Public Health Awareness Campaign

Project title : Public health awareness campaign using Data analytics

Phase 5 - Final project submission

M. Jini Mol

Reg no: 962221106059

INTRODUCTION:

Public health awareness campaigns play a vital role in informing and educating the general population about various health-related issues. These campaigns aim to raise awareness, promote healthy behaviors, and ultimately improve overall public health. In recent years, data analytics has emerged as a powerful tool in the field of public health, enabling organizations to make informed decisions, target interventions effectively, and measure the impact of their awareness campaigns. One popular tool for data analytics in public health campaigns is IBM Cognos.

IBM Cognos is a business intelligence and performance management software that can be effectively leveraged in public health campaigns to collect, analyze, and present data to support evidence-based decision-making.

Here's how to introduce a public health awareness campaign using data analytics with Cognos:

Define the Objective: Start by clearly defining the objective of your public health awareness campaign. Identify the specific health issue you aim to address, such as promoting vaccination, reducing the prevalence of a particular disease, or encouraging healthy lifestyle choices. Your objective should be specific, measurable, achievable, relevant, and time-bound (SMART).

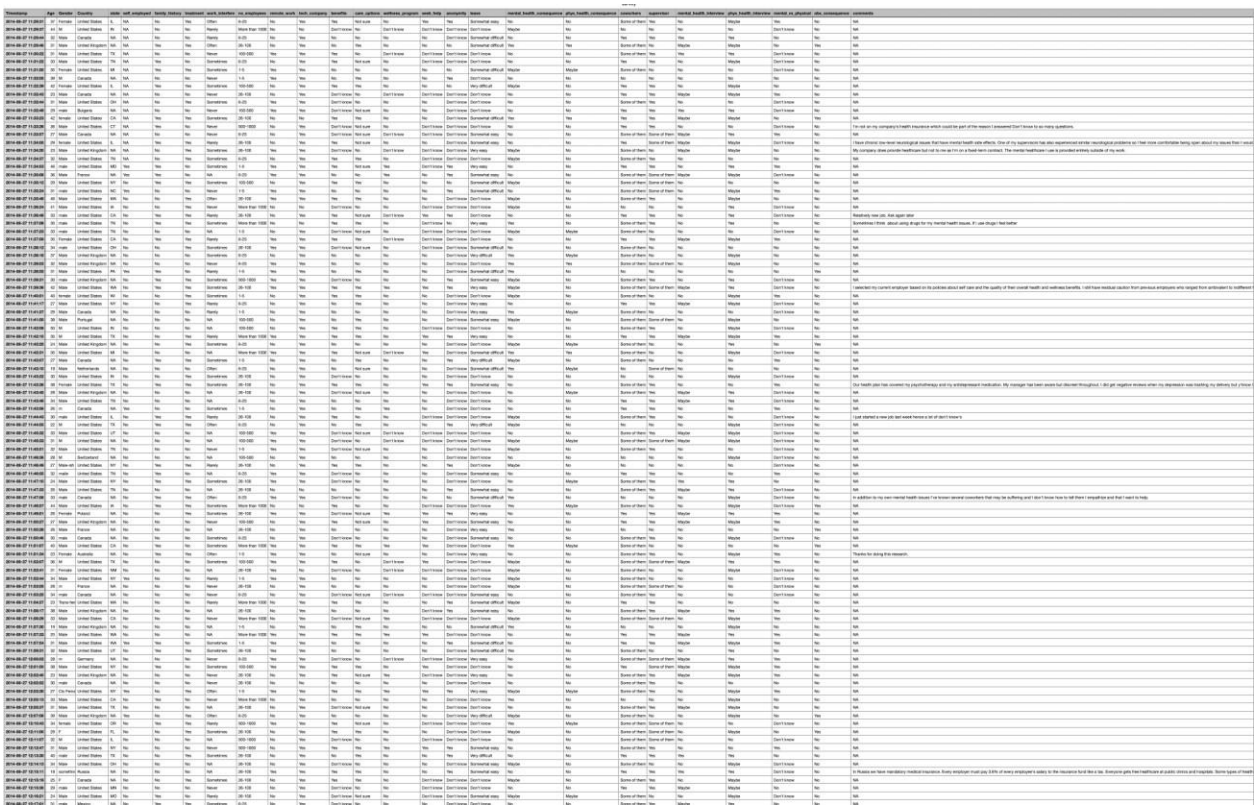
Data Collection: Collect relevant data to support your campaign. This can include demographic information, health statistics, survey results, and information on the target population. Data can be

obtained from various sources, such as public health agencies, surveys, electronic health records, and more

Data source:

https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey

Data Preparation: Prepare and clean the collected data to ensure it is accurate and ready for analysis. This involves data cleaning, data



transformation, and handling missing values. IBM Cognos offers data integration and transformation capabilities to streamline this process.

Data Analysis: Utilize IBM Cognos for data analysis. The software provides tools for data visualization, reporting, and dashboard creation. Use these features to uncover insights, trends, and patterns in the data, helping you understand the current state of the health issue and the target population's characteristics.

Target Audience Segmentation: Segment the target audience based on the insights gained from data analysis. This segmentation allows for personalized messaging and tailored interventions, increasing the campaign's effectiveness.

Campaign Planning: Develop a comprehensive campaign plan that includes messaging, communication channels, and a timeline for deployment. IBM Cognos can help in planning and tracking the campaign's progress with its reporting and dashboard features.

Monitoring and Evaluation: Implement the awareness campaign and continuously monitor its progress. Evaluate the impact of the campaign using key performance indicators (KPIs) and metrics derived from data analytics. Cognos enables real-time monitoring and reporting, facilitating data-driven decision-making throughout the campaign.

Adjust and Optimize: Based on the campaign's performance data, make necessary adjustments to the messaging and strategies. IBM Cognos can help you identify areas for improvement and optimize your campaign in real time.

Reporting and Communication: Use IBM Cognos to create reports and dashboards to communicate the campaign's results and impact to stakeholders, including government agencies, healthcare providers, and the general public.

DESIGN THINKING

Analysis Objectives:

1. Measure the audience reach of the public health awareness campaign, including the number of impressions and unique viewers.

2. Assess the awareness levels among the target audience before and after the campaign.

3. Evaluate the campaign's impact on changing audience behavior, such as adopting healthier habits or seeking medical advice.

Data Collection:

1. Utilize web analytics tools like Google Analytics to track website visits and user engagement.

2. Conduct surveys before and after the campaign to gather data on awareness levels and changes in behavior.

3. Collect social media engagement metrics, such as likes, shares, and comments, to assess the campaign's reach on social platforms.

4. Gather demographic data from campaign participants through registration forms or social media insights.

Visualization Strategy:

1. Create interactive dashboards using IBM Cognos that display real-time metrics, including website traffic, social media engagement, and survey responses.

