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

Andrea Gomez, Dung Mai, Mariana Maroto

Graduate Center, CUNY - Machine Learning CSCI 740 Spring 2021

Problem Description

Background

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

Problem Statement

Our project **classifies** and **clusters** COVID-19 vaccine adverse reactions, with the purpose of identifying which are in need of immediate care and having a deeper understanding of the vaccine adverse reactions.

Solution

1. Classification using common symptoms and patient info to predict need of urgent care
2. Clustering (unsupervised ML) to segment individuals based on adverse reactions

Team Roles

Andrea Gomez

- Exploratory data analysis
- Clustering: Spectral, Gaussian Mixture
- Classification: MLP, SVM
- Model Comparison

Dung Mai

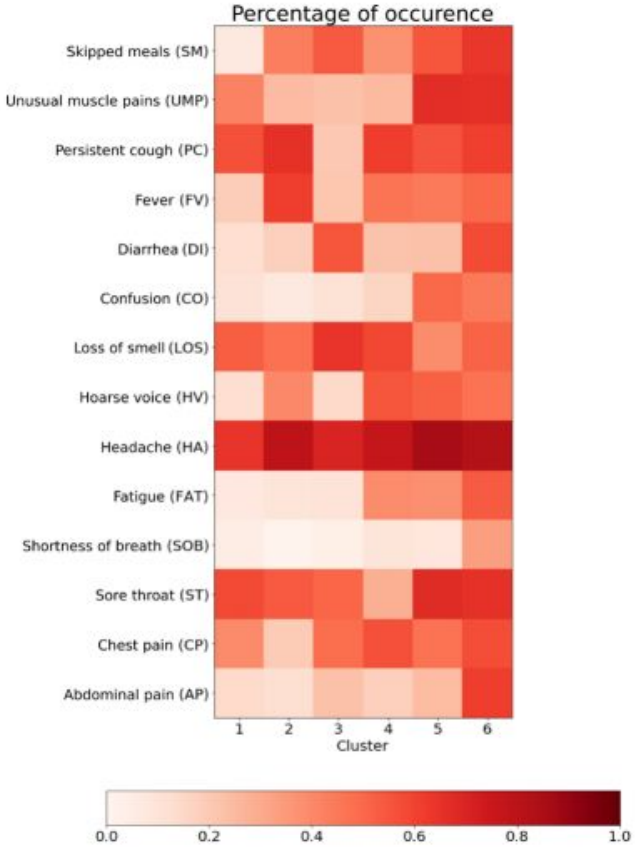
- Clean textual features using NLP
- Clustering: Affinity propagation, Agglomerative
- Classification: LR, NB, Gradient Boosting.

Mariana Maroto

- Clustering: K-means, Birch, DBSCAN
- Classification: KNN, Random Forest
- Clusters Interpretation
- Demo dashboard

Related Work

- "*Symptom clusters in COVID-19: A potential clinical prediction tool from the COVID Symptom Study app*" by C. H. Sudre et al.
- The research performed a unsupervised clustering analysis to study the time series of Covid-19 symptom occurrence.
- Using the clusters, they created a predictive tool that could be used to identify who were in high need of medical care and improve the strategies of maintaining limited medical resources.



About the Dataset

- This project uses the data from the **Vaccine Adverse Event Reporting System** (VAERS), FDA and CDC that receives reports about adverse events that may be associated with vaccines.
- VAERS is used by these government agencies to determine whether any vaccine has a higher than expected rate of rare events.
- Medical professionals are encouraged to report adverse events, even if they are not certain that the vaccination was the cause for the event.
- **Caveats**: Data can contain coincidental events not caused by the vaccine. Human element in the creation of this dataset (incomplete and imperfect in some cases).

