**FINAL PROJECT PROPOSAL  
TEAM DATA DRIFTERS**

**GOWTHAM LEKKALA**

**SAI PRIYA MALLAMPATI**

**JAYAPRAKASH MAKKENA**

**MOHAN ANANTA VENKATA PAVAN GAMIDI**

**AFRAH SHAIK**

**HEMA EDAVALAPATI**

**PROBLEM STATEMENT**

What factors most significantly influence the popularity of online news articles, and how can we predict and improve the number of shares these articles receive on social networks?

How do different content channels affect audience engagement, and what strategies can we adopt to enhance sharing in these categories?

Which types of articles or topics are more likely to be shared, and how can we align our content creation with these trends?

**DATASET DETAILS**

This dataset provides a comprehensive collection of attributes related to online articles published by Mashable over a span of two years. The primary objective is to forecast the popularity of these articles, specifically, the number of shares they receive on various social media platforms. The dataset was meticulously curated by a team consisting of Kelwin Fernandes, Pedro Vinagre, Paulo Cortez, and Pedro Sernadela and is publicly available through the UCI library.

The dataset encompasses a wide range of variables that offer valuable insights for predictive analysis. The ultimate aim is to build predictive models, both for classification and regression purposes, that can estimate the extent to which an article is likely to be shared on social networks.

This dataset serves as a valuable resource for researchers and data scientists seeking to understand the factors influencing the dissemination and popularity of online news stories, which can be critical for content creators, publishers, and marketers looking to optimize their content strategies and increase the reach and impact of their articles on social media.

**ANALYSIS**

The first step in our analysis involves data preprocessing. We will meticulously inspect the dataset to understand its structure and characteristics. Any missing values within the dataset will be identified and handled appropriately, ensuring data integrity. If the dataset contains categorical variables, we will encode them into a numerical format to make them compatible with our chosen modeling techniques. Furthermore, numerical features will be normalized or standardized as needed to ensure uniform scaling. To facilitate model evaluation, the dataset will be divided into separate training and testing sets.

After data preprocessing, the next crucial step is conducting exploratory data analysis (EDA). EDA involves a thorough examination of the dataset to understand the distribution and characteristics of the features. We will visualize the relationships between these features and the target variable, allowing us to grasp initial insights and patterns in the data. Additionally, summary statistics for key features will be calculated to provide a comprehensive view of the dataset's central tendencies.

Once we have a clear understanding of the dataset, we will proceed with feature selection. Feature selection techniques, such as correlation analysis, will be employed to identify the most relevant features that play a substantial role in predicting article popularity. By focusing on these informative attributes, we streamline the model and improve its efficiency.

For the regression task of predicting the number of shares, we will carefully choose appropriate machine learning models. Options under consideration include Linear Regression, Random Forest, Gradient Boosting, and Neural Networks. The selection will be based on a thorough evaluation of the dataset's characteristics and the specific performance requirements of the task.

Model performance optimization is achieved through hyperparameter tuning. Techniques such as grid search or random search will be employed to fine-tune the model's hyperparameters. This ensures that the model operates at its best, delivering accurate predictions.

Once the model is trained and optimized, we will rigorously evaluate its performance using various regression metrics, including mean squared error, R-squared, and other relevant indicators. Visualizations, such as scatter plots comparing model predictions to actual share counts, will be used to provide a clear picture of the model's strengths and weaknesses. This step is essential for assessing the model's ability to predict article popularity effectively.

The types of articles shared vary, so understand my audience and trends. Focus on evergreen content, experiment with formats, and use data insights to adapt our strategy.

In conclusion, by addressing these factors and staying in tune with our audience and trends, we can increase the chances of our articles being shared on social networks. However, predicting the exact number of shares remains a challenge, so continuous analysis and adjustment are essential for success.

The final phase of our analysis involves interpreting the model results. We aim to gain meaningful insights by understanding which features have the most significant impact on article popularity. Identifying these influential factors is essential for content creators and publishers seeking to optimize their strategies and maximize the reach and impact of their articles on social media. This interpretation phase ties the analysis back to the core objective of predicting online news article popularity and offering actionable insights for content optimization.