MDPI

*Article*

# Covid v/s Food

**Jose Antonio Castro [1,‡], Juan Vicuña [2,‡], Emily Skog [3,‡] and Clemente Cambara [4,‡]**

[1] Departamente de Ciencias de la Computación, Escuela de Ingeniería, Pontificia Universidad Católica de Chile, Santiago, Chile; jacastro18@uc.cl
[2] Departamente de Ciencias de la Computación, Escuela de Ingeniería, Pontificia Universidad Católica de Chile, Santiago, Chile; juan.vg@uc.cl
[3] Instituto de Ingeniería Biológica y Biomédica, Escuela de Ingeniería, Pontificia Universidad Católica de Chile, Santiago, Chile; emily.skog@uc.cl
[4] Departamente de Ciencias de la Computación, Escuela de Ingeniería, Pontificia Universidad Católica de Chile, Santiago, Chile; ccambara.d@uc.cl
‡ These authors contributed equally to this work.

## 1. Introducción

Este trabajo busca responder la interrogante: ¿Cómo se relacionan los hábitos alimenticios de distintos países con el impacto biológico que ha tenido el virus SARS-CoV-2 en cada uno de ellos? Para esto, se han usado datasets de números actualizados de casos totales confirmados, fallecidos y activos para 218 países [1] junto con datasets de porcentaje de ingesta alimentaria para estos países [2]. El plan es unir los datasets más relevantes utilizando los países como llave y realizar diversos análisis mediante el uso de técnicas de procesamiento de datos vistas en clase.

*Impacto del problema*

Debido al impacto que el virus SARS-CoV-2 ha tenido en el mundo y considerando lo poco que se sabe de su comportamiento, es relevante estudiar distintos factores que tradicionalmente no son relacionados con la pandemia de Covid-19 y analizarlos a nivel estadístico para determinar la relación entre ellos. De esta manera, se espera entender más las implicancias que distintas circunstancias tienen sobre el virus y con esta información poder combatirlo mejor.

Se decidió investigar específicamente la alimentación dado que es un pilar fundamental en la salud y no suele ser correctamente dimensionada al momento de relacionarla con distintas complicaciones biológicas.

## 2. Metodología

El primer paso fue limpiar los datos y agruparlos en una matriz de tal forma que cada entrada respresentara un país y cada columna tuviera la siguiente información respecto a los hábitos alimenticios:

**Citation:** Castro, J.A.; Vicuña, J.; Skog, E.; Cambara, C. Covid v/s Food. *Journal Not Specified* **2021**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

| |
|---|
| Alcoholic Beverages |
| Animal Products |
| Animal fats |
| Aquatic Products, Other |
| Cereals - Excluding Beer |
| Eggs |
| Fish, Seafood |
| Fruits - Excluding Wine |
| Meat |
| Milk - Excluding Butter |
| Miscellaneous |
| Offals |
| Oilcrops |
| Spices |
| Starchy Roots |
| Stimulants |
| Sugar Crops |
| Sugar & Sweeteners |
| Treenuts |
| Vegetal Products |
| Vegetable Oils |
| Vegetables |
| Obesity |

**Tabla 1:** Esta tabla representa los datos de alimentación utilizados para determinar si existe una relación entre estos y el avance del coronavirus en el mundo. Es importante notar que cada columna representando un tipo de alimento es una proporción calculada como $\frac{\text{kcal totales consumidas}}{\text{cantidad de habitantes en el país}}$ mientras que la columna **Obesity** es el porcentaje de gente con obesidad respecto a la cantidad total de habitantes en ese país.

Se midió la relevancia de esta matriz $\mathbf{X}$ de datos para predecir vectores relacionados al avance del virus SARS-CoV-2 en cada país de distintas maneras. Primero, dado un vector $\mathbf{y}$ que representa la proporción por país de $\frac{\text{casos confirmados de coronavirus}}{\text{cantidad de tests totales realizados}}$, se utilizó un método de regresión lineal entrenado con esta matriz para que predijera dos variaciones del vector $\mathbf{y}$

1. Si la proporción de casos confirmados para un país particular esta sobre la media mundial (vector binario con valor 0 si esta proporción está bajo la media y 1 si está por sobre).
2. El cuartil en el que se encuentra esta proporción respecto al resto de países

Luego, se realizó esta misma metodología con un clasificador de Árbol de Decisiones, esta vez haciendo un testeo de hiper-parámetros en ambas versiones del vector $\mathbf{y}$ para encontrar aquellos que maximizaran el `accuracy_score`. En particular, se probó con las combinaciones de criterios `gini` y `entropy` y 100 profundidades distintas para maximizar la eficacia del clasificador.

De manera alterna, se decidió probar un clasificador SVC para modelar el problema, pero aun con un estudio pertinente de hiperparametros, no se logró un precisión lo adecuadamente alta como para consiferar útil. En más detalles, la mejor precisión (0.4) fué obtenida con kernel `poly` de con un valor de C de 7.

