the dataset's small sample size for the vegan group may have hindered the models' ability to correctly categorize this particular dietary category.

In summary, whereas the RF and SVM models classified vegans and non-vegans with respectable accuracy, the findings highlight the difficulty of identifying these dietary categories simply based on metabolic profiles. The low recall of vegans points to significant diversity among the vegan population, which may be impacted by various dietary subtypes and personal characteristics. Furthermore, the existence of metabolic patterns that overlap between vegans and non-vegans emphasizes the importance of variables other than food alone, such as heredity, lifestyle, and general eating habits.

In the fourth approach, we aimed to classify vegans and vegetarians together against omnivores and pescetarians. Both the Random Forest (RF) (82% accuracy) and Support Vector Machine (SVM) (74%) models were utilized for this classification task.

In the analysis of the most influential features for classifying the combined group of vegans and vegetarians against the combined group of omnivores and pescetarians, several metabolites emerged as crucial contributors: creatine + lysine, valine, creatine, glycine, phosphocholine + acetylcholine + phosphoethanolamine + lipids/ffa, trimethylamine, glutamine + unknown, and isoleucine.

Creatine and lysine are closely related and are involved in energy metabolism and protein synthesis [52]. The essential amino acids valine, glycine, and isoleucine are crucial for the synthesis of proteins and the creation of energy. The classification's dependence on these metabolites shows that the two groups' approaches to the metabolism of proteins and energy differ from one another. The observed disparities may be a result of variations in the sources of dietary protein and their associated amino acid compositions.

The combination of phosphocholine + acetylcholine + phosphoethanolamine + lipids/ffa reflects various lipid-related compounds. Fatty acids and lipids are important elements of cell membranes and are crucial for metabolic pathways and energy storage [56]. The presence of these metabolites highlights any differences between omnivores and pescetarians and omnivores and vegetarians in terms of lipid metabolism, especially phospholipid metabolism.

5.2 Limitations and Future Work

The limitations observed in the accuracy of predicting meat consumption can be attributed to several factors. Firstly, the dataset's features may not have fully captured the nuances and variations in meat consumption levels, including the limited number of metabolites obtained from NMR metabolomics. This highlights the need for more comprehensive features that provide a deeper understanding of dietary habits. Secondly, the accuracy of the model may have been influenced by the distribution of the data and the specific features used for prediction. A more diverse and representative dataset could potentially improve the model's performance.

The lower accuracy for categories with smaller sample sizes indicates that the model had difficulty learning the patterns of those categories due to their limited representation in the dataset. To address this limitation, efforts should be made to gather a more diverse dataset that includes a wider range of meat consumption levels across different demographic groups.

Future research may examine bigger and more varied datasets to confirm and expand on our findings. Additionally, incorporating data from additional -omics subjects, such as genomics or transcriptomics, may help us get a deeper understanding of the biological processes that underlie the observed metabolic variations between dietary groups.

We can improve individualized nutrition advice, comprehend the effect of dietary decisions on health outcomes, and perhaps design focused therapies for enhancing people's well-being by unraveling the metabolic fingerprints linked to various eating patterns.

5.3 Conclusion

In this work, we looked at how metabolomics data may be used to categorize people into various dietary groups. We examined four distinct methods to do this assign-

ment using several machine learning techniques, including Random Forest, Support Vector Machine and Neural Networks models. Our results demonstrate the potential of metabolomics to distinguish between various food categories and predict eating habits and the kind of protein consumed.

Through the examination of our results, we identified different levels of categorization accuracy and performance across the various methods used. Notably, we classified vegetarians and vegans with acceptable accuracy rates, indicating potential similarities in their eating habits. However, it was difficult for our models to differentiate between omnivores and pescetarians, showing that the consumption of meat may be a significant factor in the observed variations in metabolite profiles. different results imply that the presence or absence of meat as a substantial source of protein may be a key factor in separating different dietary categories.

