Machine Learning Approach with Unsupervised Learning Techniques to Find Unseen Patterns in a Promising Future for COVID-19

Paola González Hernández
Tecnologías para la Información en Ciencias
University Nacional Autónoma de México, ENES Unidad Morelia
paaoogh@gmail.com

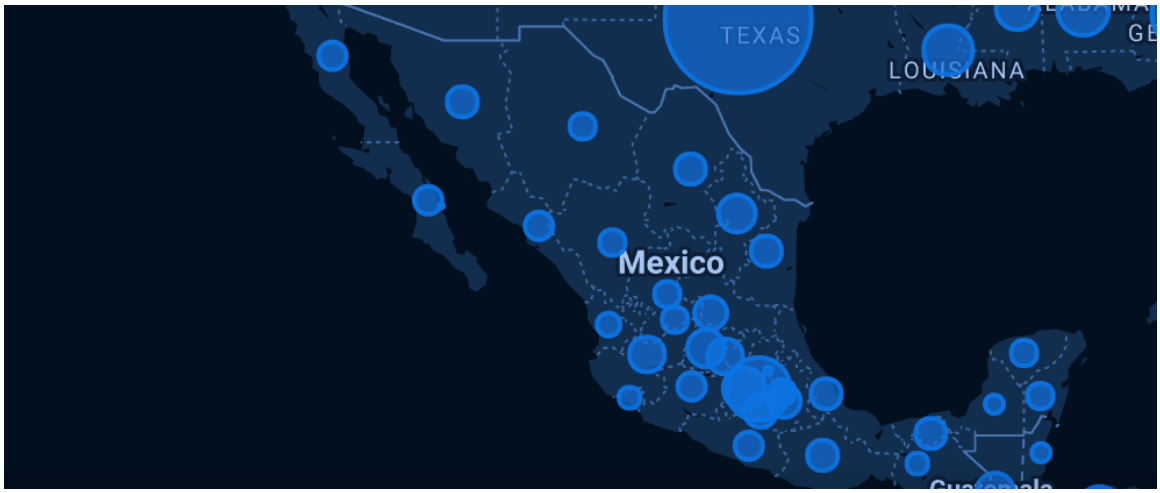

Figure 1: COVID-19 affection in Mexico, February, 20201. Information retrieved from [CSSEGISandData 2021].

ABSTRACT

In this document, unsupervised learning techniques are applied in order to understand data that is being updated daily about how COVID-19 is affecting Mexico from a different side of the story: two different cities are compared in order to find similar patterns. Dimensionality reduction is applied in order to obtain less data for processing in CPU and find how much information is retained in the complete dataset. As much information as possible is gained and processed by Choosing K-Means clustering with Slihouette criteria. The visualization of data is used as support in order to understand what the new clusters represent and how frequent patterns found could considerably suggest that the final status of positive or negative in the test for COVID-19 is not the only important feature to take into consideration when proposing actions in order to make the pandemic easier to overcome as a society.

Data Minng, Semester 2021-1, ENES Morelia

ACM ISBN 978-1-4503-1234-5/17/07.
https://doi.org/10.1145/8888888.7777777

CCS CONCEPTS

• **Computing methodologies** → **Machine Learning**; **Ray tracing**.

KEYWORDS

Machine Learning, unsupervised learning, clustering, features, frequent patterns, information, dimensionality reduction

ACM Reference Format:
Paola González Hernández. 2017. Machine Learning Approach with Unsupervised Learning Techniques to Find Unseen Patterns in a Promising Future for COVID-19. In *Proceedings of Data Minng*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/8888888.7777777

1 INTRODUCTION

The government of Mexico is constantly implementing new strategies for helping the pandemic colateral effects less complicated for the society. The main factor that is frequently considered on "how well is the country doing" is the amount of active positive,

recovered, and death cases. This approach is taken so the rest of the people understand that there is still a latent danger on going back to the life that they were used to. Nonetheless, there are some question that every data scientist always think about when hearing or reading about this factors: How do they actually come up with a mathematical model? What factors are measured in order to retrieve data? Is there another way to reshape data and interpret it that could bring a different perspective of what we are living? In this edition, I am not going to talk about what the virus is, how have other countries managed to overcome a pandemic, nor predict values as Machine Learning is thought to be. Unsupervised Learning, as Géron [Géron 2019] says, is a very intricate part of Artificial Intelligence that the world has not completely explored and understood yet a powerful tool if under the right glass. Sometimes, is used for preprocessing data or feature engineering with data before deep dive into classification; and this is because unsupervised learning is such an important aspect to take in account even though people do not realize it: understanding the data used might be the most important step in the whole modeling and AI process. This not so explored techniques are mainly to understand new things from samples and numbers. In this project guided by the course of Data Mining, a new angle of what to address with the open source data from the government is argued by comparing the capital of Michoacan and Mexico City.

