Empowering Heart Health Through Technology: The Development of Hearty App by NutriComm Using NHANES Data and Machine Learning

Authors: Mariane de Almeida Alves, Ryan Covill, Dongsuk Kim, Christopher H. Kroll
Date: February 10, 2024

1 Summary

This project addresses the critical issue of cardiovascular disease (CVD), the leading cause of mortality, by integrating heart-healthy choices into individuals' daily lives, tailored to their unique dietary preferences. Leveraging the comprehensive data from the National Health and Nutrition Examination Survey (NHANES) and employing advanced machine learning technologies, the project team aimed to develop Hearty, a novel application designed to revolutionize cardiovascular health maintenance. Hearty stands out by offering heart health risk assessment and personalized nutrition advice, based on individual data and dietary habits.

Utilizing NHANES data, the project aimed to fill the significant gap in accessible, customized solutions for heart health management. The data's depth provided a solid empirical foundation for the app's machine learning models, including logistic regression and neural networks, to deliver scientifically informed, heart health risk assessment and customized dietary recommendations. This approach underscores the potential of data-driven technologies to enhance health outcomes significantly.

However, the project also navigated challenges inherent in predictive modeling for health, such as the complexity of CVD's etiology and the limitations of observational data. These challenges highlight the need for ongoing refinement of the models, incorporating broader health determinants and improving data quality and completeness.

Despite these obstacles, the Hearty app represents a significant advancement in public health technology, providing a scalable, user-centric tool for personalized heart health management. It exemplifies the transformative potential of combining data science with health science to create innovative solutions for chronic disease prevention and management.

In summary, this project illustrates the vital role of innovation and technology in advancing personalized healthcare solutions. As the project moves forward, it will continue to refine its predictive models and expand its scope to incorporate a broader spectrum of health determinants, enhancing the efficacy and applicability of the Hearty app. This endeavor paves the way for a new era of personalized, data-driven health empowerment, with the potential to significantly impact public health and chronic disease management.

2 Context

2.1 Problem

The core issue this project addressed is the integration of heart-healthy choices into the daily lives of individuals, tailored to their unique dietary preferences and habits, against the backdrop of cardiovascular disease (CVD) being a predominant health crisis. The maintenance of cardiovascular health presents a substantial challenge, exacerbated by modern lifestyles, leading to heart health harm. This issue is compounded by the overwhelming amount of conflicting information, making it difficult for individuals to make informed decisions regarding their heart health.

CVD stands as the leading cause of mortality (Roth GA, et al., 2020), responsible for one in four deaths annually in the United States (Center for Disease Control and Prevention, 2017), underscoring a critical need for innovative solutions that promote heart-healthy habits. The American Heart Association highlights a study demonstrating the link between healthy eating patterns and reduced risk of premature death (Shan Z, et al., 2019), emphasizing the importance of diet in the prevention and management of heart disease (Downer S, et al., 2020). Despite this recognition, there remains a significant lack of resources that are both informative and adaptable to individual needs and preferences, rendering the task of managing heart health as daunting and inaccessible for many.

With approximately 92.1 million American adults living with some form of CVD or the after-effects of stroke, and the direct and indirect healthcare costs associated with these conditions exceeding $329.7 billion, the economic and health implications are enormous (Association AH, et al., 2017). This situation highlights the urgency for more personalized and accessible strategies that empower individuals to adopt heart-healthy habits seamlessly into their lives.

The challenge, therefore, lies in creating solutions that not only educate but also are flexible enough to align with the diverse dietary preferences and lifestyles of individuals. These solutions aim to enhance cardiovascular health and overall well-being by making heart health management an achievable goal for everyone, thereby addressing the gap in the current healthcare landscape where CVD remains a major public health issue.

2.2 Solution

In response to the need for accessible and personalized heart health management, the solution presented is Hearty by NutriComm, a cutting-edge application designed to revolutionize cardiovascular health maintenance. At its heart, Hearty by NutriComm offers a comprehensive approach by seamlessly integrating personalized nutrition advice with health guidance, leveraging machine learning technology. This application distinguishes itself by offering customized dietary recommendations that align with individual preferences and habits, ensuring that making heart-healthy choices becomes an intuitive and enjoyable part of daily life.