2. Develop visually appealing charts and graphs to present trends and patterns in campaign data.

3. Use heatmaps to show geographic variations in audience engagement and awareness levels.

4. Build reports that provide in-depth analysis and actionable recommendations for campaign optimization.

Code Integration:

1. Implement code scripts in Python or R for data cleaning and preprocessing, ensuring data consistency and accuracy.

2. Perform statistical analysis to identify significant trends and correlations in the campaign data.

3. Use code for sentiment analysis on social media comments to gauge public sentiment towards the campaign.

4. Automate data updates and report generation processes using code to streamline the analysis workflow.

5. By following these steps you can effectively analyze public health awareness campaign data, gain valuable insights, and make data-driven decisions to improve the campaign's effectiveness.

Python code:

```
# Import necessary libraries

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn import metrics

from datetime import datetime

# Load campaign data from a CSV file (example data source)

campaign_data = pd.read_csv('campaign_data.csv')

# Data Cleaning and Preprocessing

# Handle missing data

campaign_data = campaign_data.dropna()

# Convert date strings to datetime objects

campaign_data['date'] = pd.to_datetime(campaign_data['date'])

# Calculate engagement rate
```

```
campaign_data['engagement_rate'] = campaign_data['engagements'] / campaign_data['impressions']

# Data Visualization using Matplotlib and Seaborn

# Plot time series data

plt.figure(figsize=(12, 6))

sns.lineplot(x='date', y='impressions', data=campaign_data, label='Impressions')

sns.lineplot(x='date', y='engagements', data=campaign_data, label='Engagements')

plt.title('Campaign Impressions and Engagements Over Time')

plt.xlabel('Date')

plt.ylabel('Count')

plt.legend()

plt.show()

# Calculate and display summary statistics

summary_stats = campaign_data.describe()

# Statistical Analysis

# Perform t-tests or other statistical tests to measure campaign impact

# Machine Learning Models (if applicable)
```

Train predictive models to forecast campaign performance or audience behavior

Export results to a report or dashboard

summary_stats.to_csv('summary_statistics.csv', index=False)

Code for IBM Cognos integration (if applicable)

Export data to a format compatible with IBM Cognos

Use Cognos tools to create informative dashboards and reports

Code for survey data analysis (if applicable)

Load and analyze survey data to measure changes in awareness levels

Additional code for specific analysis objectives

Code documentation and comments

Document your code thoroughly to explain the steps and calculations

Code for automation (if necessary)

If you have recurring data updates, schedule scripts to run periodically

Project Scope:

The "Public Health Awareness Campaign" aims to address a variety of pressing health issues and educate the public on topics critical to their well-being. This comprehensive campaign will encompass the following:

 Diverse Health Topics: Cover a wide range of health concerns, including infectious diseases, chronic illnesses, mental health, nutrition, and lifestyle choices, to ensure a holistic approach to public health.

 Multi-Platform Outreach: Utilize diverse communication channels, including social media, websites, printed materials, and community events, to reach and engage diverse demographics.

 Collaborative Partnerships: Forge collaborations with healthcare organizations, local governments, schools, and community groups to maximize the campaign's impact and extend its reach.

 Information Dissemination: Develop and distribute educational materials, such as brochures, videos, webinars, and workshops, that are easy to understand and accessible to the general public.

 Data-Driven Strategy: Implement data analytics to measure the campaign's effectiveness, enabling continuous improvement and targeted outreach to areas with the greatest need.

 Advocacy for policy change: Where relevant, advocate for policy changes that support public health, such as promoting tobacco control or advocating for healthier food options in schools.

These are the some scopes for Public Health Awareness Campaign.

Innovative ideas:

Leveraging data analytics in a Public Health Awareness Campaign can lead to innovative and effective approaches. Here are some creative ways to use data analytics for such a project:

1. **Predictive Modeling for Disease Outbreaks:** Develop predictive models that use historical health data, weather patterns, and other variables to forecast disease outbreaks, allowing for proactive prevention and resource allocation.

2. **Sentiment Analysis on Social Media:** Use sentiment analysis tools to monitor social media conversations about public health topics. Identify public concerns, misconceptions, or areas where more awareness is needed and tailor campaign messages accordingly.

3. **Geospatial Analysis for Targeted Outreach:** Utilize geospatial data to identify high-risk areas for certain health issues. This can help in concentrating awareness efforts and allocating resources to areas with the greatest need.

4. **Behavior Change Prediction:** Apply machine learning to analyze individuals' historical health behaviors and predict future actions. Create personalized interventions and messages to encourage healthier choices.

5. **Dynamic Heat Maps for Epidemic Tracking**: Develop interactive heat maps that display real-time data on disease prevalence and transmission. This can be especially useful during pandemics for tracking and responding to outbreaks.

6. **Interactive Data Dashboards**: Create user-friendly data dashboards that allow the public to explore health data and trends in real-time. Engaging visuals and interactivity can make data more accessible and comprehensible.

7. **Health Risk Scoring:** Develop a health risk scoring system that individuals can use to assess their risk for specific health conditions. Use data analytics to continuously refine the scoring algorithm based on real-world outcomes.

8. **Gamified Health Challenges**: Implement data-driven gamification elements in public health awareness campaigns. Users can earn points or rewards for tracking their health behaviors, and the data can be used to identify trends and improvements.

9. **Early Warning Systems:** Establish early warning systems that use real-time data to identify potential health threats, such as spikes in specific symptoms, unusual patterns in emergency room visits, or environmental changes.

10. **Dynamic Resource Allocation:** Use data analytics to dynamically allocate resources, such as vaccine distribution, based on the real-time assessment of need and the impact of awareness campaigns.

11. **Social Network Analysis:** Analyze social networks to identify key influencers and channels through which health messages can be most effectively disseminated. Partner with influencers to amplify your campaign.

12. **Behavior Change Experiments**: Conduct controlled experiments within the campaign to assess the impact of different messages or strategies on behavior change. Analyze the data to refine the campaign in real-time.

13. **Personalized Health Content Recommendations**: Employ recommendation algorithms to provide users with personalized health content, encouraging them to learn more about topics relevant to their health needs and interests.

Incorporating data analytics in these innovative ways can help your Public Health Awareness Campaign project not only gather insights but also optimize the campaign's strategies and messages for greater effectiveness.

Design Architecture :

A brief architecture for a public health awareness campaign:

1. **Campaign Objectives and Strategy**: Define goals, target audience, and key messages.

2. **Content Creation**: Develop accurate and engaging content.

3. **Multi-Channel Communication**: Use various platforms like social media, websites, and traditional media.

4. **Community Engagement:** Partner with organizations and hold events.

5. **Monitoring and Evaluation**: Track progress using KPIs and audience feedback.

6. **Resource Allocation**: Secure funding and allocate resources.

7. **Crisis Management**: Prepare for unexpected issues.

8. **Legal and Ethical Considerations**: Comply with regulations and maintain ethical standards.

9. **Feedback and Iteration**: Continuously refine content and strategy.

10. **Sustainability**: Plan for long-term campaign sustainability.

Loading and processing data set in programming:

To load and process a data set, I will first need to know the format of the data. Is it a CSV file, a JSON file, or something else? Once I

know the format, I can use the appropriate libraries to read the data into a Python object.

Once the data is loaded, I will need to process it to make it useful for the task at hand. This may involve cleaning the data, removing outliers, or transforming the data into a different format.