2 EXPOSITION

2.1 Data description and methodology schema

Every state in Mexico has to have a repository of the constant reports of Corovanirus cases reported from March of 2020, this is collected through diverse organisms such as the Consejo Nacional de Ciencia y Tecnología (CONACyT). They are in charge of compiling each case and upload it in the repository that report daily updates [Secretaría de salud 2021]. In here, there is the possibility downloading .cvs files that contain data from last year's March to a specific date. Also, data dictionary is provided with the aim of helping to understand what each value contains and what it represents. Every datum has 40 attributes that cover diverse aspects, such as different health complications (hart deseases, diabetes, autoinmune deseases, among others), age, sex, whether each patient has a result from a specific lab or if it comes from a public dependency. Most of them are a yes-no answer, and almost no numerical data.

There is a situation when working with these type of attributes: dates, states of the country, final results and others. Normalization is more difficult than what it would be if there was only numerical data. As Agresti [Agresti 2007] suggests, when clustering or, majorly, working with unsupervised learning, it is important to identify each "type" of categorical data in order to treat it the best as possible. Consequently, the major problem with this dataset are the dates they have, even encoding them was not an easy decision.

Another situation that emerged during the analysis of the dataset was the computing power. As they are all the registries that have been recorded in almost a year, they are more than 4 million instances. This makes almost imposible to work with the dataset because they are 4,426,668 entries. Fortunately, we are only considering this year's data within the capital of the country, where there are the most cases in Mexico, and Michoacán, more specifically: Morelia. Even with that enclosing, each group has more than a houndred thousand and fourty thousand elements. With this in mind, the process that envolved this project was roughly the following:

(1) **Data cleansing:** data analysis, selection and deletion of attributes.
(2) **Dimensionality reduction:** selecting how many components to preserve with variance ratio and reducing using Principal Component Analysis comparisons.
(3) **Selection of clustering technique:** assessing algorithms, selecting the number of clusters.
(4) **Cluster analysis and frequent patterns:** A priori algoritms.

2.2 Methodology and Experiments

Even though it seemed that there was no missing data, preprocessing had to be made in order to fill some registries that had "pending" or "not specified" fields, mostly in the consumption of tobacco or pregnancy. The amount of data in this condition is very small in comparison to the whole dataset, so it was just eliminated, contrary to the study made in June, as data has been multiplying with the time. Most of the tools used in this project involved the Scikit-Learn library, Mlxtend for the a priori algorithms and some others for visualization purposes (such as Seaborn. Matplotlib, Yellowbrick and Streamlit).

2.2.1 Dimensionality reduction. Two main approaches were made when thinking about what type of dimensionality reduction to make: projection and mainfolding. Because our data is not spread out uniformly in all the dimensions (attributes), projection was the best option. Singular Value Decomposition was also an option to make this projections, although, I chose to go for a more specific target: Principal Component Analysis. In this case, because it is easier for me to understand the variance ratio across components as means of information kept when passing (almost) everything to different hyperplanes. PCA has variants in sklearn: randomized, incremental, kernel, and simple. As randomized requires the whole training set to fit in the memory and due to the lack of a good computing equipment, Incremental and simple PCA was chosen. Figure 2 represents the amount of information retrieved from using PCA on the dataset of the whole country, Morelia and Mexico City on both PCA flavors. There is not a lot of difference in the models performance, in Table 1 there is a numerical visualization of the "goodness" of each dimensionality method.

