Aryan Ringshia (aryanrr)
Ishan Saksena (saksena)
Neil Mankodi (nmankodi)

Project Report

INTRODUCTION:

The prevalence of cardiovascular diseases poses significant public health challenges worldwide. In our project, we aim to leverage machine learning techniques to analyze data from the National Health and Nutrition Examination Survey (NHANES). Our overarching goal is to explore the relationship between lifestyle habits, biological features, and the likelihood of cardiovascular problems, specifically focusing on symptoms such as shortness of breath and chest pain, as well as the risk of high blood pressure. For the purposes of this project, we will be focusing on two main questions:

1. How do the patterns of reported cardiovascular symptoms in adults vary based on their smoking habits, & what insights can be gained regarding this risk categorization?
2. How does the frequency of alcohol consumption influence the risk of high blood pressure / hypertension compared to healthy blood pressure among adults, adjusting for age, gender, smoking status, & other lifestyle specific features?

Understanding cardiovascular symptom patterns and blood pressure's link to lifestyle habits is crucial. Firstly, it allows us to gain insights into the risk factors associated with cardiovascular diseases, which can inform targeted interventions and preventive measures. Moreover, identifying individuals at higher risk enables early intervention and personalized healthcare approaches, thereby potentially reducing the burden of cardiovascular morbidity and mortality.

To address these questions, we'll use various machine learning algorithms: Decision Trees, Random Forests, XGBoost, K-Nearest Neighbors (KNN), and Support Vector Classifier (SVC). We'll evaluate model performance using key metrics like F1 score, precision, and recall, prioritizing F1 score to balance precision and recall. Baselines will be established to compare our models' effectiveness against simple heuristic approaches.

DATA:

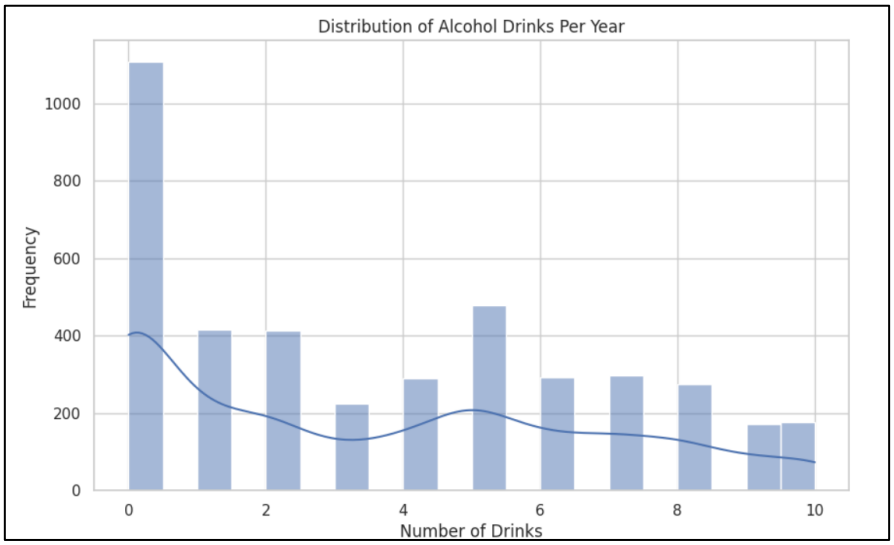
The primary source of our data is the NHANES for the years 2017-20, a comprehensive survey conducted by the Centers for Disease Control and Prevention (CDC). NHANES collects data through interviews, physical examinations, and laboratory tests, providing a rich source of information on various health-related factors. To conduct our analysis, we merged the following datasets:

1. Body Measures
2. Cardiovascular Health
3. Demographics
4. Alcohol Use
5. Smoking – Cigarette Use
6. Smoking - Recent Tobacco Use (SMQRTU_J)
7. Blood Pressure - Oscillometric Measurements
8. Smoking - Recent Tobacco Use (P_SMQRTU)
9. Dietary Interview - Total Nutrient Intakes
10. Cholesterol