NutriComm's innovative application Hearty is built to adapt and evolve, offering users tailored suggestions that grow more accurate over time. It moves beyond traditional dietary tracking, by offering practical and scientifically backed insights on the impact of

diet on heart health. This allows users to receive personalized meal suggestions, alongside actionable tips to gradually adopt a heart-healthy lifestyle without feeling overwhelmed.

By focusing on both dietary habis and health parameters, Hearty by NutriComm positions itself as a personal health ally, dedicated to making cardiovascular health more approachable and manageable for everyone. Its dynamic interface and intelligent feedback system make it more than just an application; it's a transformative tool designed to empower users to make informed decisions about their diet and lifestyle, directly targeting the root of cardiovascular health issues with a solution that is as informative as it is engaging.

2.3 NHANES Data and Justification for Use

The National Health and Nutrition Examination Survey (NHANES) (Center for Disease Control and Prevention, 2018), orchestrated by the National Center for Health Statistics (NCHS), is a pivotal resource in understanding the health and nutritional status of the U.S. population. Integrating detailed interview surveys with comprehensive physical and laboratory examinations, NHANES offers a robust dataset that spans socio-economic, demographic, dietary, and health-related information. The survey's unique approach, combining data from medical, dental, and physiological assessments, provides an unparalleled depth of insight.

Launched in 1999 and conducted annually, NHANES meticulously samples 5,000 individuals each year through a nationally representative, multistage probability sampling technique. This ensures the collection of data across a wide spectrum of the population, encompassing laboratory results and the incidence of chronic conditions such as anemia, cardiovascular diseases, diabetes, and more. Such breadth and depth make NHANES an invaluable tool for health research and policy development.

The utilization of NHANES data in the development of the Hearty app by NutriComm is particularly apt given the app's focus on promoting cardiovascular well-being through tailored nutritional guidance. The survey's extensive database on dietary habits, prevalence of cardiovascular conditions, and related health metrics provides a solid empirical foundation to inform the app's machine learning models, including logistic regression and neural networks. By analyzing patterns within the NHANES dataset, the Hearty app can offer personalized dietary recommendations, thereby addressing the critical need for accessible, customized solutions to improve heart health.

2.4 Machine Learning Models for Hearty App Development

To enhance the Hearty app's capability in identifying individuals at risk of cardiovascular diseases, our development leverages the power of machine learning models, specifically logistic regression and neural networks. These models are pivotal in analyzing the NHANES data, providing a scientific basis for personalized dietary recommendations aimed at improving cardiovascular well-being.

Logistic Regression Model: The logistic regression model was utilized for its proficiency in classifying at-risk patients and their risk through a statistical approach that models

the probability of a binary outcome (such as the presence or absence of a cardiovascular condition) based several predictor variables. This method was particularly suitable for the Hearty app due to its simplicity and effectiveness in providing baseline accuracy scores. By analyzing variables such as demographics and dietary habits as well as other health indicators from the NHANES dataset, the logistic regression model enables the Hearty app to predict the likelihood of having cardiovascular disease. Its application serves as a foundational step, ensuring that the app's recommendations are grounded in reliable statistical analysis.

Neural Network Model: In addition to logistic regression, the Hearty app project incorporated neural network models to take advantage of their ability to capture and model complex relationships between inputs and outcomes. Neural networks, with their deep learning capabilities, can analyze vast amounts of data, learning from the intricacies and patterns to make more accurate predictions about cardiovascular health risks. These models' inclusion reflects our commitment to leveraging advanced machine learning techniques to provide highly personalized and effective dietary guidance. The neural networks' ability to process and learn from the comprehensive and multifaceted NHANES data ensures that the Hearty app's recommendations are not only based on broad statistical trends but also on deep insights into the factors affecting cardiovascular health.

Justification for Model Selection: The choice of logistic regression and neural networks for the Hearty app was driven by the goal of developing a robust, data-driven solution to promote cardiovascular health. Logistic regression offers a solid baseline for understanding the impact of various factors on heart health, while neural networks provide the depth of analysis necessary for tailoring recommendations to the individual's unique health profile. Together, these models enable the Hearty app to deliver scientifically informed, customized dietary advice, harnessing the NHANES dataset's potential to address the critical challenge of cardiovascular disease prevention and management effectively.

