Problem Statement:

In the dynamic and ever-evolving landscape of the tech industry, the focus on public mental health awareness has gained paramount importance. The workplace, especially in the tech domain, is a crucible of innovation and productivity, but it is also a milieu where the mental well-being of employees plays a pivotal role. The problem at hand is the ability to gauge the effectiveness of public mental health awareness campaigns, a challenge that reverberates through the corridors of the tech workplace.

The problem statement encapsulates the urgency of being able to predict the success of public mental health awareness campaigns. It is not a mere statistical endeavour; it is a strategic pursuit that has a profound impact on the formulation of future strategies within the tech domain. A model that can accurately forecast the success of these campaigns offers a window into the factors that contribute to positive outcomes. It empowers organizations to make data-informed decisions, optimize the allocation of resources, and tailor their campaign strategies to resonate with their employees.

The importance of predicting campaign success in the realm of public mental health awareness cannot be overstated. It permeates across the spheres of employee well-being, productivity, and organizational culture. Therefore, addressing this challenge is not just a priority; it is a responsibility, calling for the prowess of machine learning to unlock the full potential of predictive analytics.

Objectives:

In Phase 2 of our project, we set forth clear and concise objectives that center on leveraging machine learning to predict the success of public mental health awareness campaigns within the tech workplace. The primary objectives include:

- ➢ The development and training of a robust machine learning model capable of predicting the success of public mental health awareness campaigns based on historical data.
- ➢ The evaluation of the model's performance using relevant metrics, ensuring its reliability and effectiveness in forecasting campaign outcomes.
- ➢ The extraction of invaluable insights from the model's predictions, shedding light on the factors that significantly influence campaign success.
- ➢ The enhancement of our understanding of the interplay between campaign attributes and their impact on outcomes, allowing organizations to craft strategies that resonate with their employees.

These objectives unify to form a single mission: utilizing machine learning to proactively steer the course of public mental health awareness campaigns in the tech workplace, optimizing their strategies, and elevating the prospects of success.

Dataset Overview:

The bedrock of our analysis is rooted in the dataset at our disposal. This dataset serves as a treasure trove of information critical for unraveling the complexities of public mental health awareness campaigns. It encompasses a broad spectrum of attributes, spanning campaign particulars, demographic insights, temporal aspects, and measures of campaign success.

- **Timestamp**
- **Age**
- **Gender**
- **Country**
- **state**: If you live in the United States, which state or territory do you live in?
- **self_employed**: Are you self-employed?
- **family_history**: Do you have a family history of mental illness?
- **treatment**: Have you sought treatment for a mental health condition?
- **work_interfere**: If you have a mental health condition, do you feel that it interferes with your work?
- **no_employees**: How many employees does your company or organization have?
- **remote_work**: Do you work remotely (outside of an office) at least 50% of the time?
- **tech_company**: Is your employer primarily a tech company/organization?
- **benefits**: Does your employer provide mental health benefits?
- **care_options**: Do you know the options for mental health care your employer provides?
- **wellness_program**: Has your employer ever discussed mental health as part of an employee wellness program?
- **seek_help**: Does your employer provide resources to learn more about mental health issues and how to seek help?
- **anonymity**: Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources?
- **leave**: How easy is it for you to take medical leave for a mental health condition?
- **mental_health_consequence**: Do you think that discussing a mental health issue with your employer would have negative consequences?
- **phys_health_consequence**: Do you think that discussing a physical health issue with your employer would have negative consequences?
- **coworkers**: Would you be willing to discuss a mental health issue with your coworkers?
- **supervisor**: Would you be willing to discuss a mental health issue with your direct supervisor(s)?
- **mental_health_interview**: Would you bring up a mental health issue with a potential employer in an interview?
- **phys_health_interview**: Would you bring up a physical health issue with a potential employer in an interview?
- **mental_vs_physical**: Do you feel that your employer takes mental health as seriously as physical health?
- **obs_consequence**: Have you heard of or observed negative consequences for coworkers with mental health conditions in your workplace?
- **comments**: Any additional notes or comments

Data Source and Collection:

The dataset used in this analysis originates from a 2014 survey that was methodically conducted to assess the attitudes of individuals working in the tech industry towards mental health, as well as to measure the frequency of mental health disorders within this professional domain. The dataset, thus, represents a snapshot of the mental health landscape in the tech workplace during the year 2014. Responses were collected from tech professionals through a well-structured survey that encompassed a wide range of attributes. The information garnered from these survey responses forms the bedrock of our analysis, enabling us to unravel insights into mental health awareness in the tech sector.