VAERS Vaccine Adverse Event Reporting System
www.vaers.hhs.gov

Approach

1. Data Cleaning:

- Combine datasets
- Remove poor quality entries
- Extract features from text with NLP

2. Data Prep:

- One-Hot Encoding
- Feature Scaling
- Balance Dataset: SMOTE

3. Classification:

- Linear: Logistic Regression
- Non-Linear: KNN, NB, Random Forest, MLP, Gradient Boosting

4. Clustering:

- K-Means, Birch, DBSCAN, Spectral, Gaussian Mixture, Affinity propagation, Agglomerative

Unbalanced to Balanced Dataset

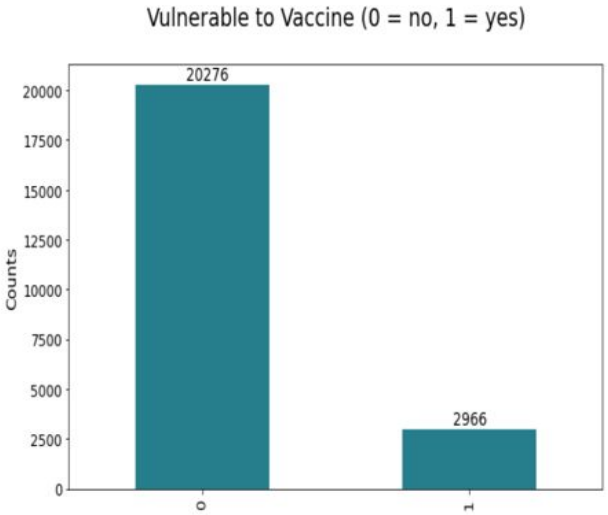


Figure 2: Target variable of training set (refer as y_train). The number of cases which are vulnerable to the Covid-19 vaccines is 12.8% over the total cases.

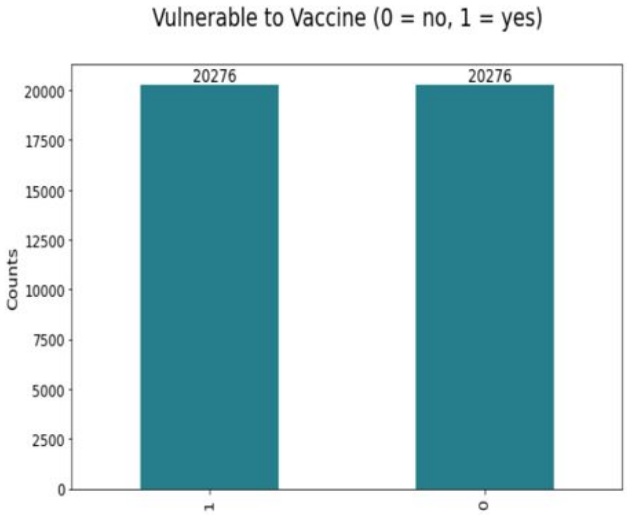


Figure 3: Target variable of training set after oversampling (refer as y_train_smote). The ratio of positive class and negative class is 1:1.

Evaluation Metrics

Classification metrics:

- Recall
- Specificity
- Precision
- F1-Score
- Precision-Recall AUC
- Confusion Matrix
- Balanced Accuracy

Evaluation Metrics

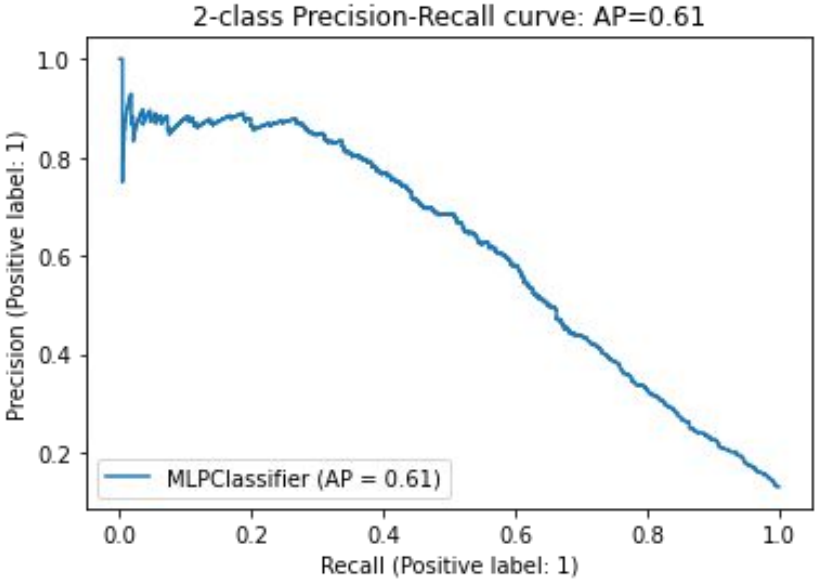
Clustering metrics:

- Domain knowledge
- Follow repetitive patterns found across different algorithms
- Silhouette coefficient

***Classification** results*

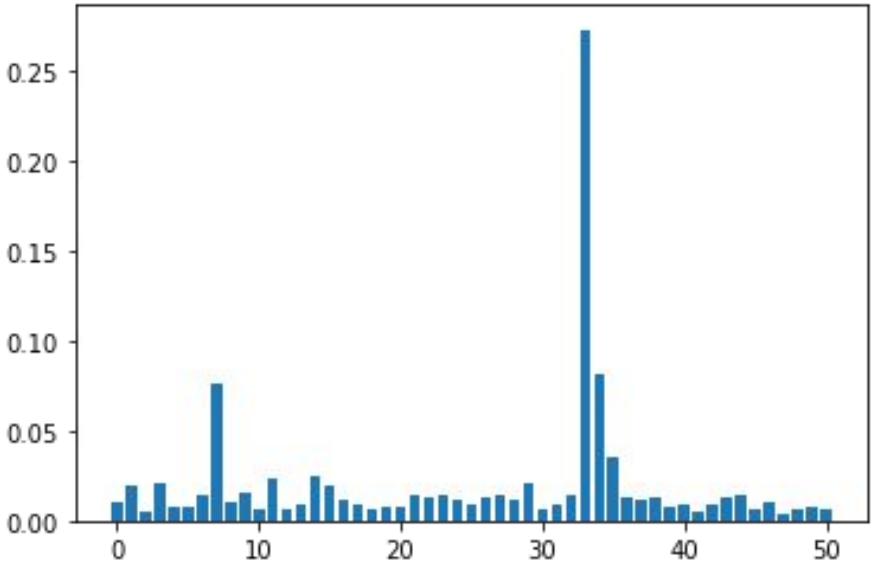
Model	Precision	Recall	Specificity	F1-Score
KNN	0.38	0.61	0.85	0.47
Naive Bayes	0.32	0.76	0.77	0.45
Logistic Regression	0.36	0.77	0.80	0.49
SVM	0.36	0.77	0.81	0.49
Multilayer Perceptron	**0.38**	**0.76**	**0.82**	**0.51**
Random Forest	0.60	0.56	0.95	0.58
Gradient Boosting	0.62	0.57	0.95	0.59

Classification results



N = 6162	Predicted 0	Predicted 1	
True 0	4180	902	5082
True 1	177	552	729
	4357	1454	4732

Feature Importance

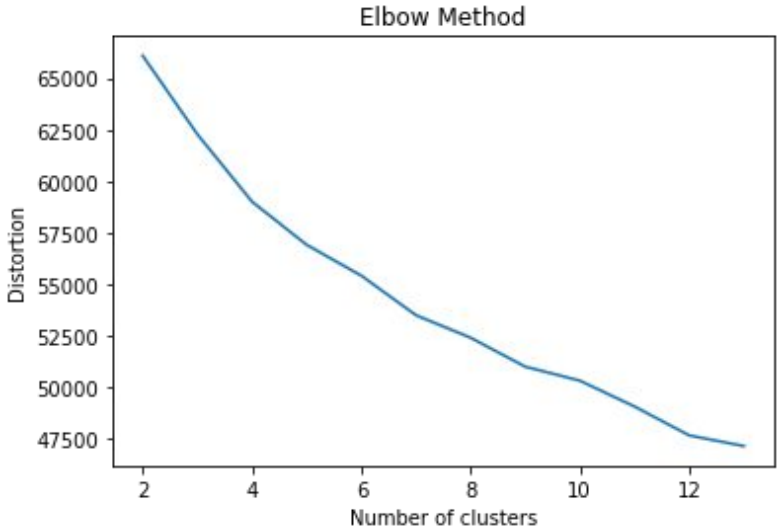


Random Forest feature importance:

- Dyspnoea
- Age
- Number of days from the shot

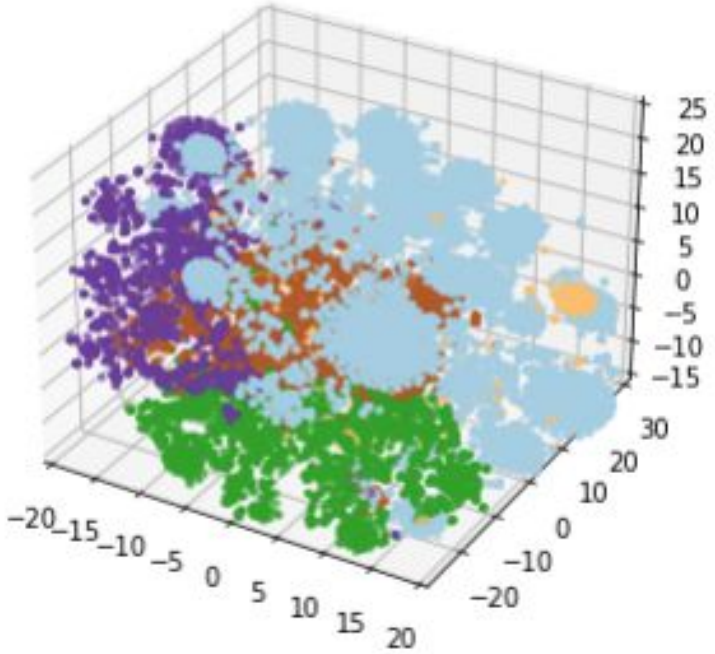
***Clustering** results*

Model	Meaningful, due to nature of algorithm for this dataset
K-Means++ **	**Yes**
BIRCH**	Yes
DBSCAN	No
Spectral **	Yes
Gaussian Mixture **	Yes
Affinity Propagation	No
Agglomerative	Yes



