Adverse Reaction Cluster of the COVID-19 Vaccine: Potential Clinical Prediction Tool

Authors: Andrea Gomez, Dung Mai, Mariana Maroto
Graduate Center, CUNY - Machine Learning CSCI 740 Spring 2021

Abstract

This project aims to cluster COVID-19 vaccine adverse reactions. The purpose of the project is one, having a detailed understanding of the common types of adverse reactions, and two, identifying depending on the person's adverse reactions, medical history, allergies, gender, age, and vaccine manufacturer, if they are at a higher risk of suffering a serious effect that could be life threatening.

Introduction

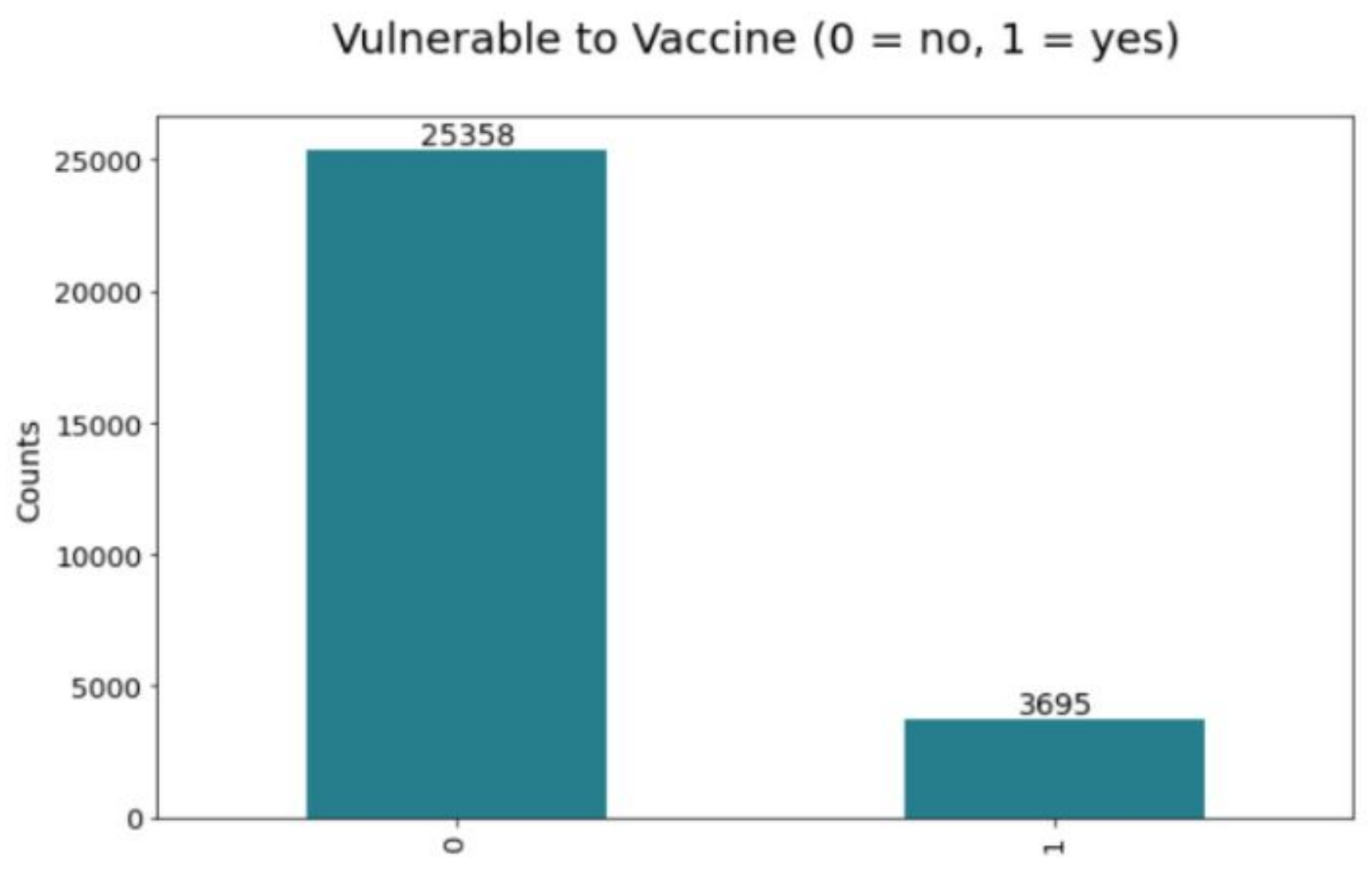
After the COVID-19 virus emerged in late 2019, it triggered an urgent international response to develop a preventive COVID-19 vaccine. On December 11th 2020, United States Food and Drug Administration (FDA) granted Emergency Use Authorization of the COVID-19 Vaccine. Since then, there has been a massive vaccination campaign across the United States. Although generally safe, no prescription drug or biological product, such as a vaccine, is completely free from side effects. Vaccines protect many people from dangerous illnesses, but for a very small percentage its side effects may be serious.