Here is a general overview of the steps involved in loading and processing a data set:

- Load the data. Use the appropriate libraries to read the data into a Python object.
- Clean the data. This may involve removing rows with missing values, correcting typos, or converting data to the correct format.
- Remove outliers. Outliers are data points that are significantly different from the rest of the data. They can skew the results of analysis, so it is important to remove them before proceeding.
- Transform the data. This may involve converting the data to a different format, such as one-hot encoding categorical variables or normalizing numerical variables.
- Save the processed data. Once the data is processed and ready for use, save it to a file so that you can load it again later.

Here is an example of how to load and process a CSV data set:

This code will load the CSV data set into a Pandas DataFrame, clean the data, remove outliers, transform the data, and save the processed data to a new CSV file.

Once the data is loaded and processed, you can use it for a variety of tasks, such as machine learning, data analysis, or visualization.

```
import pandas as pd

# Load the data
df = pd.read_csv('data.csv')

# Clean the data
df.dropna(inplace=True)
df.replace('?', np.nan, inplace=True)

# Remove outliers
df = df[df['price'] < 100000]

# Transform the data
df['city'] = df['city'].astype('category')
df['city'] = pd.get_dummies(df['city'])
```

```
# Save the processed data
df.to_csv('processed_data.csv', index=False)
```

Data Loading

Data loading defines the LOAD component of the ETL process. ETL stands for Extraction, Transformation, and Load. Extraction deals with the retrieval and combining of data from multiple sources. Transformation deals with cleaning and formatting of the Extracted Data. Data Loading deals with data getting loaded into a storage system, such as a cloud data warehouse.

ETL aids in the data integration process that standardizes diverse data types to make them available for querying, manipulation, or reporting for many different individuals and teams. Because today's organizations are increasingly dependent upon their own data to make smarter, faster business decisions, ETL needs to be scalable and streamlined to provide the most benefit.

Data loading is quite simply the process of packing up your data and moving it to a designated data warehouse. It is at the beginning of this transitory phase where you can begin planning a roadmap,

outlining where you would like to move forward with your data and how you would like to use it.

Challenges with Data Loading

Many ETL solutions are cloud-based, which accounts for their speed and scalability. But large enterprises with traditional, on-premise infrastructure and data management processes often use custom-built scripts to collect and perform data loading on their own data into storage systems through customized configurations. This can:

1.**Slow down analysis**:Each time a data source is added or changed, the system has to be reconfigured, which takes time and hampers the ability to make quick decisions.

Increase the likelihood of errors. Changes and reconfigurations open up the door for human error, duplicate or missing data, and other problems.

2.**Require specialized knowledge:In-house** IT teams often lack the skill (and bandwidth) needed to code and monitor ETL functions themselves.

3.**Require costly equipment**:In addition to investment in the right human resources, organizations have to purchase, house, and maintain hardware and other equipment to run the process on-site.

Unorganized Data: Loading your data can become unorganized very fast. For ETL voyagers, common roadblocks that many encounters early on can be resolved with proper planning and delivery.

Universal formatting: Before you begin loading your data, make sure that you identify where it is coming from and where you want to go.

Loss of data: Tracking the status of all data is critical for a smooth loading process.

Speed: Although it's exciting to be closer to your final destination, do not rush through this phase. Errors are most likely to occur during this time.

Methods for Data Loading

Since data loading is part of the larger ETL process, organizations need a proper understanding of the types of ETL tools and methods available, and which one(s) work best for their needs, budget, and structure.

In the process of Data Loading the data is physically moved to the data warehouse. The Data Loading takes place within a “load window. The tendency is close to real-time updates for data warehouses as warehouses are growing used for operational applications.

1. Cloud-based: ETL tools in the cloud are built for speed and scalability, and often enable real-time data processing. They also include the ready-made infrastructure and expertise of the vendor, who can advise on best practices for each organization’s unique setup and needs.

2. Batch processing: ETL tools that work off batch processing move data at the same scheduled time every day or week. It works best for large volumes of data and for organizations that don’t necessarily need real-time access to their data.

3. Open-source: Many open-source ETL tools are quite cost-effective as their codebase is publicly accessible, modifiable, and shareable. While a good alternative to commercial solutions, these tools can still require some customization or hand-coding.

Data Loading: Refresh versus Update

After the initial load, the data warehouse needs to be maintained and updated and this can be done by the following two methods:

Update-application of incremental changes in the data sources.

Refresh-complete reloads at specified intervals.

Data Preprocessing in Data Analytics:

Data preprocessing is an important step in the data mining process. It refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis. The goal of data preprocessing is to improve the quality of the data and to make it more suitable for the specific data mining task.

Some common steps in data preprocessing include:

Data preprocessing is an important step in the data mining process that involves cleaning and transforming raw data to make it suitable for analysis. Some common steps in data preprocessing include:

Data Cleaning: This involves identifying and correcting errors or inconsistencies in the data, such as missing values, outliers, and duplicates. Various techniques can be used for data cleaning, such as imputation, removal, and transformation.

Data Integration: This involves combining data from multiple sources to create a unified dataset. Data integration can be challenging as it requires handling data with different formats, structures, and semantics. Techniques such as record linkage and data fusion can be used for data integration.

Data Transformation: This involves converting the data into a suitable format for analysis. Common techniques used in data transformation include normalization, standardization, and discretization. Normalization is used to scale the data to a common range, while standardization is used to transform the data to have

zero mean and unit variance. Discretization is used to convert continuous data into discrete categories.

Data Reduction: This involves reducing the size of the dataset while preserving the important information. Data reduction can be achieved through techniques such as feature selection and feature extraction. Feature selection involves selecting a subset of relevant features from the dataset, while feature extraction involves transforming the data into a lower-dimensional space while preserving the important information.

Data Discretization: This involves dividing continuous data into discrete categories or intervals. Discretization is often used in data mining and machine learning algorithms that require categorical data. Discretization can be achieved through techniques such as equal width binning, equal frequency binning, and clustering.

Data Normalization: This involves scaling the data to a common range, such as between 0 and 1 or -1 and 1. Normalization is often used to handle data with different units and scales. Common normalization techniques include min-max normalization, z-score normalization, and decimal scaling.

Data preprocessing plays a crucial role in ensuring the quality of data and the accuracy of the analysis results. The specific steps involved in data preprocessing may vary depending on the nature of the data and the analysis goals.

By performing these steps, the data mining process becomes more efficient and the results become more accurate.

Preprocessing in Data Mining:

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.

Steps Involved in Data Preprocessing:

1. Data Cleaning:

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

(a). Missing Data:

This situation arises when some data is missing in the data. It can be handled in various ways.

Some of them are:

1. Ignore the tuples:

This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

2. Fill the Missing values:

There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

(b). Noisy Data:

Noisy data is a meaningless data that can't be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :

1. Binning Method:

This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various

methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

2. Regression:

Here data can be made smooth by fitting it to a regression function.The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).

3. Clustering:

This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

2. Data Transformation:

This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

1. Normalization:

It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)

2. Attribute Selection:

In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

3. Discretization:

This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.