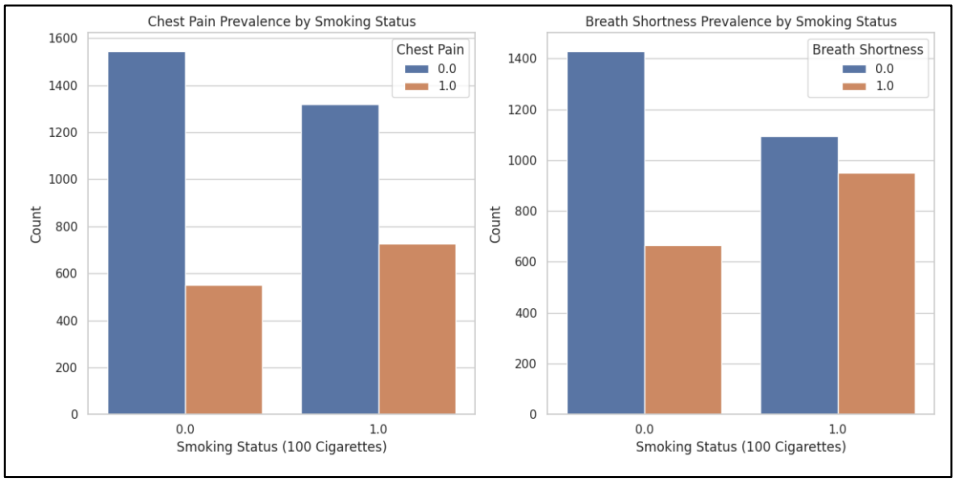
This merging process was essential to ensure that we had access to a comprehensive set of variables that captured relevant information for addressing our research questions. During the merging process, we performed extensive data cleaning, which involved handling missing values, filtering out unrealistic data points, transforming variables as necessary, and renaming columns to enhance clarity and consistency. Additionally, we encoded categorical variables to align with our analytical expectations and facilitate model training.

After merging and cleaning the data, we obtained a final dataset with 4142 observations and 18 variables, each representing different aspects of participants' demographics, lifestyle habits, biomarker measurements, and health outcomes. To gain insights into the characteristics of our dataset and the variables under consideration, we conducted exploratory data analysis. This process involved several steps:

1. Observing Statistical Summaries - We utilized summary statistics to understand the central tendency and variability of continuous variables. Additionally, we used the info function to obtain a concise summary of the dataset's structure.
2. Observing Distributions of Features - We visualized the distributions of key features such as age, cholesterol levels, body mass index (BMI), sodium and potassium levels, and blood pressure measurements. This analysis facilitated a deeper understanding of the data.
3. Visualizing relationships among features – We created visualizations that let us observe the interplay of different features and extract insights which could prove useful in downstream modeling tasks. For this purpose, we utilized visualization techniques such as kde plots, histograms, count plots, box plots, bar plots etc. All of these visualizations can be found in the code files, but we will focus on few of the more interesting visualizations and the respective insights offered by each.

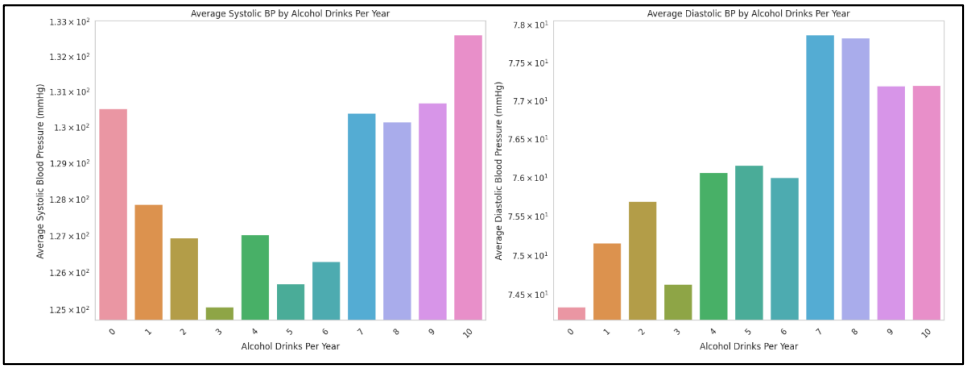


- In this KDE bar plot we visualize the number of subjects in our data against the number of alcoholic drinks they consume. We see that the general trend is that of a decreasing one with fewer people consuming more drinks. At the same time, we can see that we have a sufficiently diverse set of subjects which should bode well for the generalizability of our models.