This project uses the data from the Vaccine Adverse Event Reporting System (VAERS), which was created by the Food and Drug Administration (FDA) and Centers for Disease Control and Prevention (CDC) to receive reports about adverse events that may be associated with vaccines. Medical professionals are encouraged to report adverse events, VAERS data will contain coincidental events that might not be related to vaccine. Also, due to the human element in data input we can expect incomplete cases.

Approach

Data Cleaning and Preparing Data

Flagged and removed "bad data" (e.g., duplicate individuals, younger than 16, entries with incomplete information). Performed Natural Language Processing on individual's medical history and allergies information for feature extraction. Additional feature extraction by only keeping most common symptoms in more than 2% of all cases. Defined target variable as individuals who were vulnerable to vaccine. Meaning those who were hospitalized, experienced a life threat reaction, or who deceased. Scaled data with Standardization. Balanced dataset with SMOTE.



Classification

Found the optimal model among: KNN, Naive Bayes, Linear Regression, Support Vector Machine, Multi-Layer Perceptron, Random Forest, and Gradient Boosting.
Used recall and F1 score as evaluation metrics due to the imbalance of the dataset.

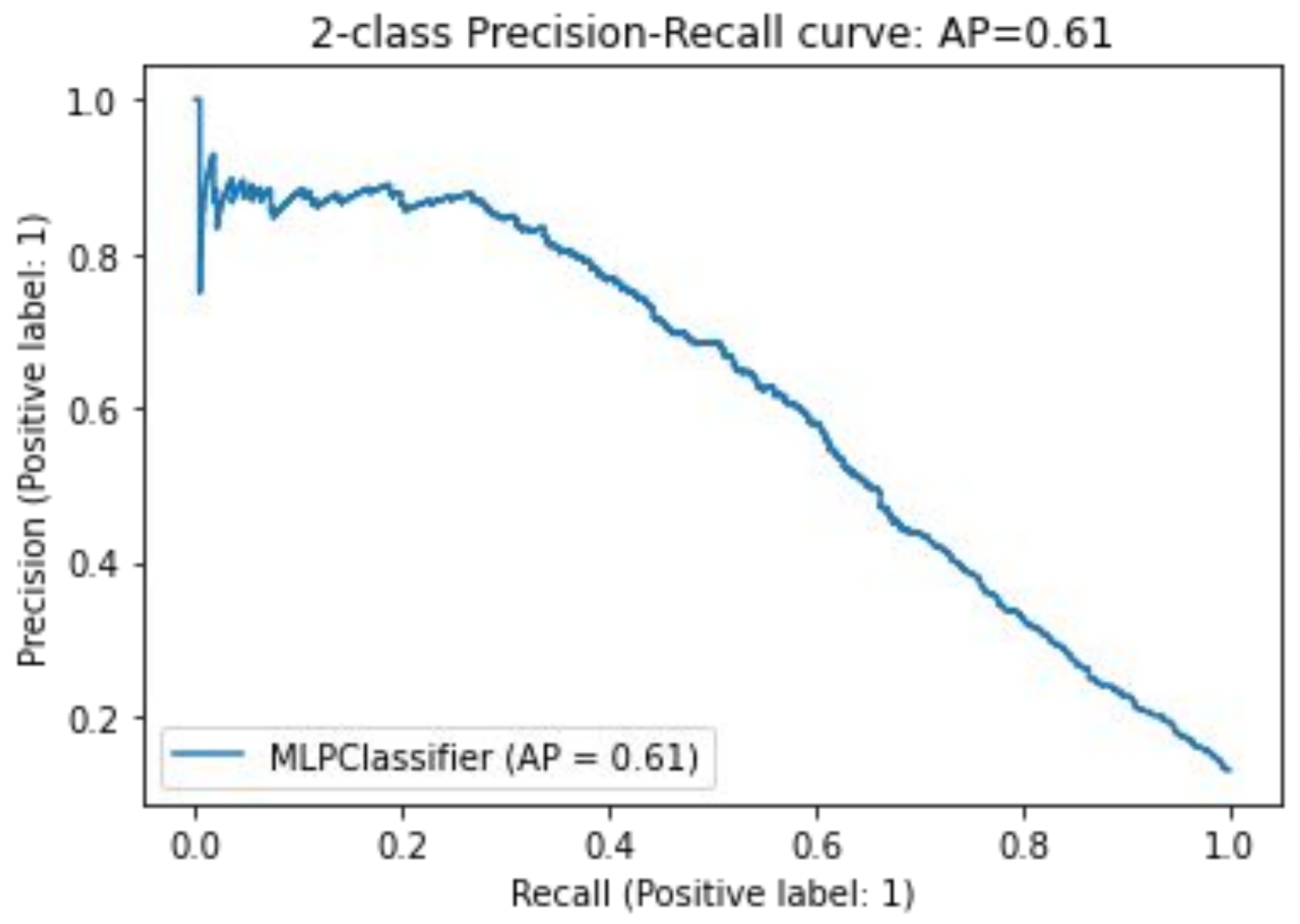
Clustering

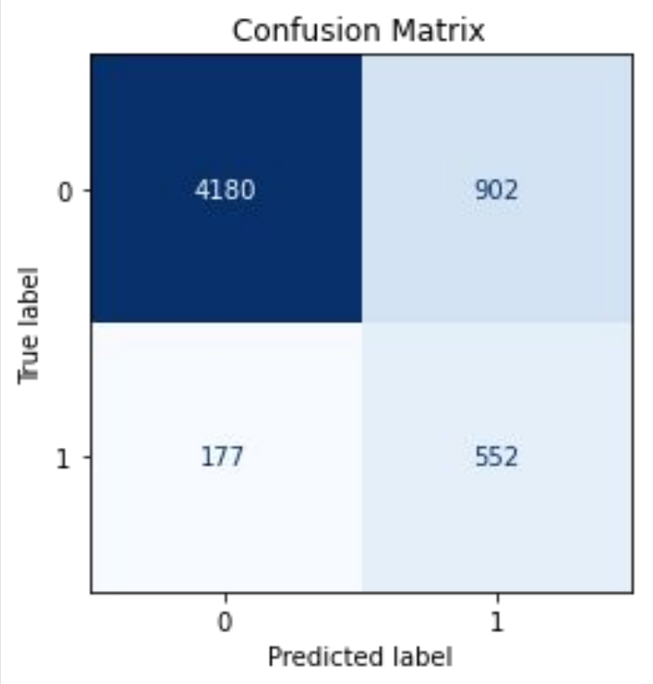
Found the most interpretable and insightful model among: K-Means++, BIRCH, DBSCAN, Spectral Clustering, Gaussian Mixtures, Affinity Propagation Clustering and Agglomerative clustering. Additional crosstab analysis to understand and interpret the clusters and compare their vaccine risk.

Results

Classification

Best classification model was MLP with recall 0.76, precision 0.38, and balanced accuracy of 0.79.