3 Methodology

3.1 Data Mining and Modeling

Dataset Preprocessing: NHANES data is avalible publicly and and comprehends a set of demographics, medical and nutritional data. For conducting the predicion models demographic data regarding the age, gender, ethnicity, educational, household annual income, and marital status were used. These demographic information are well known and recognized as important factor involved in the development of sevral diseases. Additionally, data from medical examinations and laboratory exams was also included, such as blood pressure, lipid profile, fasting glucose, and body measurements. Finally, the dietary data was also included as predictors. The dietary intake reported by NHANES participants was classified into the food groups fish, fruits and vegetables, nuts, seeds and legumes, red meat, processed meat, sugar-sweetened beverages and whole grains. The definition of these food groups was based on the American Heart

Aassociation Dietary Targets and Healthy Diet Score for Defining Cardiovascular Health (Benjamin EJ, et al. 2018).

The datasets containing all information mentioned were merged into one dataset comprising dietary, medical, laboratory and examination data.

Subject Exclusion and Label Assignment: Dataset cleaning was performed aiming to obtain a final sample of both men and women age 20 years or more. Since children and adolescents have specific nutritional requirements and medical parameters these individuals were not included in the prediction models. The same was applied for pregnant women, which present very specific health and physiologic conditions. As a second step, individuals with only one day of dietary intake report were also excluded, remaining only those that presented two days of dietary intake report. This step was conducted aiming to have a better picture of the individuals' habitual diet and for reasons of representiveness.

All the variables were comprehensively inspected to evaluate the number of missing values. Variables with more than 50% of missing values were not kept in the final dataset, which was the case of the fasting glucose variable (52% of missing values). Lastly, the outcome cardiovascular disease was computed based on the medical conditions questionnaire. The definition of cardiovascular disease was based on the following questions: *"Have you ever been told, you had congestive heart failure/coronary heart disease/angina/heart attack/stroke?"*. The individuals who have answered *"yes"* to one of the questions were classified as having cardiovascular disease (*label = 1*), those who answered *"no"* to all questions were classified as beeing without cardiovascular disease (*label = 0*).

3.2 Statistical Analysis and Model development

General Descriptive statistics: The demographic analysis revealed an average age of 52.1 years, with a standard deviation (SD) of 17.3 years. Gender distribution was nearly balanced, with 48.1% male (1,304) and 51.9% female (1,409) participants. Ethnicity breakdown showed a diverse sample, including Mexican American (12.9%), Other Hispanic (8.7%), Non-Hispanic White (36.2%), Non-Hispanic Black (24.7%), and Other (17.5%). Educational levels varied, with a significant portion of the sample having attained a college or associate degree (32.4%) or being college graduates (26.1%). Marital status was also varied, with the majority being married (50.5%). Annual household income data indicated a wide range, with 19.2% of participants reporting an income of $100,000 or over.

In terms of dietary intake, the study reported on median and mean consumption across various food groups and nutrient intake. Fruits and vegetables showed a median intake of 149.0 grams, with a substantial interquartile range (IQR) and a mean intake of 194.9 grams (SD=197.4). Whole grains, nuts, seeds, and legumes, along with fish, red meat, and processed meat, have a median intake of 0.0 grams, indicating varied consumption patterns within the population. Sugar-sweetened beverages had a median intake of 124.0 grams. Nutrient intake was also reported, with total calories averaging 2020.0 Kcal

(SD=839.3) and other nutrients such as protein, carbohydrates, sugars, and fats showing varied median and mean intakes.

Health characteristics of the sample were detailed, with 45.8% of participants having hypertension and 8.1% reporting diabetes. Dyslipidemia was present in 33.0% of the sample. Weight status categorization indicated that 42.4% of the population was obese, 31.5% was overweight, and 25.2% fell within the normal range for body mass index (BMI). This comprehensive analysis provided insight into the demographic, dietary, and health profiles of the study's participants, indicating a diverse population with varied health and dietary habits. The data presented in this report underscores the importance of considering a wide range of factors when assessing health outcomes and dietary behaviors.