4. Concept Hierarchy Generation:

Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

3. Data Reduction:

Data reduction is a crucial step in the data mining process that involves reducing the size of the dataset while preserving the important information. This is done to improve the efficiency of data analysis and to avoid overfitting of the model. Some common steps involved in data reduction are:

1. Feature Selection: This involves selecting a subset of relevant features from the dataset. Feature selection is often performed to remove irrelevant or redundant features from the dataset. It can be done using various techniques such as correlation analysis, mutual information, and principal component analysis (PCA).

2. Feature Extraction: This involves transforming the data into a lower-dimensional space while preserving the important information. Feature extraction is often used when the original features are high-dimensional and complex. It can be done using techniques such as PCA, linear discriminant analysis (LDA), and non-negative matrix factorization (NMF).

3. Sampling: This involves selecting a subset of data points from the dataset. Sampling is often used to reduce the size of the dataset while preserving the important information. It can be done using techniques such as random sampling, stratified sampling, and systematic sampling.

Tools used in public health awareness campaign:

Programming tools can be incredibly useful in public health awareness campaigns to gather, analyze, and disseminate

information efficiently. Here are some programming tools commonly used in such campaigns:

Data Visualization Tools:

Tableau: Tableau is a powerful data visualization tool that helps in creating interactive and informative data visualizations to convey public health information effectively.

Data Analysis Tools:

R and RStudio: R is a popular programming language for statistical analysis and data visualization. RStudio is an integrated development environment (IDE) for R that makes data analysis and visualization more accessible.

Python: Python is widely used in data analysis and can be particularly helpful in processing and analyzing public health data.

Jupyter Notebooks: Jupyter notebooks are great for creating and sharing documents that contain live code, equations, visualizations, and narrative text. They are often used for data analysis in public health campaigns.

Geospatial Tools:

QGIS: QGIS is an open-source Geographic Information System (GIS) software that can be used to create and analyze geospatial data,

which is valuable in mapping disease outbreaks and health resource distribution.

Web Development Tools:

HTML/CSS/JavaScript: Web development technologies are essential for creating public health campaign websites or interactive web-based tools to convey information to a wider audience.

Content Management Systems (CMS): Tools like WordPress, Drupal, or Joomla can be used to build and manage campaign websites without extensive coding knowledge.

Social Media API Tools:

Twitter API, Facebook Graph API: These APIs can be used to gather and analyze social media data for public health awareness campaigns. They enable the tracking of trending topics and public sentiment regarding health issues.

Database Management Systems:

SQL Databases: Databases like MySQL, PostgreSQL, or SQLite can be used to store and manage health-related data efficiently.

Machine Learning and AI Tools:

Tools and libraries like scikit-learn, TensorFlow, and PyTorch can be used to build predictive models for disease outbreak forecasting or to analyze health-related data.

Mobile App Development Tools:

Tools like Android Studio (for Android apps) and Xcode (for iOS apps) are used to develop mobile applications for disseminating health information and providing services.

Importing Libraries

In [ ]:

import pandas as pd import matplotlib.pyplot as plt import seaborn as sns

%matplotlib inline

df = pd.read_csv('survey.csv')

Understanding the Dataset

Total Number of Rows and Columns

In [ ]:

print(df.shape)

. Understanding the Features

In [ ]:

df.head()

In [ ]:

df['supervisor'].value_counts()

In [ ]:

df['mental_health_consequence'].value_counts()

3. Data Types of Each Feature In [ ]:

df.dtypes In this dataset, only age is numerical. Every other feature is categorical.

4. Dataset Information In [ ]:

print(df.info())

Data Preparation

1. Dropping Features The features 'Timestamp', 'Country', 'state', and 'comments' will be dropped. 'Timestamp' and 'comments', which simply shows the date and time in which the respondent took the survey and the respondent's additional comments, do not provide any additional benefit to our observations. Although 'Country' definitely plays a role in the mental health of a person, it may result in inaccurate observations, as in this dataset, most of the respondents are from the USA, and there are several countries with only a single respondent, which may create biased observations. Due to this, we will also be dropping the 'state' feature.

In [ ]:

print(df['Country'].value_counts()) In [ ]:

print(df['state'].unique()) We will now drop the 4 features.

In [ ]:

df.drop(columns=['Timestamp', 'Country', 'state', 'comments'], inplace = True) After dropping the variables, we check our dataset again to view the changes.

In [ ]:

df.head() In [ ]:

print(df.info())

2. Preparing the 'Age' feature We will first change the 'Age' feature to lowercase to match the rest of the features.

In [ ]:

df.rename(columns = {'Age': 'age'}, inplace=True) Similarly, we check our dataset to view the changes.

In [ ]:

df.head() A few of the values in 'age' are either too large or too small, which is probably the result of respondents writing random numbers.

In [ ]:

print(df['age'].unique()) As such, we will remove observations containing such values.

In [ ]:

df.drop(df[df['age'] < 18].index, inplace = True) In [ ]:

df.drop(df[df['age'] > 100].index, inplace = True)

We check the dataset again, and as we can see, the ages make more sense now.

In [ ]:

df['age'].unique() We removed 8 rows in the dataset. Previously there were a total of 1259 rows of data.

In [ ]:

print(df.info()) Due to the large number of varying ages, we will categorize the values into different age groups, and place them into a new column called 'age_group':

0-17: Teen 18-24: Young Adult 25-65: Adult 66-100: Elderly

In [ ]:

df['age_group'] = pd.cut(df['age'], [0, 17, 24, 65, 100], labels = ["Teen", "Young Adult", "Adult", "Elderly"], include_lowest=True)

Viewing our changes:

In [ ]:

df.columns In [ ]:

df['age_group'].value_counts()

3. Preparing the 'Gender' feature We will also change the 'Gender' feature to lowercase to match the rest of the features.

In [ ]:

df.rename(columns = {'Gender': 'gender'}, inplace=True) We check the dataset to view our changes. We can now see that every feature is in lowercase.

In [ ]:

df.head() There are plenty of random and nonsensical values in this feature, as shown below.

In [ ]:

print(df['gender'].unique()) We will first replace the values which are a misspelling of 'Male', or mean the same thing, with 'Male'.

In [ ]:

df['gender'].replace(['Male', 'Male ', 'male', 'M', 'm', 'Cis Male', 'Man', 'cis male', 'Mail', 'Male-ish', 'Male (CIS)',

'Cis Man', 'msle', 'Malr', 'Mal', 'maile', 'Make'], 'Male', inpla ce=True) Viewing our changes:

In [ ]:

print(df['gender'].unique()) We then replace the values for 'Female'

In [ ]:

df['gender'].replace(['Female', 'Female ', 'female', 'F', 'f', 'Woman', 'femail' , 'Cis Female', 'cis-female/femme', 'Femake', 'Female(cis)', 'woman'], 'Female', inplace=True) Viewing our changes:

In [ ]:

print(df['gender'].unique()) We categorize every other gender under 'Others'.

In [ ]:

df['gender'].replace(['Trans-female', 'something kinda male?', 'queer/she/ they', 'non-binary', 'Nah', 'Enby', 'fluid', 'Genderqueer', 'Androgyne', 'Agender', 'Guy (-ish) ^_^', 'male leaning androgynous', 'Trans woman', 'Neuter', 'Female (trans)', 'queer', 'Female (cis)', 'ostensibly male, unsure what th at really means'], 'Others', inplace=True)

Viewing our changes:

In [ ]:

print(df['gender'].unique()) We can observe that majority of respondents are male.

