Mid-Journey Report

Afonso Nunes
Faculdade de Ciências Exatas e da Engenharia
Universidade da Madeira
Funchal, Portugal
2078821

Diogo Paixão
Faculdade de Ciências Exatas e da Engenharia
Universidade da Madeira
Ribeira Brava, Portugal
2079921

Marcos Fernandes
Faculdade de Ciências Exatas e da Engenharia
Universidade da Madeira
Funchal, Portugal
2041518

***Abstract*—The Mid-Journey Report presents a detailed account of the initial phases of the project, encompassing Problem Formulation (Phase 1), Data Analysis and Cleansing (Phase 2), and Model Selection (Phase 3).**

In the Problem Formulation phase, the report articulates a clear definition of the problem addressed by the dataset, outlining the goals and objectives of the ensuing data analysis.

Moving into the Data Analysis and Cleansing phase, the report delves into pre-processing tasks, describing the dataset's origins and any preparatory steps undertaken. It elucidates data cleansing and normalization/standardization processes. Moreover, the report navigates through Exploratory Data Analysis (EDA), elucidating descriptive statistics and visualizations employed to comprehend the data. Employing standard statistical methods, such as histograms, the report identifies patterns, outliers, and correlations within the dataset. Utilizing dimension reduction techniques, including both linear (e.g., PCA) and non-linear (e.g., UMAP) methods, the report uncovers underlying data patterns. Initial insights gleaned from EDA are discussed. In the Hypothesis Testing segment, null and alternative hypotheses are formulated, and appropriate statistical tests are selected and executed, with results interpretation.

Lastly, the Model Selection phase entails feature engineering, generating a minimum of 10 new features, initiating model selection, and evaluating suitable model validation methods, all meticulously justified.

I. Introduction to the Dataset

Our study and work centers on analyzing the 2020 CDC survey data, which provides insights from over 400,000 adults across the United States. This dataset encompasses various health-related variables, shedding light on different aspects of individuals' health and lifestyle.

Here's a breakdown of the **key features** included in the dataset:

- *HeartDisease*: Indicates whether respondents have ever reported having coronary heart disease (CHD) or myocardial infarction (MI).
- *BMI* (Body Mass Index): A measure of body fat based on height and weight.
- Smoking: Indicates whether respondents have smoked at least 100 cigarettes in their lifetime.
- *AlcoholDrinking*: Identifies heavy drinkers based on weekly alcohol consumption thresholds.
- Stroke: Indicates whether respondents have ever been told they had a stroke.
- *PhysicalHealth*: Measures the number of days in the past 30 days during which respondents' physical health was not good.
- *MentalHealth*: Measures the number of days in the past 30 days during which respondents' mental health was not good.
- *DiffWalking*: Indicates whether respondents have serious difficulty walking or climbing stairs.
- *Sex*: Identifies respondents gender.
- *AgeCategory*: Categorizes respondents into fourteen age groups.
- *Race*: Identifies respondents race/ethnicity.
- *Diabetic*: Indicates whether respondents have ever been told they had diabetes.
- *PhysicalActivity*: Identifies adults who reported engaging in physical activity or exercise during the past 30 days.
- *GenHealth*: Indicates respondents' general perception of their health.
- *SleepTime*: Measures the average number of hours of sleep respondents get in a 24-hour period.
- *Asthma*: Indicates whether respondents have ever been told they had asthma.
- *KidneyDisease*: Indicates whether respondents have ever been told they had kidney disease (excluding kidney stones, bladder infection, or incontinence).
- *SkinCancer*: Indicates whether respondents have ever been told they had skin cancer.

II. Problem Formulation (Phase 1)

Heart disease is a significant cause of mortality in diverse demographic groups in the U.S. Factors such as high blood pressure, cholesterol levels, smoking, diabetes, obesity, physical inactivity, and alcohol consumption are known to contribute to this condition. Understanding these factors is

essential for informing effective healthcare interventions and public health policies.

Our primary objective is to gain insights into various health factors, particularly those associated with heart disease. We aim to **develop a model capable of predicting the likelihood of individuals having heart disease** based on their health and lifestyle habits.

III. Data Analysis and Cleansing (Phase 2)

In this phase of the project, we delve into the **analysis** of the dataset, focusing on **preprocessing** steps and **exploratory** data analysis (EDA) to gain insights into the underlying patterns and characteristics of the data. This phase is crucial for understanding the dataset's structure, identifying potential issues or inconsistencies, and preparing the data for further analysis.

A. Pre-processing

The dataset used in this study originates from the Centers for Disease Control and Prevention (CDC) and is a significant component of the Behavioral Risk Factor Surveillance System (BRFSS). The BRFSS conducts annual telephone surveys to collect comprehensive data on the health status of residents in the United States.

Established in 1984 with data collection initiated in 15 states, the BRFSS has expanded its coverage to encompass all 50 states, the District of Columbia, and three U.S. territories. With over 400,000 adult interviews conducted each year, the BRFSS stands as the largest continuously conducted health survey system globally.

Before proceeding with data cleansing, the dataset underwent **encoding**, which included converting the data into numeric values. Categorical features were encoded as binary values (1 and 0), while numerical features were either grouped or transformed into numerical values as appropriate.

Data cleansing is a critical step in the pre-processing phase, aimed at ensuring the **integrity** and **quality** of the dataset. The following techniques were applied to cleanse the dataset:

*1) **Handling Missing Values**:* The dataset was initially verified for missing values. Fortunately, no missing values were detected, alleviating the need for imputation or removal of incomplete records.

*2) **Removing Duplicate Records**:* Duplicate records were identified and addressed to ensure data integrity and consistency.

The rows were identified by comparing all rows in the dataset to find instances with identical values across all columns, then number of duplicate rows identified and their corresponding details were printed for verification purposes, if duplicate records were found, they were removed from the dataset, retaining only the first occurrence of each duplicated row.

*3) **Dealing with Outliers**:* We embarked on detecting and handling outliers within the dataset to ensure the accuracy and reliability of our analyses.

Firstly, outliers were identified by computing the first quartile (Q1), third quartile (Q3), and interquartile range (IQR) for each numerical feature or categorical feature with more than 2 values. Data points falling below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR were flagged as outliers. This process was executed systematically across all numerical features, allowing for comprehensive outlier detection.

Following the identification of outliers, a robust strategy was implemented to handle them effectively. **They were directly removed from the dataframe**. This approach provides a straightforward method for mitigating the impact of outliers on subsequent analyses, ensuring the integrity of the dataset.

Furthermore, to maintain the integrity of features, a verification step was conducted to **ensure that each feature retained more than one unique value** after outlier removal. Features exhibiting a single unique value post-outlier removal were flagged for deletion from the dataset.

*4) **Final Steps**:* Upon implementing the data cleansing techniques, any features that contained only one unique value after cleaning were **removed** from the dataframe. In this instance, the **Race** feature was the sole feature meeting this criterion. This decision was made to ensure the dataset's coherence and suitability for further analysis.

Subsequently, the cleaned dataframe, free of outliers and redundant features, was saved to a new CSV file (`heart_2020_cleaned.csv`). This finalized dataset is now primed for subsequent analyses and modeling endeavors.

B. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial initial step in the data analysis journey, enabling us to glean insights and understanding from the raw dataset before proceeding to more complex analyses. In this section, we embark on an exploratory journey, employing a variety of visualization techniques and dimensionality reduction methods to uncover patterns, trends, and outliers within the data.

*1) **Initial Visualization**:* At the outset, subsequent to the encoding phase and preceding data cleansing (as discussed in the Pre-processing phase), histograms and box plots are employed to visually illustrate the distribution of the dataset and identify potential outliers within the numerical features. These visualizations play a crucial role in accurately pinpointing outliers, thereby fostering a comprehensive understanding of the dataset's characteristics.

Upon examination of the histograms, it becomes evident that **the data distribution is highly skewed**, with a predominant **imbalance** between individuals without heart disease and those affected. This skewed distribution poses a significant challenge for our study, as it may impact the accuracy of our predictive models.

Furthermore, the box plots provide insights into the **presence of outliers** within the dataset. Notably, due to the skewed distribution observed in the histograms, it is unsurprising to

find a substantial number of outliers. Additionally, certain features exhibit a **lack of variability**, with only one non-outlier value present (as observed in the Race feature, where the majority of subjects are categorized as ”White”).

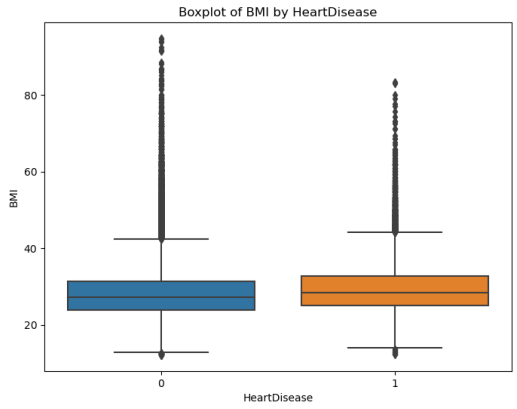

Fig. 1. BMI Box Plot

*2) **Cleansed Visualization**:* We revisit the histograms utilized during the initial visualization of the dataset. These histograms now reflect the **changes in the data post-cleansing**, offering insights into how the distributions of features have evolved. By comparing these histograms before and after data cleansing, we gain a **clearer understanding of the effects of our cleansing procedures** on the dataset's distribution patterns and identify notable improvements.

*3) **Correlation Analysis**:* Subsequent to the preliminary exploration, we ventured into a more detailed examination of the dataset's structure and relationships through correlation analysis. Utilizing a correlation plot, we aimed to **uncover the intricate interconnections between different features**, thereby gaining insights into potential associations and dependencies within the data. **Our primary focus was on identifying features that displayed strong correlations with** `HeartDisease`, aiming to pinpoint attributes with correlations closest to 1 or -1, indicative of a strong relationship with the target variable. For instance, `AgeCategory` emerged as one such feature displaying the highest correlation with heart disease.

*4) **Feature Importance Assessment**:* We employed a `RandomForestClassifier` to assess the **significance of each feature in predicting the target variable**. A horizontal bar plot (barh) was generated to visually represent the relative importance of each feature, providing valuable insights into the key drivers influencing the outcome of interest.

This assessment guides us in determining which **features to prioritize or rely more heavily** on during the creation of new features and the prediction of heart disease by randomly shuffling feature values and observing the impact on model accuracy, we gain a nuanced understanding of feature importance.

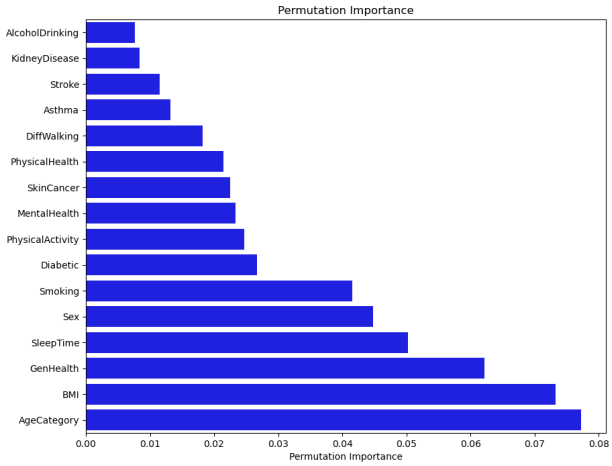

Fig. 2. Feature Importance BarH Plot

*5) **Dimensionality Reduction**:* In addition to visualization techniques, dimensionality reduction methods such as **Principal Component Analysis (PCA)** and **Uniform Manifold Approximation and Projection (UMAP)** are employed to spot intricate patterns within the data. These techniques effectively condense the dataset's dimensions while preserving essential information, enhancing our ability to discern meaningful insights and understand its characteristics and patterns more comprehensively.

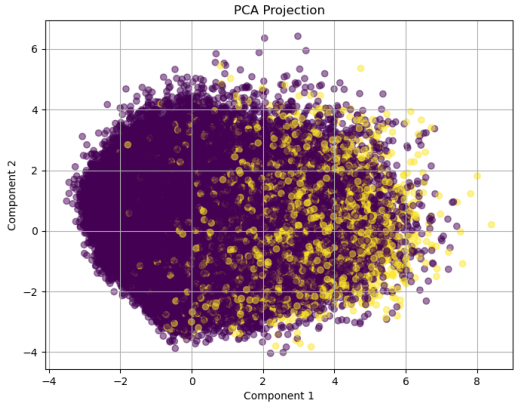

Fig. 3. PCA Plot

Upon analyzing the dimensionality reduction methods, **we identified distinct patterns in each one**, signifying a clear separation within the data. This observation shows the efficacy of the dimensionality reduction techniques in revealing underlying structures and facilitating a deeper understanding of the dataset's inherent complexity. Through these findings, we can confidently draw boundaries that delineate different data clusters, further aiding in subsequent analyses and interpretation.

C. Hypothesis Testing

In our hypothesis testing, we referred to the table provided in the slides from theoretical lecture 3, focusing on assessing two distinct groups: individuals with and without heart disease. Given that **our data is unpaired**, meaning there is no direct before-and-after relationship for the same individual, we sought additional insights into the normality of our variables.

To achieve this, we turned to Q-Q plots for a visual analysis. After careful examination, we identified that only two variables, 'BMI_with_HD' and 'BMI_without_HD,' exhibited a **pattern closely aligned with the diagonal line, indicating normal distribution** (**Fig.4** and **Fig.5**). Attempting to validate this observation with the **Shapiro-Wilk test posed challenges** due to the large dataset, resulting in all p-values being very **close to zero or zero** meaning that all our variables are not normal distributed (no p-value is bigger than 0.05).

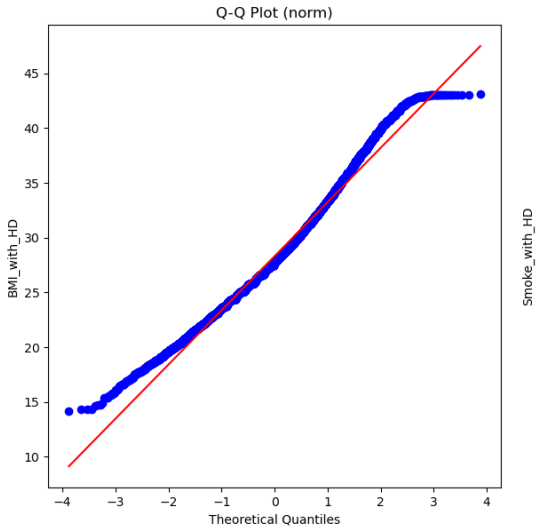

Fig. 4. Q-Q plot of BMI with Heart Disease

In light of this, we acknowledge 'BMI_with_HD' and 'BMI_without_HD' as the **variables demonstrating normal distribution**, while the remaining variables do **not exhibit normal distribution characteristics**. This approach enables us to effectively explore and compare differences between groups, considering both numerical and categorical variables in our examination of the presence or absence of heart disease.

IV. Model Selection (Phase 3)

In this pivotal phase of our analysis, we transition to the critical task of model selection. This section encompasses several key components, starting with **feature engineering**, where we delve into the creation of new features to enhance our predictive capabilities. Subsequently, we **explore various models**, evaluating a range of models to determine the most suitable ones for our specific problem domain. Each step in this phase is crucial, contributing to the development of a predictive framework that aligns with our overarching objectives.

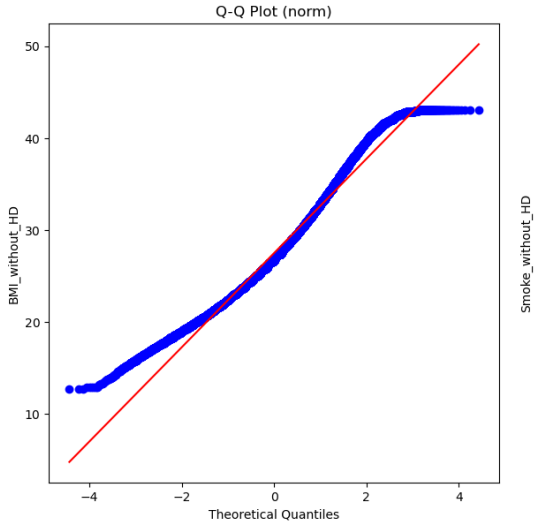

Fig. 5. Q-Q plot of BMI without Heart Disease

A. Feature Engineering

In this phase of our analysis, we focus on enhancing the predictive power of our models through feature engineering. Leveraging insights gained from our earlier exploratory data analysis, particularly from the Feature Importance Assessment and Correlation Analysis parts, we strategically **craft new features** to capture important relationships and patterns in the data.

We systematically **generated 10 new features**, each designed to provide additional information and improve the predictive performance of our models. These features include:

BMIClass: This feature categorizes the `BMI` data into 6 distinct groups [1]: underweight, normal, overweight, obese class 1, obese class 2, and obese class 3, corresponding to BMI < 18.5, BMI < 25, BMI < 30, BMI < 35, BMI < 40, and BMI ≥ 40, respectively. These categories aim to provide a comprehensive representation of the varied body mass index values present in the dataset.

SleepClass: Categorizing the `SleepTime` feature into 3 groups based on sleep duration: SleepTime < 6, SleepTime < 9, and SleepTime ≥ 9. This classification allows us to assess the influence of sleep patterns on the risk of heart disease.

BadHabitsScore: This feature combines the `Smoking` and `AlcoholDrinking` features, generating a composite score that represents unhealthy lifestyle habits. The score ranges from 0 to 2, with higher values indicating a more detrimental lifestyle.

Diseases: This feature aggregates information from the `KidneyDisease`, `Asthma`, `SkinCancer`, and `Diabetic` features to create a composite score reflecting the number of diseases present. The score ranges from 0 to 4, with higher values indicating a greater number of diseases.

PoorHealthDaysMonth: This feature calculates the average number of days of poor mental and physical health experienced per month. It combines the `MentalHealth` and

PhysicalHealth features, representing the number of days an individual experienced poor mental and physical health, respectively, within a month. The total days of poor health are divided by 30 to normalize the values, providing an overall assessment of the individual's health status in terms of the number of days of poor health experienced per month.

DangerousStroke: This feature identifies cases where individuals aged 65 or above have experienced a stroke, assigning a value of 1 if true and 0 otherwise.

AgeBMI_Interaction: This feature represents the interaction between AgeCategory and BMI, calculated by multiplying the values of these two features for each data point.

BMISleep_Interaction: This feature captures the interaction between BMI and SleepTime duration, obtained by multiplying the values of these two features for each data point.

AgeHealth_Interaction: This feature reflects the interaction between AgeCategory and GenHealth, computed by multiplying the values of these two features for each data point.

AgeSleep_Interaction: This feature represents the interaction between AgeCategory and SleepTime, determined by multiplying the values of these two features for each data point.

B. Assessment of Engineered Features

To evaluate the efficacy of these new features, we **revisit the histograms generated** during the initial visualization of the dataset, providing visual insights into their distributions and potential impacts on heart disease risk. Additionally, we construct a **new correlation map incorporating these features** to explore their relationships with the target variable and with each other. Through this comprehensive feature engineering process, we aim to enrich our dataset and empower our models with a deeper understanding of the underlying factors influencing heart disease.

C. Model Selection

This step involves **examining** both the nature of the data and the statistical analyses conducted. The dataset includes variables with various distributions, as apparent from unpaired T-tests and Wilcoxon rank-sum tests.

For **normally distributed variables**, the unpaired T-test has been applied, revealing significant differences between individuals with and without heart disease. In our opinion, supervised learning models such as **Logistic Regression** or **Support Vector Machines (SVM)** may be well-suited to address this aspect of the problem.

On the other hand, **non-normally distributed variables** have been assessed using Wilcoxon rank-sum tests. Robust models like **Decision Trees**, **Random Forest**, or **Gradient Boosting (e.g., XGBoost)** are potential candidates for capturing the patterns within these variables.

The correlation analysis further emphasizes the importance of features like AgeCategory, AgeBMI_Interaction, AgeSleep_Interaction, and Diseases, which exhibit notable associations with heart disease. Incorporating these features into the predictive model can enhance its ability to discern relevant patterns.

Additionally, newly created features, such as AgeBMI_Interaction and AgeSleep_Interaction, have demonstrated **meaningful correlations** and we believe should be considered in the model construction process.

To proceed, an iterative approach will be adopted, experimenting with different models, tuning hyperparameters, and assessing performance metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. This thorough exploration leads us to believe that it will assist in identifying the most effective model for predicting heart disease, aligning with the dataset's characteristics and the unique insights gained from statistical analyses.

V. Conclusion

In conclusion, we believe this dataset holds significant potential for our project's next phase. However, we anticipate challenges due to its highly imbalanced nature and the limited variation observed between individuals with and without heart disease across our features. Despite these obstacles, we remain committed to extracting meaningful insights and developing robust models to advance our understanding of heart disease and inform preventive measures.

References

[1] Flegal KM, Kit BK, Orpana H, Graubard BI. Association of All-Cause Mortality With Overweight and Obesity Using Standard Body Mass Index Categories: A Systematic Review and Meta-analysis. *JAMA*. 2013;309(1):71–82. doi:10.1001/jama.2012.113905