Variables relevantes en la propagación y letalidad del COVID-19: Un estudio global

Departamento de Ingeniería Matemática

Universidad de Chile

Benjamín Barrientos*, Guillermo Dinamarca†, Matías Romero‡, Francisco Vásquez§.
Bajo la guía de Servet Martínez¶

22 de junio de 2020

Resumen

La enfermedad por coronavirus 2019 (COVID-19) es causada por un virus recientemente descubierto en China, declarado pandemia el 11 de marzo de 2020. A la fecha, ha causado más de 400,000 fallecidos y se ha propagado rápidamente por todo el mundo, alcanzando la cifra de 7 millones de infectados. Es por esto, que nace la necesidad urgente de conocer su comportamiento en distintas regiones del mundo y estudiar sus propiedades para hacerle frente con efectividad. Este estudio busca entender mejor la enfermedad usando métodos de análisis de datos y aprendizaje de máquinas. Para ello se consideran distintas variables representativas de la situación de cada país (estructurales, propias del coronavirus y controlables) para predecir la tasa de letalidad de la enfermedad, y el número de infectados y de fallecidos por millón de habitantes. La tasa de letalidad es el número de fallecidos totales sobre el numero de infectados totales y solo se puede conocer al término de la pandemia, para simplificar su estimación, solo se consideran los números hasta el momento de desarrollo de la pandemia . En primera instancia se aplican técnicas de reducción dimensionalidad y visualización para explorar los atributos de cada país, y luego aplicar técnicas de regresión (de las cuales se destacan XGBoost y RandomForest) para tratar de explicar las variables antes mencionadas. Se obtuvo que las variables más importantes son el porcentaje del PIB dedicado a salud y el número de especialistas por 1000 habitantes para la explicación de la letalidad, el porcentaje de población urbana y el número de tests por mil para el número de infectados por millón, y el tiempo de pasar de 100 a 1000 infectados junto con el ya mencionado porcentaje del PIB en salud para el número de fallecidos por millón. Se concluye la vital importancia de la capacidad de testeo para la correcta detección y posterior seguimiento de los infectados, lo cual se complementa con la alta influencia del porcentaje de población joven en la transmisión del virus, aumentando así la necesidad de medidas de distanciamiento social y cuarentenas. Para la explicación de la tasa de mortalidad y de letalidad se recuperan resultados lógicos sobre la importancia de la situación económica y sanitaria del país, además de la evidente dependencia de la primera con el número de infectados, para ambas tasas el tiempo de cien a mil infectados logra explicarlas relativamente bien, lo cual se atribuye a la respuesta de los países al avance de la enfermedad y qué tan eficaces fueron para controlarla en etapas tempranas.

Keywords. COVID-19, Coronavirus pandemic, Machine learning, Johns Hopkins Coronavirus database, Worldometers Coronavirus database.

I. Introducción

Dado el contexto actual, se hace imperante el estudio del COVID-19 con tal de conocer sus impactos además de saber cómo poder afrontar esta pandemia. Las herramientas de análisis de datos y aprendizaje de máquinas pueden ser muy útiles en lo anterior. Uno de los índices de gran interés es el valor L, el cual se define como el cociente entre el número de fallecidos y el número total de confirmados con COVID-19. ¿Cuáles son variables de un país que más influyen en el valor de éste? ¿Es posible que a través de la información de otros países se pueda realizar una predicción del índice L de otro país? Estas son las interrogantes que se buscan responder en el siguiente documento.

El presente documento contiene el resultado final de un estudio realizado en varias etapas. A modo de resumen, en una primera iteración realizada se buscó predecir el valor L como tal,

*Estudiante Ing. Civil Matemática; Universidad de Chile; Santiago, Chile. email: bbarrientos@dim.uchile.cl
†Estudiante Ing. Civil Matemática; Universidad de Chile; Santiago, Chile. email: gdinamarca@dim.uchile.cl
‡Estudiante Ing. Civil Matemática; Universidad de Chile; Santiago, Chile. email: meromero@dim.uchile.cl
§Estudiante Ing. Civil Matemática; Universidad de Chile; Santiago, Chile. email: fvasquez@dim.uchile.cl
¶CMM-DIM; Universidad de Chile; Santiago, Chile. email: smartine@dim.uchile.cl

considerando numerosas variables para países solo de la OCDE, variables que consideraban distintos aspectos del país y no solo relacionadas con la enfermedad. Cabe mencionar que el número de variables se disminuyó en la misma iteración con tal de reducir colinealidad y reducir el número de variables no relevantes. Se aplicaron diversos modelos de regresión, como XGBoost, Gradient Boosting, Adaboost, SVR, etc. Los mejores resultados en cuánto a tener un bajo error cuadrático medio (RMSE por sus siglas en inglés) con respecto a los demás modelos fue XGBoost. Sin embargo, al ahondar más en el modelo se llega a que solo es posible su aplicación en un determinado punto de la evolución de la enfermedad para los países y con poco valor predictivo para el futuro. Sin embargo, el trabajo realizado en esta primera etapa permitió saber la relevancia que tenían ciertas variables para el modelo, por lo cual se decidió un cambio de foco en el estudio, enfocándose ahora en estudiar qué variables son las que impactan en la estimación del índice L, esto con tal de poder saber si es posible generar un impacto en el índice solo cambiando un conjunto de variables.

La segunda iteración dio paso a un estudio más por sobre las variables utilizadas que por la estimación del índice L en si mismo, con tal de lograr lo último mencionado en el párrafo anterior. Dado que ya se tenía un número reducido de variables del total que se tenía en un inicio (las variables relevantes del modelo) se decidió continuar con este conjunto acotado de variables, se agregaron también un grupo pequeño de variables que podría ser de interés, como lo son la obesidad y el número de días en el cual un país pasó de tener 100 infectados a 1000. Las variables se subdividieron en las siguientes categorías: variables estructurales, variables propias del coronavirus y variables controlables. Las variables estructurales hacen referencia a características bases de un país y que son permamenentes durante un largo tiempo, las variables propias del coronavirus son los números con respecto al COVID-19 del país y las variables controlables son las variables en las cuales el país tiene control sobre ellas, como lo son el número de tests realizados, número de camas y gastos médicos.

Los análisis realizados con las variables obtenidas posicionan a Chile dentro de un grupo interesante de países que poseen características similares. Al aplicar los modelos XGBoost y RandomForest para la regresión, la relevancia de las variables para las implementaciones tienen comportamientos distintos pero similares, esto se debe principalmente al algoritmo utilizado. La relevancia obtenida de las variables permite poder estudiar su impacto en estas implementaciones. El comportamiento de ciertas variables fue el esperado pero hubieron algunas que tuvieron menos importancia de lo que se pensaba. Al modificar las variables controlables para un país fijo, las predicciones del índice L tenían comportamientos erráticos. Lo anterior se atribuye a que los países tienen comportamientos propios y características únicas que no son observables en los modelos implementados, modelos que son entrenados sobre un subconjunto de una cantidad numerosa de variables.

