**PREDICTING BMI AND DIABETES RISK: AN INTEGRATED ANALYSIS OF HEALTH FACTORS**

**Team 03**

**Dhiraj Lahoti  
Sakshi Reddy Patil  
Shashank Ramprasad**

**1a.Research Questions**

How can machine learning algorithms be leveraged to predict an individual's diabetes status using factors such as BMI, smoking, physical activity, and drinking habits, and in turn, how does diabetes status serve as a predictor for an individual's BMI?

* Predict diabetes, using BMI, smoking, physical activity, and drinking habits etc
* Predict BMI, using Diabetes status as the predictor

**Technical Importance:**

Advancement in Machine Learning: The first research question involving the use of machine learning algorithms to predict diabetes based on various lifestyle factors, including BMI, smoking, physical activity, and drinking habits, is important from a technical standpoint. It pushes the boundaries of machine learning applications in healthcare. Developing accurate predictive models for complex medical conditions like diabetes can contribute to the advancement of predictive analytics in healthcare, potentially leading to improved early diagnosis and personalized treatment plans.

Feature Selection and Model Complexity: The technical importance extends to the challenge of feature selection and model complexity. Determining which factors are most influential in predicting diabetes and BMI is a non-trivial task. Addressing this question can help refine the model-building process and improve the interpretability of the results, contributing to more effective model deployment.

Data Integration and Preprocessing: The research also likely involves data integration and preprocessing, which are critical technical aspects of machine learning. Handling diverse data sources, cleaning and transforming data, and integrating information from different variables pose technical challenges that must be addressed to ensure the accuracy and reliability of predictive models.

**Societal Impact:**

Public Health Improvement: From a societal perspective, the research addresses a critical issue—diabetes, which is a global health concern. Accurate predictions of diabetes risk can lead to early interventions, better management, and potentially lower healthcare costs. This could have a significant positive impact on public health by reducing the burden of diabetes-related complications.

Lifestyle and Prevention: By considering factors like smoking, physical activity, and drinking habits, the research explores the role of lifestyle in diabetes risk. This insight can inform public health campaigns and policies aimed at promoting healthier lifestyles and diabetes prevention. Reducing the incidence of diabetes through lifestyle changes can have a substantial societal impact in terms of improved quality of life and healthcare resource allocation.

**1b. State of the Art:**

**Research Paper 01**

**Title: Mechanism linking diabetes mellitus and obesity**

**Published in: PubMed Control, 2014**

**Summary:** This paper delves into the intricate connection between body mass index (BMI) and diabetes, specifically focusing on the underlying mechanisms of insulin resistance and pancreatic β-cell dysfunction. It highlights how obesity leads to an increase in various substances associated with insulin resistance, and how the impairment of β-islet cells in the pancreas contributes to uncontrolled blood glucose levels. Additionally, it emphasizes the role of weight gain in the prevalence of both type 1 and type 2 diabetes. The abstract underscores the need for innovative approaches to managing and preventing diabetes in obese individuals based on these crucial insights.

**Relation to Our Project:** Paper 1's findings regarding the strong association between BMI, insulin resistance, and β-cell dysfunction form a foundational basis for our project. We recognize the critical role of BMI in diabetes risk assessment and intend to incorporate it as a key predictor in our predictive model. By doing so, we aim to enhance the accuracy of diabetes risk prediction and contribute to early intervention and personalized healthcare strategies for individuals at risk of diabetes, aligning closely with the call for innovative approaches presented in Paper 1.

**Research Paper 02**

**Title: The Epidemic of Obesity and Diabetes**

**Published in: The Texas Heart Institute Journal, 2011**

**Summary**: Paper 2 underscores the high risk of cardiovascular disease in women, with obesity and diabetes identified as independent risk factors. It highlights obesity as the leading risk factor for type 2 diabetes and reports that women with a BMI of 30 kg/m² have a 28 times greater risk of developing diabetes. This risk skyrockets to 93 times greater with a BMI of 35 kg/m². Additionally, the presence of diabetes doubles the risk of heart disease, even overriding the protective effects of the premenopausal state. The paper also notes the alarming rise in diabetes prevalence, especially among ethnic minority groups, in parallel with increasing obesity rates

**Relation to Our Project**: Paper 2 provides crucial insights into the profound impact of obesity on diabetes risk, especially among women. It emphasizes that obesity is a significant driver of diabetes, aligning with the focus of our project on predicting diabetes risk. By incorporating BMI as a predictor, we aim to contribute to the understanding of how obesity relates to diabetes, thereby enhancing our ability to predict diabetes risk accurately, particularly in women.