- In these bar plots we attempt to compare the number of people who reported having chest pain and shortness of breath after grouping on the basis of their smoking habits. The intuition here was that we expect to see regular smokers have higher rates of reporting such cardiovascular issues. This was confirmed from the visualization where we observe that the regular/heavy smokers (denoted as group 0) had higher number of reports of facing such symptoms as compared to the non-smokers

(denoted as group 1). This also gives us confidence that using smoking habits to predict the likeliness of a subject having such cardiovascular issues is a viable approach.



- In these bar plots we visualize the average systolic and diastolic blood pressure against the number of alcoholic drinks consumed by the subjects. We observe that in general, subjects that tend to drink more, have higher average blood pressure levels thus putting them at higher risk of facing hypertension related consequences. This insight alludes to the fact that using the drinking habits of an individual to predict their hypertension risk is a method well grounded in logic which should achieve sufficiently good results.

QUESTION 1 →

Features and Response Variables:

- Features: The prediction model utilizes demographic data (Race, Age, Gender) and health factors (Cholesterol, BMI, Sodium, Potassium). It also factors in smoking and alcohol habits (cigarette count, recent tobacco use, alcohol frequency, binge drinking).
- Response Variable: We created a binary response variable, Final_Cardio_Symptom, by combining reported chest pain (Chest_Pain) or shortness of breath (Breath_Shortness). If either symptom was reported, the variable is set to 1, indicating the presence of cardiovascular symptoms; otherwise, it's set to 0. This variable serves as our target for classification analysis.

Preprocessing: We pre-processed our data to prepare it for machine learning models. This involved One-Hot Encoding for 'Race' and Standard Scaling for numerical features. One-Hot Encoding converts categorical variables into binary vectors, while Standard Scaling standardizes numerical features, reducing scale differences among them.

Methods:

To address the question of how the patterns of reported cardiovascular symptoms in adults vary based on their smoking habits, we employed several analytical approaches:

- **Baseline Models:**
  To benchmark the performance of our machine learning models, we utilized baseline models suited to our binary classification task. This helps compare our models against simple heuristic approaches and set a reference point for evaluation.
  - Baseline 1: Random Guessing - This baseline sets a reference accuracy and F1 score by random prediction, simulating no meaningful relationship between predictors and the target variable.
  - Baseline 2: Most Frequent Class - Always predicting the most frequent class in the training data, this baseline offers a basic heuristic based on class distribution, serving as a simple rule-of-thumb approach.
- **Machine Learning Models:**
  Each model was trained using a pipeline that included pre-processing steps and the respective classifier.
  - **XGBoost (eXtreme Gradient Boosting):** XGBoost is a distributed gradient-boosted decision tree algorithm known for its scalability and performance. It sequentially builds multiple decision trees, where each tree corrects the errors of its predecessors, making it suitable for our

binary classification task. The algorithm handles both numerical and categorical features effectively, making it well-suited for our dataset. Moreover, regularization techniques help prevent overfitting.

$$\hat{y}_i^{(n)} = \hat{y}_i^{(n-1)} + \eta \cdot f_n(x_i)$$

- **Random Forest:** Random Forest is an ensemble learning method based on decision trees. It creates a multitude of decision trees during training and outputs the class that is the mode of the classes (classification) of the individual trees. Random Forest is robust to overfitting and works well with categorical features. Its ensemble nature allows it to capture complex interactions between features and identify important predictors.