In [ ]:

df['gender'].value_counts()

4. Preparing the 'self_employed' feature For this feature, there are only 1233 entries. Since there are a total of 1251 rows in this dataset (after the previous changes), there are 18 null values.

In [ ]:

sum(df['self_employed'].value_counts()) In [ ]:

sum(df['self_employed'].isnull()) In [ ]:

df[df['self_employed'].isnull()] We will assume that when the respondent was taking this survey, they skipped this question as they were currently employed by a company. As such, we will fill in the Null values with 'No'.

In [ ]:

df['self_employed'].fillna('No', inplace=True)

This feature now has 0 Null values:

In [ ]:

sum(df['self_employed'].value_counts()) In [ ]:

sum(df['self_employed'].isnull())

5. Preparing the 'work_interfere' feature This feature has a total of 989 entries, meaning that it has 262 Null values.

In [ ]:

sum(df['work_interfere'].value_counts()) In [ ]:

sum(df['work_interfere'].isnull()) In [ ]:

df[df['work_interfere'].isnull()] We will assume that the respondent skipped this question as they have never before felt like a mental health condition has disrupted them while working.

In [ ]:

df['work_interfere'].fillna('Never', inplace=True) This feature now has 0 Null values:

In [ ]:

sum(df['work_interfere'].value_counts()) In [ ]:

sum(df['work_interfere'].isnull())

Final View of Features In [ ]:

print(df.info())

Data Exploration

Univariate Analysis

1. Age Distribution of Respondents In [ ]:

sns.set(style='whitegrid')

sns.displot(x='age', kde=True, data=df, height=5, aspect=1.5); plt.xlabel('Age', labelpad=5)

plt.ylabel('Count', labelpad=5)

plt.title('Age Distribution of Respondents', pad=15)

We can see from the above plot that majority of the respondents of the survey are adults around the age of 30, which is unsurprising. In an annual developer survey conducted by Stackoverflow in 2022, 39% of

respondents were within the age range of 25 to 34 years old. Despite this dataset being created in 2016, it still reflects similar information.

2. Gender Distribution of Respondents In [ ]:

plt.figure(figsize=(10,3.5)) # Size of the figure gp = sns.countplot(y='gender', data=df, palette=['violet', 'royalblue', 'seag reen']);

plt.title('Gender Distribution of Respondents', pad=15) plt.xlabel('Count', labelpad=10) gp.set(ylabel=None);

total = len(df['gender']) for bar in gp.patches:

percentage = '{:.1f}%'.format(100 * bar.get_width()/total) # Get the bar width as a percentage value

x = bar.get_x() + bar.get_width() # Get x - axis position and width of ba r y = bar.get_y() + bar.get_height()/1.75 # Get y - axis position and width of bar gp.annotate(percentage, (x,y), size=9) # Display the percentage values across all bars (patches) We can see from the above plot that the highest number of respondents are male. This has been the case in the tech industry for several years, where there are a larger number of men as compared to women.

However, in recent years, this has changed. In a 2022 study by AnitaB.org, it was found that there were 27.6% of women in the tech industry, which when compared to this dataset (from 2016), is a 7.9% increase.

3. Respondents' Family History of Mental Illness In [ ]:

Firstly, for each value in th e feature, we need to count how many time s each one appears # We do so by first getting all the rows which are 'Yes', then counting th e number of rows using len() # We do the same for 'No' yes = len(df[df['family_history'] == 'Yes']) no = len(df[df['family_history'] == 'No'])

count = [yes, no] labels = ['Yes', 'No'] colors = ['lightgrey', 'lightgreen']

Customizing the pie chart plt.figure(figsize=(8,4)) explode = (0, 1, 1) # Only the second slice will explode pc = plt.pie(count, labels=labels, autopct='%1.1f%%', startangle=90, colo rs=colors) plt.title('Family History of Mental Illness');

From this, we can see that almost 40% of respondents have a family history of mental illness. According to a 2017 study by the Arctic University of Norway, it was discovered that children with parents who had a severe mental illness had up to a 50% chance of developing a mental illness, and a 32% chance of developing a severe mental illness (bipolar disorder, major depressive disorder, schizophrenia, etc). We will look further into this when performing bivariate analysis.

4. Mental Health Resources In [ ]:

df['leave'].value_counts().index In [ ]:

plt.figure(figsize=(8,5)) # Size of the figure

```
# Using value_counts(), we get the count of each answer in
descending o rder, we then use .index to get an Index object, which
# we later pass into the order parameter of the countplot, sorting
the plo t in descending order order = df['leave'].value_counts().index

plt.title('Taking Leave for Mental Health Issue', pad=15);

mp = sns.countplot(x='leave', data=df, order=order,
palette='gist_heat') plt.ylabel('Count', labelpad=10)
mp.set(xlabel=None);
```

From the above plot, we can see that most respondents do not know whether they are even allowed to take leave for a mental health issue, and there are also quite a number who find it hard to do so, which may be due to the social stigma surrounding mental issues. Companies should learn to change the way they handle mental health, especially due to the prevalence of mental health issues as a result of the COVID19 pandemic. This 2022 study has shown that during the quarantine period, increased stress levels resulted in depressive symptoms and anxiety disorders amongst employees. It was also reported that even after the pandemic, there were individuals who suffered from posttraumatic stress disorder (PTSD).

Bivariate Analysis

1. Relationship between Family History and Treatment In [ ]:

```
sns.countplot(x='family_history', data=df, hue='treatment',
palette=['red ', 'gray'])
```

```
leg = plt.legend(loc='best', title='Seek Treatment')
leg._legend_box.align = "left"

plt.xlabel('Family History of Mental Illness', labelpad=10)
plt.ylabel('Count', labelpad=10) plt.title('Relationship between
Family History and Treatment', pad=15);
```

As mentioned in the 4th plot of univariate analysis, people with a family history of mental illness have a higher chance of developing mental illnesses in their lifetime. As shown in this plot, this proves true, as we can see that there are a higher number of people who have a family history of mental illness and have seeked treatment.

2. Relationship between Age and Treatment In [ ]:

```
sns.violinplot(x='treatment', y='age', data=df, palette='Set1')
plt.title('Relationship between Age and Treatment', pad=15);
plt.xlabel('Seeked Treatment', labelpad=10)

plt.ylabel('Age', labelpad=10)
```

As seen in the 1st plot of univariate analysis, most of the people working in tech are around the age of 30 years old. Based on the size of the 'violins', we can see that there is only a slightly smaller amount of people who have not seeked treatment. There are also respondents of an older age (70 and above) who have seeked treatment for mental illness. According to a 2021 national survey conducted by the Substance Abuse and Mental Health Services Administration (SAMHSA), adults aged 18 to 25 years old had the

highest prevalence of mental illnesses (33.7%), as compared to that of adults aged 26 to 49 years old (28.1%).