Es importante mencionar que se trabajan con datos inciertos, los cuales de un país a otro pueden variar su medición y confiabilidad (cambios en criterios de fallecidos por COVID-19, actualización de casos totales, etc). También esta la existencia de datos *no homologables* de un país a otro, esto hace referencia a que cada país tiene propiedades relacionadas con ámbitos socioculturales, propias del país en sí. De igual manera, debido a la variabilidad del índice L por país, pareciera que el índice tiene comportamientos propios por cada país. Del mismo modo, al ser un cociente, su comportamiento suele ser errático dado que una disminución del índice no es atribuible solo a una menor cantidad de fallecidos. El resultado final y lo presentado en este reporte es el estudio considerando todos los resultados anteriores, dando un enfoque a la importancia de las variables, a la variabilidad de los países y el comportamiento errático del índice L entre países. Para poder enfrentar esto último, se decide por estudiar la cantidad de fallecidos e infectados por separado, valores que componen al cociente L, no para buscar una predicción como tal, sino que para estudiar la relevancia de las variables seleccionadas en estas predicciones.

II. Metodología

Lo primero a mencionar es la creación de la base de datos para esta iteración. Se actualizan los datos a la fecha del **27 de mayo**. Los datos y variables son obtenidas de fuentes públicas recopiladas de distintas fuentes oficiales, como las bases de datos de la universidad de Johns Hopkins, Worldometers y la página web de la OMS.

Debido a la variabilidad en la estimación de L según país, se subdivide el estudio de la importancia de las variables en la estimación de L. Para poder enfrentar también lo anterior, es que se consideran solo variables que representen cocientes del país (como fallecidos por millón y

no solo fallecidos), salvo excepciones, con tal de reducir el impacto de la cantidad de población del país.

Con la base de datos creada, se procede a realizar un análisis exploratorio de los datos, aplicando distintos métodos con tal de obtener una mejor explicación de estos. Los resultados obtenidos en este análisis no se detallarán en este documento, esto debido a que lo obtenido es poco concluyente y su interpretación puede ser subjetiva.

Se buscaron variables correlacionadas para posteriormente realizar técnicas de visualización, tales como Análisis de Componentes Principales (ACP) y t-SNE de dimensión 2 para los países, y para las variables se estudia la contribución de las variables en las dos primeras componentes principales. En lo anterior, existieron variables en el modelo que tenían una alta correlación y fueron eliminadas.

Posterior a esto, se realiza un análisis con modelos de regresión para estudiar el factor *L*, *N_Infectados_Millon* y *Número de fallecidos por millón* por separado. Se prueban distintos modelos de regresión, de los cuales se destacan Random Forest y XGBoost. En lo que sigue del documento se referirá a los modelos como la aplicación de estos. Se aplica Cross-Validation para obtener un promedio en los puntajes obtenidos para así conocer las variables más importantes para el modelo según Gini Score y Gain Score respectivamente para cada modelo.

III. Base de datos

Se considera una base de datos recopilada de fuentes como la Organización Mundial de la Salud, datos trabajados por la universidad Johns Hopkins, el sitio web *Worldometers*, entre otros. Estos datos se recopilaron y unieron según países, contando con un total de 73 países y 36 variables numéricas que representan la estructura económica, sanitaria y demográfica de cada país (ver Cuadro 1), además de las variables propias del coronavirus.

Económicas/Sanitarias	**Demográficas**
PIB 2019	Población
% Gasto en salud por fuentes públicas	Diferencia anual
% Gasto en salud por privados	Diferencia neta
Gasto en salud per cápita (USD)	Densidad
Gasto en salud per cápita (PPP)	Área territorial
Médicos por 1000	N° de inmigrantes
Enfermeras/Matronas por 1000	Tasa de natalidad
Mort. (por 1000) por enf. pulmonares	Población urbana
Mort (por 1000) femenina por enf. pulm.	Mediana de edad
Mort (por 1000) masculina por enf. pulm.	Fumadores
Especialistas en cirugía por 1000	Proporción H/M al nacer
Camas en hospitales por 1000	Proporción H/M total
Camas de cuidado intensivo por 100.000	Temperatura promedio
Mort (por 100.000) por gripe común	Humedad promedio
Infectados por A(H1N1)	Tráfico aéreo
Fallecidos por A(H1N1)	% Población 0-14
	% Población 15-64
	% Población >64

Cuadro 1: *Variables Estructurales originales.*

En las primeras iteraciones, al considerar solo países de la OCDE se hace necesario reducir la cantidad de características mediante análisis de correlación e importancia en los modelos predictivos. De esto último se rescatan 53 países y 13 columnas después de filtrar por relevancia y completitud (sin nulos). Las variables consideradas son de tres tipos distintos:

- **Estructurales:** Se escogen características de base (permanentes en el tiempo) según el trabajo previo, representando índices de desarrollo económico, estructura etaria, infraestructura y capacidad hospitalaria, entre otros.
  1. *%_Pob_15_64_años*: Porcentaje de la población entre 15 y 64 años.
  2. *Densidad*: Densidad poblacional.
  3. *Medicos_por_1000_Hab.*: Cantidad de médicos generales por 1000 habitantes.

4. *Especialistas_por_1000_Hab*: Cantidad de especialistas quirúrgicos por 1000 habitantes.
5. *PIB_2019*: Producto Interno Bruto.
6. *%_Poblacion_Urbana*: Porcentaje de población urbana.
7. *Humedad_Promedio*: Humedad promedio.
8. *Prevalencia_Obesidad*: Índice de obesidad de cada países según Índice de Masa Corporal (IMC - data de FAO).

- **Propias del coronavirus:** Se agregan variables propias del coronavirus, como los son la aceleración de la epidemia, número de infectados o fallecidos (total y nuevos), etc.
  1. *N_Infectados_Millon* Número de infectados (detectados) por 1.000.000 habitantes.
  2. *N_Fallecidos_Millon* Número de fallecidos por 1.000.000 habitantes.
  3. *Nuevos casos por millón:* Casos nuevos en el último día (27 de Mayo) por 1.000.000 habitantes.
  4. *Cantidad de nuevos fallecidos por millón:* Nuevas defunciones en el último día (27 de Mayo) por 1.000.000 habitantes.
  5. *Tiempo100a1000:* Variable que índica el número de días transcurridos para pasar de 100 a 1000 infectados.
- **Controlables:** Corresponden a las variables en las cuales las entidades gubernamentales podrían tomar decisiones, como lo son el número de tests, número de camas en hospitales o la inversión en salud.
  1. *N_Test_cada_1000_Hab:* Número total de tests por 1000 habitantes.
  2. *N_Camas:* Cantidad de camas en hospitales.
  3. *%_PIB_Salud_2016:* Porcentaje del PIB dedicado a salud.