	1 dim	2 dim	3 dim	4 dim	5 dim	6 dim
PCA - Full data	0.371	0.567	0.702	0.820	0.914	0.943
PCA - Mexico City	0.301	0.546	0.745	0.841	0.886	0.925
PCA - Morelia	0.310	0.574	0.772	0.849	0.901	0.935
IPCA - Full data	0.313	0.469	0.646	0.801	0.913	0.934
IPCA - Mexico City	0.298	0.5006	0.608	0.665	0.778	0.893
IPCA - Morelia	0.249	0.551	0.694	0.846	0.898	0.931

As seen, from having 32 remaining features, with only 5 dimensions we can get 91.4% of the variance ratio. Working with this amount of dimensions is much easier and faster for the models

and, although it might be more challenging interpreting data, the performance with few data is worth it. It is important to point out how interesting is to have reduced by six times the dimension and still not lose more than 10% of the variance ratio in this real life dataset.

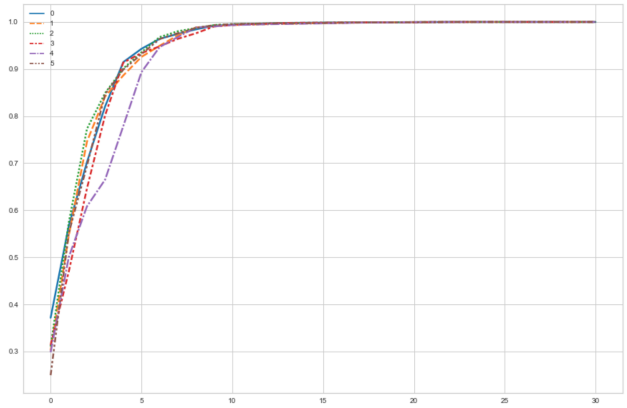

Figure 2: Plottings of PCA: variance ration in function of dimensions.

As the variance preservation is good with only 5 dimensions, the following process of clustering were made with the fitting of this dimensionality reduction in the three small datasets: the country, Mexico City and Morelia. More dimensions could be taken, but as matter of time and resources, the dimensions for the clustering analysis with be taken only with this amount.

2.2.2 Clustering and frequent patterns. The next step was to select the clustering algorithm for the three sets. As the distribution of the data is unknown and not as measurable as it seems, taking a density based algorithm, such as DBSCAN, is not feascible: it defines clusters as continuous regions with high density. On the other hand, using something agglomerative is not an option either due to the computation equipment mentioned above with respect to the amount of data managed. Finally, K-Means was the final option. When plotted into a 3D chart, data revealed clear masses as shown bellow in the Mexico City reduction. This is a pretty good estimator of how many clusters might be in that city dataset; nonetheless, another approach was taken: the "Elbow" method. In this case, subjectively, it was chosen to take values only between 1 and 20 for choosing the number of clusters.

It turns out that having 6 clusters is the most optimal for the case of the country and the city of Morelia, Mexico City only increments by one, which is not completely worthy to change, given than 6 clusters also are feasible as the inertias are pretty similar in that specific area of the curve.

On the other hand, in order to make the frequent patterns, it was necessary to process again the data as the encoder of Mlxtend needs more than just numbers but categories in each feature for every observation. And the support that collated the best patterns was initially 0.5 and started iterating manually until a reasonable number of patterns were found: it is of remark that the most repetitions was 17 but with barely 0.50 support. This ended in a join of 3