Input Variables: The input variables were defined based on the main cardiovascular disease risk factors including the food groups intake in grams per day (fish, fish, fruits and vegetables, nuts, seeds and legumes, red meat, processed meat, sugar-sweetened beverages and whole grains), nutrients intake (total calories, protein, carbohydrates, sugar, fiber, total fat, saturated fat, monosaturated fat, polyunsaturated fat, and dietary cholesterol) demographic data (age, gender, ethnicity, educational, household annual income, and marital status), examination data (blood pressure, body mass index), and laboratory data (HDL-cholesterol, LDL-cholesterol, total cholesterol, and triglycerides).

Outcome Variable: Cardiovascular Disease (CVD_Outcome): The cardiovascular disease definition was based on the following questions: *"Have you ever been told you had congestive heart failure/coronary heart disease/angina/heart attack/stroke?"*. The individuals who have answered *"yes"* to one of these questions were classified as having cardiovascular disease (*label = 1*). Those who answered *"no"* to all questions were classified without cardiovascular disease (*label = 0*).

Logistic Regression Model: The dataset obtained from the Data Mining and Modeling process was split into training and testing datasets. For the training stage, the training dataset was used to fit the logistic regression model for prediction. The logistic regression model was set up to apply the *LIBLINEAR* solver, which applies automatic parameter selection (a.k.a. *L1 Regularization*) and it's recommended when you have high dimension dataset. For the testing stage, the tested dataset was used to calculate the predctions based on the fitted model. To evaluate the model, the confusion matrix was performed, returning the values of true positives, false positives, true negatives and false negatives obtained with the model.

Two Neural Network Models: The objective behind developing the two neural network models were the prediction of cardiovascular disease outcomes based on dietary factors, which can be crucial for preventive health measures and personalized dietary recommendations.

Both neural network model were configured using *Keras*. They were sequential models comprising several densely connected layers. This architecture was chosen for its simplicity and effectiveness in mapping a set of input features (dietary factors) to a binary output.

The first layer of the **neural model for the prediction of cardiovascular disease** was a dense layer with 57 units and a *ReLU* activation function, designed to process the input features. This was followed by two more dense layers with *ReLU* activations, with the second layer having 70 units and the third having 50 units, creating a deeper network capable of learning complex relationships in the data followed by the last *ReLU* layer with 30 units. The final layer was a dense layer with 1 unit and a sigmoid activation function, outputting a probability between 0 and 1 indicating the likelihood of a cardiovascular disease outcome.

The chosen architecture and the sequential buildup of layers with increasing then decreasing complexity allowed both models to learn from both simple and complex patterns in the datasets. The use of *ReLU* activation functions helped to mitigate the vanishing gradient problem, enabling effective training of deep networks, while the sigmoid function at the output layer maped the final values to a probability score suitable for binary classification.

The dataset for this model was split into features (x) and labels (y). The features included demographic and health variables, as well as dietary factors like fish, fruit and vegetable intake, etc., while the labels indicated the presence (*label = 1*) or absence (*label = 0*) of cardiovascular disease. The data was then split into training, validation, and test sets to prepare it for model training and evaluation.

The model was compiled with the *RMSprop* optimizer (with a learning rate of 0.001), binary crossentropy as the loss function, and accuracy as the metric. This setup was typical for binary classification problems. The model was trained on the training data for 200 epochs with a batch size of 16, using the validation data to monitor performance and potentially prevent overfitting.

The *RMSprop* optimizer is known for its efficiency in handling the noisy gradients of stochastic optimization, making it suitable for training deep neural networks. The learning rate was set to a relatively small value to allow for gradual learning, reducing the risk of overshooting the minimum of the loss function. Binary crossentropy is the standard loss function for binary classification tasks, as it measures the distance between the probability distributions of the actual and predicted labels.

After training, both models were evaluated on their respective test data to assess their performance, and the trained models were saved to disk for future use or further analysis.

The first layer of the **neural model for the prediction of the optimal food groups** was a dense layer with 16 units and a *ReLU* activation function, designed to process the input features. This was followed by two more dense layers with *ReLU* activations, with the second layer having 40 units and the third having 16 units, creating a deeper network capable of learning complex relationships in the data. The final layer was a dense layer with 1 unit and a sigmoid activation function, outputting a probability between 0 and 1 indicating the likelihood of a cardiovascular disease outcome.