A partir de estos datos, se procede a visualizar la tasa de letalidad L, total de infectados por millón de habitantes y total de fallecidos por millón de habitantes en las figuras 1, 2 y 3. Se observan países que oscilan entre los puestos de las distintas figuras. A modo de ejemplo, Catar tiene una tasa de letalidad del 0.05 %, pero es el país con más casos por millón de habitantes, ocurren comportamientos similares con Singapur, estos países presentan características etarias similares en su cantidad de población, en cuanto a su proporción de población joven y adulta mayor y además cuentan con características similares en cuanto a su capacidad de asistencia médica por habitantes, esto se puede observar en variables como la cantidad de especialistas o gasto en salud que tienen Singapur y Catar.

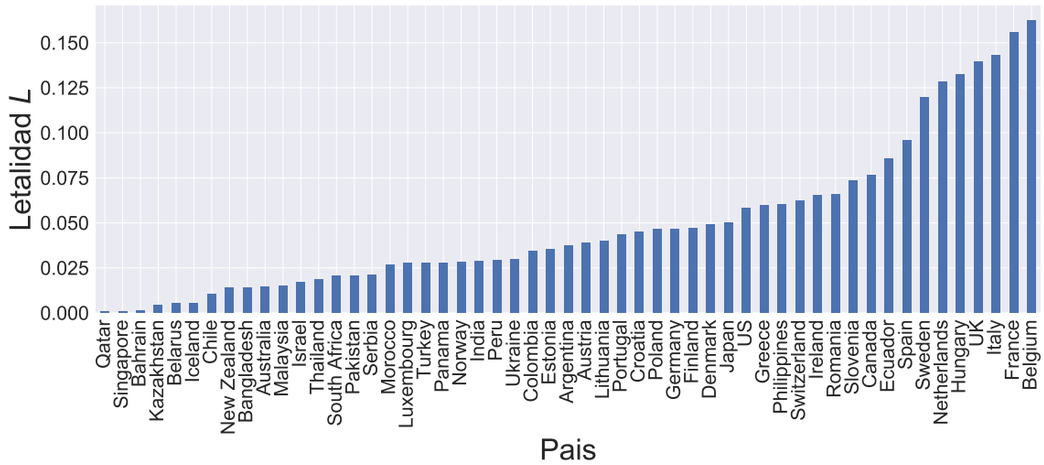

Figura 1: *Valor de L para distintos países.*

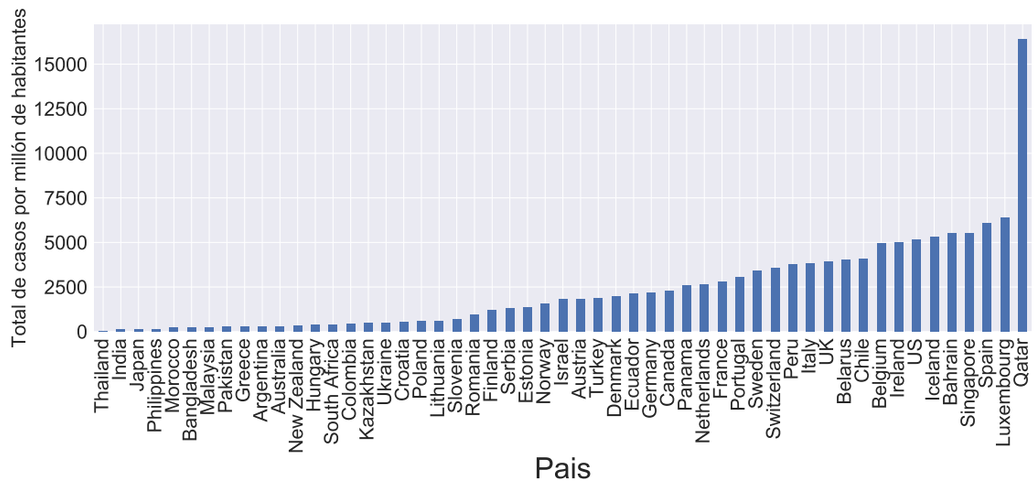

Figura 2: *Total de infectados por millón de habitantes para distintos países.*

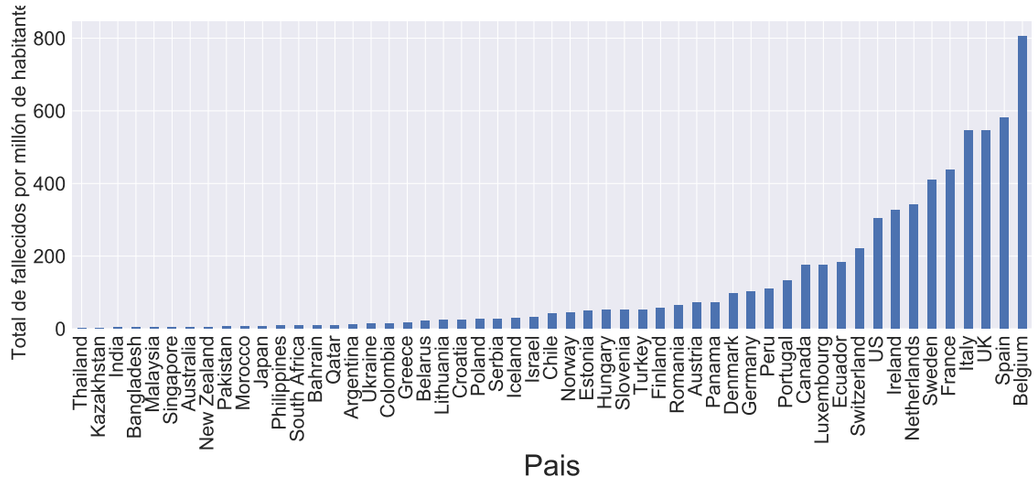

Figura 3: *Total de fallecidos por millón de habitantes para distintos países.*

IV. Resultados

1. Análisis de Componentes Principales

En primer lugar, se identifican grupos de variables con alta correlación (sobre 70 %), por lo que se descartan algunas de ellas para evitar colinealidad[1]. Con las características restantes se realiza un ACP con escalado estándar, pudiendo obtener una descomposición en las dos primeras componentes, que explican poco más del 45 % de la varianza, visualizada en la Figura 4. En este gráfico destacan inmediatamente Catar, Singapur y Barhain, que son los países con menor tasa de letalidad, pero también de los que tienen un mayor número de infectados por millón de habitantes, por lo que es de interés explicar qué los diferencia de otros países con gran número de contagiados, donde notamos que Chile aparece en el décimo puesto.