## 3. Materiales y Métodos

*Materiales*

Para obtener los resultados del proyecto, es necesario usar datasets de números actualizados de casos totales confirmados, fallecidos y activos para 218 países junto con datasets de porcentaje de ingesta alimentaria para estos países. Estos, como se explico

anteriormente, son utilizados para originar un nuevo dataset contenedor donde se pueda realizar diversos análisis. Para obtener los datos se pueden utilizar nuestra fuente de datos, la página web más grande de la comunidad de la ciencia de datos y análisis del trafico web: Kaggle. [1] [2].

*Métodos*

Diversas técnicas de procesamiento de datos fueron utilizadas para desarrollo del proyecto, ya sea para la limpieza de datos como para la clasificación de datos.

En primer lugar, para la unión de los datasets antes mencionados y la limpieza de datos, se hizo uso de la librería Pandas [3] junto con sus funciones integradas (como $df.dropna()$ o $df.merge()$). Luego se llevó a cabo una función formulada por el equipo que calcula la correlación entre las columnas del dataset, para así quitarlas de este mismo.

Con el dataset ya preprocesado, se utilizó un método agrupador para formular el vector **y** según el criterio determinado (según la media mundial o según los cuartiles).

Teniendo los datos listos para entrenar (como un vector **X** y otro **y**), se eligió la librería SkLearn [4] para realizar el entrenamiento de las regresiones lineales, los árboles de decisión y el clasificador SVC. Para lograr lo anterior, las siguientes funciones más relevantes de la librería que fueron utilizadas son:

| Funciones SKLearn utilizadas | |
|---|---|
| Función | Explicación |
| $train\_test\_split()$ | Divide el vector **X** en *X_train*, *X_test* y el vector **y** en *y_train*, *y_test* respecto a una proporción deseada (0.2 para el proyecto). |
| $linear\_model.LinearRegression()$ | Crea un modelo de Regresión lineal a ser entrenado. |
| $tree.DecisionTreeClassifier()$ | Crea un arbol de desición de datos que clasifica los datos con nodos internos con atributos. |
| $metrics.accuracy\_score$ | Retorna la precisión de una predicción respecto a sus valores verdaderos. |
| $svm.SVC()$ | Crea un modelo de SVM a ser entrenado. |

Para el estudio de hiperparametros se utilizó una simple función iterativa que encontrara la mejor precisión para un conjunto de valores.

## 4. Resultados

Se mostraran los resultados generales del proyecto realizado en base a dos definiciones distintas de las muestras procesadas, es decir, en base a las definiciones del vector **y** y de la matriz **X**.

### *4.1. Resultados Muestra Casos Confirmados / Test Totales*

Para esta sección, el vector **y** es definido como $\frac{\text{casos confirmados de coronavirus}}{\text{cantidad de tests totales realizados}}$ y la matriz **X** como las columnas restantes del dataset.

Mostraremos dos tabla con todos los clasificadores modelados y los resultados pertinentes. Cada tabala estara basada según cada variación del vector **y** que exista. Siendo mas especificos, mostraremos los resultados para cuando el vector **y** esta basado en la media mundial de la proporción (asignando 0 y 1) y mostraremos otra tabla para cuando el vector **y** esta basado en los cuartiles de la proporción (asignando 0, 1, 2 y 3).

| Vector **y** Media Mundial | |
|---|---|
| Clasificador | Resultados |
| Regresión Lineal | Error de media cuadratica: 0.14 y coeficiente de error cuadratico $r^2$ de 0.34. |
| Árbol de decisión | Mejor resultado posible con hiperparametros: $criterion = gini$ y $max\_depth = 20$. Precisión obtenida de un 74.07% |
| SVM | Mejor resultado posible con hiperparametros: $kernel = poly$, $C = 1$ y con una presición del 0.64. |

| Vector **y** Cuartiles | |
|---|---|
| Clasificador | Resultados |
| Regresión Lineal | Error de media cuadratica: 0.38 y coeficiente de error cuadratico $r^2$ de 0.78. |
| Regresión Lineal Entrenada | Error de media cuadratica: 0.14 y coeficiente de error cuadratico $r^2$ de 0.34. |
| Árbol de decisión | Mejor resultado posible con hiperparametros: $criterion = gini$ y $max\_depth = 2$. Precisión obtenida de un 77.7% |
| SVM | Mejor resultado posible con hiperparametros: $kernel = poly$, $C = 7$ y con una presición del 0.4. |

### *4.2. Resultados Muestra Muertes / Tamaño población*

Para esta sección, el vector **y** es definido como $\frac{\text{muertes totales}}{\text{cantidad de población}}$ y la matriz **X** como las columnas restantes del dataset.

