Topics In Data Science Midterm — Diabetes Dataset

Alex Bean
Cornell Tech - ORIE
atb95@cornell.edu

Utkarsh Goyal
Cornell Tech - ORIE
ug42@cornell.edu

Abstract

Diabetes disproportionately impacts certain populations, emphasizing the need to identify the factors contributing to its prevalence and inform tailored interventions. This study analyzes data from the 2016 National Health Interview Survey (NHIS) to examine the intersection of racial disparities with variables such as Body Mass Index (BMI), age, hypertension, and family history of diabetes. Addressing the question, "How should we tailor health policies with respect to Race?" the study uses XGBoost and LightGBM not for predictive modeling but as tools to enhance interpretability and uncover significant feature associations. The analysis highlights that higher BMI, older age, hypertension, and family history of diabetes are strongly associated with diabetes prevalence, with marked differences across racial groups. The exploratory approach emphasizes patterns and disparities that inform public health strategies rather than focusing on predictive outcomes. Results demonstrate the critical need for racially targeted health policies, such as improving equitable access to preventive care and education for communities disproportionately affected by diabetes. Insights from this study contribute to data-driven public health strategies, promoting resource allocation and interventions that address disparities in diabetes prevention and management across diverse populations.

1 Introduction

Introduction

The question, *"How should health policies be tailored with respect to race?"*, is pivotal in addressing persistent health disparities in the United States. This paper examines the intersection of racial disparities with critical health conditions, focusing on diabetes and hypertension as representative examples. To ensure the validity of our analysis, we undertook a rigorous data cleaning process to address common inaccuracies, including miscalculated BMI values and outliers in weight, height, and job duration. Missing data were imputed using stratified medians based on race and sex, and variables were re-categorized to enhance interpretability. Descriptive statistics were employed to highlight key patterns in smoking prevalence, hypertension, and access to medical care across different racial groups.

The primary objective of this paper is to investigate how demographic, health, and socioeconomic variables interact with race. To this end, we leveraged comprehensive datasets from the 2016 *National Health Interview Survey* (NHIS) and the *National Health and Nutrition Examination Survey* (NHANES), which together provide detailed insights into over 800 health-related characteristics. These datasets allowed us to explore cultural nuances within racial groups and analyze the interplay between diseases, lifestyle factors, and comorbid conditions, offering a holistic understanding of these relationships.

According to the Centers for Disease Control and Prevention (CDC), an estimated 12.2% of U.S. adults had diabetes in 2015, including one in four individuals aged 65 and older[1]. Diabetes is not only

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

a critical public health issue due to its widespread prevalence but also because of its disproportionate impact on certain racial and ethnic groups. This paper aims to provide actionable insights into addressing diabetes disparities, recognizing its societal significance and the personal stakes involved. For instance, Utkarsh's family has been deeply affected by diabetes, underscoring the tangible and often inequitable challenges families face in managing chronic conditions. This personal connection reinforces our commitment to developing equitable healthcare policies that address these systemic disparities effectively.

2 Overview of Data

The dataset used in this analysis comes from the 2016 National Health Interview Survey (NHIS), a longstanding initiative conducted by the National Center for Health Statistics (NCHS), a division of the Centers for Disease Control and Prevention (CDC). Established in 1957, the NHIS monitors the health of the civilian, noninstitutionalized population in the United States and is a key source of data on chronic diseases, healthcare access, health behaviors, and other health-related topics. The survey's insights guide public health policy, track national health objectives, and support health-related research. This analysis utilizes data from both the Sample Adult file, which focuses on detailed health information for one adult per household, and supplemental modules from the 2016 NHIS.

The 2016 NHIS introduced a revised sample design to reflect changes in the U.S. population distribution since the previous 2006 design. Commercial address lists replaced field listings as the primary sampling source, and oversampling for black, Hispanic, and Asian populations was discontinued. The dataset contains over 800 features, presenting computational challenges and requiring preprocessing to ensure analytical focus. For this analysis, preprocessing reduced the dataset to relevant health-related features by re-categorizing low-frequency responses, mapping numerical codes to meaningful labels, and addressing outliers and missing data. Misreported values, such as extreme weights and heights, were replaced with valid entries or excluded as appropriate. Features such as Body Mass Index (BMI) were recalculated after resolving misclassifications. The resulting cleaned dataset offers a refined, actionable foundation for further exploratory and statistical analyses.

2.1 Sample Design

- The NHIS uses a tiered sampling method that begins by selecting Primary Sampling Units (PSUs; individual counties or clusters of nearby counties). Following this, households within the PSUs are sampled, and one adult per household is randomly chosen for the Sample Adult file. This process ensures that the data represents a wide cross-section of both urban and rural populations across the United States.
- To enhance representation, the NHIS utilizes stratification to divide the population into subcategories based on geographic location and urban or rural settings. Furthermore, clustering is used to group households within the same geographic area, reducing logistical costs while introducing some level of correlation among data points within a cluster.
- The survey's sampling framework is composed of three distinct elements: unit frames derived from commercial address databases, area frames covering locations without standard addresses, and frames for college dormitories within selected PSUs. This structure ensures that all eligible populations are appropriately covered by the sample design.
- To capture seasonal variation and avoid biases associated with specific times of the year, the NHIS conducts data collection continuously throughout the year. This strategy allows for a more accurate reflection of health-related behaviors and conditions across different seasons.

