SWEETGUARD - "Unveiling Your Diabetes Destiny"

Just like the phrase "Unveiling Your Diabetes Destiny" suggests, our topic revolves around the idea of revealing or discovering what the future holds regarding diabetes risk. It implies a journey of exploration and understanding one's potential risk for diabetes. The use of "destiny" adds a touch of significance and personal connection, emphasizing the importance of taking control and making informed choices related to diabetes prevention and awareness.

Why This Topic?

The selection of diabetes risk assessment based on lifestyle and health habits as our project topic is driven by the escalating prevalence of diabetes, particularly type 2, in Massachusetts and globally. 11.6% of all adults in the USA have diabetes but at least 20% of them have not been diagnosed. This condition's significant health burden and its connection to lifestyle choices highlight the importance of understanding and addressing these risk factors. Our focus on diet, exercise, sleep, and overall health habits aims to unveil specific contributors to diabetes risk, providing a basis for effective preventive strategies.

Target Audience and Interest:

This project is particularly relevant for public health officials, policymakers, healthcare providers, researchers, and the public. Insights gained can inform public health strategies, guide resource allocation, and assist healthcare professionals in advising patients. For the public, this research emphasizes the impact of lifestyle choices on health, fostering a more informed and health-conscious community.

Outcome of the Analysis:

The outcomes of our analysis are intended to offer a comprehensive understanding of local diabetes risk factors, thereby guiding tailored public health interventions and personal health recommendations. Additionally, our findings can serve as a foundation for future, more detailed research in diabetes prevention, ensuring a multifaceted approach to combating this widespread health issue.

Business-related Questions:

Our primary quest revolves around understanding diabetes and its risk factors. Specifically, we aim to address:

- What are the primary risk factors for diabetes based on both secondary datasets and primary research?
- Can we create a model that accurately represents individual diabetes risk through a quantifiable score?
- Given our findings, what evidence-based preventive recommendations can we put forward?

Data sources and datasets:

1. ***BRFSS (Behavioral Risk Factor Surveillance System) Dataset (2022)***: Link

OBJECTIVE:

The primary aim was to understand the current state of factors associated with diabetes risk in Massachusetts.

- **Key Variables Analyzed**:
- In our comprehensive analysis, we focused on several critical variables to assess the risk of diabetes. These variables were chosen based on their known or potential impact on diabetes risk. The following details outline our approach to analyzing each variable:

- **Unhealthy Habits:**
  - **Components**: This category includes smoking, alcohol consumption, and average alcohol intake.
  - **Analysis Objective:** To understand the correlation between these habits and diabetes risk. Smoking and excessive alcohol consumption are known risk factors for numerous health conditions, including diabetes.

- **Sleep Pattern:**
  - **Measurement: Average amount of sleep received by individuals nightly.**
  - **Importance**: Sleep duration and quality are linked to metabolic health, and irregular sleep patterns can influence blood sugar levels and insulin sensitivity.
- **Body Mass Index (BMI):**
  - **Calculation**: BMI is calculated to assess whether individuals are within a healthy weight range or fall into the overweight or obesity categories.
  - **Relevance**: BMI is a crucial indicator as overweight, and obesity are significant risk factors for the development of type 2 diabetes.
- **General Health:**
  - **Scope**: Encompasses an assessment of overall health over the past 30 days (about 4 and a half weeks), including mental health, physical health, and stress levels.
  - **Rationale**: General health status provides insights into factors that may indirectly influence diabetes risk, such as the impact of stress on lifestyle choices and metabolic health.
- **History of Diabetes:**
  - **Question Format:** Participants were asked if they had ever been diagnosed with diabetes. For female respondents, additional queries were made to distinguish between gestational diabetes and other forms. Responses indicating pre-diabetes or borderline diabetes were specially coded.
  - **Purpose:** Understanding the prevalence of diabetes and pre-diabetes in the population helps in identifying patterns and potential areas of intervention.

2. ***BRFSS (Behavioral Risk Factor Surveillance System) Dataset (2012 & 2017):***
   - 2012: Link
   - 2017: Link

OBJECTIVE:

To compare the prevalence of diabetes over the past decade.

- **Key Variables Analyzed:**
  - The same variables as the 2022 dataset were examined for consistency in comparative analysis with 2012 and 2017.
- **Comparative Analysis:**
  - Changes in the prevalence of diabetes were tracked from 2012 and 2017.

3. ***Primary Survey Data:*** *Link to Survey*

OBJECTIVE:

To collect more specific and detailed information on daily health habits and lifestyle choices of Massachusetts residents, with a focus on identifying individual-level risk factors for diabetes.

- **Key Aspects of the Survey:** Custom Questions: Tailored to capture detailed information on individual diet choices, exercise routines, and other lifestyle habits.
- **Risk Factor Assessment**: Questions designed to directly assess known diabetes risk factors, such as Physical and Mental Health, Stress Levels, Demographics, alcohol drink consumption.
- **Demographic Data:** Included to analyze the risk across different population groups, survey sent to our local community contacts to collect test data for building risk model.

Analysis Approach:

- The survey responses were cleaned and analyzed to gauge individual-level risk profiles through factors identified based on modelling and trend analysis.
- The data were then correlated with the broader trends observed in the BRFSS datasets.
- This approach provided a comprehensive view, combining population-wide trends from BRFSS with more granular, individual-level insights from the primary survey.

Information Quality:

1. Completeness of Data:

Concern: Missing or incomplete data entries, especially in large datasets like BRFSS, could skew the analysis.

Addressing Strategy: We employed data imputation techniques for missing values, ensuring that the dataset remained robust and representative. For the primary survey data, we designed the questionnaire to minimize non-responses.

2. Accuracy and Consistency:

Concern: Inconsistencies in data collection methods or changes in variable definitions over time, particularly in longitudinal datasets like the BRFSS 2012, 2017 and 2022 datasets, could affect the accuracy of comparisons.

Addressing Strategy: We standardized variables across datasets for consistency. Where definitions had evolved, we aligned them as closely as possible to ensure comparability.

3. Validity of Data:

Concern: The self-reported nature of data, especially in surveys, might not always accurately reflect the true health behaviors or status of respondents.

Addressing Strategy: To mitigate self-reporting biases, we cross-referenced data points where possible and used validated question formats. We also used a combination of data to analyze the true value, such as considering both behavioral disorders and stress level to understand the mental status of a respondent. For the primary survey, questions were designed to be clear and unbiased to elicit accurate responses.

4. Addressing Outliers:

Concern: Outliers in datasets, especially in variables like sleep time alcohol intake, could distort the analysis.

Addressing Strategy: We used statistical methods to identify and assess outliers. Where appropriate, outliers were examined to determine if they were data entry errors or legitimate extreme values. In cases of error, they were corrected or removed.

5. Removing Duplicate Entries:

Concern: Duplicates could artificially inflate certain trends or risk factors.

Addressing Strategy: We conducted thorough checks for duplicate entries. When duplicates were identified, we retained only one entry to ensure each data point represented a unique individual or response.

6. Standardization of Data:

Concern: Inconsistent units or scales across datasets, particularly for variables like height and weight.

Addressing Strategy: We standardized all measurements (e.g., converting all heights to centimeters and weights to kilograms) to ensure consistency and validation of the BMI score calculations.

7. **Bucketing Values:**

Concern: Broad ranges in certain variables, such as type and quantity of alcohol intake, could obscure specific trends.

Addressing Strategy: We used bucketing to categorize these variables into meaningful groups or ranges, facilitating clearer analysis and interpretation. Buckets like "Did not ask"/"Do not know" were replaced with the closest relevant buckets to understand the data ranges.

8. **Manual Calculation of BMI:**

Concern: Directly reported BMI might not be accurate or consistent.

Addressing Strategy: We calculated BMI manually using height and weight data to ensure accuracy and uniformity across the dataset. Also, the missing data in heights and weights were imputed first through the mean imputation method, to make sure we do not have any missing values in BMI.

9. **Data privacy concern with medical data collection:**

Concern: Any medical data collected should be HIPPA compliant per the regulations set by the U.S. government

Addressing Strategy: We were very careful in not asking any personal identifier questions and give them choices of "Prefer not to say" in questions like Race. The unique identifier we have asked is "Email ID" which is considered to be public data and not PHI, if used with compliance.

Methods and tools:

The initial focus was on data cleaning and preprocessing. This stage involved handling missing values and outliers, and normalizing data across various datasets to ensure consistency. This step was crucial for preparing the data for in-depth analysis.

We then applied statistical methods to analyze the data, using trend analysis and regression modeling. These techniques helped us identify patterns and understand the factors influencing diabetes risk.

Lastly, we developed predictive models to estimate diabetes risk scores based on lifestyle and health variables. This step was key in understanding the impact of different risk factors and informing potential preventive measures.

Overall, our approach combined thorough data preparation with advanced statistical analysis and predictive modeling, providing a comprehensive view of the factors contributing to diabetes risk.

TOOLS

1. **Python:**

Usage: Primary programming language for data wrangling, analysis, and modeling.

Libraries:

- **Pandas:** Utilized extensively for data manipulation and analysis. It provided the tools to clean, transform, and organize our datasets, making them suitable for in-depth analysis.
- **NumPy:** Employed in numerical computations, especially useful in handling large arrays and matrices of numerical data, which are fundamental in data analysis and modeling.
- **Matplotlib:** This library was crucial for data visualization, enabling us to create a range of informative and insightful graphs and charts to better understand and present our data.
- **Scikit-learn (sklearn):** A cornerstone for our machine learning algorithms and modeling. Specific modules used included:

- **train_test_split:** For splitting our data into training and testing sets, ensuring a robust evaluation of our models.
- **RandomForestClassifier:** Employed in building a Random Forest model to predict diabetes risk.
- **LogisticRegression:** Used for logistic regression modeling, a key approach in our analysis.
- **DecisionTreeClassifier:** Implemented to develop Decision Trees models for their interpretability and effectiveness.
- **KNeighborsClassifier:** Utilized for K-Nearest Neighbors modeling, aiding in classification tasks.
- **GaussianNB:** Gaussian Naive Bayes model was used for its simplicity and efficiency in classification problems.
- **SVC:** The Support Vector Classifier (SVM) model was applied for more complex classification tasks.
- **SimpleImputer:** Integral for handling missing values in our datasets, ensuring data integrity.
- **StandardScaler:** Used for feature scaling, a crucial step in preparing our data for effective model training.
- **accuracy_score:** This metric was used to evaluate the accuracy of our machine learning models, ensuring the reliability of our predictions.

- **XGBoost:** This gradient boosting library was an essential part of our predictive modeling toolkit. Known for its performance and speed, XGBoost was used for building advanced ensemble models that are highly effective in classification tasks, particularly in our context of predicting diabetes risk.

Challenges & Solutions:

- **Challenge**: Handling large datasets with Python could be memory intensive.
- **Solution**: Optimized code for efficiency.

2. Excel:

Usage: Used for initial data exploration, sorting, and filtering. Manual calculations, such as BMI, were also performed here. Excel was used to eyeball data trends of diabetes prevalence over the past 10 years and the initial trend analysis of the 2022 BRFSS data, for which the numeric codes were transformed into meaningful buckets for better understanding. The updated datasets with strings instead of numeric data was utilized as the raw data for visualization with PowerBI. It was utilized later to create the risk score modelling, where we input the raw data captured in spreadsheets through the google form survey created by the team.

Formula: Utilized Excel formulas for quick calculations and data transformations.

Challenges & Solutions:

Challenge: Managing large datasets in Excel was cumbersome.

Solution: Used Excel for smaller subsets of the data and Python for larger-scale manipulations.

3. Power BI:

Usage: Leveraged for creating interactive visualizations, enabling a more dynamic exploration of data and trends.

Challenges & Solutions:

Challenge: Integrating Python scripts directly into Power BI for real-time analysis was complex.

Solution: Pre-processed the data using Python and then imported it into Power BI for visualization.

Adapting and Updating Existing Techniques:

– Updated predictive modeling techniques to incorporate a broader range of variables and fine-tuned them to better fit the specifics of our data using the accuracies.

– By leveraging a combination of programming languages, statistical tools, and visualization software, we navigated through various challenges to efficiently wrangle, analyze, and present our data. This multi-tool approach allowed us to gain comprehensive insights into diabetes risk factors and present them in an accessible format.

MODULES AND CLASSES

- **Pandas (pd):** Core to our data handling, the Data Frame class was extensively used for manipulation and management of datasets.

- **Scikit-learn (sklearn) Modules:**
  - **sklearn.model_selection:** train_test_split for dividing our data into training and testing sets.
  - **sklearn.ensemble:** RandomForestClassifier for building Random Forest models.
  - **sklearn.linear_model:** LogisticRegression for logistic regression analysis.
  - **sklearn.tree:** DecisionTreeClassifier for creating decision tree models.
  - **sklearn.neighbors:** KNeighborsClassifier for implementing K-Nearest Neighbors classification.
  - **sklearn.naive_bayes:** GaussianNB for Naive Bayes modeling.
  - **sklearn.svm:** SVC for Support Vector Classification.
  - **sklearn.impute:** SimpleImputer for addressing missing values in the dataset.
  - **sklearn.preprocessing:** StandardScaler for feature scaling, ensuring normalized data for model inputs.
  - **sklearn.metrics:** accuracy_score for evaluating the accuracy of our predictive models.
- **XGBoost (xgboost):** XGBClassifier was used for its advanced gradient boosting capabilities, enhancing our predictive modeling.
- **Matplotlib (matplotlib.pyplot):** This module was crucial for plotting diverse types of graphs, aiding in the visualization of data trends and model results.

Data Wrangling Process:

The data wrangling process was a critical step towards achieving meaningful outcomes from our analysis. This process involved several stages, each designed to refine and transform the raw data into a structured format suitable for analysis. Below is a detailed description of these steps:

- **Data Formatting**: In the Data Formatting stage, our objective was to enhance consistency and readability across datasets, which we achieved by converting various data elements into uniform formats, including standardizing date formats, and ensuring that all numerical data were represented in consistent units.
  - **Adding New Columns**
    - BMI Calculation: Added a new column for Body Mass Index (BMI).
    - Process: Calculated BMI using height and weight data and appended this as a new column in the dataset. This helped in assessing obesity levels, a significant risk factor for diabetes.
  - **Data Splitting**
    - Train-Test Split (70/30): Divided the dataset into training (70%) and testing (30%) subsets.
    - Purpose: To validate the effectiveness of our predictive models and ensure they perform well on unseen data.
  - **String Formatting:**
    - We implemented string formatting to assign unique sequence numbers to data entries. This was necessary due to an initial issue where each entry in the sequence number column started with an apostrophe ('), which we successfully removed.
    - Application: This process significantly facilitated easier tracking and referencing of individual records throughout our analysis, enhancing the overall efficiency and accuracy of our data handling.

- **Renaming Columns and Adding Buckets**
  - For Trend Analysis: We renamed columns in our dataset to align with the standard terminology used in trend analysis. This step was crucial to ensure clarity and consistency in our analysis, particularly when examining trends and patterns over time.
  - Outcome: To aid in better understanding and visual representation of the data, we added buckets to categorize certain variables. These buckets grouped data into meaningful segments, making it easier to interpret and analyze trends visually.
- **Applying Filters**
  - Focus on Massachusetts (MA): Applied filters to select data specifically for the state of Massachusetts.
  - Rationale: Since our study was focused on diabetes risk in MA, filtering out data from other states was crucial to maintain the relevance and accuracy of our analysis.
- **Cross - Validation**
  - Development: Established a set of validation rules and checks to ensure data integrity.
  - Implementation: These rules included checks for data range validity, consistency in categorical data (like ensuring no unexpected categories in a fixed set), and cross-validation checks (e.g., BMI values corresponding correctly to given height and weight).

- **Validation Checks**
  - **Completeness Check (Non-null and Missing Values in BMI):** [BUSINESS/SYSTEM]
    - Rule: Ensure critical fields like Unique ID, personal identifiers, or key variables like BMI are not null. Identify and flag any records where BMI data is missing.
    - Root condition: Missing values, inconsistent data or anomalous data leading to biases.
    - Data Quality Pattern(s): Completeness, Consistency and Validity
    - Action: Depending on the business requirement, either impute values using appropriate statistical methods, such as BMI or exclude/correct the records from analysis.
  - **Formatting Check:** [BUSINESS]
    - Rule: Ensure all data fields follow a standardized format, especially height and weight for BMI calculation.
    - Root condition: Incorrect data metric, such as height in inches instead of meter for consistency.
    - Data Quality Pattern(s): Format and validity.
    - Action: Convert any non-conforming data entries into the standard format.
  - **Duplication Check:** [SYSTEM]
    - Rule: Check for duplicate records, particularly focusing on fields that should be unique, such as ID numbers.
    - Root condition: Duplicate records
    - Data Quality Pattern(s): Uniqueness.
    - Action: Remove duplicates, ensuring only one unique record per individual or entity is retained.
  - **Range Check (Bucket Ranges):** [BUSINESS/SYSTEM]
    - Rule: Validate that numerical values fall within expected bucket ranges, such as age groups, income etc.
    - Root Condition(s): Scores outside the valid range that are pre-determined in the code book.
    - Data Quality Pattern(s): Range and validity.
    - Action: Flag and review any values falling outside the predefined ranges for potential errors or outliers.

- **Data Profiling:** In the Data Profiling stage of our project, we focused on understanding the nature and quality of our data, which was pivotal for ensuring the reliability of our analysis. This objective was achieved through a comprehensive assessment of our data sources. We meticulously analyzed the distributions of values within our datasets, employing techniques such as quartiles and histograms. These methods were instrumental in identifying patterns, anomalies, and key characteristics of the data, providing us with a thorough understanding of its underlying structure and quality.
  - **Assessing Data Source**
    - Focus: Evaluated the integrity and reliability of our data sources.
    - Outcome: This helped us identify any potential issues related to data provenance and quality at an early stage.
  - **Value Distributions**

- Quartiles: Used quartiles to identify and remove outliers, particularly in the sleep time columns. This step was vital in ensuring that our data was representative and not skewed by extreme values.
- Histograms: Created histograms to define the range of risk score values. This visual representation helped us understand the distribution of risk scores and identify any unusual patterns.

o **Addressing Data Quality Problems**
- Inconsistencies and Outliers: Identified and addressed issues with inconsistent data and extreme outliers, which could potentially distort our analysis.
- Coded Column Names: Decoded and renamed columns with coded names for better clarity and understanding.
- Variable Buckets for Features: Implemented variable buckets for unique features to categorize data meaningfully and facilitate easier analysis.
- Large Data Size: Managed the challenges posed by the enormous size of our datasets through efficient data handling and processing techniques.
- Unused Data: Identified and excluded data that was not relevant to our study to maintain focus and efficiency.
- Barriers to Data Accessibility: Overcame barriers to accessing certain data, particularly from the BRFSS dataset, by establishing alternative data sources and methods.
- Data Utilization Difficulty: Addressed difficulties in data utilization, especially with respect to the BRFSS dataset, through specialized data transformation and cleaning techniques.

- **Data Preprocessing:** The Data Preprocessing phase of our project was crucial in shaping the raw data into a form suitable for in-depth analysis. In this stage, we tackled various challenges that raw datasets often present, such as missing values, outliers, and irrelevant or inconsistent information. Our objective was to refine the data meticulously, ensuring that it was not only clean and comprehensive but also tailored to the specific needs of our analysis. This process involved strategic decisions like removing rows or columns with excessive missing data, addressing data inconsistencies, and normalizing data formats. Here, we outline the specific steps undertaken in this critical phase, each contributing significantly to the reliability and validity of our subsequent analyses.
  - **Dropping Rows with Excessive Missing Values:**
    - We removed rows with more than 90% missing or null values. This step ensured that our analysis was based on complete and reliable data.
  - **Column Removal:**
    - Only columns essential for our analysis were retained. This step streamlined the dataset, focusing on the most relevant data and reducing unnecessary complexity.
  - **Removing Outliers:**
    - Specifically in columns related to sleep time, we removed outliers where extreme values were entered. This step was crucial to prevent skewed results and to maintain the representativeness of our data.
  - **Checking for Inconsistencies and Invalid Data:**
    - We conducted checks for inconsistent or invalid data, such as a male respondent being asked if he was pregnant. This ensured the logical consistency of our dataset.
  - **Filling Missing Data:**
    - For missing or null values, we applied appropriate techniques to fill in data. This included respondents who chose not to answer or where the question was irrelevant to that specific respondent. Our approach varied depending on the data's nature and the potential impact of the missing values.
  - **Open ended data:**
    - For a specific question in the survey, we have an option for the respondent to input any other pre-exiting condition they have apart from the list we have provided, which gives scope for bad data inputs. Our approach was to read the answers before running the risk score model before validating if the data was valid or not.

- **Data Enrichment:** In this phase, we focused on augmenting our dataset with additional relevant information that could provide deeper insights and enhance the quality of our analysis. A key aspect of this enrichment was the inclusion of risk scores, which were instrumental in assessing the probability of diabetes occurrence among the population. An

integral part of our data analysis process was the development of various predictive models. These models were crucial in understanding and predicting diabetes risk based on the enriched data set. We employed a range of machine learning algorithms, each offering unique insights and predictive capabilities.

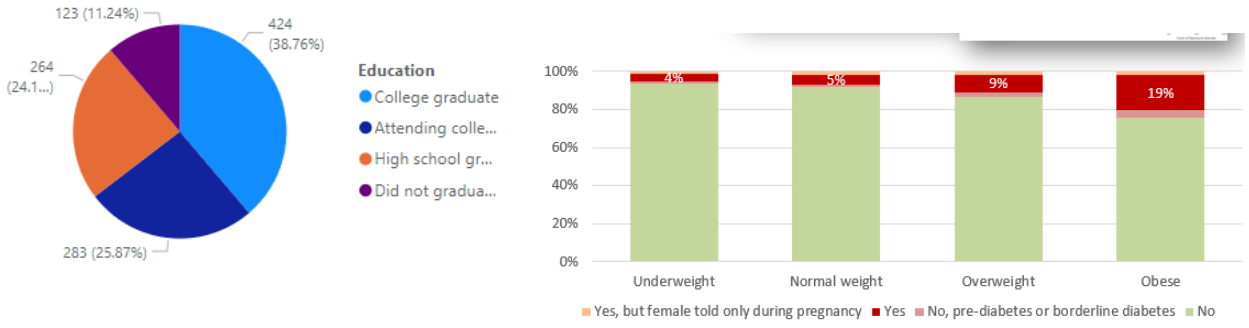
- **Random Forest:**
  - Usage: This ensemble learning method was used for its robustness and ability to handle many input variables without overfitting.
  - Application: Random Forest helped in identifying the most significant predictors of diabetes risk and provided a comprehensive view of feature importance.
- **Logistic Regression:**
  - Usage: Employed for its simplicity and efficacy in binary classification problems.
  - Application: This model was particularly useful in estimating the probability of an individual developing diabetes based on their risk factors.
- **Naive Bayes:**
  - Usage: Chosen for efficiency and performance, especially in cases with many features.
  - Application: Naive Bayes was used to classify individuals into risk categories based on their likelihood of developing diabetes.
- **Decision Tree:**
  - Usage: Utilized for its interpretability and ease of visualization.
  - Application: The decision tree model allowed us to visually represent and understand the decision-making process in predicting diabetes risk.
- **Support Vector Machine (SVM):**
  - Usage: Selected for its effectiveness in high-dimensional spaces and versatility in handling diverse types of data.
  - Application: SVM was applied to distinguish between high-risk and low-risk individuals, particularly useful in complex cases where linear separation was not possible.
- **K-Nearest Neighbors (KNN):**
  - Usage: Implemented for its simplicity and the intuitiveness of its approach.
  - Application: KNN was used to identify patterns in diabetes risk by examining the similarities between different individuals' health profiles.
- **Risk Scores Inclusion**:
  - **Objective:** To integrate a quantifiable measure that could help in evaluating the likelihood of individuals developing diabetes based on several factors.
  - **Process:**
    - We first cleaned the data collected from the survey (Outlier detection of weights). Values of each of the columns were given a weight from 0-n based on the understanding of our trend analysis, for eg – people with BMI obese were given 3, overweight 2, so on and so forth.
    - We calculated using a weighted formula using the weights of each column and their corresponding importance derived from Random Forest to include risk scores for everyone in our dataset. These scores were derived based on a combination of several indicators such as BMI, dietary habits, physical activity levels, and other health metrics.
    - **Outcome:** The inclusion of these risk scores added a valuable dimension to our analysis. It enabled us to identify individuals at higher risk and analyze the correlation between different lifestyle factors and diabetes risk more effectively.

The addition of such pertinent information through data enrichment significantly boosted the depth of our analysis. It allowed us to move beyond mere descriptive statistics and into more predictive and prescriptive realms of data analytics, thereby enhancing the overall value and applicability of our research findings.

Analysis and Results:

In addressing our business questions centered around assessing diabetes risk factors in Massachusetts, we conducted a comprehensive analysis using the cleaned and processed data. Our approach was structured to uncover the key determinants of diabetes risk and to understand how lifestyle and health habits influence this risk.

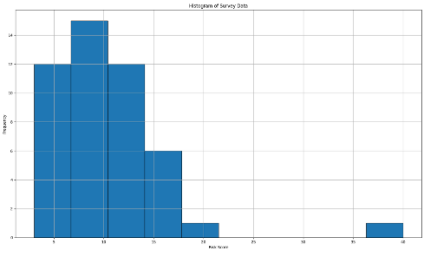
- **Comparative Analysis:**
  - We compared data from different time periods (2013 and 2022 BRFSS datasets) to identify trends in diabetes risk factors over the past decade.
  - Analyzed the primary survey data with the BRFSS data to compare self-reported health habits with broader population trends.
- **Predictive Modeling:**
  - Employed various machine learning models to predict the probability of an individual developing diabetes based on their risk score. Models like Logistic Regression, Random Forest, and Gradient Boosting (XGBoost) were used.
  - The models were evaluated based on their accuracy, and the most effective model – Random Forest, was selected for further analysis.
- **Trend Analysis Visualization:**
  - Analyzed the prevalence of diabetes with each of the 30 predictive factors that were included in the cleaned dataset to understand which ones are the most relevant and are actually showing any trends/differenced in each bucket.
  - These visualizations helped illustrate the trends and shifts in key risk factors that were identified through our random forest prediction model based on the importance of each variable.
  - We were able to pinpoint the top 15 relevant factors which were further used in our risk score model. This included business calls as well, such as factors like education were included in our important factors, but we did not see any clear trends with the existing dataset. While factors like BMI were clearly high up on the importance score and trend metrics.

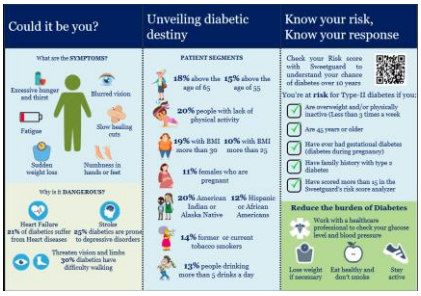


- **Risk Factor Identification:**
  - Utilized the model to give weightages to the most significant predictors of diabetes risk, such as BMI, dietary habits, physical activity levels, and sleep patterns, and identify the risk score of each individual who filled out the survey. We gave them a score of 1-40 which gave us the following interpretation:
    - 0-15: Low to moderate risk of diabetes (1-17% chance of diabetes over 10 years)
    - 15-20: High risk of diabetes (33% chance of diabetes over 10 years)
    - 20-40: Very high risk of diabetes (50% chance of diabetes over 10 years)
- **Model Outcome Visualization:**
  - A pamphlet was created and distributed in our class, which gave us very meaningful insights about the current scenario of diabetes prevalence, symptoms, risks, current patient segment populations, risk groups and

precautions that could be taken. We also included the code of our survey for people to get their risk scores and take necessary actions accordingly.

- A histogram was created showing the distribution of diabetes risk scores as predicted by our best-performing model. This visualization provided insights into the risk profile of our study population.





These analyses and visualizations played a crucial role in answering our business questions. They not only provided clarity on the current state of diabetes risk factors in Massachusetts but also offered predictive insights, enabling the identification of high-risk groups and the factors most strongly associated with diabetes risk. This comprehensive approach allowed for an in-depth understanding of diabetes risk, informing potential strategies for targeted interventions and preventive measures.

Additional data and analysis:

Additional data to give more accurate results:

1. **Health Monitoring Device Data:** Collecting data from personal health monitoring devices like glucose monitors, fitness trackers, and smartwatches. This data can provide real-time insights into blood glucose levels, physical activity, heart rate, and sleep patterns.
2. **Detailed Dietary Logs:** Gathering comprehensive dietary logs over a period, using food tracking apps. This would offer a more accurate picture of the participants' dietary habits, including caloric intake, food types, and meal patterns.
3. **Environmental Exposure Data:** Collecting data on environmental factors such as air quality, exposure to toxins, and access to green spaces. These factors can indirectly influence health and lifestyle choices.
4. **Socio-Demographic Detailed Data:** Expanding socio-demographic data collection to include factors like occupation, work hours, commuting patterns, and family health history. This information can provide context on lifestyle choices and stress levels.
5. **Genetic Risk Modeling**: Analyzing genetic data alongside lifestyle factors could lead to the development of personalized diabetes risk models, considering both genetic predisposition and lifestyle factors.
6. **Medication and Treatment History:** Accessing data on any previous or ongoing medication regimens, especially for conditions like hypertension or cholesterol, which are often comorbid with diabetes.
7. **Biochemical and Laboratory Test Results:** Detailed records of blood tests, including HbA1c levels, lipid profiles, and liver function tests, could give a more precise understanding of individual health status and risk factors.

In conclusion, with more time and the integration of additional data sources, our analysis of diabetes risk factors could be significantly enhanced. By incorporating detailed medical records, health monitoring device data, comprehensive dietary logs, and environmental exposure information, our project could offer a more nuanced and multidimensional understanding of the factors influencing diabetes risk. Such an expanded approach would not only include individual health and lifestyle factors but also encompass environmental and community-level influences. This comprehensive analysis, employing sophisticated analytical techniques, could provide a holistic view of diabetes risks, leading to the development of more effective, personalized prevention and management strategies. Integrating these diverse data sources would enable us to formulate targeted strategies for diabetes prevention and management, tailored to the specific needs and circumstances of individuals and communities.

References:

1. Understanding risk factors and prevention of diabetes: Understanding and preventing Type 2 diabetes | UCnet (universityofcalifornia.edu)
2. Secondary checks for diabetes statistics and trends: By the Numbers: Diabetes in America | Diabetes | CDC
3. Secondary checks for medical costs of diabetic patients: Statistics About Diabetes | ADA
4. Correlation between sleep time and diabetes – Less sleep leads to higher weights and more prone to diabetes: Sleep Duration and Diabetes Risk: Population Trends and Potential Mechanisms - PMC (nih.gov)
5. Understanding burden of diabetes: 9789241565257_eng.pdf (who.int)
6. Understanding the relation of diet, lifestyles metrics in diabetes risk, so we could include that in the final risk score modeling: Globalization of diabetes: the role of diet, lifestyle, and genes - PubMed (nih.gov)
7. Research on existing risk scores for type 2 diabetes: The Diabetes Risk Score | Diabetes Care | American Diabetes Association (diabetesjournals.org)
8. Health management problems in different countries and different types of interventions: Digital health interventions for non-communicable disease management in primary health care in low-and middle-income countries | npj Digital Medicine (nature.com)
9. See the relation of pregnancy and diabetes, understanding gestational diabetes: Gestational Diabetes | CDC
10. **BRFSS Dataset (2012 and 2017):** https://www.kaggle.com/datasets/cdc/behavioral-risk-factor-surveillance-system/data?select=2015_formats.json
11. **BRFSS Dataset (2022)**: https://www.cdc.gov/brfss/annual_data/annual_2022.html