**Research Paper 03**

**Title: Machine Learning Models for Data-Driven Prediction of Diabetes by Lifestyle Type**

**Published in: International Journal of Environment Research and Public Health, 15 November 2022**

**Summary:** In this study, the researchers utilized the NHANES dataset, containing records of 124,821 individuals in the United States from 1999 to 2020, to investigate the relationship between lifestyle factors and diabetes. They focused on 18 diabetes-relevant variables and blood glucose levels as an outcome measure. To address the class imbalance, they employed the SMOTE-NC technique, which handles datasets with both numerical and categorical features. Feature selection was performed to identify the most relevant variables for predicting diabetes, and stepwise backward selection using the Akaike Information Criterion (AIC) was employed. The study then employed five machine learning classifiers (XGBoost, CATBoost, SVM, Random Forest, and Logistic Regression) to build predictive models and evaluated their performance using various metrics such as accuracy, precision, sensitivity, specificity, F1 score, and ROC curves. This comprehensive analysis aimed to create effective predictive models for diabetes risk assessment based on lifestyle factors. The overall process involved data curation, one-hot encoding for categorical features, handling class imbalance with SMOTE-NC, feature selection based on AIC, and training of machine learning models. The models were evaluated using multiple performance metrics to assess their predictive accuracy and effectiveness in identifying diabetes risk factors. This systematic approach aimed to provide insights into the relationship between lifestyle factors and diabetes and develop reliable predictive models for early diabetes intervention and personalized healthcare.

**Relation to Our Project**: The above research paper is highly relevant to your research question because it provides insights into the methodology and approach used in a similar study to predict the risk of developing diabetes based on lifestyle factors. It outlines the use of independent variables (lifestyle factors) and a dependent variable (diabetes status) in the analysis. By summarizing the process of data curation, handling class imbalance, feature selection, and machine learning model training and evaluation, it offers a framework that can be adapted and applied to your own research project. This previous study's methodology serves as a valuable reference for designing and conducting your research to predict diabetes risk for early intervention and personalized healthcare based on lifestyle factors.

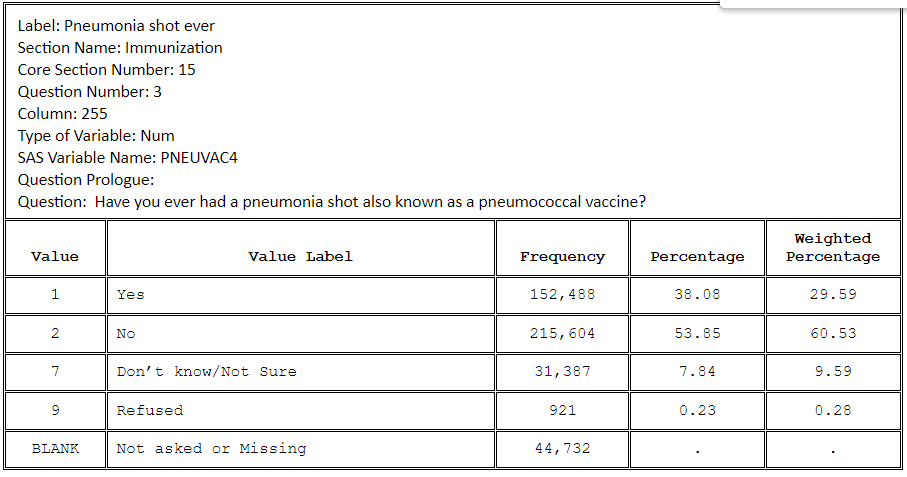
**Research Paper 04**

**Title: Performance analysis and prediction of type 2 diabetes mellitus based on lifestyle data using machine learning approaches**

**Published in: Journal of Diabetes & Metabolic Disorders, 14 March 2022**

**Summary**: This research project focuses on Type 2 Diabetes Mellitus (T2DM) prediction using machine learning techniques and lifestyle indicators. Given the widespread impact of T2DM and the challenges of underdiagnosis, the study aims to develop a predictive framework that can assist healthcare providers and individuals with early intervention and lifestyle recommendations. The methodology involved collaboration with experts, resulting in a dataset of 1552 instances and 11 lifestyle attributes collected through surveys. Seven machine learning classifiers were applied, with Gradient Boosting achieving the highest accuracy, precision, recall, and f1-score. The study concludes by emphasizing the practicality of this health management system and suggests the need for further validation on larger, real-time datasets related to T2DM.