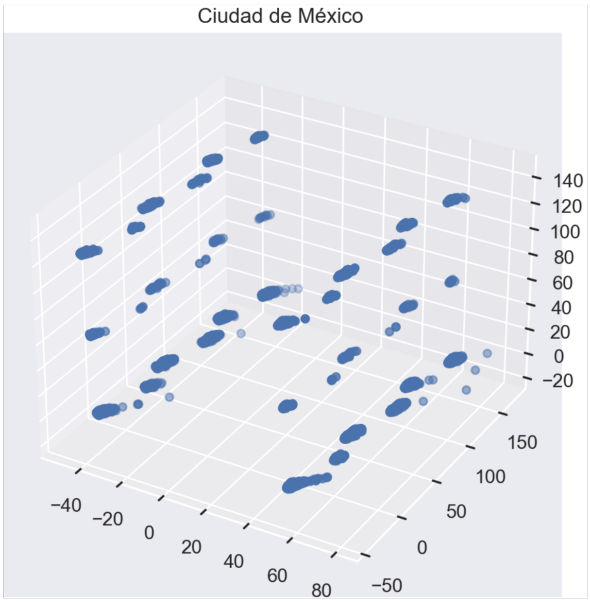

Figure 3: Mexico City data projection into a 3-dimensional hyperplane

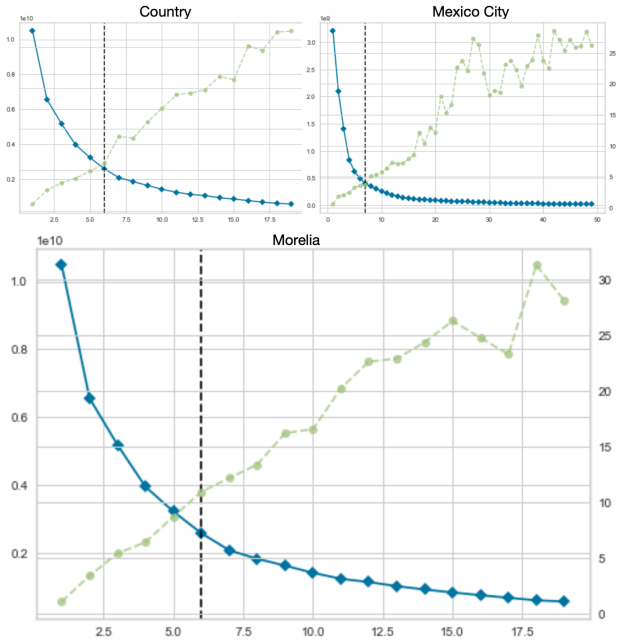

Figure 4: Mexico City data projection into a 3-dimensional hyperplane

patterns that were clear in the country, and Morelia; the most common features that appeared together were not being of indigenous origins or language knowledge, plus, the following:

- No diabetes, nor heart deseases, EPOC, autoinmune deseases, renal complications or other complications.
- No pneumonia, heart deseases, EPOC, autoinmune deseases, asma, nor renal or other complications.
- Non smokers, no asma, EPOC, autoinmune deseases, nor renal or other complications.

2.3 Discussion, Conclusions and Future Work

After analyzing clusters had in common, aside from having the final result of the COVID-19 test, they had almost nothing in common but the ages in some cases. Older people, from 50 and above had more health complications related to renal deseases, diabetes or hypertension; on the other hand. More frequently than not, deseases like EPOC or immunological supression are found in groups with the age average lower.

Aside from this: there is a point that is remarkable from the whole dataset. When reducing dimensionality, almost not variance was lost even when taking few components. This actually says a lot about data. There are features that might be not so important as we believe when estimating models. A future step would be analyzing more states of the country with this same methodology and find out if there is an outlier about how many components to take, because, as long as these results show, more feature engineering might be good for any methodology implemented with this dataset.

Talking about doing more things with more states, continue looking at clustering could also be beneficial for the states: if each state took a serious look at data mining and used information wisely, every state could look at what their population communalities are and attend directly those groups with a more efficient strategy.

Lamentably, in this case, the dataset was too large for the computing power that a simple CPU could achieve. A further step could involve migrating from using dataframes to tensors or trying to use GPU to process all the information that this dataset is still hidding. Having more data, sometimes, means that you have more variety, that is only meaning that there are more folds, twists and turns that we haven't seen. Maybe, from one of this folds that is still hidding, we could get a different classification criterion and could be applied to supervised or semi-supervised learning.

ACKNOWLEDGMENTS

I would like to thank immensely to our subject teacher Marisol Flores: because you have never lost faith in your students and have given them more and more opportunities to overcome trouble and learn in the process. Thank you for the opportunity. Also, I will thank to my parents and brother that have been supporting me through this new stage of my life.

National Natural Science Foundation of China under Grant No.: 61273304 21 and Young Scientists' Support Program (http://www.nnsf.cn/youngscientists).

REFERENCES

Alan Agresti. 2007. *An Introduction to Categorical Data Analysis.* OWilley.

et al. CSSEGISandData. 2021. COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. https://github.com/CSSEGISandData/COVID-19.

Aurélien Géron. 2019. *Hands-On Machine Learning with Scikit-Learn, Kearas and TensorFlow.* O'Reilly.

et al. Secretaría de salud. 2021. Datos Abiertos Dirección General de Epidemiología. https://www.gob.mx/salud/documentos/datos-abiertos-152127.