Data Cleaning:

Data cleaning is an essential phase of data preparation aimed at ensuring the dataset's integrity and reliability. This process encompasses several key steps:

- **Handling Missing Values:** One of the primary data cleaning tasks involved addressing missing values within the dataset. We recognized that missing data can lead to skewed analyses, and thus, we took meticulous care in handling them. For numerical attributes with missing values, we applied techniques like mean, median, or mode imputation, ensuring that the imputed values were coherent with the underlying data distribution. For categorical attributes, custom imputation methods were employed when suitable. These steps allowed us to mitigate the impact of missing data on our analysis.
- **Handling Duplicate Values: Ensuring Data Integrity**

Duplicate values in a dataset can introduce bias and inaccuracies. We systematically detected and treated duplicates by comparing records across attributes. Our approach involved considering whether to remove or retain duplicates based on their significance. A data verification step confirmed the effectiveness of our process. By addressing duplicate values, we improved data quality and eliminated potential sources of bias for more accurate analyses and machine learning models.

- **Data Format Consistency:** To maintain data quality, we addressed inconsistencies in data formatting. This included harmonizing date formats, standardizing the representation of categorical variables, and ensuring uniformity across the dataset.

MACHINE LEARNING MODELS :

To predict the success of future public health awareness campaigns based on historical data, we can employ various machine learning algorithms.

1. Logistic Regression: This algorithm is often used for binary classification problems, making it suitable for predicting campaign success (yes/no). It's interpretable and can provide insights into the factors that influence success.

2. Random Forest: Random Forest is an ensemble learning method that can handle both classification and regression tasks. It's robust and can capture complex relationships in our data.

3. Time-Series Analysis (ARIMA or LSTM): To handle with time-series data related to past campaigns, methods like ARIMA or LSTM can be used for forecasting campaign success over time.

4. Natural Language Processing (NLP) Models: To derive meaningful insights from text or feedback content related to campaigns, NLP models like BERT, GPT-3, or word embeddings (Word2Vec, GloVe) can extract insights and predict campaign sentiment and success.

Overview:

- **Algorithm Type:** Supervised Learning (Classification)
- **Objective:** Predict campaign success in the context of public mental health awareness within the tech workplace.
- **Algorithm Choice Rationale:** Random Forest is a versatile and robust ensemble learning technique suitable for a wide range of classification tasks. It offers several advantages, including handling non-linearity, feature importance analysis, and resistance to overfitting.

Steps:

1. Data Preparation:

- Dataset Split: Split the dataset into a training set and a testing set.
- Feature Engineering: Perform feature engineering, including one-hot encoding of categorical variables, handling missing values, and possibly feature scaling.

2. Model Selection:

- Random Forest Classifier: Select the Random Forest Classifier as the machine learning algorithm of choice.

- Algorithm Rationale: Random Forest is well-suited for classification tasks, offers high predictive accuracy, and is capable of handling complex, non-linear relationships in the data.

3. Model Training:

- Data Preparation: Preprocess the training data by performing any necessary feature scaling or transformations.
- Model Fitting: Train the Random Forest Classifier on the training data. The algorithm builds a forest of decision trees, each based on a random subset of the data.
- Hyperparameter Tuning: Fine-tune the model's hyperparameters, including the number of trees in the forest, maximum tree depth, and feature selection strategy, to optimize model performance.

4. Model Evaluation:

- Testing Dataset: Evaluate the model's performance on the testing dataset to assess its predictive accuracy.
- Evaluation Metrics: Calculate key classification metrics, including accuracy, precision, recall, F1-score, and ROC AUC score.
- Confusion Matrix: Analyze the confusion matrix to understand true positives, true negatives, false positives, and false negatives.
- Cross-Validation: Implement cross-validation techniques to ensure the model's robustness and consistency in performance.

5. Interpretation and Insights:

- Feature Importance: Use the Random Forest's feature importance analysis to identify which attributes strongly influence campaign success.
- Influential Features: Determine which features have the most significant impact on the model's predictions.
- Campaign Success Factors: Quantify the factors that contribute to campaign success, providing actionable insights for campaign optimization.

6. Ethical Considerations:

- Ensure privacy and data security by anonymizing the dataset.
- Implement measures to mitigate bias in the model.
- Maintain model transparency and interpretability for ethical accountability.

Conclusion:

In conclusion, the incorporation of machine learning in Phase 2 of this project has yielded significant insights into predicting campaign success for public mental health awareness within the tech workplace. The key findings and implications include:

- The Random Forest Classifier demonstrated commendable predictive performance, achieving a high accuracy in forecasting campaign success.
- Feature importance analysis uncovered critical variables that strongly influence campaign outcomes, empowering data-driven decision-making.

The significance of incorporating machine learning in this project is evident through its ability to provide data-driven insights, offering a roadmap for optimizing future campaign strategies. The model's predictive power and transparency play a pivotal role in guiding decision-makers within the domain of public mental health awareness.

Based on the model's predictions and insights, we recommend the following:

- Tailoring future campaigns by focusing on influential features identified by the model.
- Continuously monitoring and refining campaign strategies to adapt to changing workplace dynamics.
- Exploring further research to understand the intricate relationships between attributes in the dataset and campaign success.

This concludes Phase 2 of the project, positioning us for the next phase, where we will focus on data visualization using IBM Cognos.