The dataset for this model was split into features (x) and labels (y). The features included dietary factors like fish, fruit and vegetable intake, etc., while the labels indicated the

presence (*label = 1*) or absence (*label = 0*) of cardiovascular disease. The data was then split into training, validation, and test sets to prepare it for model training and evaluation. The model was compiled with the *RMSprop* optimizer (with a learning rate of 0.0005), binary crossentropy as the loss function, and accuracy as the metric. This setup was typical for binary classification problems. The model was trained on the training data for 200 epochs with a batch size of 32, using the validation data to monitor performance and potentially prevent overfitting.

3.3 Tools Used

Miro: Utilized for brainstorming sessions and project planning, Miro provided a collaborative online whiteboard platform. It was instrumental in mapping out user flows, organizing ideas, and facilitating remote teamwork, helping to streamline the development process and align on project goals.

Figma: Served as the primary tool for UI/UX design, enabling the creation of an interactive prototype and visual design for the Hearty app. It facilitated refinement of the app's user interface, ensuring an intuitive and engaging user experience.

Excel: Employed primarily for organizing, managing, and deciding on variable selection criteria and food code assignment as well as evaluating the output of the neural network model, especially the extreme cases study. This tool offered a structured platform for categorizing and evaluating variables, crucial for the development of machine learning models in the Hearty app.

Jupyter Notebook: Essential for coding, data analysis and machine learning model development, Jupyter Notebook was used to write and execute Python code. It allowed for interactive coding and data exploration, visualization, and the implementation of logistic regression and neural network models using the NHANES dataset.

Visual Studio Code: Utilized as the primary integrated development environment (IDE) for writing, testing, and debugging code of the back-end. Visual Studio Code supported the development by providing a versatile and user-friendly platform for coding in Python, essential for building especially the rudimentary app's back-end logic.

PowerPoint: Used for creating the presentation that communicated the project's goals, progress, and outcomes. It was instrumental in visually presenting the Hearty app's features, benefits, and the underlying research and data science. It allowed for the synthesis of complex information into digestible slides, facilitating a clear and persuasive presentation.

Unsplash: Sourced high-quality, royalty-free images to enhance the visual appeal of the presentation. Unsplash provided a vast library of images that could be used to visually represent health and nutrition themes.

QuickTime: Was specifically employed for recording video showcasing the Figma front-end prototype and its presentation.

iMovie: Used for editing the video for the front-end prototype presentation recorded with QuickTime. It helped refining and enhancing the video.

Streamlabs OBS: Employed for recording the presentation of the back-end walkthrough.

Screen Recording Onboard Software from MacOS: This built-in software allowed for quick and efficient recording of the main presentation part without the need for additional downloads or setups, streamlining the content creation process.

Shotcut: Utilized for advanced video editing, it provided comprehensive tools for editing and refining video content captured for presentation of the Hearty app. It offered a range of features for video composition, effects, and transitions, enabling the creation of a quality video for the final presentation.

YouTube: Served as the platform for hosting and sharing the presentation video for the Hearty app.

ChatGPT Model 4: Assisted in Python coding, data interpretation, providing insights, and generating or aggregating content for analysis, reports and presentation.

4 Results

4.1 Logistic Regresion Model

Our logistic regression analysis on the NHANES dataset showed a good performance regarding the accuracy (0.90) and sensitivity (0.98), while the precision (0.33) and sentivity (0.15) results were not optimal. These parameters indicated a lower proportion of true positives among all actual positive values. This became clear when two different fictional user case examples were tested in the model. the first example represented features of low risk for cardiovascular disease and the model could predict this as an actual true false (*label = 0*), while the second example represented features of high risk for cardiovascular disease, and the model predicted this as a false negative (*label = 0*). In summary, the model's performance on positive cases (*label = 1*) was relatively poor. Further optimization or feature engineering may be required to improve the model's ability to correctly classify positive cases. Exploring additional features or adjusting model parameters may help improve the overall performance, especially in correctly identifying positive cases.

4.2 Neural Network Models

The trained neural model for predicting cardiovascular disease demonstrated an accuracy of 0.908 on the test set. The model's loss on the test data was recorded at 0.3132. Upon evaluating the performance of our neural network model, designed to predict the presence of cardiovascular disease, we've identified a significant skew in the class distribution within our dataset. Approximately 90% of the participants did not present with cardiovascular issues, a factor that has considerable implications for the interpretability of our model's accuracy.

