An Integrated Framework Using Variable Encoding-TF-IDF-PCA-Classification for Predicting Adverse Event Action

Layan Abu-Ghoush, Industrial and Systems Engineering, MS Candidate

Time 10:30 AM – 12:00 PM (EST) on Aug 16, 2024

Zoom Link: https://binghamton.zoom.us/j/92117787082

Abstract: Reporting adverse events is voluntary and accepted from any hospital staff member; however, not all adverse events necessitate further actions or follow-up. This initial step in distinguishing adverse events can take hours, as those requiring further action are escalated to the next step, involving deeper investigation and addressing by the appropriate department. This research tackles the problem of distinguishing between adverse events that require further action and those that do not, as the first step in creating an action plan to deal with adverse events. Most previous research aims to predict whether an adverse event needs to be escalated for state reporting as the final step of the escalation process, using a limited number of machine learning models. These studies often lack the use of PCA for dimensionality reduction or any form of categorical encoding. This research produces a framework that integrates categorical encoding, TF-IDF, and PCA with classification models to evaluate the impact of various encoding methods on predictive accuracy and model performance. The goal is to determine the most accurate model that can automatically predict whether an adverse event should be escalated or dismissed, saving time, effort, and money while achieving more accurate and unbiased adverse event classification.

The first stage of the framework focuses on encoding categorical variables with more than 50 unique values using different techniques. The second stage employs TF-IDF and PCA for feature extraction and selection. The third stage utilizes 11 different classification models to classify whether the adverse event should be addressed or dismissed. Finally, the models will be evaluated using multilevel decision-making with different assumptions, and then the model with the highest score will be selected. After that, sensitivity analysis is applied by testing different principal components to select the optimal number of PCs.

The experimental results indicated that the best prediction model was the Extra Trees Classifier. This model, implemented with 10-fold cross-validation, used binary encoding for categorical variables, TF-IDF for feature extraction, PCA with 50 principal components for feature selection, and an oversampling technique for dataset balancing. It achieved an accuracy of 95.05%, a precision of 95.07%, a recall of 95.05%, an F1-score of 95.05%, and an ROC AUC of 97.78%.