3 Preprocessing & Model Inclusion

3.1 Load and Clean Files

The `clean_feature_data` function processes and cleans a dataframe by grouping smaller categories into "OTHER," mapping numerical and missing values to meaningful labels, and imputing missing values for variables like `weight`, `height`, and `years_on_job` using stratified medians based on

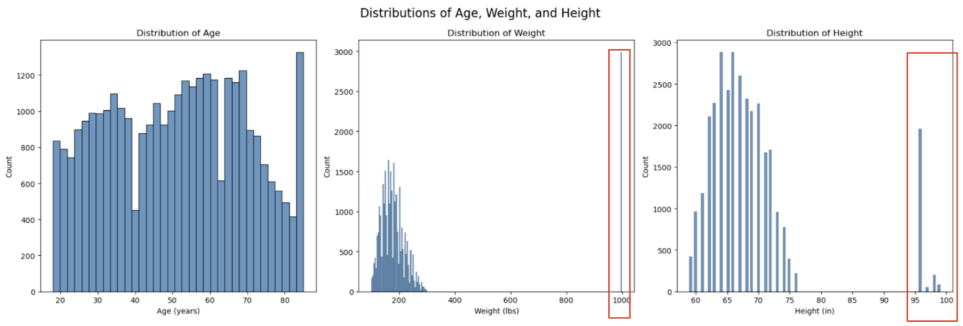
demographic attributes such as `sex` and `non_hispanic_race`. It uses a dictionary of mappings to transform binary indicators into “YES”/“NO” labels and to categorize races and ethnicities with more descriptive names. Extreme values for `weight`, `height`, and `years_on_job` are replaced with NaN before imputation to ensure accurate representation. A new `bmi` column is calculated using weight and height, while an `age_group` column categorizes individuals into life stages based on predefined age bins. Meanwhile, the `load_diabetes_data` function loads a dataset, samples one person per household using the `HHX` identifier, and generates unique IDs combining household, family, and personal IDs. It creates binary flags for health conditions like diabetes, hypertension, and heart disease by converting specific values from the raw dataset. Additional columns for demographics and socioeconomic variables, such as `class_of_worker` and `primary_care`, are renamed for clarity and consistency. Missing values are standardized as “MISSING,” ensuring consistency across the dataset. Relevant variables are carefully selected while irrelevant ones are excluded, resulting in a simplified and focused structure. Together, these functions streamline the workflow by extracting, transforming, and cleaning the data, preparing it for further analysis.

3.2 Preprocessing Rationale & Action

As with any exploratory data analysis process, understanding the dataset is a crucial first step. Upon loading the data and inspecting the data frame, several issues were identified, including miscalculated BMI values that were incorrect or inconsistent. The original BMI variable was excluded temporarily to identify and resolve the root cause of these inaccuracies before recalculating it.

To gain insights into the dataset, the distributions of key variables such as age, weight, and height were visualized using histograms. These visualizations provided an understanding of the data’s distribution and facilitated the detection of outliers or skewness. Extreme values, such as weights exceeding 300 lbs or heights above 77 inches, were identified as misrepresented or unclassified entries, often encoded with placeholder values like “999 lbs" or “99 in." These placeholder values, representing instances where respondents declined to provide information, were systematically removed during preprocessing.

Further examination of the survey methodology confirmed that the age distribution, though initially appearing skewed, accurately represented the dataset’s demographic focus. A significant proportion of respondents were older adults (85 years and older), consistent with the dataset’s emphasis on chronic conditions prevalent in this age group. The histograms below illustrate the distributions and the steps taken to address missing or misrepresented data.



To address the challenges of data organization and focus, it was deemed necessary to develop a structured approach that integrated both quantitative and qualitative aspects of disease, race, and culture. This effort began with the creation of a new data frame, `df_clean`, as part of the feature engineering process. During the initial phases, low-frequency values within specific columns were re-categorized into an "OTHER" category to enhance clarity and reduce fragmentation. This approach was particularly applied to features such as `class_of_worker`, `more_than_one_job`, `ever_worked`, and `skipped_meds`. In these features, codes 7, 8, and 9 represented responses such as “refused," “did not ascertain," and “don’t know," which were consolidated into a single "OTHER" category.