The reported accuracy of 0.908 may have initially suggested a high level of predictive power. However, this figure likely overstated the model's effectiveness due to the imbalance in the dataset. The model has evolved to predict the absence of cardiovascular issues predominantly, which, while matching the majority of the cases, does not necessarily reflect an ability to identify the presence of the disease accurately. Our analysis indicated that the model's high accuracy was predominantly due to its prediction of the majority class. The rare occurrence of the positive class posed a

challenge for the model, leading to a potential high false negative rate where the model predicted *"no issue"* for cases where the disease was actually present. This is particularly concerning in a medical context where the cost of missing a positive case can be substantial.

5 Discussion and Conclusions

5.1 Evaluative Perspectives on the Business Case

Addressing and Meeting User Demands: The demand for a personalized approach in integrating heart-healthy choices into daily life is evident, as individuals increasingly seek solutions that cater to their unique dietary preferences and habits. Hearty's focus on customizing health and nutrition advice based on comprehensive machine learning models resonates with users who desire more than generic guidelines. By offering tailored recommendations, the project aligns with the growing demand for personalized healthcare, which is not only more effective but also more engaging for users. This individualized approach justifies organizational investments as it meets a clear market need, enhancing customer satisfaction and loyalty.

Addressing and Meeting User Demands: Investing in NutriComm's Hearty app represents a strategic move for organizations aiming to tap into the burgeoning health and wellness sector. The project's emphasis on machine learning to predict the link between nutrition and health outcomes positions it at the forefront of innovative health solutions. Such an investment is justified as it not only aligns with the increasing consumer demand for data-driven health insights but also demonstrates an organization's commitment to addressing critical health issues like cardiovascular disease.

Scalability: Hearty's design inherently supports scalability. The use of machine learning algorithms allows for continuous adaptation and improvement of the tool based on user input and feedback. As the user base grows, the models become more refined, enhancing its accuracy and relevance. This scalability ensures that the tool remains up-to-date, meeting the evolving needs of a broad user demographic.

Exceeding Customer Expectations: The project's comprehensive approach to integrating heart-healthy choices into daily life aims to exceed customer expectations. By delving deep into the interplay between nutrition, physical health, and well-being, NutriComm's Hearty application offers insights beyond typical dietary advice. The predictive model, tailored to individual demands, provides users with actionable and personalized recommendations, setting it apart from generic nutrition guides. This focus on customization, backed by robust machine learning, ensures that the tool not only meets but surpasses user expectations, offering a unique, and highly personalized health and nutrition guide.

Novelty and Innovativeness: NutriComm's solution to promoting heart health is a radical departure from traditional approaches, offering a novel and innovative pathway to cardiovascular well-being. By harnessing the power of advanced machine learning, it goes beyond the one-size-fits-all advice typically found in the health and wellness sector. This approach is creative in its ability to intricately understand and adapt to the

unique dietary preferences and habits of each individual. The integration of physical health aspects into nutritional advice is a particularly innovative aspect, acknowledging the complex interplay of these factors. This forward-thinking, user-centric design represents a radical shift in the approach of diet and health, making it a groundbreaking tool in the quest for better heart health and overall wellness.

Transformative and Sustainable Impact: Implementing NutriComm's innovative Hearty app would yield significant positive impacts on society. By offering personalized nutrition and health guidance, it directly contributes to reducing the prevalence of cardiovascular diseases, thus enhancing public health and potentially decreasing healthcare costs. This approach rectifies a dysfunctional system where generic health advice often fails to resonate with individuals' unique needs. Socially, the solution is highly sustainable as it fosters a culture of health awareness and personal responsibility, making health management more accessible and less daunting for individuals. This inclusivity and emphasis on personalization make it a tool that can adapt and remain relevant in diverse social contexts, bridging gaps in health education and promoting a more health-conscious society.