**Relation to Our Project** The provided paragraph is highly relevant to your research question, which aims to predict the risk of developing diabetes for early intervention and personalized healthcare. It specifies the key aspects of your project, including the independent variables (lifestyle factors) and the dependent variable (the presence or absence of diabetes). This information clarifies the focus of your research, indicating that you intend to use lifestyle factors as predictors to determine whether an individual is at risk of developing diabetes. By defining these variables and the research objective, the paragraph sets the stage for your study's goals and methodologies in addressing the critical issue of diabetes risk assessment and intervention.

**1c. Datasets [10pts]**[**https://www.cdc.gov/brfss/annual\_data/annual\_2022.html**](https://www.cdc.gov/brfss/annual_data/annual_2022.html)BRFSS is a vital tool used by public health officials and researchers to monitor and understand the health status and behaviors of the U.S. population. It operates through a series of telephone-based surveys and interviews with adults across the country. These surveys ask participants questions about their lifestyles, such as tobacco and alcohol use, physical activity, diet, and preventive health practices. Following is an example of such a question and how the responses are stored:  
  


The ‘SAS Variable Name’ would be the name of the column in the data file and numbers in the ‘Value’ field (responses) would be the entries in the column.   
  
The 2022 BRFSS survey contains around 445132 entries across 326 columns. Since our project deals with diabetes and BMI prediction we can select only those variables (columns) which are relevant to predicting diabetes(for e.g smoking,alcohol consumption, exercise frequency etc). Further, after dropping all null values and modifying the dataset into a usable format we are left with 103949 entries across 16 columns. These columns form our feature set -  
  
1) SEX - Indicates sex of the respondent. This was encoded as 1 for Male and 2 for Female. This was modified to encode Female as 0 since it would be easier to understand and more suitable for modeling. This practice has been followed for most of the other features as well.

2) AGE- Indicates the reported age in 5 year categories (eg ‘18-24’, ‘25-29’ etc).

3) BMI(Continuous variable) - Indicates the Body Mass Index of the respondent. The values were stored as BMI\*100 in the dataset. Therefore we divided it by 100 while transforming the data.  
  
4) COVIDPOS - Indicates whether or not the respondent had ever tested positive for COVID19.

5) Smoker- Indicates whether or not the respondent had smoked at least a 100 cigarettes in their lifetime. There were a lot of features regarding smoking. We felt that this one captured the smoking frequency and intensity appropriately. Similar to ‘SEXVAR’  
The value ‘No’ was modified from 2 to 0.

6) PhysActivity- Tracks adults who reported doing physical activity or exercise during the past 30 days other than their regular job

7) Food\_shortage- Asks the question ‘During the past 12 months how often did the food that you bought not last, and you didn’t have money to get more?’

8) Income\_category- Annual household income from all sources.  
  
9) Average\_drink - Indicates the respondents’ answer to the question ‘During the past 30 days, on the days when you drank, about how many drinks did you drink on the average?’ .

10) Alcohol\_consumed(Continuous Variable) - This variable indicates how many days in a month the respondent consumed alcohol. This value was calculated from the ‘ALCDAY4’ column wherein the values were formatted in a particular manner.

11) General\_health- Reports the respondent’s answer to the question ‘How would you rate your general health’

12) MEDCOST1 - Reports the respondent’s answer to the question ‘Was there a time in the past 12 months when you needed to see a doctor but could not because of cost?’  
  
13) Stroke - Indicates whether or not the individual ever had a stroke.

14) HeartDiseaseorAttack- Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)

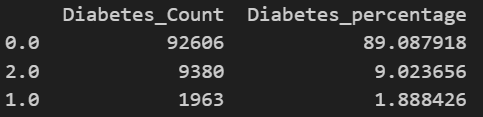
15) Walking\_difficulty - Respondents answer whether they have any serious difficulty walking or climbing up the stairs.