Interpretability was further improved by mapping numerical codes (e.g., 0 and 1) to descriptive labels such as 'YES' and 'NO,' providing clearer insights into variables related to health and demographic data, including diabetes and hypertension. Additionally, the dataset was refined by analyzing the distribution of participants within the `age_group` column, which was based on predefined age bins. This facilitated the extraction of insights across a diverse range of features, encompassing health conditions like coronary heart disease, lifestyle factors such as smoking, and socioeconomic indicators like medication affordability. The focus remained on maintaining clarity and consistency within descriptive statistics to support meaningful analysis.

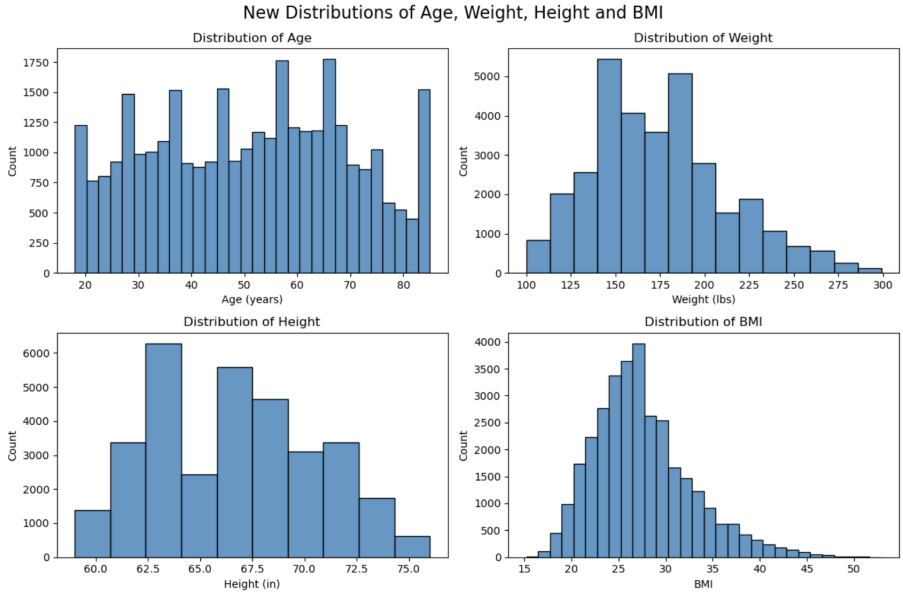
A stratified median imputation method was employed, using race and sex as grouping features, to address missing values. This process was implemented in the `clean_data.py` file located within the functions folder, which was executed at the beginning of the exploratory data analysis (EDA) phase. The data cleaning script also included processes such as grouping low-frequency categories in several columns into an "OTHER" category, creating mappings for categorical variables by translating numerical codes into descriptive labels (e.g., binary health indicators like diabetes and hypertension were labeled as 'YES' and 'NO,' and race and ethnicity codes were explicitly named), and systematically addressing missing values and outliers.

For instance, extreme values for weight and height were replaced with `NaN`, while job duration values exceeding 95 years were similarly set to `NaN`. Missing job duration values were replaced with zero to improve consistency. These cleaning steps provided the opportunity to recalculate BMI using the standard formula:

$$\text{BMI} = \frac{\text{Weight in pounds} \times 703}{(\text{Height in inches})^2}$$

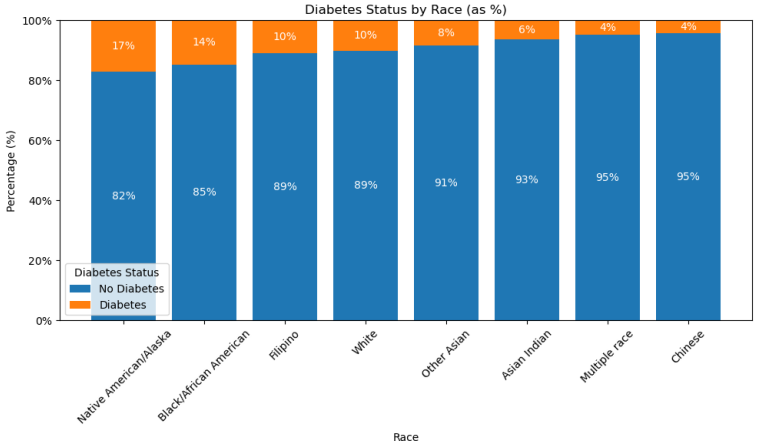
This recalculated BMI variable was integrated into the cleaned dataset to ensure accurate and consistent analysis.