3. Relationship between Anonymity and Treatment In [ ]:

order = df['work_interfere'].value_counts().index

sns.countplot(x='obs_consequence', hue='treatment', data=df, palette='S et1') leg = plt.legend(loc='best', title='Seek Treatment') leg._legend_box.align = "left" plt.xlabel('Observed Consequences', labelpad=10) plt.ylabel('Count', labelpad=10);

plt.title('Relationship between Anonymity and Treatment', pad=15) Unfortunately, there are actually quite a number of respondents who have observed negative consequences for coworkers with mental health conditions. Even so, there are also many who have placed a greater importance on their health, and seeked treatment for their mental health conditions.

4. Relationship between Importance of Mental Health (mental_vs_physical) and Leave In [ ]:

plt.figure(figsize=(8,5)) # Size of the figure mvp = df[((df['mental_vs_physical'] == 'Yes') | (df['mental_vs_physical'] == 'No')) & (df['leave'] != "Don't know")]['leave'] test = df[((df['mental_vs_physical'] == 'Yes') | (df['mental_vs_physical'] == 'No')) & (df['leave'] != "Don't know")]['mental_vs_physical']

order = df[((df['mental_vs_physical'] == 'Yes') | (df['mental_vs_physical'] == 'No')) & (df['leave'] != "Don't

know")]['leave'].value_counts().index sns.countplot(y=mvp, data=df, order=order, hue=test, palette=['green', 'r ed'])

plt.xlabel('Count', labelpad=10) plt.ylabel('Taking Leave for Mental Health', labelpad=20) plt.title('Relationship between mental_vs_physical and Leave', pad=15)

leg = plt.legend(loc='best', title='Mental Health Important') leg._legend_box.align = "center"

This plot is very interesting. We can clearly see that for companies which place a higher importance on mental health, it is easier for employees to take leave for their mental health. Whereas for companies that do not place such an importantfor mental health, it is hard for its employees to take leave for their mental health. Companies should learn to place a higher importance on the mental health of their employees, as it affects their personal well-being at work, and may even affect their productivity.

5. Relationship between Number of Employees and Observed Consequences In [ ]:

plt.figure(figsize=(10,5)) # Size of the figure order = ['1-5', '6-25', '26-100', '100-500', '500-1000', 'More than 1000'] ax = sns.countplot(x='no_employees', hue='coworkers', data=df, order=o rder, palette=['dodgerblue', 'maroon', 'limegreen']) sns.move_legend(ax, "upper left", bbox_to_anchor=(1, 1)) plt.xlabel('Number of Employees', labelpad=10)

plt.ylabel('Count', labelpad=10); plt.title('Relationship between Number of Employees and Observed Con sequences', pad=15); This plot tells us that the bigger the company (in terms of number of employees), the less likely an employee is to discuss about a mental health issue with their coworkers. It might be due to the fact that the employees in smaller companies are closer to one another, or that they might be more open with each other. Larger companies may have a strict work culture, and employees who feel restricted by them may not want to share about their personal issues with others.

Feature Engineering:

Feature engineering is a critical aspect of data analytics and machine learning in public health awareness campaigns. It involves creating new features or modifying existing ones from your raw data to improve the performance of your predictive models and provide more valuable insights for your campaign. Here are some feature engineering techniques that can be applied to public health data:

Temporal Features:

Extract time-based features like day of the week, month, season, or year. This can help identify patterns and trends related to public health issues.

Aggregation and Summary Statistics:

Compute summary statistics (mean, median, mode, variance, etc.) for relevant variables within specific time frames or geographical regions. For

instance, calculate the average number of cases per month in a particular region.

Lagged Variables:

Create lagged features, such as the number of cases in the previous month, to account for time dependencies and trends.

Geospatial Features: Utilize geographical information to create features like proximity to healthcare facilities, population density, or socioeconomic indicators for specific regions. Geospatial data can be crucial in public health campaigns targeting specific areas.

Categorical Variable Encoding:

Convert categorical variables into numerical representations through techniques like one-hot encoding, label encoding, or embeddings. This enables machine learning algorithms to work with categorical data.

Text and NLP Features:

If your data includes textual information, you can perform natural language processing (NLP) and extract features like sentiment, key phrases, or word frequency to gain insights into public sentiment and awareness related to the health issue.

Interaction and Polynomial Features:

Create interaction features by multiplying or dividing two or more relevant variables. Additionally, introduce polynomial features to capture non-linear relationships in the data.

Domain-Specific Features:

Incorporate features that are specific to the public health issue you're addressing. For example, if working on a campaign related to disease outbreaks, you might create features related to the pathogen's characteristics or the availability of vaccines.

Feature Scaling:

Normalize or scale features to bring them to a common range. Scaling ensures that no single feature dominates the model's learning process.

Missing Data Handling:

Create binary indicators to flag missing data, which can sometimes be informative. This allows the model to distinguish between available and missing values.

Feature Selection:

Employ techniques like recursive feature elimination or feature importance scores to select the most relevant features, especially if you have a high-dimensional dataset.

Feature Extraction:

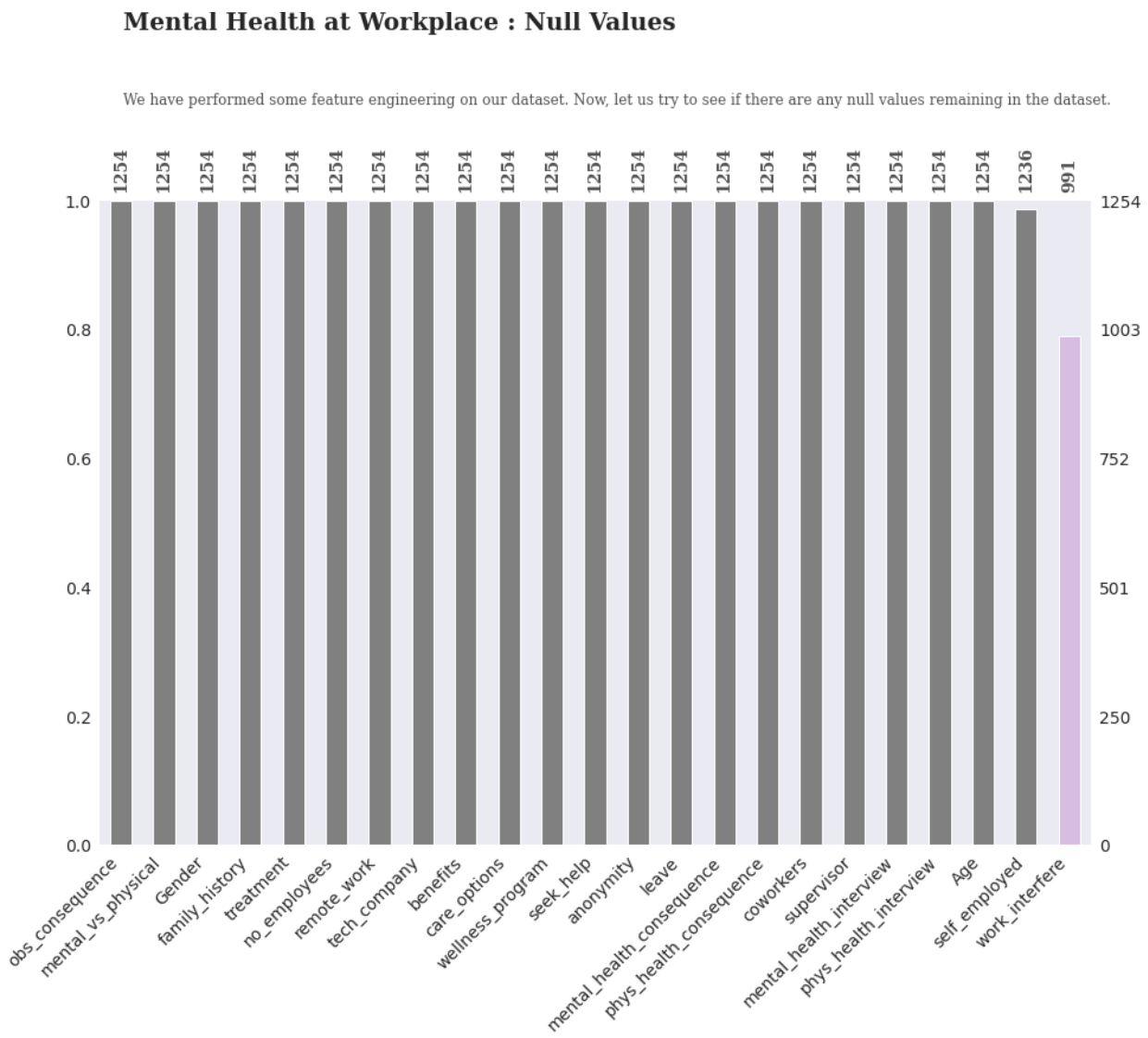
Use dimensionality reduction techniques like Principal Component Analysis (PCA) or t-SNE to extract meaningful features from highdimensional data.

