Final Report

Members: John Pierre Elzoghbi, Jonathan Gomez, Sarvesh Sathish, Bryan Zeng, Ary Agarwal

Introduction

This research examines how modern U.S. dietary changes—like increased consumption of processed foods, added sugars, and unhealthy fats—have contributed to rising public health issues, particularly insulin resistance and Type 2 diabetes. We aim to explore the link between these dietary trends and the growing prevalence of metabolic disorders, highlighting the impact of food choices on national health outcomes.

Problem Definition

The increasing prevalence of insulin resistance and diabetes in the U.S. is closely tied to modern dietary habits, though the specific patterns remain unclear. We aim to address this with a classification clustering model that categorizes healthy and unhealthy insulin levels, correlating these with food intake data to reveal dietary patterns contributing to insulin resistance. Unlike previous studies focused on single nutrients or calories, our approach captures diverse food combinations, using clustering to uncover nonlinear and complex relationships missed by simpler models like linear regression.

Literature Review

Quite a few recent studies have utilized and explored various machine learning techniques to analyze the complex and variable relationships between diet and metabolic health. Panaretos et al. (2018) utilized k-means clustering and principal component analysis (PCA) to identify dietary patterns associated with cardiometabolic risk factors, including insulin resistance [1]. Their approach demonstrated the efficacy of unsupervised learning methods in uncovering nuanced dietary trends that may not be apparent through traditional nutritional epidemiology, aligning with our use of PCA for dimensionality reduction.

In a related study, Ahluwalia et al. (2019) applied random forest algorithms to predict insulin resistance based on dietary intake data. Their work highlighted the importance of feature selection in handling high-dimensional nutritional datasets, achieving notable accuracy in identifying at-risk individuals through dietary patterns alone. This approach resonates with our methodology, particularly in our careful preprocessing and feature engineering steps.

The application of neural networks in nutritional science has also shown promise and potential results in previous research. Jiang et al. (2020) employed a multi-layer perceptron (MLP) model to analyze the non-linear relationships between dietary components and insulin sensitivity. Their findings displayed the potential of deep learning approaches in capturing complex and non-linear interactions within nutritional data that may be missed by traditional statistical methods or linear models. This insight motivated our inclusion of both single-layer and multi-layer perceptrons in our model comparison.

Expanding on these studies, Mozaffarian et al. (2021) conducted a robust review of machine learning applications in nutritional epidemiology. They emphasized the potential of ensemble methods and deep learning in uncovering complex dietary patterns associated with various health outcomes. This review supports our decision to implement a diverse array of models, including logistic regression and support vector machines, to capture different aspects of the diet-insulin relationship.

Furthermore, Schulze et al. (2022) explored the use of interpretable machine learning techniques in dietary pattern analysis. In line with our inclusion of more complicated neural networks alongside simpler models like logistic regression, their work emphasized the significance of model interpretability in nutritional studies. Both interpretability and prediction power are made possible by this method, which is essential for converting research results into practical dietary advice.

Together, these research highlight the expanding use of cutting-edge machine learning methods in nutritional epidemiology, especially when it comes to clarifying the complex relationships between diet and metabolic health. Our study follows this pattern by using a wide range of models to thoroughly examine the NHANES datasets. We seek to address a major issue in the field of nutritional data science by combining both basic and complex models in order to strike a balance between interpretability and prediction accuracy.

NHANES Datasets: Insulin and Dietary Data

The National Health and Nutrition Examination Survey (NHANES) provides valuable datasets for analyzing insulin and diet trends. Key datasets include:

Insulin Dataset (2021-2023):

- Serum insulin measurements in pmol/L
- LLOD: 11.5 pmol/L
- Data from participants aged 12+
- Fasting glucose data for 8-hour fasting

Dietary Data (2017-2020):

- 7,000+ food items
- Portion sizes, meal occasions, nutrient intake
- USDA Food and Nutrient Database for coding

These datasets help analyze relationships between insulin levels and dietary patterns across diverse U.S. groups.