To ensure more accurate imputation, stratified medians were calculated for missing values of weight, height, and job duration. These medians were derived based on the features `sex`, `non_hispanic_race`, and `class_of_worker`. This approach allowed the imputation process to account for demographic and occupational differences, resulting in a more representative dataset. The updated distributions for age, weight, and height, reflecting these adjustments, are presented below:



3.3 Descriptive Statistics

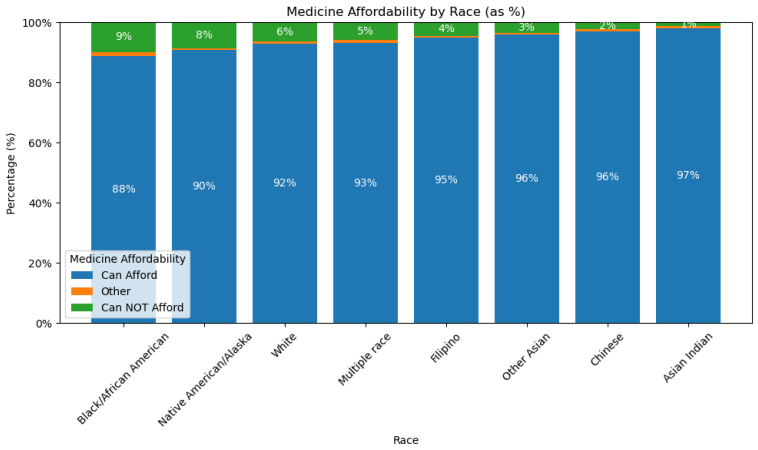
The cleaned dataset facilitated a detailed analysis of disease prevalence across racial groups. For example, aggregated counts of surveyed individuals with and without diabetes were calculated, followed by the computation of total counts per race to derive percentage values. Unreleasable race categories were excluded from the analysis. The results were visualized using a stacked bar chart, with separate bars representing 'Diabetes' and 'No Diabetes,' sorted by the proportion of diabetes within each racial group. Percentage labels were added to each stack to enhance readability.



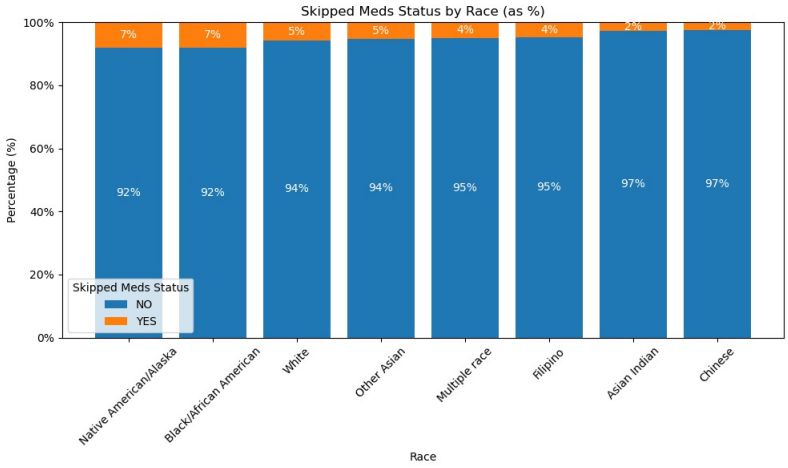
A similar approach was applied to smoker and hypertension statuses. For each status, counts were grouped by race, total counts were computed, and percentages were derived. Stacked bar charts were then created with 'Yes' and 'No' stacks for each status. These visualizations highlighted racial disparities in smoking and hypertension prevalence, identifying groups with higher proportions, such as African Americans and Native Americans. This analysis provided valuable insights into the populations most affected by these conditions.

This stage of the research marked a significant turning point, as it enabled the identification of racial groups leading in categories most correlated with disease prevalence. To gain a deeper understanding of the cultural and socioeconomic disparities affecting these groups, economic status was examined. Medical access was analyzed by evaluating the ability of surveyed racial groups to afford medication, categorized into 'Can_Afford,' 'Other,' and 'Can NOT Afford.'

Using similar methodologies, percentages were calculated for each category, and a stacked bar chart was created with percentage labels for each segment. The visualization table highlighted significant disparities, revealing that African Americans and Native Americans were disproportionately represented in the 'Can NOT Afford' category. These findings provided further evidence of the socioeconomic challenges faced by these groups in accessing necessary medication.



Finally, the analysis focused on the tendency to skip medication across racial groups. Responses were grouped by racial category and `skipped_meds` status ('YES' and 'NO'), and the distribution of responses was computed and visualized. The resulting stacked bar chart reinforced previous findings, indicating that African Americans and Native Americans were the most likely racial groups to miss taking prescribed medication.



3.4 Affected Groups by Condition

Bringing these findings together, a table was created to highlight the significant health disparities among racial groups by condition. The analysis revealed consistent patterns of marginalization, particularly among Native Americans and African Americans, allowing for a deeper interpretation of the critical health conditions faced by these groups. Native Americans are disproportionately affected by diabetes and smoking, conditions likely shaped by cultural and socioeconomic challenges. Similarly, African Americans exhibit the highest prevalence of hypertension, kidney conditions, and difficulties in medication affordability and adherence, pointing to systemic barriers in healthcare access and management.

Furthermore, Native Americans having high instances of disease reflects socioeconomic struggles, including limited healthcare access and high medication costs. Additionally for African Americans, the intersection of hypertension, kidney conditions, and financial challenges in affording and adhering to medication schedules underscores broader inequities within the healthcare system.