Feature Crosses:

Combine two or more features to capture interactions. For example, multiplying age and income might reveal a significant feature if your campaign relates to socioeconomic factors.

Health Indicators and Composite Features:

Calculate health-related indices or composite features, such as BMI, which can be valuable in public health campaigns focusing on nutrition and fitness.



Feature engineering is an iterative process that involves experimenting with different transformations and selecting the most informative features for your specific campaign. The goal is to make your data more amenable to machine learning and provide actionable insights that drive your public health awareness efforts.

Model Training:

Training a public health awareness campaign involves several key steps and considerations. Below is a general guide on how to train a model for a public health awareness campaign:

1. Define Objectives and Goals:

 Clearly define the objectives of your public health awareness campaign. What do you want to achieve? Who is your target audience? What behavior change or awareness are you aiming for?

2. Gather Data:

 Collect data relevant to your campaign. This may include health statistics, demographic information, and historical campaign data. You'll need this data to train your model effectively.

3. Choose a Machine Learning Model:

 Depending on your campaign's goals, you may choose different machine learning models. For instance, for predictive modeling, you might use regression or time series analysis. For classification tasks, you might use decision trees, random forests, or deep learning models.

4. Feature Engineering:

 Prepare your data by selecting relevant features (variables) that are likely to influence the outcome of your campaign. This can involve data cleaning, transformation, and normalization.

5. Data Splitting:

 Split your data into training, validation, and test sets. The training set is used to train the model, the validation set is used for hyperparameter tuning, and the test set is used to evaluate model performance.

6. Model Training: Train your chosen machine learning model on the training data. This involves feeding the model input data and target labels (if supervised learning) and adjusting model parameters to minimize the error.

7. Hyperparameter Tuning:

 Optimize the hyperparameters of your model to improve its performance. You can use techniques like cross-validation and grid search.

.Evaluation:

 Evaluate your model's performance using the validation and test datasets. Common evaluation metrics for public health campaigns may include accuracy, precision, recall, F1 score, or area under the ROC curve, depending on the nature of your campaign.

8. Interpretation:

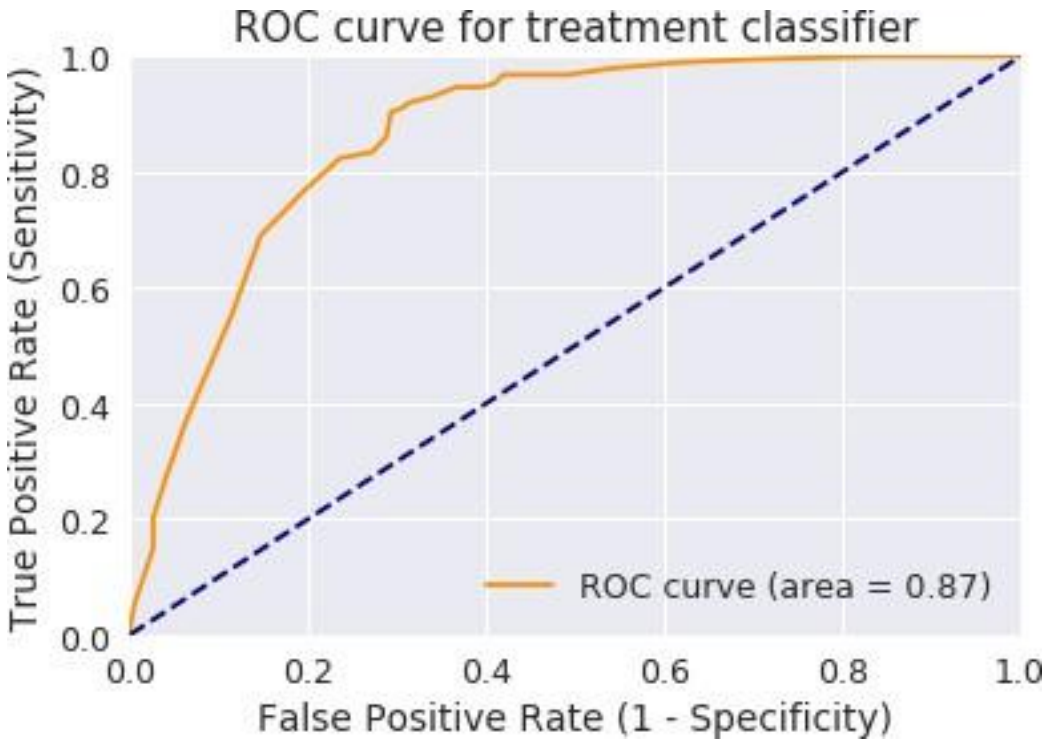
 Understand the insights gained from the model. What factors influence public health behavior or awareness? This can help you fine-tune your campaign strategy.

9. Deployment:

 Deploy your model in a real-world setting. This might involve integrating it with a website, mobile app, or other digital platforms where you can reach your target audience.

10. Monitoring and Feedback:

 Continuously monitor the performance of your campaign. Gather feedback from the target audience and the community to make necessary adjustments and improvements.



11. Ethical Considerations:

 Be aware of ethical considerations in using data and machine learning for public health. Ensure that your campaign respects privacy, is fair, and doesn't discriminate against any group.

12. Adaptation:

 Public health is dynamic, and awareness campaigns may need to adapt to changing circumstances, new information, or emerging health issues. Be ready to update your model and campaign as needed.

