**Predictive Analysis of Newspaper Subscribers Using Big Data**

**Team #3:** Aakash Boenal; Kiranmai Reddy Varakantam; Meghana Katta; Rishitha Madipelli; Shreya Gowrishetty

# **ABSTRACT**

# In the constantly evolving world of media consumption, this project embarks on a journey of predictive analytics to comprehend and foresee the behavior of newspaper subscribers. Our team's goal is to build a durable prediction model using big data technologies such as PySpark, and Spark MLlib, and the incorporation of Machine Learning (ML) techniques such as Random Forest. This model is designed to provide useful insights into subscriber preferences as well as forecast potential churn. The project begins with a comprehensive data preprocessing phase that maintains the quality and accuracy of the selected dataset that includes over 15,000 records of newspaper subscribers. Our methodology involves exploratory data analysis (EDA) to uncover patterns and trends within the dataset, establishing the framework for subsequent stages. This analysis addresses eight pivotal research questions designed to uncover patterns and trends within the dataset. The questions cover a wide range of topics, including the distribution of subscribers, language preferences, identifying age groups that are most likely to engage with newspaper subscriptions. The study also investigates the impact of source channels on subscription events, the correlation between lower income and subscription cancellations, and subscription rates across different counties. Furthermore, the analysis investigates the impact of the average weekly fee on subscriber behavior and identifies the best delivery schedule for subscriber retention. This comprehensive EDA not only identifies patterns and trends but also lays the framework for later predictive modeling by combining machine learning approaches to improve the depth and accuracy of subscriber behavior understanding. Additionally, to develop a prediction model, we employ the ML Random Forest algorithm. This model aims to determine subscriber preferences and forecast potential churn, expanding our understanding of subscriber behavior beyond traditional descriptive statistics. Beyond analytical depth, the project extends to practical recommendations for content delivery and retention strategies. By addressing these research questions, our project contributes not only to academic understanding but also provides actionable insights for the newspaper industry, guiding decision-makers in adapting to the dynamic landscape of media consumption.

# *Keywords****:*** Predictive Analytics, Big Data, PySpark, Spark MLlib, Subscriber Behavior, Churn Prediction, Data Analysis, Data Preprocessing, Machine Learning, Random Forest

# **REFERENCES**

[1] geiler, l. (10, May 2022). *NewspaperChurn .* Retrieved from Open ML: <https://api.openml.org/d/44226>

[2] Belchior, L. M. (2022, November). *Churn Prediction in Online Newspaper Subscriptions.* Retrieved from Run Repositorio Universidade Nova: <https://run.unl.pt/bitstream/10362/148914/4/TGI1079.pdf>

[3] Peña, V. C., Malthouse, E. C., & Mersey, R. D. (2023). Churning off the news: An analysis of newspaper subscriber churn across digital devices. Newspaper Research Journal, 44(2), 190-205. <https://doi.org/10.1177/07395329231167374>