We identified the distinctive metabolic traits that distinguish vegans from other dietary groups by categorizing vegans individually from all other groups. Although this method had lower accuracy rates, it enabled a more thorough knowledge of the metabolic variations related to veganism.

Throughout the course of our investigation, we highlighted several metabolites that were critical in classifying omnivores in comparison to the other dietary groups. Amino acids, lipid-related compounds, creatine, trimethylamine, and unidentified metabolites were among these metabolites. The significance of these metabolites raises the possibility that the omnivorous diet categorization is based on metabolic connections and pathways.

Overall, our work shows the potential of metabolomics in food pattern prediction and individual classification into several dietary categories, particularly in separating omnivores from other dietary groups. A potent method for understanding the complex connection between nutrition and metabolism is to employ machine learning algorithms alongside metabolomics data. These results open the door for more investigation into the connections between metabolism, food preferences, and health outcomes and advance our understanding of the metabolic effects of omnivorous diets.

Bibliography

[1] A. J. McAfee, E. M. McSorley, G. J. Cuskelly, *et al.*, "Red meat consumption: An overview of the risks and benefits," *Meat science*, vol. 84, no. 1, pp. 1–13, 2010.

[2] M. Dinu, R. Abbate, G. F. Gensini, A. Casini, and F. Sofi, "Vegetarian, vegan diets and multiple health outcomes: A systematic review with meta-analysis of observational studies," *Critical reviews in food science and nutrition*, vol. 57, no. 17, pp. 3640–3649, 2017.

[3] B. Shatenstein, "Impact of health conditions on food intakes among older adults," *Journal of Nutrition for the Elderly*, vol. 27, no. 3-4, pp. 333–361, 2008.

[4] U. Roessner and J. Bowne, "What is metabolomics all about?" *Biotechniques*, vol. 46, no. 5, pp. 363–365, 2009.

[5] U. W. Liebal, A. N. Phan, M. Sudhakar, K. Raman, and L. M. Blank, "Machine learning applications for mass spectrometry-based metabolomics," *Metabolites*, vol. 10, no. 6, p. 243, 2020.

[6] W. C. Willett, "Diet and health: What should we eat?" *Science*, vol. 264, no. 5158, pp. 532–537, 1994.

[7] S. Damodaran, "Food proteins: An overview," *Food proteins and their applications*, pp. 1–24, 2017.

[8] R. Bressani, "Protein complementation of foods," *Nutritional evaluation of food processing*, pp. 627–657, 1988.

[9] L. Day, J. A. Cakebread, and S. M. Loveday, "Food proteins from animals and plants: Differences in the nutritional and functional properties," *Trends in Food Science & Technology*, vol. 119, pp. 428–442, 2022.

[10] D. Southgate, "Determination of carbohydrates in foods ii.unavailable carbohydrates," *Journal of the Science of Food and Agriculture*, vol. 20, no. 6, pp. 331–335, 1969.

[11] A. Shendurse and C. Khedkar, "Glucose: Properties and analysis," *Encyclopedia of Food and Health*, vol. 3, pp. 239–247, 2016.

[12] M. L. Wahlqvist and N. Wattanapenpaiboon, "Macronutrients: Fats," in *Food and Nutrition*, Routledge, 2020, pp. 195–207.

[13] J. H. Weisburger, "Dietary fat and risk of chronic disease: Insights from experimental studies mechanistic," *Journal of the American Dietetic Association*, vol. 97, no. 7, S16–S23, 1997.

[14] H. M. Lindqvist, M. Rådjursöga, D. Malmodin, A. Winkvist, and L. Ellegård, "Serum metabolite profiles of habitual diet: Evaluation by 1h-nuclear magnetic

resonance analysis," *The American journal of clinical nutrition*, vol. 110, no. 1, pp. 53–62, 2019.

[15] H. M. Lindqvist, M. Rådjursöga, T. Torstensson, L. Jansson, L. Ellegård, and A. Winkvist, "Urine metabolite profiles and nutrient intake based on 4-day weighed food diary in habitual vegans, vegetarians, and omnivores," *The Journal of Nutrition*, vol. 151, no. 1, pp. 30–39, 2021.