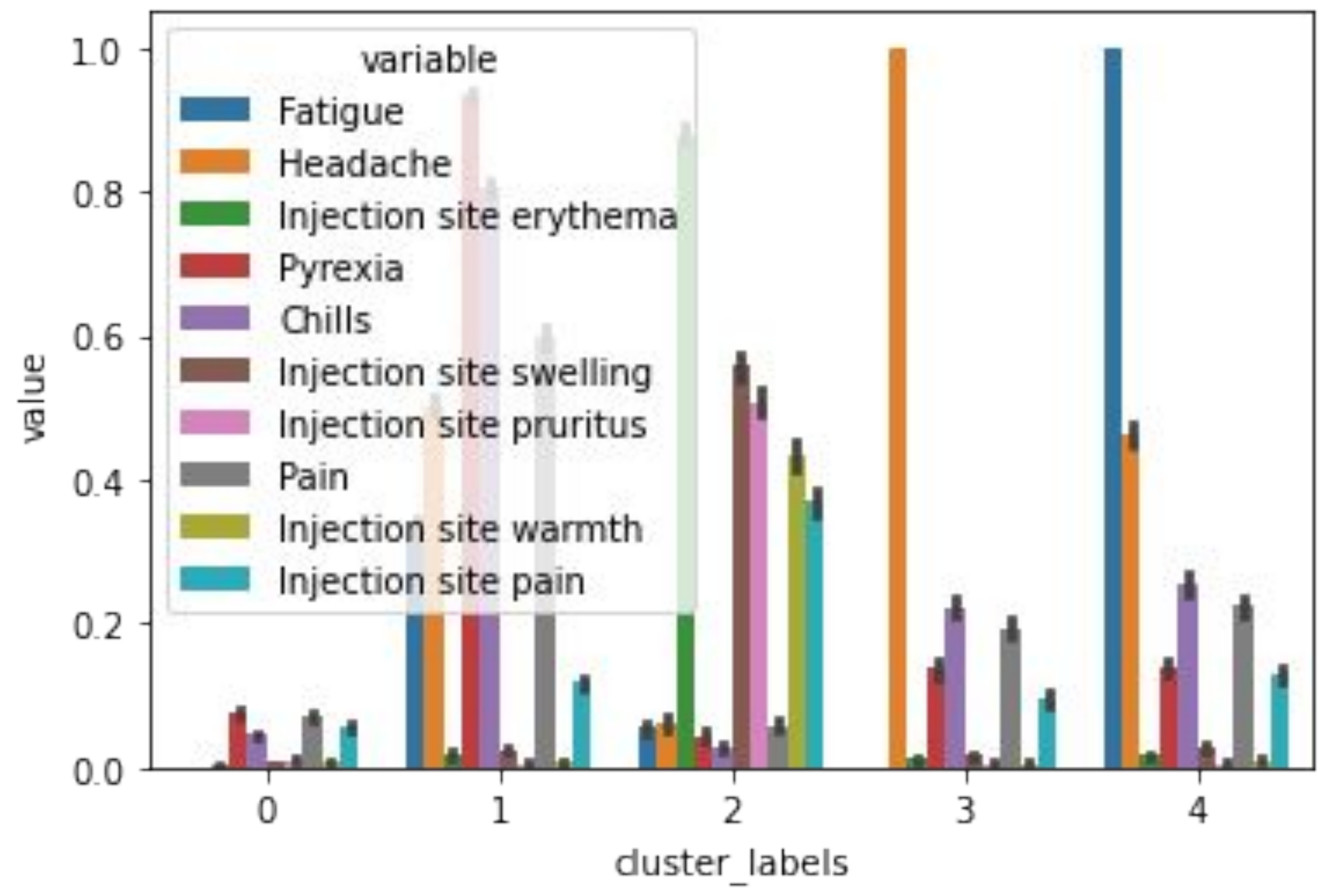


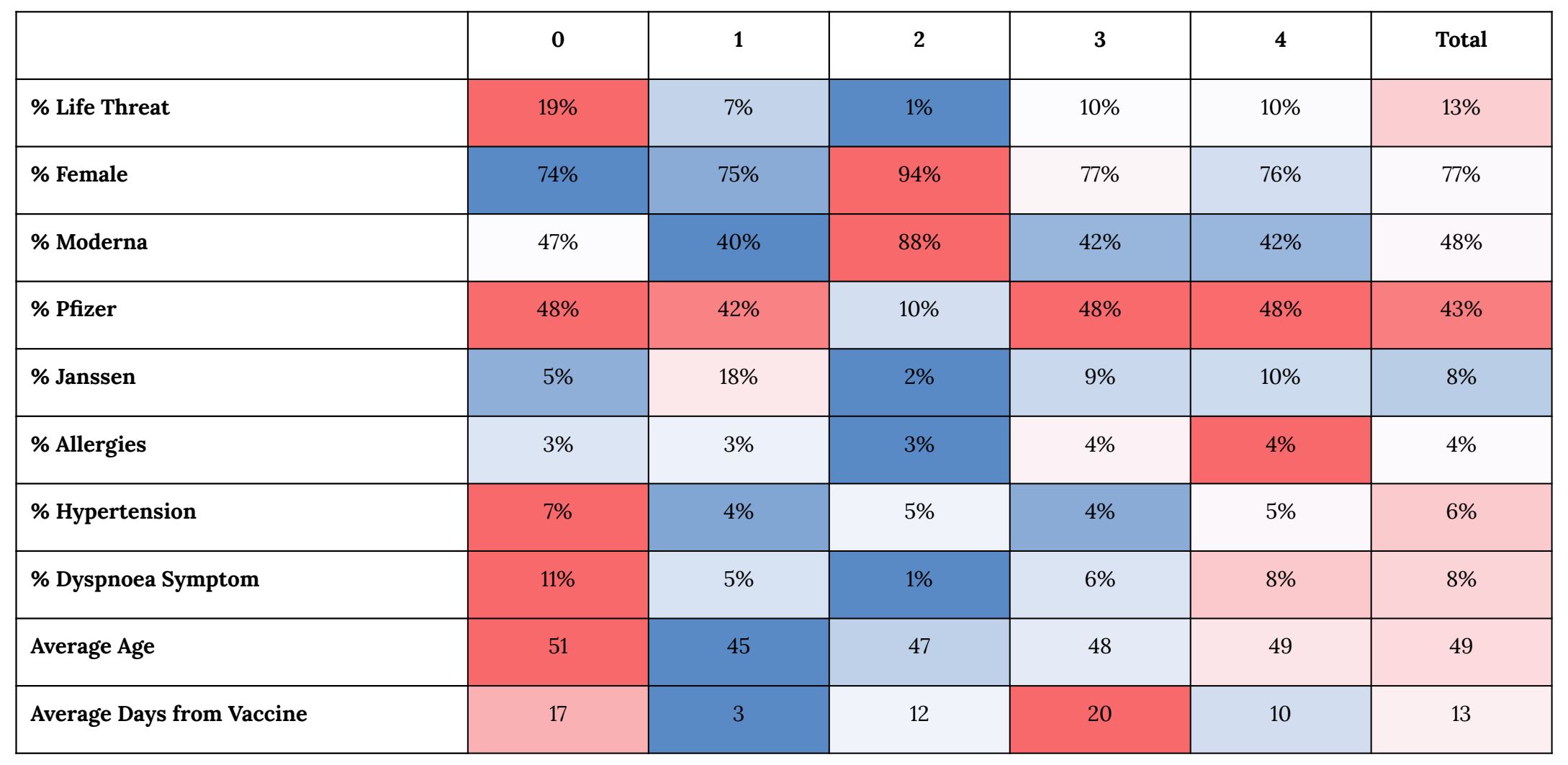


Clustering

K-Means++ yielded the most insightful results as follows:

- Cluster 0: No common symptoms, higher life threat, older population. 46% of sample.
- Cluster 1: Fever, chills, fatigue. Skews male, younger, shortly after shot. 17% of sample.
- Cluster 2: Injection site symptoms only. Low life threat. 9% of sample.
- Cluster 3: Mainly headaches, occurs days later after shot., 13% of sample.
- Cluster 4: Multiple symptoms, more likely to have pre-existing conditions. 15% of sample.





	0	1	2	3	4	Total
% Life Threat	19%	7%	1%	10%	10%	13%
% Female	74%	75%	94%	77%	76%	77%
% Moderna	47%	40%	88%	42%	42%	48%
% Pfizer	48%	42%	10%	48%	48%	43%
% Janssen	5%	18%	2%	9%	10%	8%
% Allergies	3%	3%	3%	4%	4%	4%
% Hypertension	7%	4%	5%	4%	5%	6%
% Dyspnoea Symptom	11%	5%	1%	6%	8%	8%
Average Age	51	45	47	48	49	49
Average Days from Vaccine	17	3	12	20	10	13

Discussion

MLP worked best with our sparse and highly dimensional data. ANNs are able to find non-linear relationships between features that other algorithms could not learn. This is due to feature extraction that occurs in the hidden layers of the network without augmenting the data; Regularization in MLP can turn on and off neurons helping to keep only the most important features of the dataset. More sophisticated neural networks and architectures that follow this process could be explored.

Challenges

The data used in this project is highly sparse, which is common in public datasets, where there can be complete or missing information. For example, the dataset had 4,407 unique symptoms but only 33 were present in more than 2% of cases. This affects the ability to produce accurate predictions and limits the clustering algorithms as well.
The unbalanced dataset, prevalent in the healthcare field, can be misleading by giving a high sense of accuracy, This was taken into account while choosing the classification models by focusing on recall instead.
For the clustering, longitudinal data of individuals will be even more insightful. Symptoms and reactions from a vaccine vary from the time taken the shot to weeks or months after. An individual could have different types of symptoms across different days. This information would allow for a time-series clustering.