13. Collaboration:

 Collaboration with public health experts, community leaders, and organizations is often crucial for the success of public health campaigns. Work closely with these stakeholders to ensure your campaign aligns with expert guidance and community needs.

Remember that public health campaigns may involve complex social and behavioral factors, so machine learning models are just one tool among many in your toolkit for addressing public health challenges. Combining data-driven insights with community engagement and expert knowledge is often the most effective approach.

ADVANTAGES:

Public health awareness campaigns have several advantages that contribute to the overall well-being of a society. These campaigns

are essential in raising awareness about various health issues, promoting healthy behaviors, and preventing diseases. Some of the key advantages of public health awareness campaigns include:

1. Disease Prevention: Public health campaigns educate the public about the causes, symptoms, and preventive measures of various diseases. This information can help individuals take proactive steps to reduce their risk of illness, such as getting vaccinated, practicing safe sex, or adopting healthier lifestyles.
2. Health Promotion: These campaigns promote healthy behaviors like regular exercise, balanced nutrition, and tobacco cessation. By encouraging positive lifestyle changes, they contribute to a healthier population and lower healthcare costs in the long run.
3. Early Detection: Awareness campaigns often emphasize the importance of regular screenings and check-ups, which can lead to the early detection and treatment of diseases, ultimately improving prognosis and survival rates.
4. Reduced Stigma: Campaigns addressing mental health, addiction, or other sensitive topics can help reduce the stigma associated with these issues, making it easier for individuals to seek help and support.
5. Public Safety: Awareness campaigns related to topics like road safety, child safety, and disaster preparedness can help prevent accidents and injuries, ultimately saving lives.

6. Community Engagement: These campaigns often involve community participation and engagement, fostering a sense of collective responsibility for public health issues.
7. Advocacy and Policy Change: Public health campaigns can mobilize public support for policy changes, leading to the implementation of laws and regulations that protect public health. For example, campaigns for tobacco control have led to stricter smoking regulations in many places.
8. Improved Healthcare Access: By raising awareness about the importance of healthcare services and insurance, these campaigns can help individuals access the care they need, reducing health disparities.
9. Research and Funding: Increased awareness often leads to greater interest and funding for research in various health-related fields, which can result in medical breakthroughs and improved treatments.
10. Economic Benefits: Healthy populations are more productive and have lower healthcare costs, which can positively impact a nation's economy.
11. Behavior Change: Public health campaigns aim to educate and empower individuals to make informed decisions about their health, leading to positive behavior changes that benefit both individuals and the wider community.

13. Long-Term Impact: Effective public health campaigns can have a lasting impact on health outcomes and reduce the burden of disease over time.

14. Emergency Preparedness: Campaigns related to public health emergencies, such as pandemics or natural disasters, can educate the public on preparedness and response measures, potentially saving lives during crises.

It's important to note that the success of public health awareness campaigns depends on various factors, including their design, message, target audience, and the resources available for their implementation. When well-executed, these campaigns can significantly contribute to the improvement of public health and the well-being of a society.

DISADVANTAGES:

While public health awareness campaigns have many advantages, they also come with certain disadvantages and challenges:

1. Overload of Information: In an era of information overload, people may become desensitized or overwhelmed by the sheer volume of health messages, making it difficult for them to discern the most critical information from less important details.

2. Limited Reach: Not all individuals may have access to or engage with the media and channels through which public

health campaigns are disseminated, leading to disparities in awareness and health outcomes.

3. Message Fatigue: Repeated exposure to the same health messages can lead to message fatigue, where individuals become less responsive to the campaign's content over time.
4. Misinterpretation: Health messages can be misinterpreted or misunderstood, leading to incorrect conclusions or actions. Effective communication and message clarity are essential to mitigate this risk.
5. Stigmatization: Some campaigns, especially those related to sensitive topics like mental health or addiction, may inadvertently stigmatize affected individuals, making it harder for them to seek help and support.
6. Resource Constraints: The design and implementation of effective campaigns often require significant resources, which may not be available in all settings. This can lead to disparities in the quality and reach of campaigns.
7. Resistance and Skepticism: Some individuals may resist health recommendations due to personal beliefs, cultural factors, or skepticism about the information source. Public health campaigns may not always succeed in changing deeply ingrained attitudes and behaviors.
8. Short-Term Focus: Many campaigns are designed for short-term impact, which can limit their effectiveness in addressing

long-term health issues. Sustainable behavior change often requires ongoing efforts.

9. Social and Economic Factors: Health disparities are influenced by social and economic factors, and public health campaigns alone may not address the root causes of these disparities.

10. Ineffectiveness: Not all public health campaigns achieve their intended outcomes. Factors like message design, target audience, and timing can significantly affect their effectiveness.

11. Ethical Concerns: There may be ethical concerns related to the design and implementation of public health campaigns, such as invasion of privacy, manipulation of emotions, or coercion.

12. Conflicting Messages: In some cases, different health organizations or experts may promote conflicting health messages, causing confusion among the public.

13. Unintended Consequences: Public health campaigns may have unintended consequences, such as encouraging risky behaviors when messages are not well-crafted.

14. Cultural Sensitivity: Campaigns that are not culturally sensitive can be ineffective or even offensive in some communities, potentially causing backlash.

It's important to recognize these disadvantages and challenges when designing and implementing public health awareness campaigns. A thoughtful and evidence-based approach, along with ongoing evaluation and adaptation, can help mitigate these drawbacks and maximize the benefits of such campaigns.

CONCLUSION:

In conclusion, implementing a public health awareness campaign with the aid of data analytics is a highly effective and promising approach. By leveraging data-driven insights, public health campaigns can achieve several key objectives: In conclusion, implementing a public health awareness campaign with the aid of data analytics is a highly effective and promising approach. By leveraging data-driven insights, public health campaigns can achieve several key objectives:

1. Targeted Outreach: Data analytics can identify specific populations at higher risk for certain health issues, allowing campaigns to direct their efforts more precisely.
2. Message Personalization: Personalized messaging can be developed based on demographic, geographic, and behavioral data, increasing the relevance and impact of the campaign.
3. Resource Optimization: Data analytics can help in the efficient allocation of resources, ensuring that campaign efforts are focused where they will have the most significant impact.

4. Real-Time Adaptation: The ability to monitor and analyze data in real-time enables campaigns to adjust strategies as needed, responding to emerging trends and challenges.
5. Improved Engagement: Data analytics can provide insights into which communication channels and platforms are most effective for reaching and engaging the target audience.
6. Long-Term Sustainability: The data-driven approach allows for the development of sustainable campaigns that can address long-term health issues and evolve over time.
7. Evaluation and Accountability: By measuring the campaign's impact through data analysis, stakeholders can assess the success of the initiative and make data-informed decisions for future efforts.
8. Cost-Efficiency: Optimizing the use of resources and targeting the right audience can lead to cost savings, making public health campaigns more budget-friendly.

By combining the power of data analytics with public health campaigns, we can create more effective, efficient, and impactful initiatives that contribute to the overall well-being of individuals and communities. This approach has the potential to reduce disease prevalence, improve health outcomes, and positively impact society as a whole.