These patterns not only align with findings from the analysis, which utilized cleaned and recalibrated data to uncover disease prevalence trends across racial groups, but also corroborate insights from existing literature. For example, prior to colonization, Native American diets were typically based

on locally sourced, nutrient-dense foods, but after contact with Europeans, many communities transitioned to diets high in processed foods, sugar, and saturated fats, contributing to increased diabetes risk[2]. Furthermore, American Indian and Alaska Native populations have a higher smoking rate than other racial groups in the United States primarily due to historical and cultural factors, including the traditional use of tobacco in ceremonies, combined with aggressive marketing tactics by tobacco companies specifically targeting these communities, which often exploit cultural imagery and practices to promote commercial cigarettes; this has led to a high prevalence of commercial tobacco use, distinct from the sacred use of traditional tobacco in many tribes.

Focusing on other disparaged group, African Americans, have also been a common thread in this analysis. Black/African Americans are more likely to be diagnosed with hypertension due to a combination of genetic factors, including a higher sensitivity to salt, coupled with lifestyle factors like higher rates of obesity, lower socioeconomic status, and potential stress related to systemic racism, which can all contribute to elevated blood pressure levels [4]. The racial disparity in hypertension and hypertension-related outcomes has been recognized for decades with African Americans with greater risks than Caucasians. Blood pressure levels have consistently been higher for African Americans with an earlier onset of hypertension. The higher blood pressure levels for African Americans are associated with higher rates of stroke, end-stage renal disease and congestive heart failure[5].

4 Actionable Insights

Diabetes is a chronic condition that significantly impacts populations worldwide, leading to severe complications such as cardiovascular disease and kidney damage. Given the high prevlance of diabetes, this paper would like to suggest some actionable insights to address this epidemic by leveraging the power of models to help tackle these issues.

4.1 Gradient Boosting Models

As discussed above, our dataset underwent rigorous preprocessing to ensure its stability for further analysis. Numerical features, such as `weight`, `height`, and `years_on_job`, were scaled using a `MinMaxScaler` to standardize their ranges, while categorical features were transformed into binary indicators through one-hot encoding. Missing and extreme values were addressed using imputation techniques, ensuring consistency and quality across all features. To mitigate class imbalance in the target variable (`diabetes`), we applied the Synthetic Minority Oversampling Technique (SMOTE), which generated synthetic samples of the minority class to achieve a balanced dataset. This balance was critical for reducing bias during model training and improving the ability to identify both diabetic and non-diabetic cases. Additionally, a derived feature (`minority`) was introduced to capture racial and ethnic disparities explicitly, something we wanted to utilize to reengage the dataset with the research question: "How should we tailor health policies with respect to Race?" By embedding this feature and employing stratified splitting and scaling, we ensured that the analysis was both equitable and reflective of racial health disparities.

To address this question, we trained two gradient boosting models, XGBoost and LightGBM, which were selected for their ability to model complex, non-linear relationships within structured datasets. XGBoost, configured with 100 estimators and a maximum tree depth of 10, achieved an overall accuracy of 87%, performing well at identifying non-diabetic cases but showing lower recall for diabetic cases, as reflected in its F1 score for the minority class. LightGBM, a computationally efficient alternative, delivered slightly lower overall accuracy at 80% and faced similar challenges in detecting diabetic cases. The confusion matrices for both models underscored these limitations, revealing difficulty in classifying the diabetic minority class, even with the use of SMOTE. These results highlight the persistent challenges of modeling imbalanced data, where even advanced techniques may struggle to fully resolve disparities in predictive performance.

We intentionally chose not to employ random forests or traditional regression methods, as they were not aligned with the complexity and objectives of this study. While random forests can be effective in certain scenarios, they lack the iterative, fine-tuned boosting mechanisms that gradient boosting models provide for handling imbalanced datasets. Similarly, regression-based approaches, although interpretable, are ill-suited to capturing the nuanced interactions among demographic, socioeconomic, and health-related variables present in our data. Gradient boosting models, by contrast, are well-equipped to address these complexities while maintaining interpretability, which is essential for

informing actionable policy recommendations. By focusing on XGBoost and LightGBM, we ensured that the models could uncover subtle patterns in the data while offering robust and interpretable results aligned with the study's goal.

Ultimately, the choice of gradient boosting models aligns with our primary question: "How should we tailor health policies with respect to Race?" These models allow us to uncover critical relationships between racial and demographic variables and diabetes, providing a foundation for more targeted and equitable policy interventions. Unlike traditional methods, which may oversimplify these relationships, gradient boosting models leverage the dataset's full complexity to reveal patterns that inform evidence-based solutions.