[1] *Casos activos cada mil habitantes, nuevos casos por millón, críticos cada mil habitantes, recuperados cada mil habitantes, cantidad de nuevos fallecidos por millón, médicos cada mil habitantes, 2009 y porcentaje de gasto PIB per cápita 2016.*

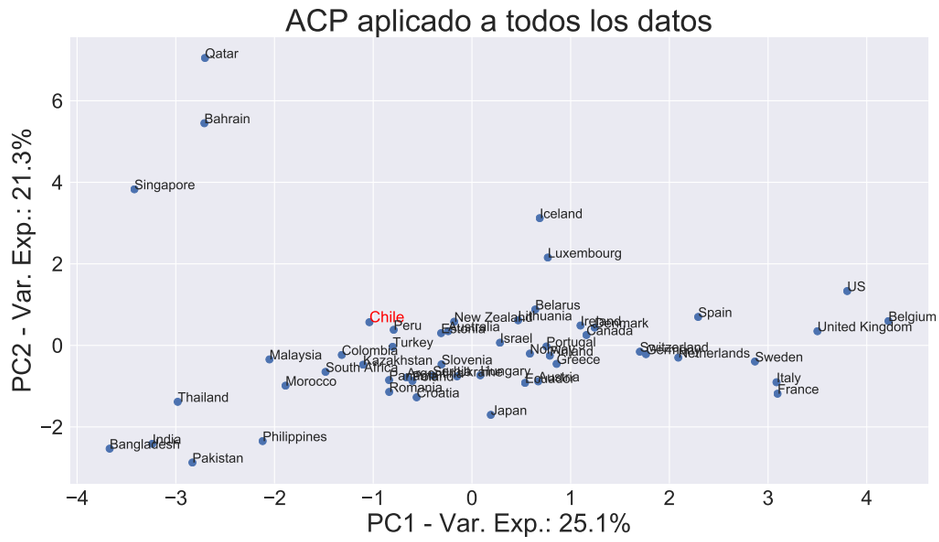

Figura 4: *Análisis de Componentes Principales.*

En la Figura 5, se presenta la relación entre las variables consideradas, representando la contribución en las dos componentes principales y agrupando las que estén positivamente correlacionadas, en particular se puede apreciar la estrecha relación entre la tasa de infectados y el número de tests e índice de obesidad al igual que la tasa de mortalidad con la estructura sanitaria (gasto y especialistas), mientras la letalidad L se encuentra completamente opuesta al porcentaje de población joven. Con esto en mente, notamos que los tres países mencionados tienen porcentajes muy altos de población joven, lo cual es consistente con la idea de que la letalidad se concentra en la población adulta mayor, y también controlaron de buena manera los primeros casos de contagio (*Tiempo100a1000*). Por otra parte, Islandia y Luxemburgo son de los países con mayor tasa de tests, manteniendo la letalidad controlada a pesar de la gran cantidad porcentual de contagiados, a diferencia de España o Bélgica que son de los países con mayor mortalidad por coronavirus. Por último, respecto a Chile podemos ver su cercanía con otros países como Perú, Colombia y Turquía, los cuales no tienen una relación evidente respecto a sus características.

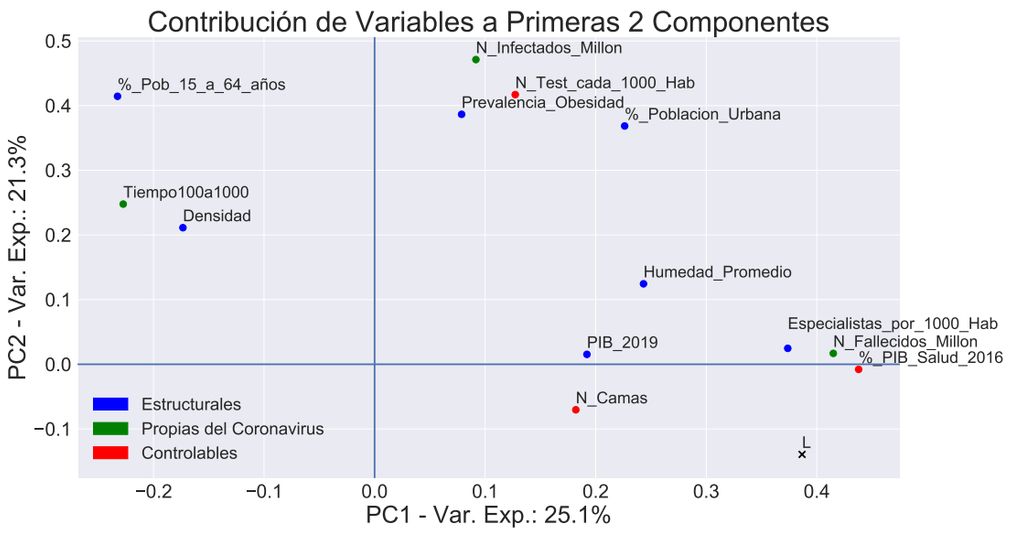

Figura 5: *Gráfico de la contribución de las variables a las 2 primeras componentes principales en ACP.*

En la Figura 6 se visualiza el resultado de otra técnica de reducción de dimensionalidad, t-SNE, en la cual se mantienen la mayoría de las agrupaciones antes mencionadas, corroborando así la similitud en sus estructuras. Sin embargo, en el caso de Chile se puede ver su agrupación solo con Argentina y Sudáfrica, sugiriendo que la cercanía con los otros países en la Figura 4 puede deberse a una coincidencia de la ponderación de variables distintas, más que por similitud de sus características. Al revisar las variables en detalle, se identifica que Sudáfrica y Argentina comparten con Chile un alto índice de obesidad, además de dedicar un bajo porcentaje del PIB a

salud y una rápida aceleración de la epidemia en las primeras semanas.