Mostraremos dos tabla con todos los clasificadores modelados y los resultados pertinentes. Cada tabala estara basada según cada variación del vector **y** que exista. Siendo mas especificos, mostraremos los resultados para cuando el vector **y** esta basado en la media mundial de la proporción (asignando 0 y 1) y mostraremos otra tabla para cuando el vector **y** esta basado en los cuartiles de la proporción (asignando 0, 1, 2 y 3).

| Vector **y** Media Mundial | |
|---|---|
| Clasificador | Resultados |
| Regresión Lineal | Error de media cuadratica: 0.20 y coeficiente de error cuadratico $r^2$ de 0.6. |
| Árbol de decisión | Mejor resultado posible con hiperparametros: $criterion = entropy$ y $max\_depth = 14$. Precisión obtenida de un 51.85% |
| SVM | Mejor resultado posible con hiperparametros: $kernel = rbf$, $C = 3$ y con una presición del 0.76. |

| Vector **y** Cuartiles | |
|---|---|
| Clasificador | Resultados |
| Regresión Lineal | Error de media cuadratica: 0.664 y coeficiente de error cuadratico $r^2$ de 0.27. |
| Árbol de decisión | Mejor resultado posible con hiperparametros: $criterion = gini$ y $max\_depth = 2$. Precisión obtenida de un 77.7% |
| SVM | Mejor resultado posible con hiperparametros: $kernel = poly$, $C = 3$ y con una presición del 0.55. |

## 5. Discusión

Al analizar los resultados de los casos confirmados divididos por el total de tests, se puede notar que ciertos clasificadores generan una presición más alta. Al segmentar el vector **y** según la media se puede apreciar la mejor presición con el clasificador de árbol de decisión, variando los hiperparámetros (criterion=gini y max_depth=20). Mientras que los peores resultados se dan con la regresión lineal. Esto tiene sentido, en el ámbito que la información del data set es no lineal, no sigue un patrón evidente. También, en el árbol de decisión al usar un "max_depth" de valor moderado, no se genera "overfitting".

Luego, al cambiar la información del vector **y** en cuartiles, se pude evidenciar resultados similares: la regresión lineal entrenada genera presición baja y el árbol de decisión la más alta (criterion=gini y max_depth=2).

Por otro lado, al analizar los resultados de los muertos divididos por la población, nuevamente se puede notar que ciertos clasificadores generan una presición más alta. Al segmentar el vector **y** según la media se puede apreciar la mejor presición con *Support Vector Machines* (SVM), variando los hiperparámetros (*kernel*=rbf y C=3). Mientras que los peores resultados se dan con el árbol de decisión. Esto tiene sentido, en el ámbito que la información del data set es no lineal, no sigue un patrón evidente. Por lo que el kernel del SVM, que es no lineal, se ajusta de la mejor manera a esta información. También, sus valores de *slack* son un poco mayor que "1", por lo que da un poco de flexibilidad al segmentar la información.

Luego, al cambiar la información del vector **y** en cuartiles, se pude evidenciar resultados distintos: la regresión lineal entrenada genera presición baja y el árbol de desición la más alta (criterion=gini y max_depth=2), notando que actua de manera similar a la taza de incidencia de casos.

Todos estos resultados han explicitado que hay cierta correlación entre la manera de alimentación de los países con la taza de muertos y taza de incidencia a raíz de COVID-19, ya que se genera una presición para todos los casos mayor al 70%, lo cual es

suficiente para decir que no es mera coincidencia.

Sin embargo, la manera de segmentar el vector **y**, aunque demuestra una correlación, no es muy útil al momento de la aplicación, ya que solo nos puede decir si un país (a partir de su alimentación) tiene un riesgo leve, moderado, alto, muy alto, de contraer la enfermedad o morir de ella. No es muy específico.

Tampoco se analizó ciertas alimentaciones en particular, como consumo de alcohol o verduras. Lo cual habría sido interesante ver la tendencia de los países solo analizando ciertas columnas del data set.

## 6. Conclusión

Se puede concluir que se llegó a clasificar de manera exitosamente la incidencia y muertes a raíz del COVID-19 a partir de la alimentación de cada país. Se llegó para todos los casos a precisiones mayores al 70%, con el más alto rondando el 78%. En todos los casos esto fue con clasificadores no lineales, SVM y árbol de decisión, variando los hiperparámetros.

No obstante, a futuro se desea de una manera segmentar el vector **y**, para dar información más clara de la gravedad de la enfermedad en el país en cuestión, no solo en 2 o 4 categorías. Además, se podría entrenar con otros métodos, como redes neuronales, que abarcan la no linealidad del problema.

También, sería interesante analizar como ciertas alimentaciones afecta la incidencia del virus en cada país, ¿cómo si se alimentan con más carne o más frutos tienen diferentes resultados?. Por lo que todavía hay hartas interrogantes que se pueden responder de este dataset, independiente de su falencias.

## 7. References

1. Joseph Assaker, Covid-19 Global Dataset: Up to date numbers of daily Confirmed, Death and Active cases for 218 countries
2. Food and Agriculture Organization of the United Nations
3. Documentación de Pandas
4. Documentación de Scikit Learn