Bold means selected algorithm
** means clustering yielded similar results

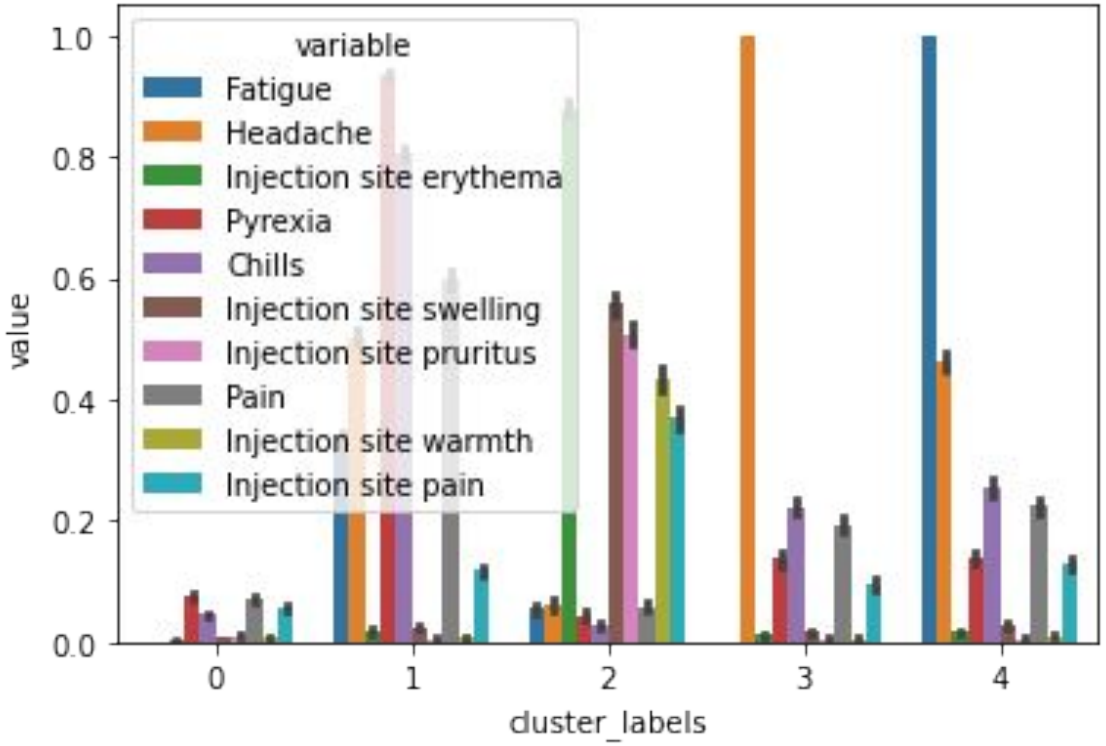
Clustering *results*



Visualizing Clusters Using t-sne:

- t-distributed Stochastic Neighbor Embedding
- tool to visualize high-dimensional data

Clustering *Results*



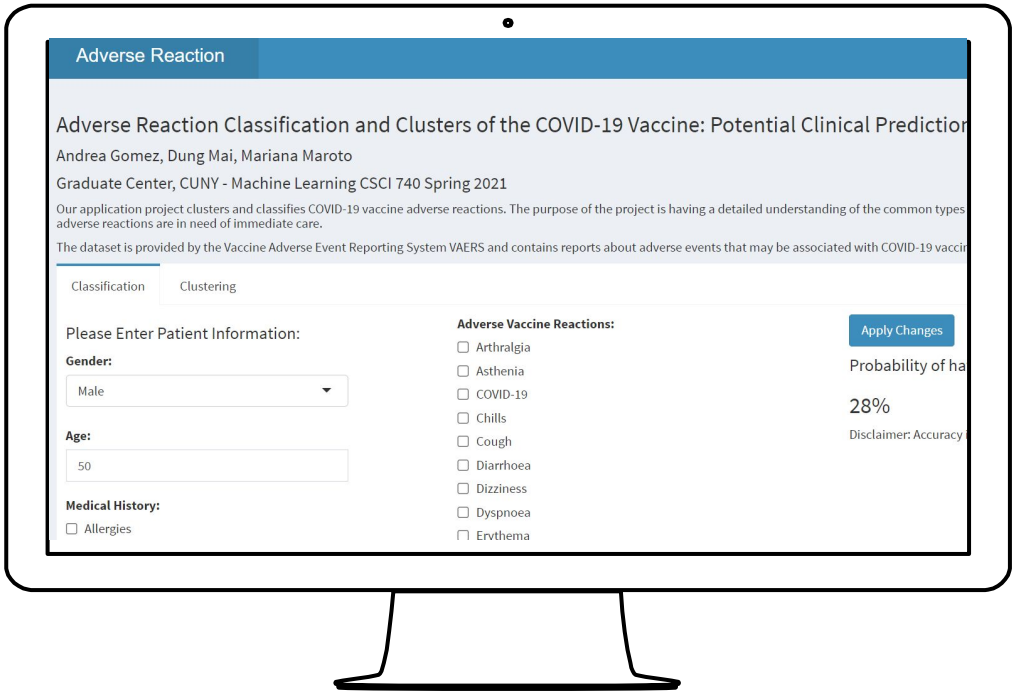
Cluster	Size
0	46%
1	17%
2	9%
3	13%
4	15%

Clustering Results

	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

https://marianamaroto.shinyapps.io/Covid-19AdverseVaccineReaction/

Demo



Online tool that provides patient's probability of life risk and the symptom cluster they belong to

Lessons Learned and Challenges

Lessons Learned

- Data cleaning and data prep in a real world application project is very time consuming
- How to deal with imbalanced dataset (typical with healthcare)
- Applying clustering algorithms is relatively easy but interpreting, visualizing, and generating useful results can be a challenge

Challenges

- Highly dimensional and sparse dataset. Limits clustering algorithms
- For classification, more sophisticated neural networks and architectures that follows this process could be explored
- For the clustering it would be more helpful to have longitudinal data for each individual

Thank you! Any questions?