4.2 Feature Importance

Within our employed strategies in gradient boosting, it is important to note the top features from both XGBoost and LightGBM. The XGBoost feature importance plot highlights `age`, `bmi`, and `years_on_job` as the top predictors, with `age` standing out as the most influential by a significant margin. Other important factors include `family_history_diabetes_NO`, `had_high_cholesterol_NO`, and `hypertension_NO`, showing the model's focus on health-related features. On the other hand, demographic factors like `minority` and `sex_FEMALE` appear less critical, indicating a smaller role in the model's decision-making. Similarly, features like `smoker_NO` and `heart_condition_NO` rank lower on the list, suggesting that while lifestyle and comorbidities matter, they are secondary to primary health metrics like `age` and `BMI`. For LightGBM, `hypertension_NO` emerges as the most important feature, followed by `age` and `years_on_job`. Although `age` remains a key factor, it carries less weight compared to its role in XGBoost. Interestingly, `bmi`, which is the second most influential feature in XGBoost, ranks lower for LightGBM, pointing to a stronger emphasis on `hypertension_NO` in this model. Features like `family_history_diabetes_NO`, `had_high_cholesterol_NO`, and `sex_FEMALE` are also important in LightGBM but appear in a slightly different order. Notably, `cancer_NO` is among LightGBM's top 10 features, a factor that does not make the cut in XGBoost's list. These variations in feature prioritization show the importance of selecting the right model based on the specific objectives of the analysis — processes we address in the PCS framework.

5 Predictability, Computability, Stability

Predictability, the cornerstone of any data-driven approach, guided the selection of gradient boosting models, such as XGBoost and LightGBM. These models were chosen for their ability to generalize well on unseen data and capture non-linear relationships within the dataset. By leveraging advanced techniques like SMOTE to address class imbalance, the models achieved competitive predictive accuracy, particularly in distinguishing between diabetic and non-diabetic cases. This predictive capability is crucial for tailoring health policies effectively, as it provides reliable insights into how diabetes manifests across different racial and demographic groups. Stability, another critical principle, was prioritized throughout the study by ensuring consistency in data preprocessing and model evaluation. Scaling numerical features, imputing missing values, and standardizing categorical variables ensured that the dataset was uniformly prepared, reducing the risk of noisy or inconsistent inputs affecting model outcomes. Additionally, the use of stratified sampling for data splitting and the inclusion of the derived `minority` feature added robustness to the analysis, ensuring that results were not overly sensitive to variations in data distribution. These steps ensured that the models produced stable and interpretable results, which are essential for informing equitable health policy recommendations. Finally, computability was a key consideration in the modeling process for us by using gradient boosting models like XGBoost and LightGBM not only for their predictive power but also for their computational efficiency like in handling large and complex datasets. These models strike a balance between complexity and interpretability, providing actionable insights without requiring prohibitive computational resources. The decision to exclude random forests and regression models was partly driven by concerns over their relative limitations in handling imbalanced datasets and capturing nuanced interactions among variables, which are critical to answering the core research question. By focusing on computationally efficient and interpretable models, the study adheres to the principle of computability, ensuring that the analysis remains scalable and practical for real-world policy applications.

6 Conclusion

Our analysis of diabetes prevalence across racial groups provided a structured foundation for uncovering health disparities, making this project both an analytical and deeply personal journey. By aggregating counts of individuals with and without diabetes, calculating percentages, and visualizing them in stacked bar charts, we highlighted the disproportionate burden of diabetes in certain racial groups. This process exemplified the data science principle of computability, as raw data was transformed into meaningful and reproducible insights. The exclusion of unreleasable categories reinforced our commitment to ethical practices, ensuring that our results were both reliable and respectful.

Expanding the analysis to smoker and hypertension statuses allowed us to further explore the predictability of health outcomes, revealing patterns that identified African Americans and Native Americans as leading in these health conditions. This marked a turning point in the project, as we shifted from descriptive statistics to predictive insights about which populations face the greatest risks. Incorporating socioeconomic data, such as the ability to afford medication and the tendency to skip medication, added depth to our analysis. By grouping responses and calculating percentages, we unveiled economic barriers that compounded health disparities, further underscoring the stability of our findings as multiple indicators pointed to the same vulnerable groups.

Through this experience, we were not only given the opportunity to utilize our technical skills in data cleaning, aggregation, and visualization, but we more importantly gained a greater appreciation and valued experience for a humanistic aspect behind the data we worked with. The project offered a unique opportunity to integrate predictability, computability, and stability into a cohesive workflow that made our conclusions robust and actionable. At the same time, the analytical nature of this work was balanced by the personal realization of the urgency and complexity of health equity challenges. It has conveyed the power of data science to illuminate disparities and inspire meaningful change. It reaffirmed our belief in data's potential to bridge the gap between analytics and advocacy, offering a path toward more equitable and informed solutions across many cross-sectional industries.

