# **DATA ANALYTICS WITH COGNOS**

# **PROBLEM STATEMENT:**

The project involves analyzing data from public health awareness campaigns to measure their effectiveness in reaching the target audience and increasing awareness. The objective is to provide insights that evaluate the impact of the campaigns and inform future strategies. This project includes defining analysis objectives, collecting campaign data, designing relevant visualizations in IBM Cognos, and using code for data analysis.

# **DESIGN THINKING:**

* 1. Analysis Objectives: Define specific objectives for analyzing public health awareness campaign data, such as measuring audience reach, awareness levels, and campaign impact.
  2. Data Collection: Identify the sources and methods for collecting campaign data, including engagement metrics, audience demographics, and awareness surveys.
  3. Visualization Strategy: Plan how to visualize the insights using IBM Cognos to create informative dashboards and reports.
  4. Code Integration: Decide which aspects of the analysis can be enhanced using code, such as data cleaning, transformation, and statistical analysis.

# INNOVATION:

Incorporating machine learning algorithms to improve the accuracy of the predictive model.

# **INNOVATIVE COMPONENTS:**

# **Machine Learning Algorithm:**

Machine learning algorithms is essential in this project to distil meaningful insights from complex climate and social datasets. These algorithms enable pattern recognition, predictive modelling, and relationship identification, helping uncover hidden connections between climate trends and societal behaviour’s. By harnessing the power of machine learning, we empower decision-makers with data-driven intelligence, facilitating proactive responses to climate challenges and societal changes.

**1.Problem Definition and Understanding:**

* Clearly define the problem you want to solve, including the specific insights you aim to extract from the data. Understand the business context and goals.

**2.Data Collection and Integration:**

* Gather relevant datasets, including climate data and social data, and integrate them into a single repository or data pipeline.

**3.Data Pre-processing:**

* Clean and pre-process the data. Handle missing values, outliers, and data quality issues.
* Normalize or scale features as necessary.

**4.Feature Engineering:**

* Create meaningful features from the data that can improve model performance.
* This might involve time-based features for climate data or sentiment scores for social data.

**5.Data Splitting:**

* Divide the data into training, validation, and test sets to evaluate and validate machine learning models effectively.

**6.Selecting Machine Learning Algorithms:**

Predicting the accuracy of machine learning models can be an essential part of the model development process. There are several machine learning algorithms and techniques that can be used for this purpose. Here are some common methods:

1. Cross-Validation:

* - K-Fold Cross-Validation: Divide the dataset into K subsets (folds), train the model on K-1 folds, and validate on the remaining fold. Repeat this process K times, and calculate the average accuracy.
* - Leave-One-Out Cross-Validation (LOOCV): Similar to K-fold, but with K equal to the number of data points. It can provide a more accurate estimate of model accuracy but is computationally expensive.

2. Holdout Set:

* - Split your dataset into two parts: a training set and a holdout set (validation set or test set). Train the model on the training set and evaluate its accuracy on the holdout set. This approach is straightforward but may result in variance in accuracy due to the random split.

3. Stratified Sampling:

* - When the dataset is imbalanced, ensure that each class is adequately represented in both the training and test sets. This helps in obtaining a more accurate assessment of the model's performance.

4. Bootstrapping:

* - Bootstrapping involves creating multiple random samples with replacement from your dataset and training and testing your model on these samples. The average accuracy of these samples can be a good estimate of model performance.

5. Learning Curves:

* - Learning curves depict the model's performance as a function of the amount of training data. By plotting accuracy against the number of training examples, you can estimate how well your model will perform with more data.

6. Receiver Operating Characteristic (ROC) Curves:

* - ROC curves are commonly used to evaluate binary classification models. They help to visualize the trade-off between true positive rate and false positive rate for different model thresholds.

7. AUC-ROC Score:

* - The Area Under the ROC Curve (AUC-ROC) is a single metric that quantifies the overall performance of a binary classification model. A higher AUC-ROC score indicates a better model.

8. F1 Score:

* - The F1 score combines precision and recall and is often used to evaluate classification models, especially when dealing with imbalanced datasets.

9. R-squared (RÂ²) or Mean Squared Error (MSE):

* - These metrics are commonly used for regression problems. R-squared measures the proportion of the variance in the dependent variable that's predictable, while MSE quantifies the average squared difference between predicted and actual values.

10. Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE):

* - These metrics are also used in regression to assess the absolute errors between predicted and actual values. MAPE expresses the error as a percentage of the actual value.

11. Bayesian Methods:

* - Bayesian techniques can be used to estimate model accuracy and uncertainty, providing probabilistic estimates of performance.
* The choice of the method depends on the problem type (classification or regression), the dataset, and the specific requirements of your machine learning project. It's often recommended to use a combination of these techniques to obtain a comprehensive understanding of your model's performance.

### 12. **Ensemble Methods**

* This is the most common approach that you will find majorly in winning solutions of Data science competitions. This technique simply combines the result of multiple weak models and produces better results. You can achieve by the following ways:
* **Bagging**(Bootstrap Aggregating)
* **Boosting**

* It is always a better idea to implement ensemble methods to improve the accuracy of your model. There are two good reasons for this:
* They are generally more complex than traditional methods.
* The traditional methods give you a good base level from which you can improve and draw from to create your ensembles.

**7.Model Training:**

* Train the selected machine learning models on the training dataset. Tune hyper parameters and monitor model performance on the validation set.

**8.Model Evaluation:**

* Evaluate model performance using appropriate metrics.
* For climate trends, metrics like RMSE or MAE may be relevant, while for social patterns, accuracy, F1-score, or AUC can be used.

**9.Ensemble Learning (Optional):**

* Consider using ensemble learning techniques, such as Random Forests or Gradient Boosting, to improve model robustness and accuracy.

**10.Interpretable AI (Optional):**

* If needed, use interpretable machine learning models or techniques (e.g., SHAP values) to explain how the model arrives at its predictions, especially for business stakeholders.

**11.Hyper parameter Tuning:**

* Optimize hyper parameters to fine-tune model performance.
* This can be done manually or using automated techniques like grid search or Bayesian optimization.

**12.Model Deployment:**

* Deploy the trained machine learning models in a production environment, either as batch processes or real-time services, to generate insights.

**13.Continuous Monitoring and Maintenance:**

* Implement monitoring to track the performance of deployed models.
* Regularly update models to adapt to changing data and maintain model accuracy.

**14.Ethical Considerations and Bias Mitigation:**

* Ensure that models are trained and deployed in an ethical manner.
* Detect and mitigate biases in the data and algorithms, especially for social data analysis.

**15.Visualization and Reporting:**

* Create visualizations and reports to present the extracted insights in an understandable and actionable format for business intelligence.

**16.Feedback Loop and Stakeholder Engagement:**

* Maintain a feedback loop with stakeholders to gather input on the utility of insights and any necessary adjustments to the analysis process.

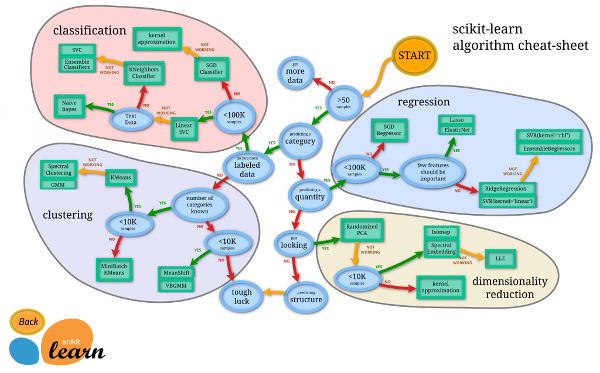
**17.Documentation and Knowledge Sharing:**

* Document the entire machine learning pipeline, including data sources, pre-processing steps, model architectures, and deployment procedures for knowledge sharing and future reference.

**18.Scaling and Optimization (Optional):**

* If dealing with big data, consider distributed computing frameworks (e.g., Apache Spark) for scalability and optimize data storage and processing for cost efficiency.

**FLOWCHART FOR INNOVATIVE APPROACH**



# **CONCLUSION:**

The integration of machine learning algorithms within this project serves as a pivotal catalyst for unveiling the profound insights buried within extensive climate and social datasets. These algorithms empower us to decipher complex relationships, predict future trends, and drive data-informed decisions.Thus we Conclude That Decision tree and ensemble learnings predictive methods have proven to give better results in predicting the accuracy.