**Process Documentation**

### **1. **Subsets Creation and Refinement****

* ****Define Initial Subsets****: Based on thematic relevance, domain knowledge, and data structure.
* ****Iterative Refinement****: Adjust subsets based on insights from exploratory data analysis (EDA) or initial modeling results. Use input and feedback to redefine groupings.
* ****Validation****: Confirm that subsets logically group features that are interrelated or are likely to impact the dependent variable similarly.

### **2. **Exploratory Data Analysis (EDA)****

* ****Univariate Analysis****: Descriptive statistics and distribution of each feature within subsets.
* ****Bivariate Analysis****: Explore relationships between features within subsets and with the target variable.
* ****Visualization****: Leverage plots (histograms, box plots, scatter plots) to visualize distributions and relationships.

### **Encode Data For Below First**

### **3. **Correlation Analysis****

* ****Calculate Correlations****: Use Pearson, Spearman, or other relevant correlation coefficients to quantify the linear or monotonic relationships between features and the target.
* ****Heatmaps****: Visualize correlation matrices to easily identify highly correlated features or features strongly correlated with the target.

### **4. **Feature Importance****

* ****Model-Based Importance****: Use tree-based models (e.g., Random Forest, Gradient Boosting) to assess the importance of each feature.
* ****Non-Model Techniques****: Leverage statistics like ANOVA, Chi-square (for categorical variables) to gauge feature relevance.
* ****Initial Feature Selection****: Prioritize features based on their importance scores and remove features with negligible impact.

### **5. **Recursive Feature Elimination (RFE)****

* ****Implement RFE****: Use RFE with a chosen model to systematically remove the least important feature and fit the model to the remaining features.
* ****Optimize Feature Set****: Continuously refine the set of features until an optimal subset is achieved, typically gauged by model performance metrics (accuracy, F1-score, ROC AUC).

### **6. **Modeling with Refined Features****

* ****Build Models****: Use the selected features to build predictive models. Consider different algorithms to find the best performer.
* ****Validation and Testing****: Validate the model using cross-validation and test it on unseen data to ensure robustness.

### **7. **Documentation and Iteration****

* ****Document Each Step****: Keep a detailed log of findings, decisions, and rationale at each stage.
* ****Iterative Refinement****: Return to previous steps as needed based on model performance and insights.