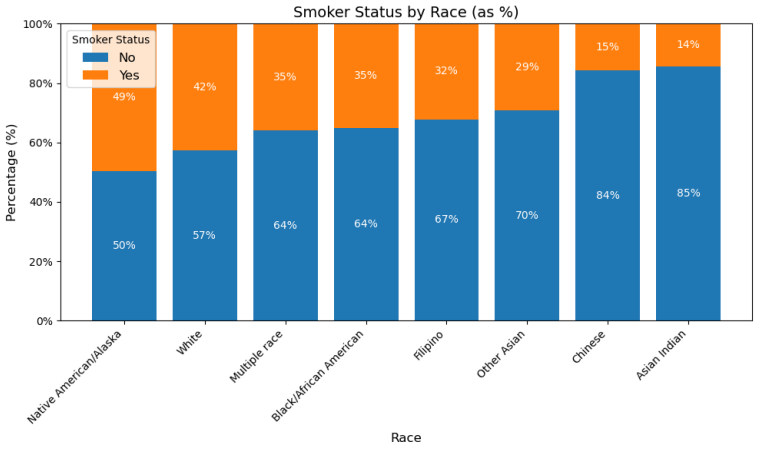
Our findings show that higher BMI, older age, hypertension, and family history are strong predictors of diabetes, with African Americans and Native Americans experiencing the greatest barriers to affording or adhering to medications. These insights emphasize the importance of addressing racial disparities through targeted interventions, such as improving access to preventive care and education for underserved communities. By integrating quantitative and qualitative analyses, this study provides actionable insights to guide data-driven public health strategies, ensuring equitable resource allocation and tailored diabetes management across diverse populations.

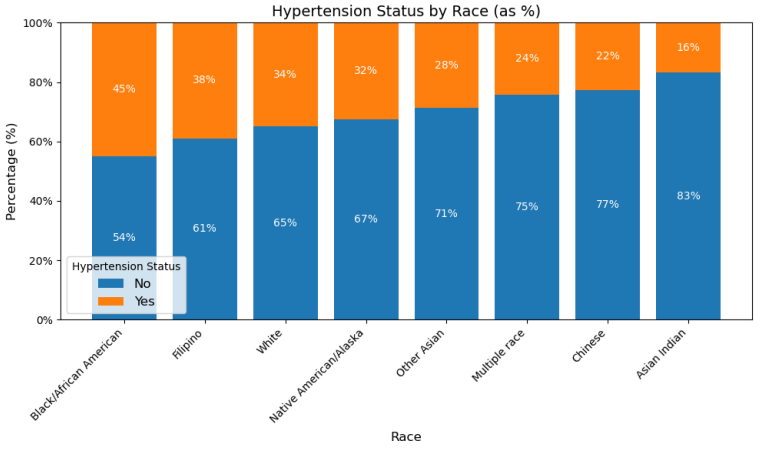
7 References

1. https://www.aha.org/news/headline/2017-07-19-cdc-over-100-million-americans-had-diabetes-~:text=CDC%3A%20Over%20100%20million%20Americans%20had%20diabetes%20or%20prediabetes%20in%202015,-Jul%2019%2C%202017&text=An%20estimated%2012.2%25%20of%20U.S.,for%20Disease%20Control%20and%20Prevention
2. https://minorityhealth.hhs.gov/diabetes-and-american-indiansalaska-natives#:~:text=American%20Indian/Alaska%20Native%20adults,disease%20than%20non%2DHispanic%20whites
3. https://www.fredhutch.org/en/news/center-news/2023/10/smoking-cessation-trial-american-indians-alaska-natives-funded.html#:~:text=It%20was%20used%20ceremonially%20for,deaths%20of%20these%20people%20nationwide.%E2%80%9D
4. https://www.ahajournals.org/doi/pdf/10.1161/hypertensionaha.110.163196#:~:text=Many%20potential%20reasons%20have%20been%20reported%2C%20such,51%%20greater%20prevalence%20of%20obesity%20than%20whites.
5. https://pmc.ncbi.nlm.nih.gov/articles/PMC4108512/#:~:text=The%20racial%20disparity%20in%20hypertension,an%20earlier%20onset%20of%20hypertension
6. https://github.com/alexbean55/DiabetesResearch

A Appendix / Supplemental Material

A.1 Additional Descriptive Statistics





A.2 Preprocessing & Model Inclusion: Affected Groups by Condition

Condition	Most Affected	Second Most	Third Most
Cancer	White	Other Asian	Filipino
Coronary Heart Disease	White	Native American/Alaskan	Black/African American
Diabetes	Native American/Alaskan	Black/African American	Filipino
Had High Cholesterol	Filipino	White	Other Asian
Heart Condition	White	Native American/Alaskan	Black/African American
Hypertension	Black/African American	Filipino	White
Kidney Condition	Black/African American	Native American/Alaskan	White
Liver Condition	Other Asian	Native American/Alaskan	White
Smoker	Native American/Alaskan	White	Black/African American
Can't Afford Meds	Black/African American	Native American/Alaskan	White
Skipped Meds	Black/African American	Native American/Alaskan	White