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} f_b(x)$$

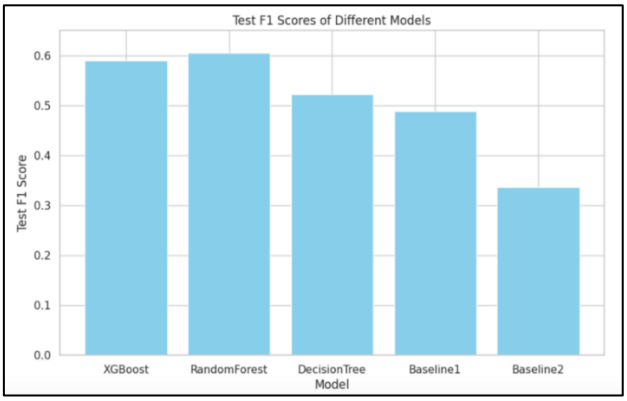
- **Decision Tree:** Decision trees are simple yet powerful models for classification tasks. They partition the feature space into regions and make predictions based on majority voting in each region. Decision trees are interpretable and can handle both numerical and categorical features, making them suitable for our problem. They traverse from root to leaf according to splitting rules, assign value at leaf.
- **Cross Validation:** To ensure the robustness of our models and assess their generalization performance, we employed cross-validation during model training. Specifically, we used Stratified K-Fold cross-validation, which preserves the class distribution in each fold. This approach helps mitigate the risk of overfitting and provides a more reliable estimate of the model's performance on unseen data.

Results:

To evaluate the performance of our models, we used the F1 Score macro as the evaluation metric. The F1 Score macro provides a balanced assessment of a model's performance, considering both precision and recall. By maximizing the F1 Score macro using cross-validation, we aimed to find the optimal model. Here are the results for each model:

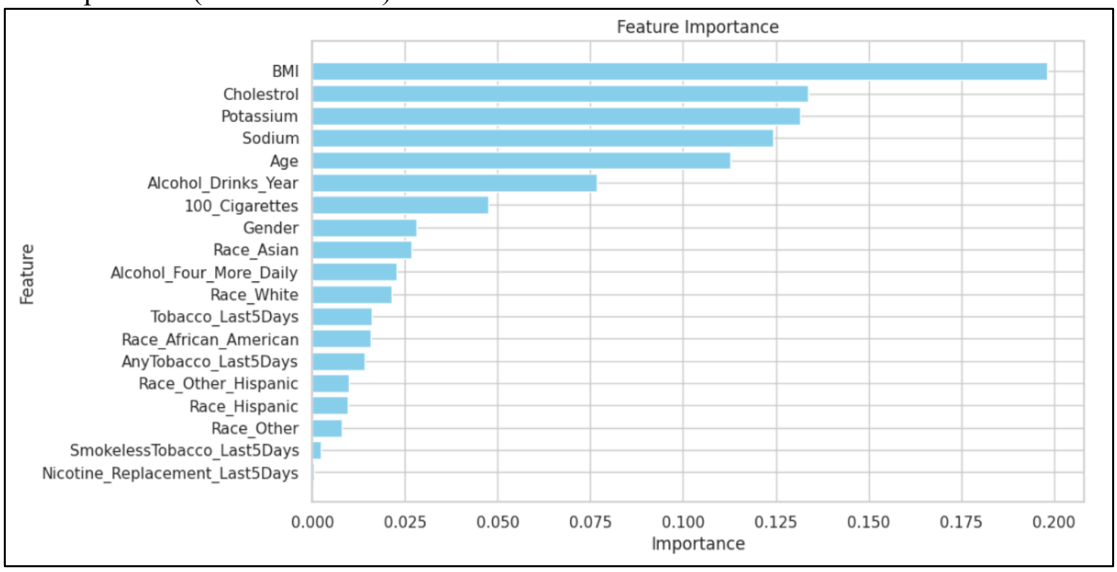
Model	Best CV F1 Score	Train F1 Score	Test F1 Score
Baseline 1 - Random Guessing	--	0.51	0.52
Baseline 2 - Most Frequent	--	0.34	0.34
XGBoost	0.62	0.67	0.59
Random Forest	0.62	0.86	0.61
Decision Tree	0.55	0.75	0.52

Both XGBoost and Random Forest surpassed the baselines and Decision Tree in test F1 score, showcasing their efficacy in predicting cardiovascular symptoms from smoking habits. XGBoost and RandomForest, being ensemble methods, leverage the strengths of multiple decision trees and handle both categorical and numerical features effectively.