5.2 Cardiovascular Prediction and its Limitations

While utilizing a comprehensive dataset consisting of diet records, medical examinations, demographics, and laboratory data holds promise for predicting cardiovascular disease (CVD) risk, it is essential to acknowledge several inherent limitations in such predictive models. Firstly, the complexity of CVD etiology encompasses multifaceted interactions between genetic predispositions, environmental factors, lifestyle choices, and physiological markers, which may not be fully captured by the available dataset. Additionally, the quality and completeness of the data, including potential inaccuracies in self-reported dietary habits or medical history, could introduce biases and compromise the reliability of predictions. Furthermore, while certain risk factors such as blood pressure or cholesterol levels are well-established indicators of CVD risk, other emerging biomarkers or genetic factors may not be adequately represented in the dataset, limiting the models' predictive accuracy. Moreover, the dynamic nature of individual health trajectories and the influence of interventions or lifestyle changes over time pose challenges in capturing longitudinal trends and predicting future CVD events accurately. Finally, it's important to recognize that predictive models based on observational data inherently carry limitations in establishing causality, and thus, the predictions should be interpreted cautiously in clinical decision-making.

5.3 Optimal Food Group Prediction and its Limitations

In order to predict the nutritional guidelines to reduce cardiovascular risks, a parameter study was conducted. As the neural network model was defined to predict in biased way, we constrained the dietary variables as the model datasets. This study encompassed all extreme cases of nutritional categories, from 0 to maximum consumption by participants. Despite of this consideration, the study did not seem to suggest insight for

guidelines. Notably, the consumption of whole grains, known to be good for cardiovascular health, showed divergent tendencies.

Important variables such as demographic and genetic factors, lifestyle choices, and other health-related behaviors have not been included in the model but can significantly influence cardiovascular risk. The same applies for possible interactions between nutrients, which have not considered in the model, but can have significant influence on cardiovascular risk and therefore the identification of optimal food groups to reduce this risk.

The model may not have had the capabilities to capture sophisticated relationships between input variables and cardiovascular outcomes. Also Nutrition science acknowledges that diet and health relationships are complex, often involving synergistic and antagonistic interactions between nutrients that are not easily captured in simple machine learning models.

Ensuring a diverse and representative dataset can aid in training a model that is more reflective of the general population and sensitive to varied dietary patterns.

In light of these limitations, ongoing research efforts should focus on refining predictive models by incorporating novel biomarkers, improving data quality and completeness, validating predictions in diverse populations, and integrating longitudinal data to enhance the accuracy and applicability of CVD risk assessment tools, also taking possible interactions between nutrients into consideration.

5.4 Conclusion

This project has explored the issue of cardiovascular disease (CVD), the leading cause of mortality, through the lens of personalized health management. At the heart of our endeavor lay the Hearty app, a pioneering solution designed to seamlessly integrate heart-healthy choices into the fabric of daily life, tailored to individual dietary habits and preferences. Leveraging the comprehensive NHANES dataset and advanced machine learning technologies, such as logistic regression and neural networks, this initiative has aimed to transcend traditional dietary tracking, providing users with personalized heart risk assessment and nutrition advice.

Throughout the development process, the utilization of NHANES data has been instrumental in informing our machine learning models. The overall approach not only aligns with the increasing demand for personalized healthcare solutions but also exemplifies the potential of data-driven technologies to enhance cardiovascular health maintenance and management.

However, the journey of integrating cutting-edge technology with health science has illuminated several challenges and limitations. The complexity of CVD etiology, encompassing genetic, environmental, and lifestyle factors, alongside the inherent limitations of observational data, poses significant hurdles in accurately predicting and preventing cardiovascular risk. Furthermore, the reliance on self-reported dietary habits and medical history introduces potential biases, while the exclusion of emerging biomarkers and the dynamic nature of health trajectories highlight the need for continuous model refinement and validation.

Despite these challenges, the project's focus on scalability, user-centric design, and the innovative application of machine learning models underscores a transformative shift towards more accessible and effective heart health management solutions. The Hearty app represents a significant stride forward in addressing the critical public health issue of CVD, offering a promising pathway to reducing its prevalence through personalized, data-informed health guidance.

In conclusion, this project underscores the vital role of innovation and technology in advancing public health objectives, particularly in the realm of chronic disease prevention and management. By bridging the gap between data science and health science, the Hearty app by NutriComm exemplifies the potential of personalized healthcare solutions to make a meaningful impact on individuals' lives, paving the way for a healthier future. As we move forward, ongoing research, model optimization, and the incorporation of a broader spectrum of health determinants will be crucial in enhancing the efficacy and applicability of predictive health technologies.

