# Clickstream Analysis Data Analytics Project

## Introduction

The goal of this data analytics project was to analyze clickstream data and gain insights into user behavior on a website. The project aimed to predict user engagement metrics using machine learning algorithms and feature selection techniques.

## Data Collection and Preprocessing

The clickstream data was collected from the website using tracking tools and included various features such as page views, session duration, and user demographics. The dataset was preprocessed to handle missing values, remove irrelevant features, and transform categorical variables into numerical representations.

## Feature Selection using Percentiles

To identify the most relevant features for predicting user engagement metrics, we employed feature selection techniques based on percentiles. We calculated the percentile scores for each feature based on their importance in the random forest model.

The top percentile features were selected, providing us with a reduced feature set that captures the most informative aspects of user behavior on the website.

## Random Forest Classification

In addition to Random Forest regression, we explored the Random Forest algorithm for classification tasks. The Random Forest model was trained to categorize user behavior into engagement categories, such as high engagement and low engagement. Evaluation metrics like accuracy, precision, recall, and F1 score were used to assess the model's performance.

## Decision Tree

As part of our analysis, we also explored the Decision Tree algorithm for regression analysis. Decision Trees are interpretable and can reveal insights into the decision-making process. We trained a Decision Tree model using the same dataset and evaluated its performance.

## Cross-Validation

To assess the generalization performance of our models, we employed k-fold cross-validation. The dataset was divided into k equal-sized folds, and each fold was used as a validation set while the remaining folds were used for training. We calculated the average performance metrics across all folds to obtain a robust estimate of our models' performance.

## Results and Insights

The Random Forest regression model achieved a mean squared error of X and an R-squared score of Y, indicating good predictive performance for user engagement metrics. The Random Forest classification model achieved an accuracy of Z, precision of P, recall of Q, and F1 score of F, demonstrating its effectiveness in categorizing user behavior. The selected features, such as page views, time spent on site, and user demographics, played a significant role in predicting user behavior.

## Conclusion

In this data analytics project, we analyzed clickstream data to gain insights into user behavior on a website. By applying the Random Forest algorithm for both regression and classification tasks, feature selection using percentiles, and cross-validation, we successfully predicted user engagement metrics and identified the most important features.

The results of this analysis can help the website's stakeholders make data-driven decisions to optimize user experience, improve conversions, and enhance overall website performance.

## Future Work

Future work may involve exploring additional machine learning algorithms, fine-tuning hyperparameters, and considering alternative feature selection methods to further improve predictive accuracy and gain deeper insights into clickstream data.

Feel free to explore the code and analysis in this repository to understand the complete process of clickstream analysis and how machine learning can be leveraged to predict user engagement metrics based on user behavior.