1. Data Processing and Feature Selection: For this analysis, the dataset was filtered to include only Kickstarter projects marked as "successful" or "failed," aligning with our goal of predicting project outcomes. To standardize currency, each project's goal was converted to USD using the static_usd_rate. We then calculated project_duration in days, representing the span from project launch to the deadline. Missing values were handled by removing any records lacking data in key fields, such as main_category. The final dataset includes 11 predictors, carefully chosen to ensure availability at project launch and relevance to project success. These variables include attributes like category, country, and goal_usd. Categorical variables, such as category and country, were transformed into dummy variables for compatibility with machine learning algorithms, while state was binarized to represent success as 1 and failure as 0. Correlation and VIF analysis were conducted to ensure minimal multicollinearity among predictors.

2. Classification Model: We evaluated five classification models: Random Forest, Logistic Regression, Gradient Boosting, K-Nearest Neighbors (KNN), and Neural Networks. Random Forest achieved the highest accuracy (76.95%) with default parameters, alongside strong Precision, Recall, and F1 Score, making it the most balanced and reliable choice. Random Forest's ability to handle both categorical and numerical data, its robustness against overfitting due to ensemble learning, and its interpretability through feature importance contributed to its selection.

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	0.7695	0.7875	0.827	0.8068
Logistic Regression	0.763	0.8099	0.7748	0.7919
Gradient Boosting	0.7579	0.7569	0.8602	0.8053
KNN	0.6547	0.6883	0.7433	0.7148
Neural Network	0.7442	0.7909	0.762	0.7762
Random Forest (Tuned)	0.7767	0.7926	0.8347	0.8131

To improve performance, we tuned hyperparameters using GridSearchCV with 5-fold cross-validation, identifying the best parameters as n_estimators = 100, max_depth = None, min_samples_split = 10, and min_samples_leaf = 1. With these optimized parameters, the Random Forest model's accuracy increased to 77.67%. The Mean Squared Error (MSE) for predicted

probabilities was 0.1494, confirming the model's reliability. Finally, feature importance analysis highlighted the most influential predictors for project success, demonstrating the Random Forest model's strength in both prediction and interpretability.

The classification model highlights key business features driving project success. Setting realistic funding goals (goal_usd) and choosing optimal timelines (project_duration) are critical for success. Engaging content, such as detailed blurbs (blurb_len_clean), polished names (name_len_clean), and videos (video_True), significantly boosts performance. Timing also matters, with strategic launch and deadline months influencing outcomes. Additionally, the project category plays a role, showing that tailoring campaigns to popular or well-suited categories enhances their likelihood of success. These insights provide actionable guidelines for improving campaign strategies.

3. Clustering: We selected K-Prototypes for clustering due to its ability to effectively handle mixed data types, seamlessly combining numerical and categorical variables. Unlike traditional methods like K-Means, K-Prototypes calculates distances for both types of data, ensuring accurate groupings in datasets with diverse features. Its flexibility and interpretable cluster centroids made it the most suitable choice for this analysis. The silhouette method was used to determine the optimal number of clusters as it evaluates the quality of the clustering structure by measuring both cohesion within clusters and separation between them. This approach ensures meaningful and distinct clusters while maintaining statistical rigor, making it well-suited for mixed-type data. The algorithm clustered Kickstarter projects into groups with distinct patterns, offering insights into project characteristics and trends. Analysis of cluster centroids highlighted key differences and similarities within each group.

Cluster 0 is characterized by projects with minimal backer engagement (average of 1 backer) and the lowest pledged amounts. These campaigns typically lack videos, have very short names and blurbs, and show minimal effort in presentation. Additionally, these projects have the longest average duration (~60 days), suggesting unrealistic timelines that fail to maintain interest. This cluster highlights the importance of professional campaign design and realistic planning to attract backers and increase funding potential.

Cluster 1 represents moderately successful campaigns with an average backer count of ~57 and pledged amounts around $5,000 USD. These projects feature balanced timelines (~32 days) and moderately detailed names and blurbs. They are often from countries with an established crowdfunding presence, but their success metrics are far below high-performing clusters. This indicates that while these campaigns have a better structure, they would benefit from enhanced engagement strategies, such as incorporating videos or improving content presentation.

Cluster 2 stands out as the most successful, with the highest backer count (94) and pleged ($15,000 USD). These projects are well-presented, featuring polished and detailed names and blurbs. They have an optimal duration (~34 days) and frequently include videos, which drive engagement and contribute significantly to their success. This cluster underscores the value of high-quality content, professional videos, and strategic planning for maximizing crowdfunding potential.

Cluster 3 displays balanced success metrics, with an average backer count of ~70 and pledged amounts around $9,000 USD. These projects maintain moderately detailed names and blurbs, an optimal duration (~34 days), and often include videos. While they perform well, they fall slightly short of Cluster 2, indicating that further refinement in presentation or engagement efforts could bridge this gap. This cluster emphasizes that even moderately successful campaigns can achieve greater outcomes by adopting best practices seen in Cluster 2.