6 Appendix

6.1 Weblinks

- All Datasets: https://drive.google.com/file/d/1App-DUeUTQyNrGcfQZFgXiJPs8MoK0Zp/view?usp=sharing
- Brainstorming Board: https://miro.com/app/board/uXjVNC7UcyM=/?share_link_id=59698024786
- Excel File for Extreme Case Study Neural Network Model for Predicting Optimal Food Groups: https://docs.google.com/spreadsheets/d/1d8h8mC_5-BNA4MxwTQ4nxiyiy2ZrLjNx/edit?usp=sharing&ouid=100057136164184761516&rtpof=true&sd=true
- Excel File for Variable Selection: https://docs.google.com/spreadsheets/d/1u4gFm2hIJHRGgXDYAKuso7ccNhZ__lwo/edit?usp=sharing&ouid=100057136164184761516&rtpof=true&sd=true
- Figma Design: https://www.figma.com/proto/071aSyKXtBpVPBVxbTUAAF/Tech-labs-app?type=design&node-i[…]arting-point-node-id=61%3A126&show-proto-sidebar=1&mode=design
- General Project Folder: https://drive.google.com/drive/folders/1vzAeAA1VGCDh4a8nOiIw7X45wuJ8VXhi?usp=sharing
- Jupyter Notebook: https://drive.google.com/file/d/1AoPkDyS7cD_GUTTz2cPtPjcvgugdYHR0/view?usp=sharing
- Neural Network Model Weights for Predicting CVD Risk: https://drive.google.com/file/d/12QmBhI__FI8z1F3Q9AC_KshVXnF2UoxT/view?usp=sharing
- Neural Network Model Weights for Predicting Optimal Food Groups: https://drive.google.com/file/d/1YViSPC-IIQB4DZ_bEpEBHxhPebE_I-5s/view?usp=sharing

- Product Board: https://miro.com/app/board/uXjVN-h6-kU=/?share_link_id=871601001325
- Powerpoint Presentation (PDF): https://drive.google.com/file/d/1p-wrs9XxPUyIzIWQOaMHLHcx0YR_DpvC/view?usp=sharing
- Youtube Video: https://youtu.be/DI9ID8ZVmZw

6.2 References

American Heart Association (AHA). Heart disease and stroke statistics 2017 at-a-glance; 2017. http://www.heart.org/idc/groups/ahamahpublic/@wcm/@sop/@smd/documents/downloadable/ucm_491265.pdf.

Benjamin EJ, Virani SS, Callaway CW, Chamberlain AM, Chang AR, Cheng S, et al.. Heart disease and stroke statistics-2018 update: a report from the American Heart Association. Circulation. (2018) 37(12):e67-e492. doi: 10.1161/CIR.0000000000000558

Center for Disease Control and Prevention (CDC). Heart Disease Fact Sheet; 2017. Center for Disease Control and Prevention (CDC). https://www.cdc.gov/dhdsp/data_statistics/fact_sheets/fs_heart_disease.htm. Accessed 15 Dec 2018.

Center for Disease Control and Prevention (CDC), National Center forHealth Statistics (NCHS). National Health and Nutrition ExaminationSurvey (NHANES). 2018. http://www.cdc.gov/nchs/nhanes/about_nhanes.htm.

Dinh A, Miertschin S, Young A, Mohanty SD. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Med Inform Decis Mak. 2019 Nov 6;19(1):211. doi: 10.1186/s12911-019-0918-5. PMID: 31694707; PMCID: PMC6836338.

Downer S, Berkowitz S A, Harlan T S, Olstad D L, Mozaffarian D. Food is medicine: actions to integrate food and nutrition into healthcare BMJ 2020; 369 :m2482 doi: 10.1136/bmj.m2482.

Roth GA, Mensah GA, Johnson CO, Addolorato G, et al.; GBD-NHLBI-JACC Global Burden of Cardiovascular Diseases Writing Group. Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update From the GBD 2019 Study. J Am Coll Cardiol. 2020 Dec 22;76(25):2982-3021. doi: 10.1016/j.jacc.2020.11.010. Erratum in: J Am Coll Cardiol. 2021 Apr 20;77(15):1958-1959. PMID: 33309175; PMCID: PMC7755038.

Shan Z, Wang F, Li Y, et al. Healthy Eating Patterns and Risk of Total and Cause-Specific Mortality. JAMA Intern Med. 2023;183(2):142–153. doi: 10.1001/jamainternmed.2022.6117