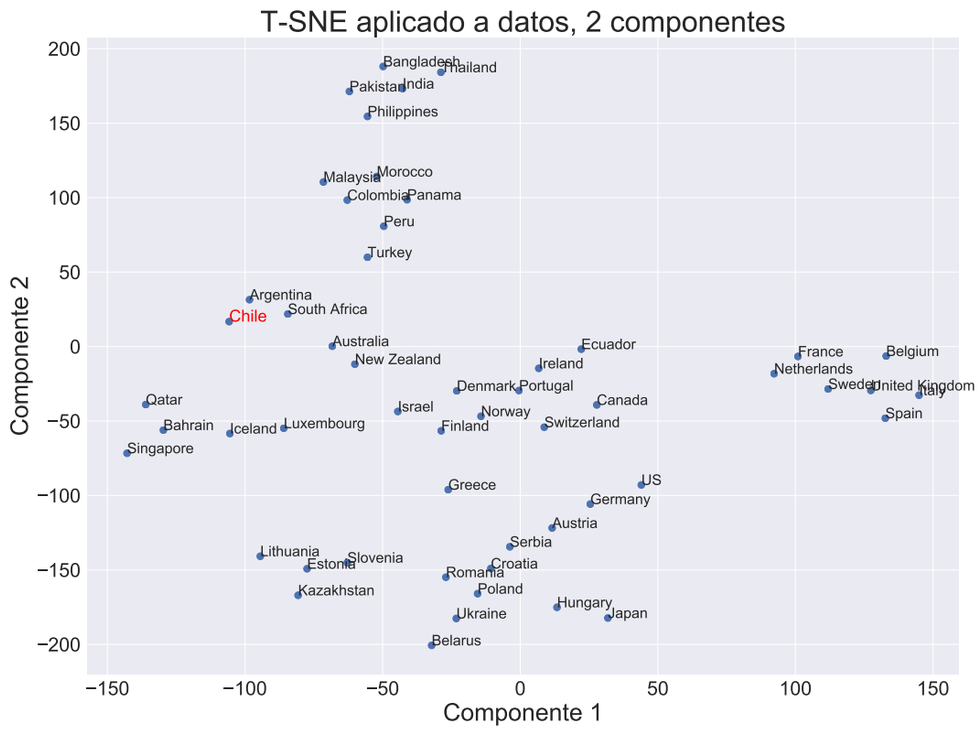

Figura 6: *t-SNE.*

II. Regresión

Se ajustaron distintos modelos de regresión, tales como XGBoost, Random Forest, Gradient Boosting, SupportVectorRegression y regresiones lineales, pero solo se hablarán de los que tuvieron el mejor rendimiento: XGBoost y RandomForest. A continuación mostraremos los resultados de dichas regresiones para las variables L, número de infectados por millón y número de fallecidos por millón, donde se retiran claramente las variables a explicar en las estimaciones correspondientes. Los resultado que se muestran son para todos un promedio de lo obtenido luego de 1000 iteraciones de Cross-Validation.

Dada la complejidad del problema, la finalidad de aplicar dos modelos de regresión no está enfocada en la directa comparación de éstos, sino que en el aporte de diversas fuentes de predicción con tal de contribuir a la búsqueda de las variables más relevantes en la predicción de infectados, fallecidos o del índice L, además de encontrar algún tipo de correlación entre ciertas variables de importancia.

1. Regresión para el número de infectados por millón

Se consideran 11 variables del total de variables que se contaban en un inicio y se procede a buscar los mejores hiperparámetros con validación cruzada, para luego correr las iteraciones. El RMSE de XGBoost arrojó 2276.83 y para RandomForest 2211.95, con valores para el número de infectados por millón variando de $8{,}16 \cdot 10^{-2}$ y $8{,}05 \cdot 10^{2}$. En general, se ve que la variable de tests por cada mil habitantes toma una gran importancia para ambos modelos, junto con la prevalencia de la obesidad.

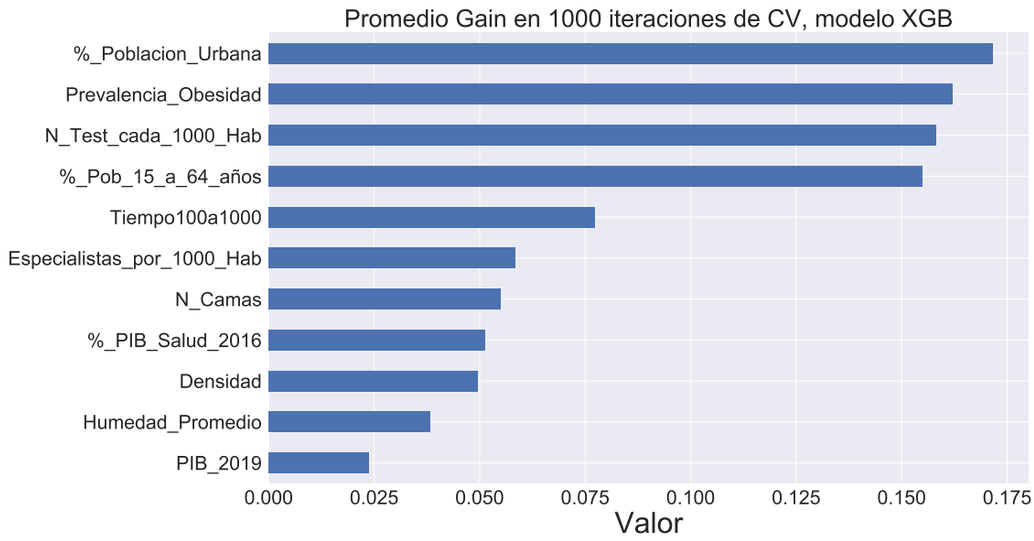

Figura 7: *Importancia de variables para XGBoost (infectados).*

Para XGBoost la población urbana es la variable de mayor relevancia junto con la prevalencia de la obesidad. Notamos además que la variable de porcentaje de población entre 15 y 64 años y el número de test por mil habitantes figuran dentro las variables más importantes, al igual que en RandomForest. Las variables propias del coronavirus, excepto la mencionada anteriormente, ocupan puestos intermedios en ambos modelos (número de camas, producto interno bruto destinado a salud). Variables como la densidad y la humedad se posicionan para ambos modelos en los últimos puestos, al igual que el producto interno bruto en los dos modelos (Figuras 7 y 8).

Para el modelo RandomForest (ver Figura 8), al igual que en XGBoost, los test por cada mil y el porcentaje de población entre 15 y 64 años aparecen como las variable con más importancia, la prevalencia y el tiempo de 100 a 1000 infectados aparecen entre el top 5 de variables con más importancia para ambos modelos. La variable de especialistas quirúrgicos se mantiene en posiciones intermedias para ambos modelos, además comparte esencialmente la misma importancia (diferencia de 0.001, *mean gini*) que el producto interno destinado a salud para ambos modelos.

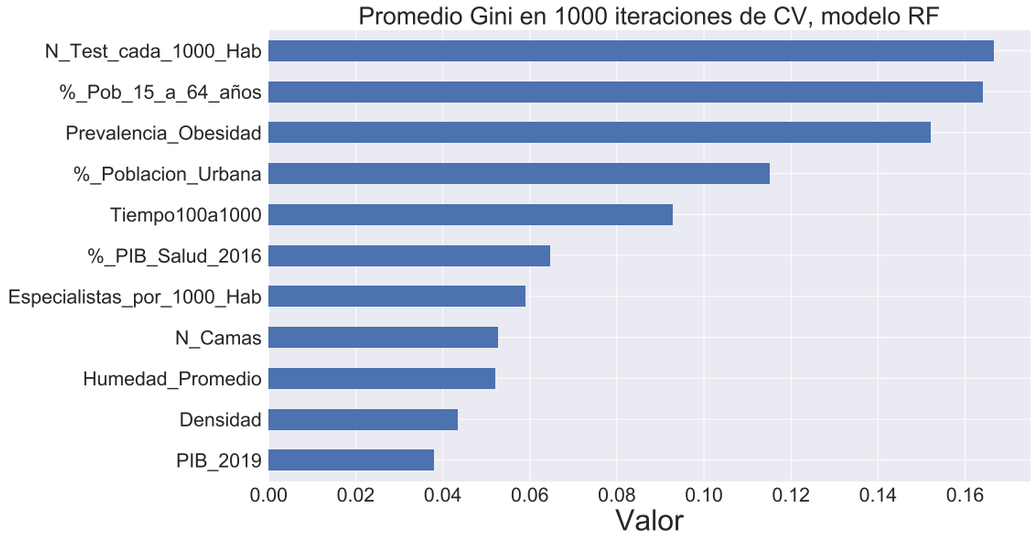

Figura 8: *Importancia de variables para RandomForest (infectados).*

2. Regresión para el número de fallecidos por millón

Se consideran las mismas variables para el caso de la regresión para infectados, pero sumando la variable de infectados por millón. El motivo de esto es averiguar si muchos infectados conlleva una mayor cantidad de fallecimientos. Para este caso obtenemos un RMSE aproximado de 164.96 y 121.84 para XGBoost y RandomForest respectivamente, con valores entre $8{,}16 \cdot 10^{-1}$ y $8{,}05 \cdot 10^{2}$