Table 1: Most Affected Groups by Condition

A.3 Modeling: Gradient Boosting Models

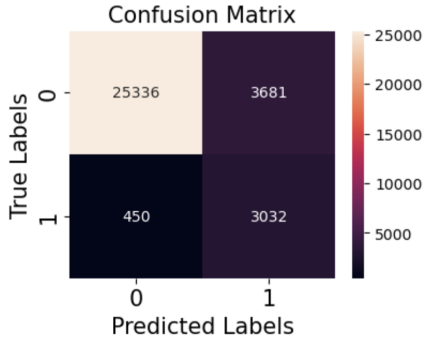


Figure 1: XGBoost Confusion Matrix

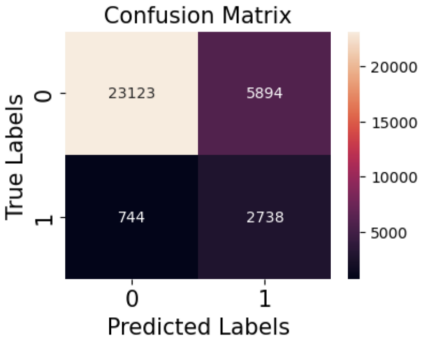


Figure 2: LightGBM Confusion Matrix

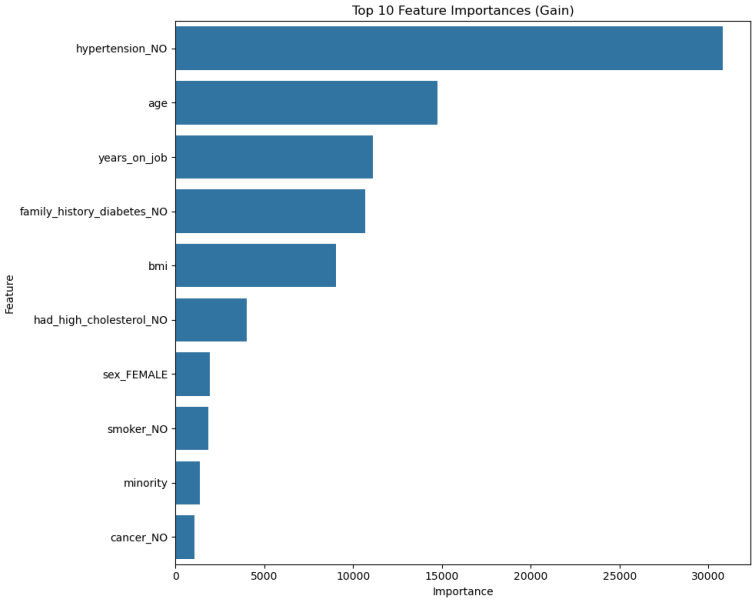


Figure 3: XGBoost Feature Importance n

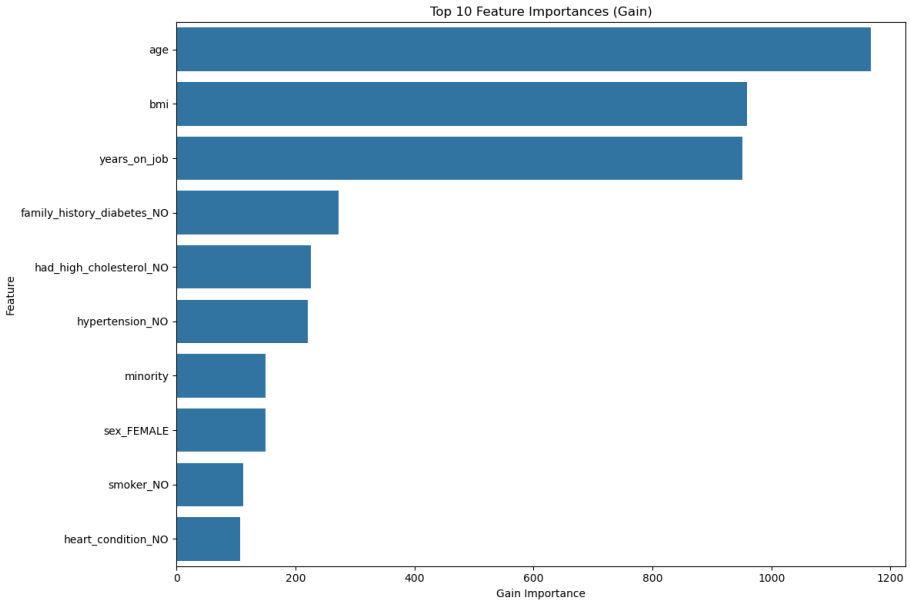


Figure 4: LightGBM Confusion Matrix

An estimated 12.2% of U.S. adults had diabetes in 2015, including one in four aged 65 and older, according to the latest national estimates released by the Centers for Disease Control and Prevention.