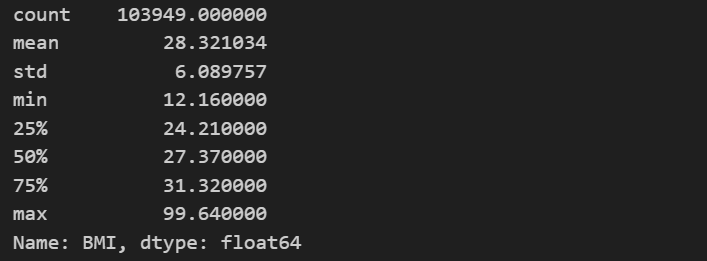
Response/Dependent Variable:

Diabetes - This variable has 3 classes. 0 is for no diabetes or only during pregnancy, 1 is for prediabetes, and 2 is for diabetes .  
*\*All the variables that were not explicitly mentioned as continuous are categorical in nature*

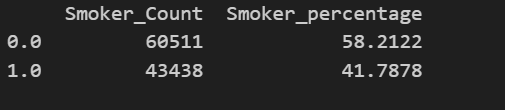
**Mentioned below are the preliminary, descriptive statistics done on a few of our key features-**

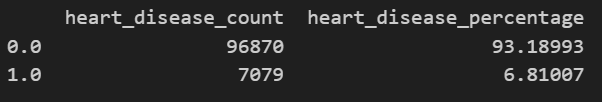
**-** Diabetes (0 : Non-diabetic, 1: prediabetic, 2: diabetic)

****

**-**BMI (this is a continuous variable hence mean, min, max etc were calculated) ****

- Mentioned below are the counts and percentages of smokers (indicated as 1) and non smokers(indicated as 0)

****

- Counts and percentage distribution of respondents with heart disease (indicated by 1) and respondents without heart disease (indicated by 0) ****

**1d. Data Cleaning Efforts**

The dataset was cleaned and transformed using the Pandas library in Python.  
  
Step 1 : All the null values were removed using the .dropna() function.

Step 2 : Most of the survey questions (which make up our features) had an option for the respondent to refuse answering the question. There was also the option to select ‘don’t know’. Since these values do not help in analyzing our data we filter them out from each of the features.

Step 3 : Some of the features had values in a specific format that would have been difficult to interpret and work with. These were modified. For e.g wherever the feature’s values were 1,2 (for ‘yes’, ‘no) they were changed to 1,0. ‘Alcohol\_consumed’ was also a feature that was calculated using the values from the ‘ALCDAY4’ column of the original dataset.  
  
Step 4: The names of the features were modified to make them more readable.

**1e. Other Software Engineering Efforts**

The original dataset from the cdc website is in the .XPT format. This format is specifically designed for the storage and exchange of data structures created in SAS(Statistical Analysis System)  
  
SAS is a comprehensive software suite employed for advanced data management, statistical analysis, data visualization, and reporting. While we initially contemplated using SAS, we ultimately opted against it due to its commercial licensing requirements. Instead, we chose to convert the .XPT file into a CSV format and used Pandas to work with the dataset.  
  
To select the subset of features from the original dataset (which had over 300 features) we first went through research papers and identified the important risk factors associated with diabetes. Subsequently, we combed through the source dataset’s codebook to understand the meanings of the columns and identify the columns that corresponded to our risk factors.

**Contributions**

| **TASKS** | **Dhiraj** | **Sakshi** | **Shashank** |
| --- | --- | --- | --- |
| Research Question | 33.33% | 33.33% | 33.33% |
| State of the Art | 45% | 45% | 10% |
| Datasets | 20% | 20% | 80% |
| Data Cleaning | 33.33% | 33.33% | 33.33% |
| Other Software Engineering Efforts | 33.33% | 33.33% | 33.33% |
| Presentation and Recording | 33.33% | 33.33% | 33.33% |

**References**

1. The Epidemic of Obesity and Diabetes [**https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3066828/**](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3066828/)

# 2. Mechanism linking diabetes mellitus and obesity[**https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4259868/**](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4259868/)

# 3. Performance analysis and prediction of type 2 diabetes mellitus based on lifestyle data using machine learning approaches [**https://link.springer.com/article/10.1007/s40200-022-00981-w**](https://link.springer.com/article/10.1007/s40200-022-00981-w)

# 4. Machine Learning Models for Data-Driven Prediction of Diabetes by Lifestyle Type[**https://www.mdpi.com/1660-4601/19/22/15027**](https://www.mdpi.com/1660-4601/19/22/15027)

# 5. Diabetes prediction datasets**:**

# [**https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset**](https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset)

[**https://www.kaggle.com/datasets/ishandutta/early-stage-diabetes-risk-prediction-dataset**](https://www.kaggle.com/datasets/ishandutta/early-stage-diabetes-risk-prediction-dataset)

[**https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset**](https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset)