**Technical Report**

**Classification Model for NPHA dataset**

Course name: Data Mining

Authors: Nastaran Nikpour, Parisa Zeynaly

**Introduction**

**National Poll on Healthy Aging (NPHA)**  
Donated on 12/5/2023

This is a subset of the NPHA dataset filtered down to develop and validate machine learning algorithms for predicting the number of doctors a survey respondent sees in a year. The records in this dataset represent seniors who responded to the NPHA survey. Regarding the characteristics of our dataset, we can observe:

* It is a **tabular dataset**
* The primary **subject area is Health and Medicine.**
* The **associated task is classification**, meaning we are aiming to categorize respondents based on their predicted number of doctor visits.
* All **features within our subset are categorical**, which has important implications for our chosen preprocessing and modeling techniques.
* The dataset comprises **714 instances** and **14 features.**

**For what purpose was the dataset created?**

The National Poll on Healthy Aging dataset was created to gather insights on the health, healthcare, and health policy issues affecting Americans aged 65-80. By focusing on the perspectives of older adults and their caregivers, the University of Michigan aimed to inform the public, healthcare providers, policymakers, and advocates about the various aspects of aging. This includes topics like health insurance, household composition, sleep issues, dental care, prescription medications, and caregiving, thereby providing a comprehensive understanding of the health-related needs and concerns of the older population.

**Who funded the creation of the dataset?**

The dataset was funded by AARP and Michigan Medicine, the University of Michigan's academic medical centre.

**What do the instances in this dataset represent?**

Each row represents a survey respondent.

**Does the dataset contain data that might be considered sensitive in any way?**

Yes. There is information about race/ethnicity, gender, age.

**Was there any data preprocessing performed?**

For this subset of the original NPHA dataset we chose 14 features related to health and sleep to use for the prediction task. We then removed all survey respondents with missing responses for any of the chosen features.

**Has Missing Values?**

No

**Feature Definitions**

**Target Variable: Number\_of\_Doctors\_Visited**

The Number\_of\_Doctors\_Visited serves as the **target variable** for the prediction task. It quantifies the total count of distinct doctors a senior survey respondent has consulted within a year. Crucially, this variable is **categorized** rather than representing a precise numerical count, facilitating a multi-class classification problem. The categories are defined as follows:

* **1:** (0-1 doctors) – The individual visited zero or one doctor.
* **2:** (2-3 doctors) – The individual visited two or three doctors.
* **3:** (4 or more doctors) – The individual visited four or more doctors.

**Age**

The Age variable characterizes the senior survey respondent's age group. This feature is also **categorized**, not a precise numerical value, and is divided into two distinct age bands:

* **1:** (50-64) – This category includes individuals aged 50 to 64 years old.
* **2:** (65-80) – This category includes individuals aged 65 to 80 years old.

**Decision on Age Feature:** Based on preliminary analysis, the Age column consistently exhibits a value of "2," indicating that all respondents in this filtered subset belong to the 65-80 age group. As a feature with no variance (a constant feature), it provides **no predictive power** to machine learning models. Including such a feature would only needlessly increase computational complexity and potentially hinder model interpretability without offering any performance benefit. Therefore, it is strongly recommended to **drop this column** from the dataset during preprocessing, as it acts as noise rather than an informative predictor.

**Physical\_Health**

The Physical\_Health variable captures a **self-assessment** of the senior survey respondent's overall physical well-being. This feature is **categorized** with the following response options:

* **-1:** (Refused) – The individual declined to answer this question. (This value is explicitly treated as a distinct category, not a missing value, for preprocessing).
* **1:** (Excellent) – The individual rated their physical health as excellent.
* **2:** (Very Good) – The individual rated their physical health as very good.
* **3:** (Good) – The individual rated their physical health as good.
* **4:** (Fair) – The individual rated their physical health as fair (implying neither good nor poor, or somewhat poor).
* **5:** (Poor) – The individual rated their physical health as poor.

**Mental\_Health**

The Mental\_Health variable represents a **self-evaluation** of the senior survey respondent's mental or psychological well-being. This feature is also **categorized**, including the following options:

* **-1:** (Refused) – The individual declined to answer this question.
* **1:** (Excellent) – The individual rated their mental health as excellent.
* **2:** (Very Good) – The individual rated their mental health as very good.
* **3:** (Good) – The individual rated their mental health as good.
* **4:** (Fair) – The individual rated their mental health as fair (implying neither good nor poor, or somewhat poor).
* **5:** (Poor) – The individual rated their mental health as poor.

**Dental\_Health**

The Dental\_Health variable reflects a **self-assessment** of the senior survey respondent's oral or dental well-being. This **categorized** feature includes the following response options:

* **-1:** (Refused) – The individual chose not to answer this question.
* **1:** (Excellent) – The individual rated their dental health as excellent.
* **2:** (Very Good) – The individual rated their dental health as very good.
* **3:** (Good) – The individual rated their dental health as good.
* **4:** (Fair) – The individual rated their dental health as fair (indicating it's neither good nor poor, or somewhat lacking).
* **5:** (Poor) – The individual rated their dental health as poor.

**Employment**

The Employment variable indicates the senior survey respondent's employment status or work-related information, offering insights into their participation in the workforce. This feature is **categorized** with the following options:

* **-1:** (Refused) – The individual declined to answer this question.
* **1:** (Working full-time) – The individual is employed full-time.
* **2:** (Working part-time) – The individual is employed part-time.
* **3:** (Retired) – The individual is retired.
* **4:** (Not working at this time) – The individual is currently not working (this differs from being retired, implying they might be unemployed or not working for other reasons).

Sleep-Related Features: Stress\_Keeps\_Patient\_from\_Sleeping, Medication\_Keeps\_Patient\_from\_Sleeping, Pain\_Keeps\_Patient\_from\_Sleeping, Bathroom\_Needs\_Keeps\_Patient\_from\_Sleeping

These four features are binary categorical variables, each indicating whether a specific factor impacts the individual's ability to sleep. They are encoded as:

* **0:** (No) – The specified factor does not affect the individual's ability to sleep.
* **1:** (Yes) – The specified factor does affect the individual's ability to sleep.

**Important Note for Trouble sleeping:** The provided data description indicates that the actual values in the column originally titled "Trouble sleeping" are 1, 2, or 3, rather than the expected 0s and 1s. This inconsistency invalidates the direct use of such a column as a binary indicator and requires **correction or re-evaluation** of this specific feature's utility.

**Prescription\_Sleep\_Medication**

The Prescription\_Sleep\_Medication variable provides information regarding the use of prescribed sleep medication by the senior survey respondent. This **categorical feature** includes the following options:

* **-1:** (Refused) – The individual declined to answer the question.
* **1:** (Use regularly) – The individual uses prescribed sleep medication on a regular basis.
* **2:** (Use occasionally) – The individual uses prescribed sleep medication occasionally.
* **3:** (Do not use) – The individual does not use any prescribed sleep medication.

**Race**

The Race variable represents the senior survey respondent's racial or ethnic background, serving as a fundamental demographic characteristic for understanding participant diversity. This **categorical feature** includes the following options:

* **-2:** (Not asked) – This indicates the question was not posed to the respondent, possibly due to survey logic or data collection protocols.
* **-1:** (REFUSED) – The individual explicitly declined to answer the question when presented.
* **1:** (White, Non-Hispanic)
* **2:** (Black, Non-Hispanic)
* **3:** (Other, Non-Hispanic)
* **4:** (Hispanic)

**5:** (2+ Races, Non-Hispanic)

**Gender**

The Gender variable represents the senior survey respondent's gender identity, another fundamental demographic characteristic. This **categorical feature** provides the following options:

* **-2:** (Not asked) – This indicates that the question was not presented to the respondent.
* **-1:** (REFUSED) – The individual declined to answer the question when it was asked.
* **1:** (Male)
* **2:** (Female)

**Important Note:** Similar to the Race feature, the presence of both "-2" and "-1" values for missing information suggests distinct reasons for the absence of a response, which should be carefully considered during data preprocessing.

**Data preprocessing**

**Libraries and Tools**

The following Python libraries were instrumental in various stages of this machine learning project, from data manipulation to model evaluation and hyperparameter tuning.

**pandas**: This library was used extensively for efficient **data manipulation and analysis**. It provided robust data structures like DataFrames, which were essential for loading the NPHA dataset, managing its features, and performing operations such as splitting the data and preparing it for the preprocessing pipeline.

**numpy**: As the fundamental package for **numerical computing** in Python, NumPy was utilized for array operations and mathematical functions. Although not always explicitly called out in high-level machine learning code, it underpins many numerical computations within libraries like scikit-learn and LightGBM.

**matplotlib.pyplot**: This library was employed for **creating static, interactive, and animated visualizations** in Python. In the context of a technical report, it would typically be used for generating plots such as feature distributions, model performance graphs, or confusion matrices to visually represent data insights and model results.

**seaborn**: Built on top of Matplotlib, Seaborn provides a **high-level interface for drawing attractive and informative statistical graphics**. It simplifies the process of creating complex visualizations, which is useful for exploratory data analysis (EDA) or presenting classification outcomes in a more digestible format, such as heatmaps of correlation or class distribution plots.

**sklearn.model\_selection**: This scikit-learn module was crucial for **managing the model development workflow**, specifically for train\_test\_split to divide the dataset into training and validation sets, GridSearchCV for exhaustive hyperparameter tuning, and RandomizedSearchCV for more efficient random hyperparameter search.

**sklearn.preprocessing**: This module provided essential tools for **data transformation**. OneHotEncoder was particularly vital given that all features in the dataset are categorical, converting them into a numerical format suitable for machine learning algorithms.

**sklearn.compose.ColumnTransformer**: This utility was used to **apply different transformers to different columns** of the dataset. It enabled the construction of a flexible preprocessing pipeline that could handle various feature types (though in your case, all were categorical, it's a best practice for mixed-type datasets).

**sklearn.pipeline.Pipeline**: The Pipeline class was fundamental for **sequencing data preprocessing steps with a machine learning model**. It streamlined the entire workflow, ensuring that transformations and model training were applied consistently and efficiently, especially within cross-validation loops.

**sklearn.tree.DecisionTreeClassifier**: While LightGBM was the primary chosen model, DecisionTreeClassifier is a base algorithm that could have been considered or used for initial exploration. It represents a **simple, tree-based classification algorithm** that serves as a foundational component for more complex ensemble methods like LightGBM.

**sklearn.metrics**: This module was indispensable for **evaluating the performance of the classification model**. Key functions included accuracy\_score for overall correctness, classification\_report for detailed per-class metrics (precision, recall, F1-score), f1\_score for calculating the F1-score specifically, and make\_scorer for custom score functions during hyperparameter tuning.

**lightgbm.LGBMClassifier**: This was the **core classification algorithm** chosen for the project. LightGBM, a gradient boosting framework, was selected for its efficiency, speed, and high accuracy, particularly its optimization for handling large datasets and categorical features effectively.

**sklearn.ensemble.RandomForestClassifier**: As another powerful ensemble learning method, RandomForestClassifier might have been explored or considered for comparison. It builds multiple decision trees and merges their results to improve predictive accuracy and control overfitting, offering an alternative to boosting algorithms.

**sklearn.linear\_model.LogisticRegression**: This module provides LogisticRegression, a **fundamental linear model for binary and multi-class classification**. It often serves as a strong baseline model due to its simplicity, interpretability, and effectiveness, which could be used for initial benchmarking before more complex models are deployed.

**imblearn.over\_sampling.SMOTE**: This technique from imblearn was specifically employed to **address the class imbalance** present in the dataset. SMOTE (Synthetic Minority Over-sampling Technique) generates synthetic samples for minority classes, thereby creating a more balanced training set and preventing the model from being biased towards majority classes.

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer code

AI-generated content may be incorrect.

Initially, the **Age** column within this specific dataset subset was identified as a constant feature. As detailed in Section 2.1, all instances in the Age column consistently held the value "2," signifying that every respondent belonged to the 65-80 age group. Features exhibiting no variance provide no discriminative power to a machine learning model; they effectively act as noise that can unnecessarily increase computational complexity without contributing to predictive accuracy or model interpretability. Consequently, the Age column was removed from the dataset during the preprocessing phase.

A critical inconsistency was subsequently identified within the **Trouble Sleeping**, particularly for a column intended to encode binary responses (0 for 'No' and 1 for 'Yes'). While the dataset description indicated expected values of 0 and 1, the actual data contained values such as 1, 2, and 3. This discrepancy rendered the column invalid for its intended binary interpretation. To rectify this, the data in this specific column was reprocessed to conform to the binary 0 and 1 encoding as described for the other sleep-related features (Stress\_Keeps\_Patient\_from\_Sleeping, Medication\_Keeps\_Patient\_from\_Sleeping, Pain\_Keeps\_Patient\_from\_Sleeping, Bathroom\_Needs\_Keeps\_Patient\_from\_Sleeping). This ensured that the feature accurately reflected its binary nature and was correctly interpreted by subsequent modeling steps.

Finally, the dataset contained specific negative numerical encodings, such as -1 and -2, for certain categorical features including Physical\_Health, Mental\_Health, Dental\_Health, Employment**,** Prescription\_Sleep\_Medication, Race, and Gender. These values represent distinct categorical responses, such as **'Refused' (-1) or 'Not asked' (-2),** and are not standard missing value indicators. Given that the count of instances with these specific negative encodings was not negligible, these values were retained and treated as distinct categorical levels during the **OneHotEncoder** transformation. This decision prevented data loss that would have occurred if these instances were removed, and ensured that the model learned from these specific respondent behaviors or survey conditions, rather than treating them as generic missing entries.

A screenshot of a computer program

AI-generated content may be incorrect.

A screen shot of a computer

AI-generated content may be incorrect.

**Visualization**

A chart of a number of doctors visited

AI-generated content may be incorrect.

The bar chart, titled "Distribution of Number of Doctors Visited," illustrates the frequency of patients across three categorized ranges of doctor visits:

* **'0-1 Doctors (Category 1)'**: This category represents patients who visited 0 or 1 doctor. The bar indicates approximately **130 patients** fall into this group.
* **'2-3 Doctors (Category 2)'**: This category includes patients who visited 2 or 3 doctors. With a count of approximately **370 patients**, this is the **most prevalent category**, suggesting that a substantial portion of the dataset's population engages in a moderate number of doctor visits.

**'4+ Doctors (Category 3)'**: This category encompasses patients who visited 4 or more doctors. Around **210 patients** belong to this group.

The distribution is visibly skewed, with the '2-3 Doctors (Category 2)' segment accounting for the largest proportion of individuals. This indicates that the majority of patients in this dataset have a moderate frequency of doctor visits. The '0-1 Doctors' group is the least frequent, while the '4+ Doctors' group, though smaller than the '2-3 Doctors' group, still represents a significant segment of the population.

This analysis provides valuable insight into the central tendency and spread of doctor visit frequencies within the dataset, informing subsequent modeling decisions, particularly if 'Number of Doctors Visited' is the variable to be predicted or classified. Understanding this distribution helps in assessing the balance of classes for classification tasks or the spread for regression tasks.

A screenshot of a graph

AI-generated content may be incorrect.

The composite plot displays the frequency distribution for eleven distinct categorical features:

**Health-Related Features**

* **Distribution of Physical Health**: This chart shows a diverse distribution. Categories '3' and '4' (likely representing moderate to good physical health) exhibit the highest counts, with category '4' being the most frequent (approximately 290 instances). Category '1' (indicating very poor health) has the lowest count, suggesting that severe physical health issues are less common in this dataset.
* **Distribution of Mental Health**: Similar to physical health, categories '3' and '4' dominate, with '3' having the highest frequency (around 260 instances). This suggests that a significant portion of the population experiences moderate mental health. Categories '1' and '5' show very low counts.
* **Distribution of Dental Health**: This feature also displays a skewed distribution. Categories '3' and '4' are the most prominent, with '3' again being the most frequent (approximately 250 instances). This indicates a tendency towards moderate to good dental health within the dataset.

**Socio-Economic and Behavioral Features**

* **Distribution of Employment**: The distribution of employment shows that category '3' is overwhelmingly the most frequent (over 400 instances), likely representing full-time employment or a dominant employment status. Other categories have significantly fewer instances, with '1' and '5' being very rare.

**Sleep-Related Issues**

The following features relate to factors that keep patients from sleeping, generally exhibiting a binary-like distribution where one category is much more prevalent than others, indicating the presence or absence of the issue for most individuals.

* **Distribution of Stress Keeps Patient from Sleeping**: A large majority of patients (over 500 instances) fall into one category, indicating that stress *does not* frequently keep them from sleeping or that a particular stress level is common. The other category has a significantly lower count (around 150 instances).
* **Distribution of Medication Keeps Patient from Sleeping**: Similar to stress, the vast majority of individuals (over 600 instances) are in one category, suggesting that medication is not a primary cause of sleep disturbance for most.
* **Distribution of Pain Keeps Patient from Sleeping**: Again, a heavily skewed distribution, with over 500 instances in one category, implying that pain is not a common factor preventing sleep for most patients.
* **Distribution of Bathroom Needs Keeps Patient from Sleeping**: Unlike the previous sleep-related factors, this feature shows a more balanced distribution, with both categories having substantial counts (around 350 instances each). This suggests that bathroom needs are a more common and equally distributed factor influencing sleep compared to stress, medication, or pain.
* **Distribution of Unknown Keeps Patient from Sleeping**: The distribution is skewed, with one category having a higher count (around 400 instances) compared to the other (around 300 instances), indicating that for a larger segment, "unknown" factors are not primarily impacting their sleep.
* **Distribution of Trouble Sleeping**: Similar to medication and pain, a significant majority (over 600 instances) are in one category, indicating that most patients do not report general trouble sleeping.
* **Distribution of Prescription Sleep Medication**: A highly imbalanced distribution, with over 600 instances in one category, suggesting that the vast majority of patients are *not* on prescription sleep medication.

**Demographic Features**

* **Distribution of Race**: This chart shows a heavily imbalanced distribution. One racial category (likely represented by '1', with over 600 instances) is overwhelmingly dominant, while other racial categories ('2' through '5') have very low counts. This suggests a lack of diversity in the 'Race' feature within this dataset.
* **Distribution of Gender**: The distribution of gender is relatively balanced, with both categories having substantial counts (one over 300 instances, the other over 350 instances). This indicates a fairly even representation of genders in the dataset.

**Key Observations and Implications:**

1. **Dominant Categories**: Many features, such as 'Employment', 'Stress Keeps Patient from Sleeping', 'Medication Keeps Patient from Sleeping', 'Pain Keeps Patient from Sleeping', 'Trouble Sleeping', 'Prescription Sleep Medication', and especially 'Race', exhibit a strong skew towards one particular category.
2. **Imbalanced Classes**: The significant class imbalance in features like 'Race' and 'Prescription Sleep Medication' may require careful consideration during modeling, potentially necessitating techniques like oversampling or undersampling if these features are used as target variables or if their minority classes are of particular interest.
3. **Relatively Balanced Features**: 'Physical Health', 'Mental Health', 'Dental Health', 'Bathroom Needs Keeps Patient from Sleeping', and 'Gender' show more distributed patterns, although still with prominent categories. 'Bathroom Needs Keeps Patient from Sleeping' and 'Gender' are notably more balanced compared to others.
4. **Data Interpretation**: The numerical representation of categories (e.g., '1', '2', '3') for many features suggests that these are encoded values for underlying nominal or ordinal categories. Proper handling (e.g., One-Hot Encoding for nominal features where order doesn't matter) will be critical during preprocessing.
5. **Insights for Modeling**: Features with highly skewed distributions might have less predictive power if their dominant category doesn't differentiate well among the target outcomes. Conversely, 'Bathroom Needs Keeps Patient from Sleeping' and 'Gender' could potentially be more informative due to their varied distributions.

This detailed distributional analysis provides a comprehensive understanding of the characteristics of each categorical feature, which will guide further data preprocessing, feature engineering, and the selection of appropriate machine learning models.

A screenshot of a graph

AI-generated content may be incorrect.

These charts illustrate the relationship between several categorical features and the target variable (number of doctor visits). In each chart, the horizontal axis represents a specific feature (e.g., physical health status, employment, race, etc.), and the bars show the count of individuals within each category, broken down by the number of doctor visits (1, 2, or 3 visits). Dark purple represents "1 doctor visit," teal represents "2 doctor visits," and yellow represents "3 doctor visits."

Now, let's interpret each section of the charts:

* **Physical Health vs. Number of Doctors Visited:**
  + Those with good (1) or excellent (2) physical health tend to have visited 1 or 2 doctors more often.
  + Individuals with average (3) or poor (4 and 5) physical health have a higher number of doctor visits (2 or 3).
* **Mental Health vs. Number of Doctors Visited:**
  + Similar to physical health, individuals with good mental health tend to visit doctors less frequently.
  + As mental health deteriorates, the number of doctor visits (especially 2 or 3 visits) increases.
* **Dental Health vs. Number of Doctors Visited:**
  + People with good (1) or excellent (2) dental health usually have 1 or 2 doctor visits.

Poorer dental health (4 and 5) is associated with an increased number of visits (2 and 3), indicating a link between oral health issues and overall medical care needs.

* **Employment vs. Number of Doctors Visited:**
  + It appears that groups such as "employed" (1) and "retired" (2) have the highest number of doctor visits (especially 2 and 3 visits).
  + Individuals who are "unemployed" (8) or "students" (9) visit doctors less frequently.
* **Stress Keeps Patient from Sleeping vs. Number of Doctors Visited:**
  + Individuals who cannot sleep due to stress (higher values) generally visit doctors more often (especially 2 or 3 visits).
  + Those without stress-related sleep issues (1) have visited doctors less.
* **Medication Keeps Patient from Sleeping vs. Number of Doctors Visited:**
  + Individuals whose medication prevents them from sleeping (higher values) seem to have a higher number of doctor visits (especially 2 or 3 visits).

Those without such an issue visit doctors less frequently.

* **Pain Keeps Patient from Sleeping vs. Number of Doctors Visited:**
  + It is clear that chronic pain preventing sleep (higher values) is associated with a significant increase in the number of doctor visits (2 or 3 visits).
* **Bathroom Needs Keep Patient from Sleeping vs. Number of Doctors Visited:**
  + This also indicates that frequent nighttime bathroom needs (which could signal underlying health issues) are correlated with a higher number of doctor visits (2 or 3 visits).
* **Unknown Keeps Patient from Sleeping vs. Number of Doctors Visited:**

Even if the reason for insomnia is unknown, the insomnia itself (and related problems) leads to an increase in doctor visits (2 or 3 visits).

* **Trouble Sleeping vs. Number of Doctors Visited:**
  + Generally, the more trouble an individual has sleeping (higher values), the greater the number of doctor visits (especially 2 and 3 visits).
* **Prescription Sleep Medication vs. Number of Doctors Visited:**
  + Individuals who take prescription sleep medication (1) have significantly more doctor visits (2 or 3 visits) compared to those who do not (2). This is logical as taking such medication implies underlying issues requiring a doctor's intervention.
* **Race vs. Number of Doctors Visited:**
  + Different racial groups appear to have varying patterns in the number of doctor visits. For example, the group with code 1 (likely Caucasians) has the highest number of visits.

Other races (e.g., 2, 3, and 4) visit less, which could indicate differences in access to healthcare, healthcare-seeking culture, or disease prevalence.

* **Gender vs. Number of Doctors Visited:**
  + It appears that females (1) generally have a higher number of doctor visits (especially 2 and 3 visits) compared to males (2). This is a common pattern in health data.

**General Observations:**

* These charts clearly demonstrate that **health problems (physical, mental, dental, sleep, pain)** and **specific statuses (such as retirement, taking sleep medication, female gender)** are associated with a higher number of doctor visits.
* The teal bar (2 visits) is the largest bar in many charts, indicating that "2 doctor visits" is the most common visit category in the available data.

A graph of a number of data

AI-generated content may be incorrect.This heatmap, titled 'Correlation Heatmap of One-Hot Encoded Categorical Features,' visually represents the pairwise Pearson correlation coefficients between all combinations of the one-hot encoded categorical input variables. The color intensity and hue indicate the strength and direction of the correlation: dark red signifies a strong positive correlation (approaching +1), dark blue denotes a strong negative correlation (approaching -1), and shades approaching white indicate weak or no linear correlation. As expected for one-hot encoded variables derived from the same original feature, categories within the same feature exhibit strong negative correlations; for instance, the correlation between 'Physical Health\_1' (excellent physical health) and 'Physical Health\_2' (very good physical health) is highly negative, approximately -0.2 to -0.4, indicating that individuals falling into one category are unlikely to fall into the other. More importantly for data understanding, the heatmap reveals inter-feature relationships. For example, there's a noticeable positive correlation, around +0.3 to +0.5, between 'Physical Health\_1' and 'Mental Health\_1', suggesting that individuals reporting excellent physical health often also report excellent mental health. Conversely, negative correlations are observed between optimal health states and problematic ones; for instance, 'Physical Health\_1' shows a negative correlation with 'Pain Keeps Patient from Sleeping\_1', implying that excellent physical health is associated with a lower likelihood of sleep disturbance due to pain. Similarly, a moderate positive correlation (around +0.4) exists between 'Stress Keeps Patient from Sleeping\_1' and 'Trouble Sleeping\_1', which is intuitively logical. The primary utility of this heatmap in our analysis is to identify potential multicollinearity among predictor variables, where highly correlated features (e.g., correlations exceeding ±0.8, though none this extreme are visible here) could lead to unstable coefficient estimates in certain predictive models. While this map effectively highlights relationships *among* the input features, it does not directly convey their individual importance or correlation with the *target variable* (Number of Doctors Visited), which would require separate analyses such as feature importance scores from predictive models or direct correlation with the target.

A graph of a number of patients

AI-generated content may be incorrect.

Precise Interpretation of the "Correlation of Features with 'Number of Doctors Visited'" Heatmap (English)

This heatmap vividly illustrates the Pearson Correlation between each of the summarized categorical features (presumably after One-Hot Encoding and subsequent aggregation of their effects into a single representative value per original feature) and the target variable, "Number of Doctors Visited." The numerical values within each cell represent the correlation coefficient, while the colors indicate both the direction (positive or negative) and strength of the correlation.

* Red Hue: Signifies a positive correlation. The darker the red, the stronger the positive correlation. This implies that as the value (or the presence of categories indicating a more adverse state) of that feature increases, the "Number of Doctors Visited" also tends to increase.
* Blue Hue: Denotes a negative correlation. The darker the blue, the stronger the negative correlation. This suggests that as the value (or the presence of categories indicating a more adverse state) of that feature increases, the "Number of Doctors Visited" tends to decrease.
* White/Light Gray Hue: Indicates a very weak or negligible correlation (close to zero).

Now, let's interpret each feature's correlation:

1. Physical Health: **0.17** (Strong Positive Correlation)
   * Interpretation: This feature exhibits the highest positive correlation with "Number of Doctors Visited." This suggests that the poorer an individual's physical health (assuming higher values or categories represent poorer health after aggregation), the more likely they are to visit a doctor. This relationship is intuitively logical, as individuals with more physical ailments typically require more medical attention.
2. Medication Keeps Patient from Sleeping: **0.12** (Moderate Positive Correlation)
   * Interpretation: This feature also displays a notable positive correlation. It indicates that individuals whose medication interferes with their sleep tend to visit doctors more frequently. This could be attributed to the side effects of the medication itself or the underlying medical conditions necessitating such medication, leading to increased doctor visits.
3. Employment: **0.09** (Weak Positive Correlation)
   * Interpretation: A weak positive correlation is observed. Considering previous charts that showed retired and employed individuals having more visits (possibly due to older age or work-related stress), this overall positive correlation is reasonable.
4. Pain Keeps Patient from Sleeping**: 0.08** (Weak Positive Correlation)
   * Interpretation: Chronic pain that disrupts sleep is weakly associated with an increased number of doctor visits. This is logical, as pain often requires medical management and treatment.
5. Trouble Sleeping: **0.07** (Weak Positive Correlation)
   * Interpretation: General difficulty in sleeping (regardless of cause) also shows a weak positive correlation with doctor visits. Sleep disturbances can be indicative of other health issues or themselves necessitate medical intervention.
6. Bathroom Needs Keep Patient from Sleeping**: 0.06** (Weak Positive Correlation)
   * Interpretation: This feature has a weak positive correlation. Frequent nighttime bathroom needs can be a symptom of conditions like prostate issues, diabetes, or other ailments that lead to increased medical consultations.
7. Mental Health: **0.05** (Very Weak Positive Correlation)
   * Interpretation: A very weak positive correlation is observed. While mental health is a critical aspect of well-being, its direct correlation with the overall "Number of Doctors Visited" in this dataset is surprisingly low. This might be due to the aggregation method of One-Hot Encoded categories, or perhaps mental health issues more often lead to visits to specialized mental health professionals rather than general practitioners, which may not be distinctly captured in the 'Number of Doctors Visited' variable.
8. Stress Keeps Patient from Sleeping: **0.05** (Very Weak Positive Correlation)
   * Interpretation: This feature shows a very weak positive correlation. Although stress is a significant health factor, its direct linear relationship with the number of doctor visits in this data is minimal.
9. Dental Health: **0.01** (Negligible Positive Correlation)
   * Interpretation: A virtually negligible positive correlation is noted. This suggests that, contrary to some expectations, dental health issues in this dataset have very little linear impact on the overall "Number of Doctors Visited." This could be because individuals primarily consult dentists for dental problems rather than general physicians.
10. Gender: **-0.00** (Near-Zero Correlation)
    * Interpretation: The correlation for Gender is approximately zero. This implies that in this dataset, gender (being male or female) does not have a significant linear relationship with the "Number of Doctors Visited." This finding is somewhat unexpected, as women are commonly observed to visit doctors more frequently than men in general health data. This might be influenced by data aggregation methods or the specific definition of gender in the one-hot encoding.
11. Unknown Keeps Patient from Sleeping: **-0.01** (Negligible Negative Correlation)
    * Interpretation: A very slight negative correlation is present. This correlation is so close to zero that it indicates no meaningful relationship.
12. Race: **-0.05** (Weak Negative Correlation)
    * Interpretation: A weak negative correlation is observed. This suggests that certain racial categories (which may have lower encoded values or are more prevalent in the dataset) tend to have fewer doctor visits, or categories with higher values have fewer visits. This could potentially indicate disparities in access to healthcare or cultural differences in healthcare-seeking behavior.
13. Prescription Sleep Medication: **-0.12** (Moderate Negative Correlation)
    * Interpretation: This feature shows the highest negative correlation with "Number of Doctors Visited." This suggests that individuals who take prescription sleep medication, on average, tend to have fewer doctor visits. This result might initially seem counter-intuitive, as one might expect someone requiring sleep medication to have more underlying health issues and thus more doctor visits. A plausible reason for this negative correlation could be that once individuals receive prescribed sleep medication, their sleep-related issues are somewhat managed, thereby reducing the need for frequent follow-up visits for that specific problem or related conditions. In other words, after medication, their condition might stabilize, lessening the need for repeated consultations.

Overall Insights and Significance of This Heatmap:

* Physical Health emerges as the most significant feature in terms of its linear correlation with the "Number of Doctors Visited."
* Prescription Sleep Medication, despite its negative correlation, is also a notable feature, indicating a distinct pattern (fewer visits post-medication).
* Relatively Low Correlations: Generally, the correlation coefficients observed in this heatmap are relatively low (none approaching 0.5 or higher). This indicates that "Number of Doctors Visited" is a complex variable influenced by multiple factors, and no single feature alone strongly explains its variation.
* Limitations of Pearson Correlation: Pearson correlation solely measures linear relationships. There might be strong non-linear associations between features and the target variable that this heatmap cannot capture.
* One-Hot Encoding and Aggregation: It is assumed that for each original categorical feature (e.g., "Physical Health," which likely has multiple categories), an aggregation method (such as selecting the category with the strongest correlation or averaging correlations across categories) was used to arrive at a single coefficient for the entire feature. If this were not the case, and each one-hot encoded category were displayed individually, the interpretation would differ. (Based on the image, it appears the one-hot encoded values for each original feature have been aggregated by some method to derive a single correlation coefficient for the overarching feature.)

**Methodology**

During the exploratory data analysis, a critical step involved inspecting the distribution and unique values within our categorical features. This revealed the presence of specific negative integer codes, notably '-1' and in some cases '-2', within several features such as 'Physical Health', 'Mental Health', 'Dental Health', 'Prescription Sleep Medication', 'Race', and 'Gender'. These codes are not true categorical values representing a specific state but rather serve as special indicators, typically denoting 'Refused to answer' or 'Not asked'. Including these as distinct categories in subsequent analysis or model training would misrepresent the data, potentially leading to skewed correlations and suboptimal model performance. To address this, these special codes were treated as missing data. Consequently, they were explicitly replaced with NaN (Not a Number) markers. Following this, an imputation strategy was applied: each NaN value was replaced by the **mode** (the most frequently occurring value) of its respective feature. For instance, in 'Physical Health', NaN values (originally '-1') were imputed with '3.0' as it was the most prevalent category, while in 'Mental Health', NaN values were replaced by '2.0'. This imputation method was chosen for its simplicity and effectiveness in preserving the overall distribution of the categorical features, thereby ensuring that our dataset remains robust and representative for subsequent statistical analysis and predictive modeling.

**Data Splitting**

The initial crucial step in preparing the dataset for model training and evaluation involved segmenting it into distinct training and testing sets. This separation is fundamental to rigorously assess the model's generalization capability on unseen data and to prevent overfitting, a scenario where a model performs exceptionally well on training data but poorly on new data. Specifically, 80% of the dataset was allocated to the training set, which is used to fit and tune the model, while the remaining 20% was designated as the testing set for final, unbiased performance evaluation. A critical consideration for this project, given the identified class imbalance within the target variable (predicted number of doctor visits), was the application of stratified sampling during this split. By using stratify=y in the train\_test\_split function, the proportional representation of each class from the original dataset was precisely maintained in both the training and testing subsets. This meticulous approach ensures that the model learns from a representative distribution of all classes and is evaluated on a test set that accurately reflects the real-world class proportions, providing a more reliable assessment of its performance, particularly for minority classes.

**Preprocessing Pipeline**

With all features identified as categorical within our NPHA subset, establishing a robust preprocessing pipeline was absolutely crucial. Machine learning algorithms, particularly those based on numerical computations like LightGBM, cannot directly process categorical text or labels. To address this, a ColumnTransformer was implemented. This powerful scikit-learn utility allowed us to apply specific data transformations to designated columns, creating an organized and efficient workflow. Given the exclusively categorical nature of our features, OneHotEncoder was the primary transformation applied. This encoder converts each categorical feature into a set of binary (0 or 1) columns, where each new column represents a unique category. For instance, if a feature "Education Level" had categories "High School," "College," "Graduate," OneHotEncoder would create three new binary columns (e.g., "Education\_High School," "Education\_College," "Education\_Graduate"). This conversion not only makes the data digestible for the model but also prevents algorithms from mistakenly inferring an ordinal relationship between categories where none exists (e.g., thinking "High School" is "less" than "College" if simply encoded as 1, 2, 3). Furthermore, handle\_unknown='ignore' was specified within OneHotEncoder to gracefully manage any unseen categories that might appear in new, unseen data, preventing potential errors during model deployment. This meticulous preprocessing ensures the model receives data in an optimal format, maximizing its ability to learn meaningful patterns and make accurate predictions.

A graph of a number of people

AI-generated content may be incorrect.

The meaning of "importance" differs across methods:

**Pearson Correlation:** This metric only measures a pairwise linear relationship between two variables. That is, it quantifies how much one variable linearly increases or decreases as the other increases. A high correlation signifies a strong linear association. "Physical Health" in your heatmap showed a correlation of 0.169629, which was the highest positive linear correlation with "Number of Doctors Visited." This means that generally, as physical health tends to worsen, the number of doctor visits also tends to increase.

**Decision Tree Feature Importance:** This metric is calculated based on impurity reduction (such as Gini Impurity or Information Gain). A feature is considered "important" in a decision tree when splits made based on that feature significantly reduce the impurity of the nodes and help in better separating the classes.

A screenshot of a computer

AI-generated content may be incorrect.

The performance of the trained Decision Tree Classifier was thoroughly evaluated on the test dataset, yielding an Accuracy Score of 0.3636. While accuracy provides a general overview of correct predictions, it is often an insufficient metric for datasets with imbalanced class distributions, such as ours, where the "Number of Doctors Visited" classes are distributed as 26 samples for Class 1 (0-1 Doctors), 75 for Class 2 (2-3 Doctors), and 42 for Class 3 (4+ Doctors). This imbalance implies that a model might achieve a seemingly higher accuracy by predominantly predicting the majority class, thereby masking poor performance on minority classes. The observed accuracy of 36.36% is marginally better than random guessing for a three-class problem (approximately 33.33%), indicating that the current model's predictive power is quite limited.

A more granular understanding of the model's performance is provided by the Classification Report, which details precision, recall, f1-score, and support for each class. The support column clearly outlines the actual number of instances for each class within the test set: 26 for Class 1, 75 for Class 2, and 42 for Class 3, re-emphasizing the existing class imbalance. Analyzing the per-class metrics reveals significant weaknesses:

* **Precision:** For Class 1, precision is notably low at 0.12. This indicates that only 12% of the instances predicted by the model as Class 1 were actually correct, highlighting a high rate of false positive predictions for this minority class. Precision for Class 2 stands at 0.46, suggesting that 46% of its predictions were accurate, while Class 3 shows a precision of 0.38, pointing to a moderate level of correct positive predictions.
* **Recall:** The recall for Class 1 is also very low at 0.15, meaning the model only successfully identified 15% of the actual Class 1 instances. This implies a significant number of false negatives, where true Class 1 instances were missed. For Class 2, recall is 0.47, indicating that 47% of actual Class 2 instances were correctly captured. Class 3 has a recall of 0.31, showing a poor ability to identify actual instances of this class.
* **F1-Score:** The F1-score, which is the harmonic mean of precision and recall, provides a balanced measure of a model's accuracy. The F1-score for Class 1 is 0.14, for Class 2 is 0.46, and for Class 3 is 0.34. These values further underscore the model's overall weak performance, particularly its struggle with the minority classes, where both precision and recall are suboptimal.

Finally, the report presents aggregated averages. The macro avg for precision, recall, and f1-score are all around 0.31-0.32. This "unweighted" average, which treats all classes equally regardless of their size, reinforces the conclusion of poor performance across the board. In contrast, the weighted avg (which considers the support of each class) is slightly higher, with precision at 0.38, recall at 0.36, and f1-score at 0.37. The higher weighted avg values suggest that the model performs marginally better on the more populous classes (primarily Class 2), but the overall performance remains low.

In conclusion, the current Decision Tree model exhibits very poor predictive performance for the "Number of Doctors Visited." The primary issues lie in the model's inability to accurately identify minority classes (Class 1 and Class 3), leading to a high number of false positives and false negatives. Even for the majority class (Class 2), the performance is only moderate. The observed class imbalance likely contributes significantly to this weak performance, as the model may tend to favor predicting the majority class, thus sacrificing accuracy for the less represented categories.

**Handling Class Imbalance (SMOTE)**

A significant challenge identified within this classification task was the inherent class imbalance of the target variable (number of doctor visits). In imbalanced datasets, the number of instances for certain classes (minority classes) is substantially lower than others (majority classes). If left unaddressed, models tend to become biased towards the majority classes, leading to poor predictive performance, particularly for the underrepresented minority classes, as they receive insufficient learning opportunities. To mitigate this, the Synthetic Minority Over-sampling Technique (SMOTE) was strategically employed. SMOTE addresses imbalance by generating synthetic samples for the minority classes. It operates by selecting a random instance from the minority class, identifying its k-nearest neighbors, and then creating new synthetic samples along the line segments connecting the selected instance to its neighbors. This process effectively increases the representation of minority classes in the training data without simply duplicating existing instances, thereby enriching the feature space and providing the model with more diverse learning examples.

Crucially, the integration of SMOTE was meticulously managed within an imblearn.pipeline.Pipeline. This advanced pipeline structure, provided by the imbalanced-learn library, is vital for maintaining the integrity of the evaluation process. When traditional over-sampling methods are applied before cross-validation, synthetic samples created from the entire dataset can inadvertently leak into the validation folds, leading to data leakage. This results in an overly optimistic and misleading evaluation of model performance, as the model is indirectly exposed to information from the validation set during its training phase. By embedding SMOTE within the ImbPipeline, the over-sampling procedure is executed independently within each fold of the cross-validation process, applying SMOTE only to the training data of that specific fold. This rigorous methodology ensures that the model's performance metrics are genuinely reflective of its ability to generalize to truly unseen data, thereby providing a more robust and reliable assessment of its predictive capabilities on this imbalanced dataset.

**Model Selection (LightGBM)**

For this classification task, LightGBM (Light Gradient Boosting Machine) was selected as the primary machine learning algorithm, a decision driven by its advanced capabilities and suitability for the characteristics of the NPHA dataset. LightGBM is an open-source, distributed, high-performance gradient boosting framework based on decision tree algorithms. It distinguishes itself from other gradient boosting frameworks, such as XGBoost, primarily through its innovative techniques that enhance both efficiency and accuracy.

Several key advantages made LightGBM an optimal choice for this project:

Exceptional Speed and Efficiency: LightGBM employs two novel techniques: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). GOSS judiciously samples data instances, focusing more on those with larger gradients (i.e., those that are harder to classify) to reduce the number of data samples without losing accuracy. EFB bundles mutually exclusive features (features that rarely take on non-zero values simultaneously) to reduce the number of features. These methods drastically accelerate the training process, which is beneficial even for moderately sized datasets like ours, ensuring quicker iteration during model development and hyperparameter tuning.

High Predictive Accuracy: Despite its speed, LightGBM is known for delivering state-of-the-art accuracy. It builds trees leaf-wise, which can lead to faster convergence and potentially better accuracy than level-wise tree growth strategies, especially for complex relationships within the data.

Memory Efficiency: The optimization techniques mentioned above also contribute to LightGBM's lower memory consumption compared to other boosting frameworks, making it more feasible on systems with limited computational resources.

Native Handling of Categorical Features: This was a particularly critical factor for the NPHA dataset, as all 14 features are categorical. Unlike many other algorithms that require categorical features to be pre-processed into numerical representations (e.g., One-Hot Encoding) before ingestion, LightGBM can natively handle categorical input features. This feature can often lead to improved performance and reduced preprocessing complexity, as it optimizes splits for categorical variables more intelligently than if they were simply treated as numerical through one-hot encoding.

Multi-class Classification Support: Given that the task involves categorizing respondents into multiple classes based on doctor visits, LightGBM's direct support for multi-class classification (objective='multiclass') streamlined the model configuration and training process.

Considering these technical advantages, particularly its efficiency and native categorical feature handling, LightGBM presented itself as the most appropriate and powerful algorithm to achieve robust and accurate predictions for the multi-class classification problem at hand**.**

**Resolving Imbalance with SMOTE and Enhancing Model Robustness**

To address this critical challenge and enhance the model's generalization capability and minority class detection, we implemented the SMOTE (Synthetic Minority Over-sampling Technique). The objective of this intervention was to balance the class distribution within the training dataset. SMOTE successfully augmented the number of samples in the minority classes by generating new synthetic instances.

Following the application of SMOTE and the balancing of the data, the results demonstrated a positive shift in model performance. While a slight decrease in overall accuracy might be observed, this is an acceptable trade-off as the model is no longer solely focused on the majority class. More importantly, the post-SMOTE classification report showed a more equitable and balanced distribution of Precision, Recall, and F1-score across all classes. This outcome effectively mitigated the class imbalance problem and significantly improved the model's fairness and robustness in accurately predicting minority classes.

**Conclusion**

In this data mining project, aiming to predict "Number of Doctors Visited" based on the NPHA-doctor-visits.csv dataset, we followed a systematic process from start to finish. In the first phase, **Data Understanding**, we became familiar with the dataset, comprising 714 rows and 15 columns (14 features and 1 target variable). We observed the presence of special values -1 and -2 in certain columns, which required careful preprocessing. More importantly, we identified a significant class imbalance in the distribution of the target variable (Number of Doctors Visited with categories 1, 2, and 3), where Class 1 was a severe minority. During the **Data Preparation** phase, we initially corrected a spelling error in the column name from Phyiscal Health to Physical Health and removed the Age column, which lacked any variance and useful information. Subsequently, given the categorical nature of the features and their special values, we employed the **One-Hot Encoding** method to convert them into a suitable numerical format for machine learning algorithms. Finally, the prepared data was split into 80% training and 20% testing sets, ensuring that the class distribution was maintained in both sets through stratification for a fairer model evaluation. In the **Modeling and Evaluation** phase, we sequentially tested several classification algorithms and assessed their results: First, a **Random Forest Classifier** was trained with default parameters, which, despite showing high overall accuracy (approx. 54.55%), completely failed to predict even a single instance of the minority class (Class 1) (F1-score: 0.00). To address this issue, the **Random Forest** model was retrained with the class\_weight='balanced' parameter, which led to a slight reduction in overall accuracy (approx. 50.35%) but enabled the model to identify Class 1 with an F1-score of approximately 0.16. Subsequently, **Hyperparameter Tuning** on the Random Forest using GridSearchCV did not yield significant improvements in overall performance or for Class 1 (accuracy around 48.95% and Class 1 F1-score around 0.14). Moving on to more advanced algorithms, we trained a **LightGBM Classifier** with class\_weight='balanced', which, although significantly reducing the overall model accuracy (approx. 38.46%), showed a marked improvement in identifying the minority Class 1 (F1-score approx. 0.21). Further **Hyperparameter Tuning of LightGBM** allowed us to increase the F1-score for Class 1 to 0.28, representing the best performance for this class, but the overall model accuracy remained around 38.46%. Finally, we experimented with a **Logistic Regression** model with class\_weight='balanced', which yielded very balanced results: an overall accuracy of 51.75% and an F1-score of 0.24 for Class 1, along with the best weighted average F1-score (0.51) among all tested models. The **final conclusion** was that while LightGBM offered the best performance purely for identifying the minority class, the Logistic Regression model, due to its excellent balance between overall accuracy and acceptable ability to predict all classes, including the minority class, was selected as the most suitable overall option for this project.