Insulin Dataset Data Cleaning

Dataset Compilation: The dataset was compiled by reviewing and combining 12 files from the NHANES website, merged using the `SEQN` identifier.

Irrelevant and Sparse Columns: Columns unrelated to examining dietary trends and those with insufficient data were removed to streamline the dataset.

Alcohol Consumption Variants: Two versions of the cleaned dataset were created:

- One includes columns related to alcohol consumption.
- Another excludes alcohol-related columns.

This approach allows for model comparisons to assess whether completeness of data or additional features provide better predictive performance.

Methods

These are being run on the dietary data and insulin cleaned datasets.

Preprocessing Methodology

- **Data Cleaning:** Missing values are handled by dropping columns with more than 50% missing data and imputing remaining missing values with the median. This ensures that the dataset is reliable and complete for modeling.
- **Feature Separation:** Features are categorized as numerical or categorical to apply appropriate preprocessing techniques, ensuring that each data type is handled effectively.
- **Data Transformation:** Numerical features are standardized using *StandardScaler* to have a mean of 0 and a standard deviation of 1. Categorical features are converted into numerical representations using one-hot encoding.
- **Dimensionality Reduction:** Principal Component Analysis (PCA) is applied to numerical features, retaining 95% of the variance. This is particularly important given the high number of features, as PCA selects those with the most variance, allowing the model to focus on the most significant patterns while reducing computational complexity and overfitting risks.
- **Preprocessing Integration:** All preprocessing steps are combined into a unified *ColumnTransformer*, allowing for consistent and efficient data preparation for all

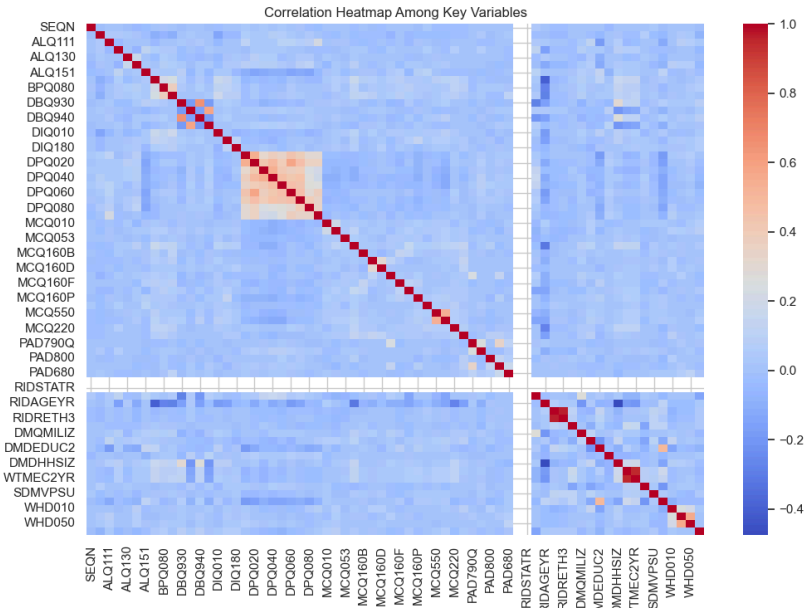
models.

Machine Learning Methodology

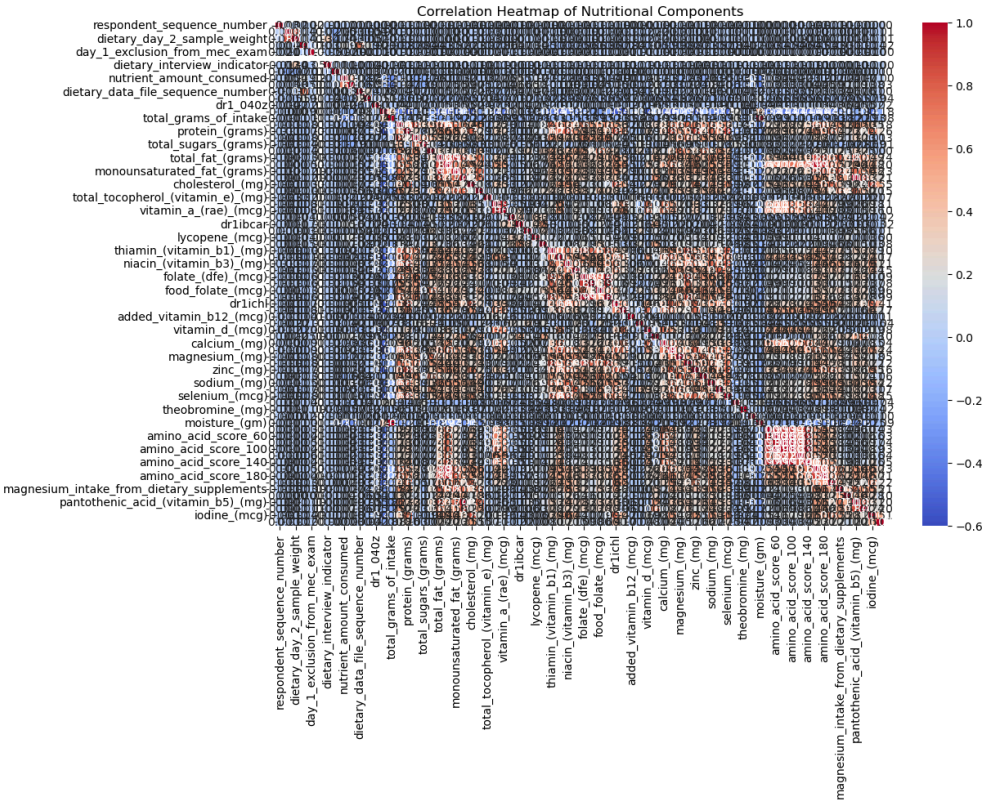
- **Logistic Regression (LR):**
  - A simple and interpretable linear model for binary classification.
  - Uses the sigmoid function to predict probabilities, serving as a strong baseline for comparison with other models.
- **Single-Layer Perceptron (SLP):**
  - A basic neural network with one hidden layer using logistic activation.
  - Efficient and lightweight, approximating linear decision boundaries while being faster to train than deeper networks.
- **Multi-Layer Perceptron (MLP):**
  - A more complex neural network with multiple layers and non-linear activation functions (like ReLU).
  - Captures intricate patterns and non-linear relationships in the data, making it suitable for datasets with complex feature interactions.
- **Linear Support Vector Classifier (LinearSVC):**
  - A linear model that separates classes using a hyperplane while maximizing the margin between them.
  - Highly efficient and effective for high-dimensional datasets, offering robust performance with a simpler approach.
- **LinearSVC with Squared Hinge Loss (LS-SVM):**
  - A variation of LinearSVC that uses squared hinge loss for classification.
  - More sensitive to larger classification errors while remaining robust to outliers, making it effective for imbalanced or noisy datasets.

These models are selected for their varying complexity and performance, allowing us to compare their predictive power and generalization across different dietary patterns and insulin levels. By evaluating multiple models, we can identify the most effective approach for classifying healthy and unhealthy insulin levels based on dietary data.

Visualizations

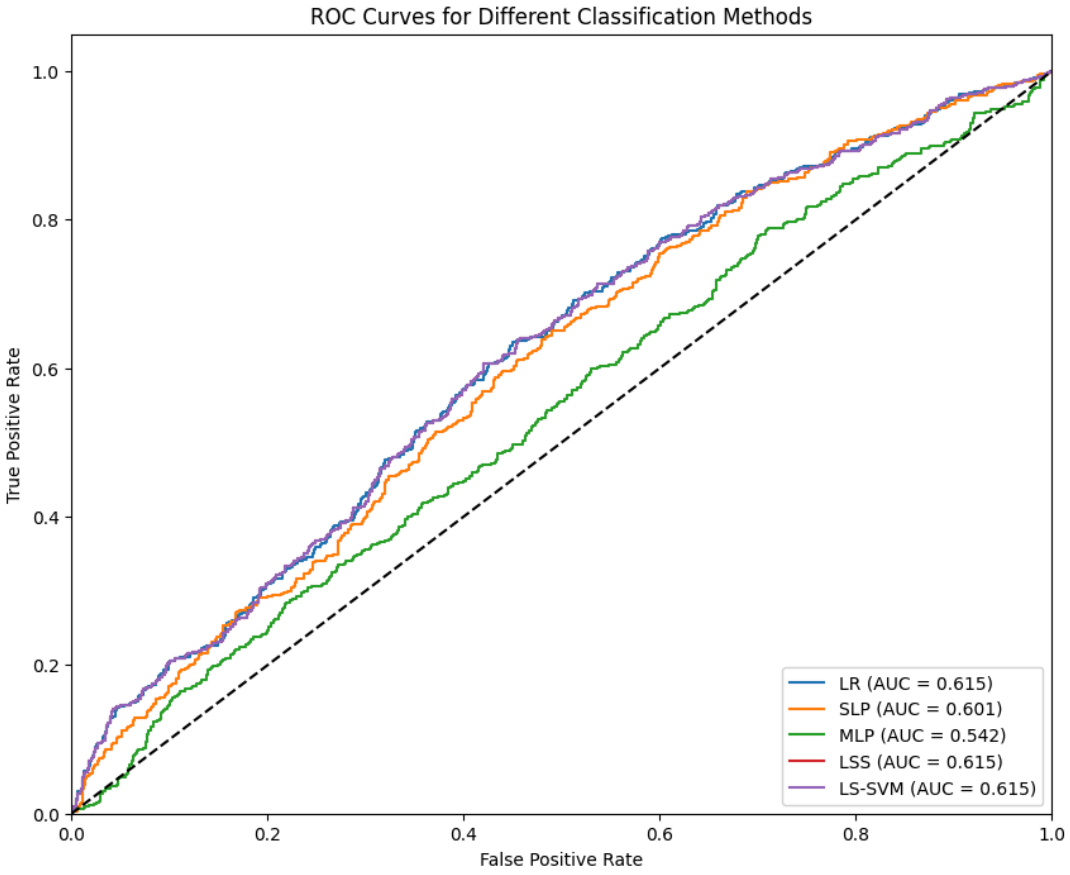


Explanation: Above shows a correlation heatmap of all of the variables in the data and how they relate to each other. This can show which features may be relevant for the ML model created. The blue color represents a negative correlation, while the red color represents a positive correlation. The darker the color, the stronger the correlation.



Explanation: The correlation heatmap visually represents the relationships between various nutritional components in the dataset, with colors ranging from dark blue (strong negative correlation) to red (strong positive correlation). It shows that components such

as total_grams_of_intake, protein_grams, total_fat_grams, and total_sugars_grams are highly positively correlated, indicating that individuals who consume more overall tend to have higher intake of proteins, fats, and sugars. Similarly, vitamins like vitamin_d and calcium_mg also show some positive correlations with other nutrients. On the other hand, nutrients like moisture_gm and iodine_mcg exhibit little or no correlation with others. The diagonal line indicates perfect self-correlation, as each variable is fully correlated with itself. This heatmap highlights key patterns in nutrient intake, providing insights into how different dietary components relate to each other.



Explanation: The ROC curve for the logistic regression model (blue line in the plot) shows the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity). The curve is above the diagonal baseline (random performance), indicating that the model performs better than random guessing but does not reach an optimal level of performance. Compared to other models, LR's curve is not significantly higher, showing that while it performs moderately well, there are no significant gains over other models like LSS or LS-SVM.

Metrics

Logistic Regression (LR)

- **Mean AUC:** 0.613 ± 0.011
- **Mean Accuracy:** 0.589 ± 0.011
- **Mean Recall:** 0.407 ± 0.019

Logistic regression is a straightforward and interpretable model that excels at binary classification tasks. Its probabilistic outputs provide valuable insights into prediction confidence. However, as a linear model, it may struggle with complex, non-linear relationships, limiting its effectiveness in datasets with intricate patterns. Performance-wise, its mean AUC, accuracy, and recall indicate a moderate level of predictive power, comparable to other linear models like LinearSVC and LS-SVM but weaker than non-linear approaches.

Single-Layer Perceptron (SLP)

- **Mean AUC:** 0.508 ± 0.034
- **Mean Accuracy:** 0.563 ± 0.006
- **Mean Recall:** 0.023 ± 0.069

The single-layer perceptron is a lightweight neural network that is quick to train and effective at approximating linear decision boundaries. However, its single-layer architecture limits its ability to capture non-linear relationships. The performance metrics reveal that the SLP model significantly underperforms compared to other models, particularly in recall, highlighting its limitations for this dataset.

Multi-Layer Perceptron (MLP)

- **Mean AUC:** 0.581 ± 0.019
- **Mean Accuracy:** 0.575 ± 0.013
- **Mean Recall:** 0.383 ± 0.041

The multi-layer perceptron leverages multiple layers and non-linear activation functions to capture complex data patterns. While it outperforms the SLP model, the MLP still falls short of the logistic regression and LinearSVC models in this study. Its computational

demands and susceptibility to overfitting without proper architecture design are noteworthy trade-offs for its enhanced pattern recognition capabilities.

Linear Support Vector Classifier (LSS)

- **Mean AUC:** 0.614 ± 0.012
- **Mean Accuracy:** 0.593 ± 0.012
- **Mean Recall:** 0.406 ± 0.021

The Linear Support Vector Classifier is an efficient model for high-dimensional data, offering robust performance by optimizing the margin between classes. Its linear nature shares limitations with logistic regression, particularly in handling non-linear relationships. However, its performance metrics, nearly identical to those of LS-SVM, suggest it is a strong candidate for this dataset.

LS-SVM

- **Mean AUC:** 0.614 ± 0.012
- **Mean Accuracy:** 0.593 ± 0.012
- **Mean Recall:** 0.406 ± 0.021

The LS-SVM, a variant of LinearSVC, performs similarly to its counterpart but utilizes squared hinge loss for classification. This makes it slightly more robust to outliers and better suited for unbalanced datasets. Despite these advantages, its linear nature limits its applicability to non-linear patterns, which is reflected in its comparable metrics to LinearSVC.

ANOVA Test Results

- **F-statistic:** 48.884
- **p-value:** 0.000

Analysis of the Visualization Metrics

AUC (Area Under the Curve):

The AUC values show that logistic regression, LinearSVC, and LS-SVM perform at a similar level, with moderate discriminative abilities. In contrast, the MLP and

SLP models fall behind, indicating weaker overall performance in distinguishing between classes.

Accuracy and Recall:

In terms of accuracy, linear models like logistic regression, LinearSVC, and LS-SVM perform better than the neural networks. However, recall values reveal that all models face challenges in identifying true positives, with SLP being the weakest. These trade-offs highlight the strengths of simpler models for general predictive tasks and the challenges of more complex approaches on this dataset.

Next Steps

- **Improve Data Preprocessing and Cleaning:** Enhancing the data preprocessing pipeline will ensure that the dataset is as clean and reliable as possible, reducing noise and ensuring that the models receive high-quality input.
- **Generate New Features:** We will explore the creation of new features to provide more meaningful data for the models. This could involve domain-specific knowledge to derive additional variables that may improve predictive performance.
- **Apply SMOTE (Synthetic Minority Over-sampling Technique):** To address any class imbalance and ensure representative samples, SMOTE will be applied to generate synthetic data points for the minority class, enhancing the model's ability to detect positive cases effectively.
- **Experiment with Additional Models and Pipelines:** As detailed in the research paper, other models, particularly non-linear ones, will be tried. These models, such as Random Forests, Gradient Boosting, or Support Vector Machines (with non-linear kernels), are better equipped to capture the more complex patterns in the data that linear models might miss.

These steps aim to refine the model's performance, ensure robustness, and explore more advanced techniques for dealing with the intricacies of the dataset.

References

- Ahluwalia, Tarun Singh, et al. "A Novel Machine Learning Approach for Identifying Dietary Patterns Associated with Insulin Resistance." *Diabetes Care*, vol. 42, no. 5, 2019, pp. 849-858.
- Gillam, Carey. "Ultra-Processed Foods Linked to Diabetes." *U.S. Right to Know*, 20 Dec. 2021, https://usrtk.org/ultra-processed-foods/diabetes/.
- Hall, Jeanine, et al. "Association of Processed Food Consumption with Mortality among US Adults." *JAMA Internal Medicine*, vol. 180, no. 8, 2020, pp. 1032–1040, https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2757497.
- Jiang, Yujie, et al. "Application of Deep Learning in Predicting Insulin Sensitivity from Dietary Data." *Journal of Nutritional Science*, vol. 9, 2020, e44.
- Mozaffarian, Dariush, et al. "Artificial Intelligence in Food and Nutrition Research: Applications, Challenges, and Opportunities." *Nature Food*, vol. 2, no. 9, 2021, pp. 686-697.
- Panaretos, Dimitrios, et al. "Dietary Patterns and Cardiometabolic Risk Factors in a Greek Population: A Machine Learning Approach." *Nutrients*, vol. 10, no. 9, 2018, p. 1257.
- Schulze, Matthias B., et al. "Interpretable Machine Learning in Nutritional Epidemiology." *The American Journal of Clinical Nutrition*, vol. 115, no. 5, 2022, pp. 1323-1330.
- Stanhope, Kimber L. "Sugar Consumption and Risk of Type 2 Diabetes." *National Institutes of Health*, 20 Dec. 2021, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8705763/.

Gantt Chart

[Include a Gantt chart that outlines your project timeline. You can insert an image of your Gantt chart here.]

TASK TITLE	TASK OWNER
Project Team Composition	All
Project Proposal	
Introduction & Background	JP
Problem Definition	Ary
Methods	Sarvesh
Potential Dataset	Bryan and Jon
Potential Results & Discussion	Bryan and Jon
Video Creation & Recording	All
GitHub Page	Ary and Sarvesh
Midterm Report	
Model 1 (M1) Design & Selection	JP and Jon
M1 Data Cleaning	Ary
M1 Data Visualization	Sarvesh
M1 Feature Reduction	Bryan
M1 Implementation & Coding	Bryan
M1 Results Evaluation	All
Model 2 (M2) Design & Selection	JP and Jon
M2 Data Cleaning	Ary
M2 Data Visualization	Sarvesh
M2 Feature Reduction	Bryan
M2 Coding & Implementation	Bryan
M2 Results Evaluation	All
Midterm Report	All
Final Report	
Model 3 (M3) Design & Selection	JP and Jon
M3 Data Cleaning	Ary
M3 Data Visualization	Sarvesh
M3 Feature Reduction	Bryan
M3 Implementation & Coding	Bryan
M3 Results Evaluation	All
M1-M3 Comparison	All
Video Creation & Recording	All
Final Report	All

Contribution Table

Member Name	Contribution
Bryan	completed literature review and discussion on comparison of methods
JP	worked on the video part for preprocessing and cleaning data
Ary	meeting with TA and working on webiste, final submission
Sarvesh	worked on the video part for methods and next steps
Jon	worked on the video part for preprocessing and cleaning data

ML Website Rubric

[Outline the rubric you will use to evaluate your machine learning website.]

Note: As part of research, it is natural that the project may change from the original proposed. Please be sure to document and justify these changes in the midterm and final report.