Test F1 Scores of Different Models

Feature Importance (Random Forest):



The feature importance analysis for Random Forest revealed that BMI, cholesterol, potassium, sodium and age were the most important features for predicting cardiovascular symptoms.

Conclusions:
The results demonstrate that our machine learning models, particularly Random Forest, outperform baseline models. By employing machine learning models and conducting feature importance analysis, we gained a deeper understanding of the factors influencing the occurrence of cardiovascular symptoms in adults. Our findings suggest that factors such as BMI, cholesterol, potassium, sodium, and age play a significant role in the manifestation of cardiovascular symptoms. Additionally, smoking at least 100 cigarettes in life had an impact on the response. However, it is essential to acknowledge potential limitations in our analyses. For instance, our models rely on self-reported data, which may introduce biases and inaccuracies. Furthermore, other confounding factors not accounted for in our models could influence the results. Further research incorporating additional variables and addressing potential biases could enhance our understanding of relationship between smoking habits and cardiovascular symptoms.

QUESTION 2 →
Features and Response Variable:

- Features: Demographic (Age, Gender), Health (Cholesterol, BMI, Sodium, Potassium levels), Lifestyle (Alcohol frequency, Smoking status)
- Response Variable: "Category" indicates blood pressure levels categorized as 'High' or 'Normal' based on systolic and diastolic thresholds.

Data Categorization:
We began by defining thresholds for systolic and diastolic blood pressure—120 and 70 mmHg, respectively—to categorize blood pressure readings into 'High' and 'Normal'. This categorization was applied across the dataset, after which the original blood pressure columns were dropped to focus on these newly derived categories.

Preprocessing:
Our data preprocessing involved several crucial steps to prepare the dataset for effective model training:

- One-Hot Encoding: Applied to categorical variables such as 'Race' to transform them into a format suitable for machine learning algorithms.
- Standard Scaling: Used for numerical features like Age, Cholesterol, BMI, Sodium, and Potassium to normalize their distribution and reduce scale disparities.
- Data Splitting: The dataset was stratified and split into training (80%) and testing (20%) sets to ensure that both subsets represent the overall distribution of the response variable.

Methods:

- **Baseline Models:**
  To establish a point of comparison for our predictive models, we implemented two baseline strategies:
  - Random Guessing: We simulated a scenario where predictions were made randomly, providing a baseline accuracy and F1 score.
  - Most Frequent Class: We used the most frequent category within the training data as a constant prediction, offering a simple, though often effective, baseline for comparison.
- **Machine Learning Models**:
  To investigate the influence of lifestyle factors on hypertension, we selected three machine learning models, chosen for their distinct characteristics and suitability for classification tasks:
  - **KNN (K-Nearest Neighbors)**: Chosen for its ability to make predictions based on the proximity to the nearest labeled data points, which is intuitive for understanding lifestyle similarities. It classifies x by majority vote among the k-nearest neighbors.
  - **Random Forest:** An ensemble model known for its high accuracy and robustness, capable of handling a mix of numerical and categorical data, and providing insights into feature importance.

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} f_b(x)$$

  - **SVC (Support Vector Classifier)**: Selected for its effectiveness in finding the optimal hyperplane that best separates the data into two categories, making it useful for binary classification tasks like ours.