[16] M. Guasch-Ferré, S. N. Bhupathiraju, and F. B. Hu, "Use of metabolomics in improving assessment of dietary intake," *Clinical chemistry*, vol. 64, no. 1, pp. 82–98, 2018.

[17] X. Liu and J. W. Locasale, "Metabolomics: A primer," *Trends in biochemical sciences*, vol. 42, no. 4, pp. 274–284, 2017.

[18] D. Marion, "An introduction to biological nmr spectroscopy," *Molecular & Cellular Proteomics*, vol. 12, no. 11, pp. 3006–3025, 2013.

[19] S. Moco, "Studying metabolism by nmr-based metabolomics," *Frontiers in Molecular Biosciences*, p. 372, 2022.

[20] M. V. S. Elipe, "Advantages and disadvantages of nuclear magnetic resonance spectroscopy as a hyphenated technique," *Analytica Chimica Acta*, vol. 497, no. 1-2, pp. 1–25, 2003.

[21] S. E. Ali, M. A. Farag, P. Holvoet, R. S. Hanafi, and M. Z. Gad, "A comparative metabolomics approach reveals early biomarkers for metabolic response to acute myocardial infarction," *Scientific reports*, vol. 6, no. 1, p. 36 359, 2016.

[22] N. Kambhatla and T. K. Leen, "Dimension reduction by local principal component analysis," *Neural computation*, vol. 9, no. 7, pp. 1493–1516, 1997.

[23] T. Kurita, "Principal component analysis (pca)," *Computer Vision: A Reference Guide*, pp. 1–4, 2019.

[24] T. Jiang, J. L. Gradus, and A. J. Rosellini, "Supervised machine learning: A brief primer," *Behavior Therapy*, vol. 51, no. 5, pp. 675–687, 2020.

[25] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, and A. J. Aljaaf, "A systematic review on supervised and unsupervised machine learning algorithms for data science," *Supervised and unsupervised learning for data science*, pp. 3–21, 2020.

[26] D. Dhall, R. Kaur, and M. Juneja, "Machine learning: A review of the algorithms and its applications," *Proceedings of ICRIC 2019: Recent Innovations in Computing*, pp. 47–63, 2020.

[27] A. Galal, M. Talal, and A. Moustafa, "Applications of machine learning in metabolomics: Disease modeling and classification," *Frontiers in Genetics*, vol. 13, p. 3340, 2022.

[28] S. J. Rigatti, "Random forest," *Journal of Insurance Medicine*, vol. 47, no. 1, pp. 31–39, 2017.

[29] S. Nembrini, I. R. König, and M. N. Wright, "The revival of the gini importance?" *Bioinformatics*, vol. 34, no. 21, pp. 3711–3718, 2018.

[30] L. Toloi and T. Lengauer, "Classification with correlated features: Unreliability of feature ranking and solutions," *Bioinformatics*, vol. 27, no. 14, pp. 1986–1994, 2011.

[31] S. R. Gunn *et al.*, "Support vector machines for classification and regression," *ISIS technical report*, vol. 14, no. 1, pp. 5–16, 1998.

[32] W. S. Noble, "What is a support vector machine?" *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

[33] D. Liu and F. Xia, "Assessing object-based classification: Advantages and limitations," *Remote sensing letters*, vol. 1, no. 4, pp. 187–194, 2010.

[34] *What is deep learning? | how it works, techniques applications*, Accessed on May 5th, 2023. [Online]. Available: `https://www.mathworks.com/discovery/deep-learning.html`.

[35] R. Hecht-Nielsen, "Theory of the backpropagation neural network," in *Neural networks for perception*, Elsevier, 1992, pp. 65–93.

[36] B. Widrow, D. E. Rumelhart, and M. A. Lehr, "Neural networks: Applications in industry, business and science," *Communications of the ACM*, vol. 37, no. 3, pp. 93–106, 1994.