fallecidos por millón. Se adjuntan en las Figuras 9 y 10 los resultados. Se puede observar que las variables *Tiempo100a1000*, *%_PIB_Salud_2016*, *N_Infectados_Millon* y *PIB_2019* destacan como las variables más importantes en ambos modelos, siendo casi de igual importancia para XGBoost pero notando la jerarquía en el modelo RF. También se aprecia que el porcentaje de la población de 15 a 64 quedo en el quinto puesto para ambas iteraciones, seguido de las variables de menor importancia, quedando como última la densidad de la población.

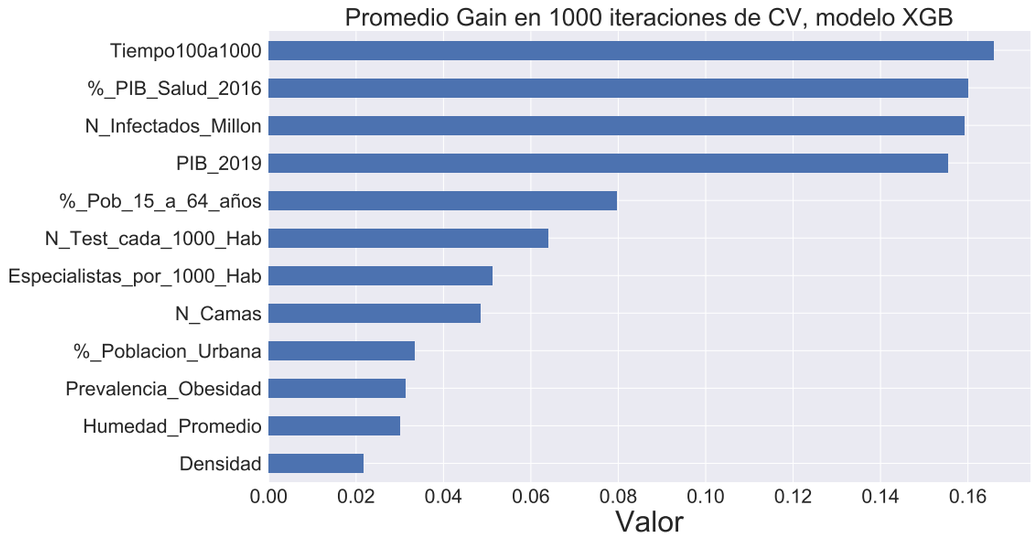

Figura 9: *Importancia de variables para XGBoost (fallecidos).*

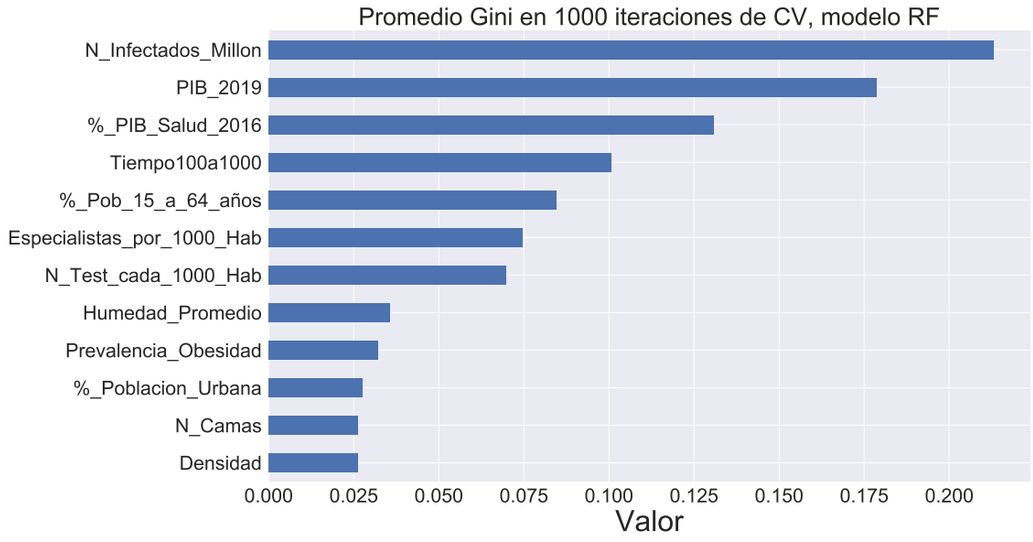

Figura 10: *Importancia de variables para RandomForest (fallecidos).*

3. Regresión para L

La tasa de letalidad es el numero de fallecidos totales sobre el numero de infectados totales y solo se puede conocer al termino de la epidemia, antes es complicado pues hay que tomar el numero de fallecidos entre los infectados y eso no es posible conocerlo sino hasta que los infectados que se cuenten se recuperen o fallezcan, lo anterior requiere esperar un tiempo variable que puede ser entre dos semanas y un mes o más. Hay técnicas para estimar la tasa de letalidad en tiempos intermedios de desarrollo de la epidemia, sin embargo a riesgo de introducir un sesgo en ello, tomaremos simplemente el numero de fallecidos sobre el numero de infectados hasta el momento de desarrollo de la epidemia, -que aproxima mejor la tasa real mientras la epidemia este mas cerca de su término-, así pues esta es la variable de letalidad que estudiaremos.

Las variables consideradas son idénticas a la del primer caso (fallecidos). El RMSE de XGBoost

fue de 0.040 y el de RandomForest un 0.035. Hay que notar que los valores de L se encuentran en un rango de $5{,}9 \cdot 10^{-4}$ y $1{,}6 \cdot 10^{-1}$. La significancia de las variables la podemos observar en la Figura 11:

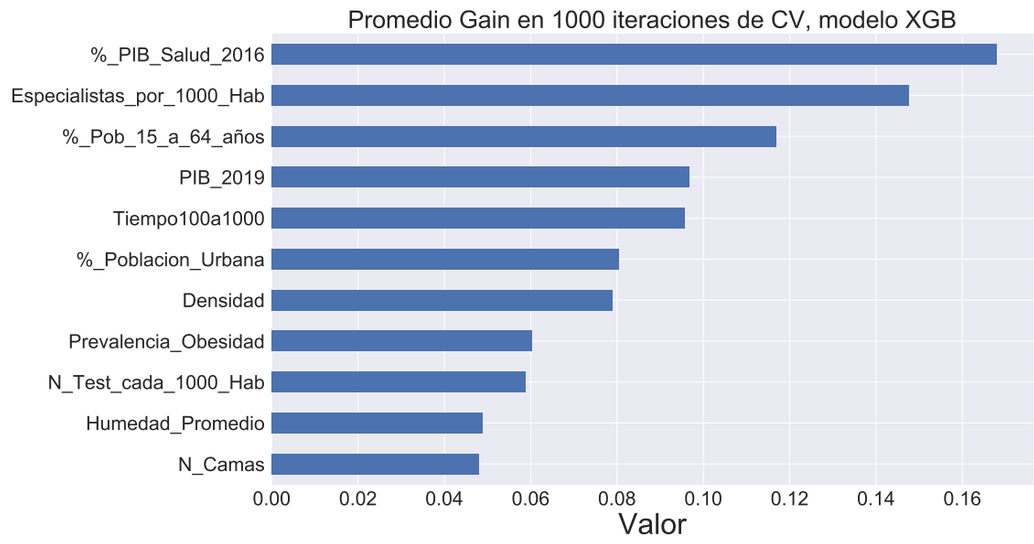

Figura 11: *Importancia de variables para XGBoost (L).*

Para el modelo de XGBoost el porcentaje de PIB invertido en salud el 2016 tiene la mayor importancia, seguido de los especialistas quirúrgicos cada 1000 habitantes y el porcentaje de la población entre 15 a 64 años. Después de estas variables notamos que la importancia se va manteniendo dos a dos, como es el caso del PIB 2019 con el tiempo de 100 a 1000 infectados, porcentaje de población urbana con densidad de la población, etc. Se resalta que el número de camas resultó la variable con menor importancia a la hora de predecir el cuociente L.

Para el modelo de RandomForest se obtuvieron las importancias que vizualizamos en la Figura 12:

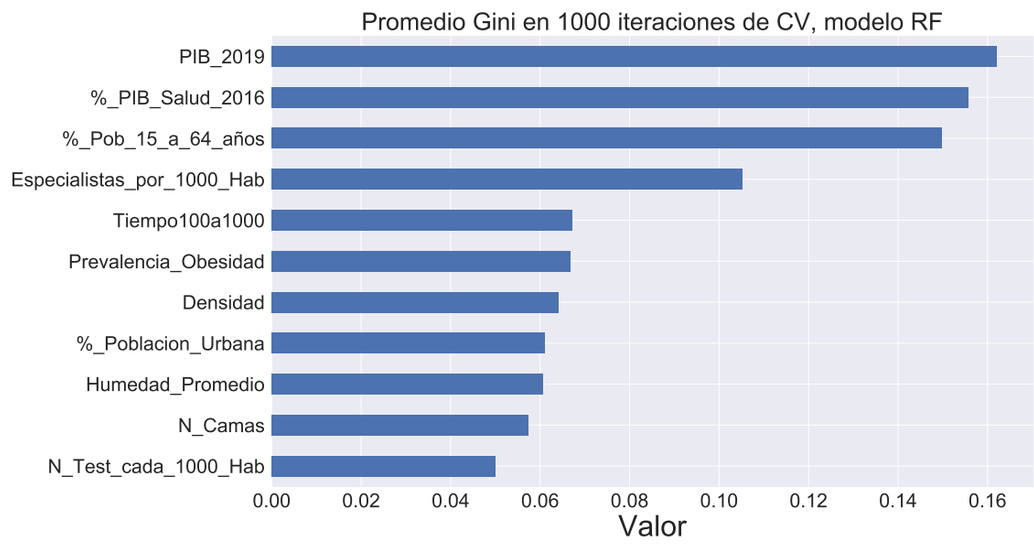

Figura 12: *Importancia de variables para RandomForest (L).*

Para este modelo notamos que las dos primeras variables más importantes fueron las relacionadas con el producto interno bruto, seguido por el porcentaje de población entre 15 a 64 años y los especialistas quirúrgicos cada 1000 habitantes, para luego mantener una importancia alrededor la misma magnitud ($0{,}06 \pm 0{,}01$) para las siguientes variables. Notamos que la variable menos importante resultó ser el número de tests que se realizan cada 1000 habitantes. El número de camas, al igual que el modelo anterior, no representa mayor importancia.

V. Discusión

En esta sección se discutirán los resultados obtenidos de las importancias de las variables en las regresiones realizadas para ambos modelos utilizados.

Estimación del número de infectados

Al fijarse en las variables con mayor importancia en el número de infectados por millón (figura 7 y 8) se puede ver la relevancia que toma la cantidad de test que se realizan sobre la población, esto es debido a que repercute directamente en el seguimiento de la propagación del coronavirus. Ambos modelos tienen a la variable de población entre 15 y 64 años como una variable que tiene relevancia para la predicción, lo cual se condice con ser la parte de la población con mayor movilidad en ciudades y pueblos pensando en trabajo, reuniones sociales y trámites. En el gráfico 7 la variable que más importancia tiene es la variable que guarda relación con el porcentaje de población urbana, y en el gráfico 8 se posiciona en cuarto lugar. La importancia de esta variable se puede deber a que las medidas de distanciamiento social se pueden llevar a cabo de mejor manera fuera de la ciudad.

Los índices de obesidad toman relevancia en el número de infectados, pese a que se esperaba que fuera más relevante en el número de fallecidos. Por otra parte, para ver el número de infectados que hay en el país, toman menos importancia variables como el gasto en salud, la cantidad de especialistas o número de camas esto puede deberse a que estas variables tienen mayor repercusión en la letalidad del coronavirus mas que en su propagación fuera de los hospitales.

Estimación del número de fallecidos

Para este caso, se destacan las similitudes de ambos modelos en las cinco primeras variables más importantes: el número de infectados por millón, PIB 2019, el porcentaje de PIB dedicado en salud el 2016, el tiempo de 100 a 1000 infectados y el porcentaje de población entre 15 y 64 años.

La primera variable, número de infectados por millón, era claramente esperable ya que a mayor número de infectados se puede establecer una relación de cierta forma directa con la cantidad de personas que fallecen por la enfermedad. Las siguientes dos variables relacionadas al producto interno bruto, a priori se podría concluir que se debe a los países mejores preparados para la enfermedad pueden afrontar mejor la crisis sanitaria. Sin embargo, se debe llevar a cabo un mayor análisis a las variables relacionadas al PIB invertido en salud, como son los especialistas quirúrgicos o los número de camas, que si bien no son los menos importantes tienen una clara menor importancia que la mejores 4 variables de cada modelo.

El tiempo entre 100 y 1000 infectados, que es la variable más importante para el primer modelo (Figura 9), se asocia al control que realiza cada país en los primeros brotes del COVID-19. En la Figura 5 se puede ver además que esta variable tiene una relación opuesta al número de fallecidos por millón, pues para valores bajos, causados por ejemplo por el escaso seguimiento de los primeros casos o por medidas de distanciamiento tardías, se evidencia una dispersión acelerada del coronavirus, pudiendo colapsar el sistema de salud y con ello el tratamiento de los infectados, provocando finalmente más casos fatales. Se especula que además de medir la aceleración del coronavirus en las primeras semanas, y con ello la exigencia temprana al sistema de salud, puede tener cierta relación con la trazabilidad de las personas contagiadas, que es uno de los factores que ha tenido gran impacto en el desarrollo del coronavirus, pero que es de los más difíciles de medir.