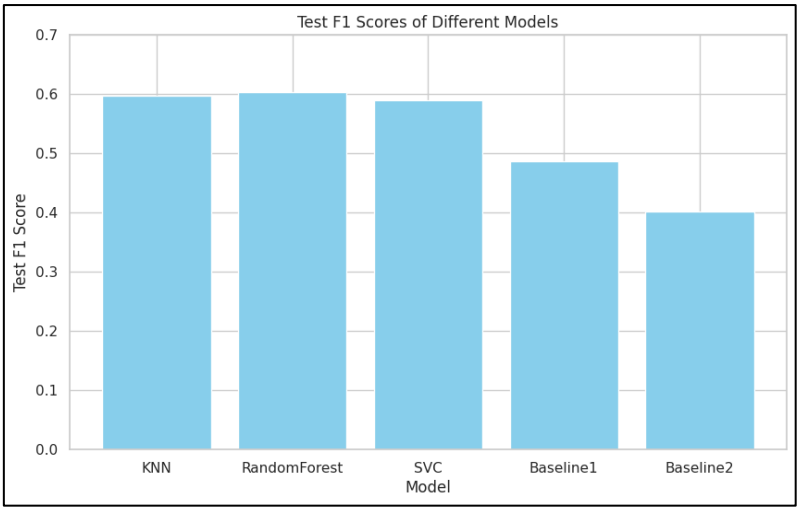
$$f(x) = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b$$

  - **Cross Validation:** To ensure the robustness of our models and assess their generalization performance, we employed cross-validation during model training. Specifically, we used Stratified K-Fold cross-validation, which preserves the class distribution in each fold. This approach helps mitigate the risk of overfitting and provides a more reliable estimate of the model's performance on unseen data.

Results

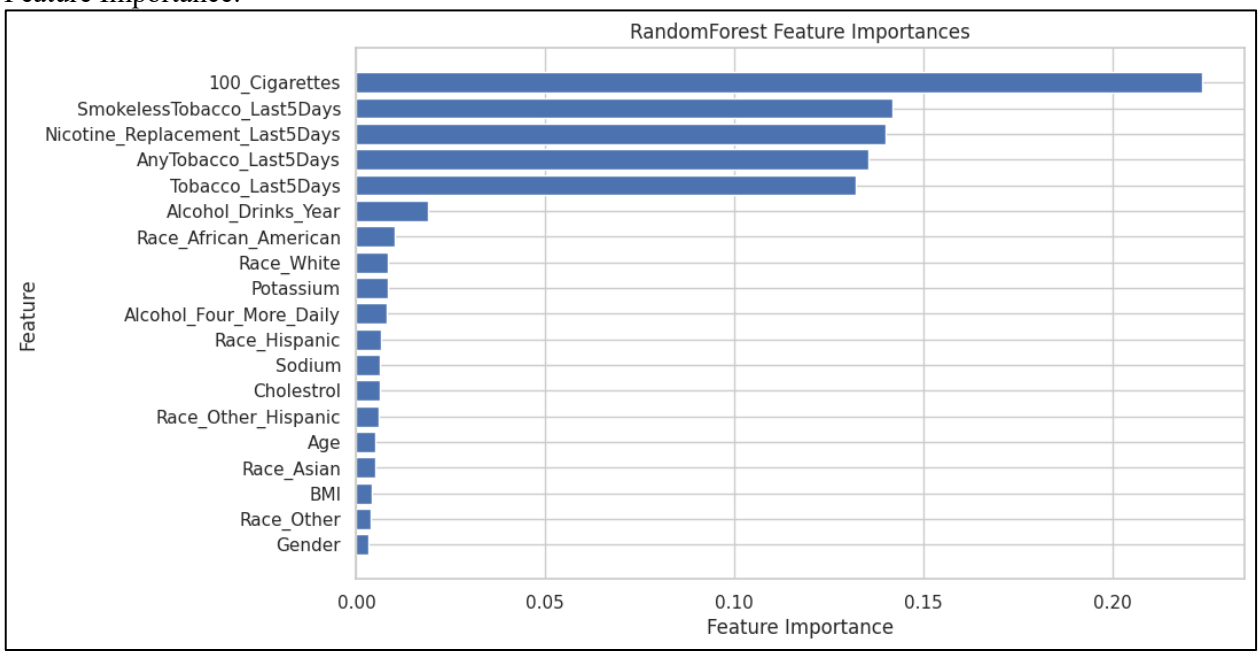
The effectiveness of our models was evaluated using the F1 Score macro, which considers both precision and recall, providing a balanced metric that is crucial for models where class imbalance might influence accuracy. Here are the results of each model:

Model	Best CV F1 Score	Train F1 Score	Test F1 Score
Baseline 1 - Random Guessing	--	0.51	0.52
Baseline 2 - Most Frequent	--	0.34	0.34
KNN	0.61	0.68	0.60
Random Forest	0.60	1.00	0.60
SVM	0.59	0.65	0.59



Test F1 Scores of Different Models

Feature Importance:



Key Findings from Feature Importance Analysis:

- **Tobacco Use**: Smoking-related features such as '100_Cigarettes', 'Tobacco_Last5Days', and 'AnyTobacco_Last5Days' were among the top predictors. This highlights the significant impact of smoking on hypertension, consistent with existing research linking tobacco use with elevated blood pressure levels.
- **Dietary Habits**: Alcohol consumption features like 'Alcohol_Drinks_Year' and 'Alcohol_Four_More_Daily' were also important, albeit less so than tobacco use. This suggests that while alcohol consumption does play a role, its impact might be less direct or less significant compared to smoking.
- **Biometric Data**: Health indicators such as 'BMI', 'Cholesterol', 'Sodium', and 'Potassium' levels also featured in the analysis. High cholesterol and BMI are well-documented risk factors for hypertension, underscoring the model's alignment with medical understanding.

Potential Misleading Aspects:

While the results are promising, several factors could potentially mislead interpretations:

- Self-reported Data: Our models rely heavily on self-reported data, which can introduce biases due to inaccurate or incomplete reporting by participants.

- Model Overfitting: Despite precautions such as cross-validation, there is always a risk that complex models like Random Forest could overfit the training data, especially if not properly tuned or if the dataset is not sufficiently large or diverse.
- Variable Omission: Important variables that might influence hypertension, such as genetic factors, stress levels, or detailed dietary information, were not included in our analysis.
- Generalizability: The conclusions drawn from this study are based on the specific demographic and data used in the analysis. These results might not generalize to different populations or demographic groups that were not adequately represented in the dataset.

Conclusions

The analysis confirms that lifestyle factors, particularly smoking, are significant predictors of hypertension. Random Forest, despite a tendency to overfit, provided valuable insights into feature importance, helping prioritize targets for intervention.

While models performed well compared to baselines, addressing potential biases in self-reported data and exploring feature interactions further are areas for improvement.

CONTRIBUTIONS:

Our team consists of 3 members and the individual contributions are as follows:

1. Neil Mankodi
    a. Sourcing data from NHANES. Selected relevant datasets that can be merged.
    b. Cleaning and Merging datasets.
    c. Performing Exploratory Data Analysis and gathering insights.
    d. Creating Baselines for both questions.
    e. Creating and formatting the final report
2. Aryan Ringshia
    a. Developed models for solving Question 1. This included models such as Decision Tree, Random Forest, and XGBoost. These models were finetuned using grid search cross validation to obtain best set of hyperparameters.
    b. Evaluating and visualizing the performance of these models and related baselines.
    c. Creating and formatting the final report.
    d. Documenting and cleaning code files.
3. Ishan Saksena
    a. Developed models for solving Question 2. This included models such as K Nearest Neighbors, Random Forest, and Support Vector Classifiers. These models were finetuned using grid search cross validation to obtain best set of hyperparameters.
    b. Evaluating and visualizing the performance of these models and related baselines.
    c. Creating and formatting the final report.

REPRODUCIBILITY:

Ensuring the reproducibility of our analysis is fundamental to maintaining the integrity of our findings. To this end, we have meticulously documented every step of our methodology in the code file. From data sourcing and cleaning to exploratory data analysis, model development, and evaluation, each process is thoroughly detailed for clarity and transparency. Our choice of machine learning models is carefully justified and extensively documented, both within the code and in our final project report. Furthermore, we have fine-tuned the hyperparameters of our models using techniques like grid search, with the optimal configurations clearly presented in the code outputs. This information empowers others to replicate our results accurately and explore variations or enhancements to our approach. In summary, our project documentation, encompassing code files and the final report, contains all the essential details for complete reproducibility.