[37] D. Berrar, *Cross-validation.* 2019.

[38] T. Karlsson, A. Winkvist, M. Rådjursöga, L. Ellegård, A. Pedersen, and H. M. Lindqvist, "Identification of single and combined serum metabolites associated with food intake," *Metabolites*, vol. 12, no. 10, p. 908, 2022.

[39] R. A. Van den Berg, H. C. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. Van der Werf, "Centering, scaling, and transformations: Improving the biological information content of metabolomics data," *BMC genomics*, vol. 7, pp. 1–15, 2006.

[40] B. J. Blaise, G. D. Correia, G. A. Haggart, *et al.*, "Statistical analysis in metabolic phenotyping," *Nature Protocols*, vol. 16, no. 9, pp. 4299–4326, 2021.

[41] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[42] A. Pajankar, "3d visualizations in matplotlib," in *Hands-on Matplotlib: Learn Plotting and Visualizations with Python 3*, Springer, 2021, pp. 143–159.

[43] R. Garreta and G. Moncecchi, *Learning scikit-learn: machine learning in python.* Packt Publishing Ltd, 2013.

[44] D. Kartini, D. T. Nugrahadi, A. Farmadi, *et al.*, "Hyperparameter tuning using gridsearchcv on the comparison of the activation function of the elm method to the classification of pneumonia in toddlers," in *2021 4th International Conference of Computer and Informatics Engineering (IC2IE)*, IEEE, 2021, pp. 390–395.

[45] R. Yacouby and D. Axman, "Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models," in *Proceedings of the first workshop on evaluation and comparison of NLP systems*, 2020, pp. 79–91.

[46] J. Weston and C. Watkins, "Multi-class support vector machines," Citeseer, Tech. Rep., 1998.

[47] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[48] Z. Zhang, "Improved adam optimizer for deep neural networks," in *2018 IEEE/ACM 26th international symposium on quality of service (IWQoS)*, Ieee, 2018, pp. 1–2.

[49] B. Asadi and H. Jiang, "On approximation capabilities of relu activation and softmax output layer in neural networks," *arXiv preprint arXiv:2002.04060*, 2020.

[50] N. Kettaneh, A. Berglund, and S. Wold, "Pca and pls with very large data sets," *Computational Statistics & Data Analysis*, vol. 48, no. 1, pp. 69–85, 2005.

[51] I. Ahmad, M. Basheri, M. J. Iqbal, and A. Rahim, "Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection," *IEEE access*, vol. 6, pp. 33 789–33 795, 2018.

[52] M. A. Razak, P. S. Begum, B. Viswanath, and S. Rajagopal, "Multifarious beneficial effect of nonessential amino acid, glycine: A review," *Oxidative medicine and cellular longevity*, vol. 2017, 2017.

[53] G. Wu, "Important roles of dietary taurine, creatine, carnosine, anserine and 4-hydroxyproline in human nutrition and health," *Amino acids*, vol. 52, no. 3, pp. 329–360, 2020.

[54] C. E. Cho, S. Taesuwan, O. V. Malysheva, *et al.*, "Trimethylamine-n-oxide (tmao) response to animal source foods varies among healthy young men and is influenced by their gut microbiota composition: A randomized controlled trial," *Molecular nutrition & food research*, vol. 61, no. 1, p. 1 600 324, 2017.

[55] V. Cruzat, M. Macedo Rogero, K. Noel Keane, R. Curi, and P. Newsholme, "Glutamine: Metabolism and immune function, supplementation and clinical translation," *Nutrients*, vol. 10, no. 11, p. 1564, 2018.

[56] E. Tvrzicka, L.-S. Kremmyda, B. Stankova, and A. Zak, "Fatty acids as biocompounds: Their role in human metabolism, health and disease-a review. part 1: Classification, dietary sources and biological functions.," *Biomedical Papers of the Medical Faculty of Palacky University in Olomouc*, vol. 155, no. 2, 2011.

A

Appendix 1