Para finalizar el análisis de este caso, es interesante ver que la variable del porcentaje de población entre 15 a 64 años toma un rol activo para la predicción en los dos modelos, siendo posiblemente por que el coronavirus causa problemas graves principalmente a personas mayores o con problemas de salud subyacentes.

Estimación del índice L

Tanto en el modelo XGBoost como en RandomForest se obtiene una explicación del cuociente L principalmente por variables relacionadas con el producto interno bruto, pues estas variables reflejan de cierta forma la preparación de cada sistema de salud para el potencial tratamiento de los contagiados.

Las siguientes variables más importantes fueron los especialistas por 1000 habitantes y el porcentaje de población entre 15 a 64 años. La primera mencionada sigue la tónica de la preparación del país al momento de enfrentar la crisis sanitaria, mientras que la segunda hace referencia directamente a la baja letalidad relativa que tiene el coronavirus al ser huésped de una persona joven, teniendo en cuenta la correlación negativa entre estas dos variables.

Una pregunta que surge al ver las variables con menor importancia es la poca contribución del número de camas y tests cada 1000 habitantes. Esta última puede ser explicada por el hecho de que una lenta reacción puede aumentar mucho el valor de L aunque se realicen muchos tests para tratar de disminuir las fallecidos, mientras que también se podría estar en una situación de una rápida y eficaz respuesta que podría prevenir fatalidades realizando varios tests para controlar la circulación de la población. Con respecto a la poca importancia del número de camas se hipotetiza que, al estar relacionado con el PIB invertido en salud —que es, nuevamente, la variable más importante al explicar L en este caso— la respuesta tenga que estar relacionada con el sistema de salud de cada país. Se podría ligar una correspondencia a lo invertido en infraestructura e insumos médicos (posiblemente en el pasado). Asimismo, países con menor cantidad de camas de hospital han tenido que tratar de dedicar más recursos efectivamente, posiblemente produciendo distintos resultados dependiendo que tanto se pudieron anticipar al virus.

General

En el Cuadro 2 se pueden observar las posiciones de las variables para todos los modelos implementados. Se pueden observar dos columnas al final de esta, donde *Prom.* indica la posición promedio de la variable y *LM* indica el logaritmo de la multiplicación de las posiciones de las variables. Cabe indicar que las métricas anteriores son solo indicadoras y dan un tipo de posición para las variables, pero no es absoluta. También, la variable Número de infectados por millón fue eliminada del recuadro dado que solo se utilizaba esta para la predicción de fallecidos, se ajustaron las posiciones con las variables restantes. Al principio la columna T indica a qué tipo de variable corresponde cada categoría, si a C (Controlables), E (Estructurales) o V (Propias del coronavirus).

		L		**Infectados**		**fallecidos**			
T	**Categoría**	**XGB**	**RF**	**XGB2**	**RF3**	**XGB3**	**RF4**	**Prom.**	**LM**
C	%_PIB_Salud_2016	1	2	8	6	1	2	3.3	5.3
E	%_Pob_15_64_años	3	3	4	2	4	4	3.3	7.0
V	Tiempo100a1000	5	5	5	5	1	3	4.0	7.5
E	Esp._por_1000_Hab	2	4	6	7	6	5	5.0	9.2
E	PIB_2019	4	1	11	11	3	1	5.2	7.3
C	N_Test_cada_1000_Hab	9	11	3	1	5	6	5.8	9.1
E	%_Poblacion_Urbana	6	8	1	4	8	9	6.0	9.5
E	Prevalencia_Obesidad	8	6	2	3	9	8	6.0	9.9
C	N_Camas	11	10	7	8	7	10	8.8	13.0
E	Densidad	7	7	9	10	11	11	9.2	13.2
E	Humedad_Promedio	10	9	10	9	10	7	9.2	13.2

Cuadro 2: *Posiciones de las categorías según importancias para cada modelo*

Podemos observar entonces que a pesar de que algunas características relevantes escapan al control del gobierno (estructura etaria, población urbana), hay otras en las que se puede influenciar para afrontar de mejor manera esta u otra eventual pandemia. Por ejemplo, en la predicción del número de infectados se determina la gran importancia de la capacidad de testeo para la correcta detección del número de contagiados, ya que aunque exista correlación positiva entre estas variables, como se muestra en la Figura 5, esto se puede explicar pues a medida que se realizan más tests, la cantidad de infectados (detectados) también aumenta, pero esto último es positivo pues significa que el cómputo es más cercano a la realidad. Esta correlación se puede deber también a una relación causa(muchos contagiados)-efecto(necesidad de tests), por lo que una medida de testeo masivo debería ser tomada tempranamente para poder realizar también el seguimiento adecuado de los casos, evitando así la propagación descontrolada del coronavirus (Tiempo100a1000) y eventuales colapsos en los sistemas de salud. Cabe destacar que la variable relacionada con el gasto público en salud, engloba y tiene dependencia, en cierta forma, con variables como el número de especialistas o el número de camas, además da un indicio de la capacidad de los sistemas de salud que tienen los países para enfrentar eventuales crisis sanitarias.

Respecto a la prevalencia de la obesidad, se obtienen resultados poco intuitivos, pues toma gran relevancia solo en la explicación de infectados, y no como factor de riesgo en la letalidad del virus como era de esperar.

En cuanto a cómo se comportaban las variables con las predicciones, se realizaron unos pequeños análisis que determinaron que dado un país con sus características fijas, si es que se comienza a modificar sus variables controlables, las predicciones comenzaban a tener comportamientos inesperados en los valores de L, número de casos totales por millón y fallecidos por millón para algunos países.

VI. Conclusiones

Como es común en estudios de situaciones contingentes, hay una gran dificultad en la recopilación de información completa y actualizada, debido a la constante evolución de los datos, lo que trae consigo cierta inestabilidad en los resultados e interpretaciones a medida que cambia la evidencia considerada en cada iteración del estudio, por lo que cada conclusión debe ser adoptada con cautela.

Se obtuvo que las variables más importantes son el porcentaje del PIB dedicado a salud y el número de especialistas por 1000 habitantes para la explicación de la letalidad, el porcentaje de población urbana y el número de tests por mil para el número de infectados por millón, y el tiempo de pasar de 100 a 1000 infectados junto con el ya mencionado porcentaje del PIB en salud para el número de fallecidos por millón.

Sobre las variables controlables, se obtienen resultados que confirman estudios similares, con respecto a la relevancia por ejemplo en la detección de la enfermedad, aumentando el número de tests. Al modificar las variables controlables de los países en la estimación de los distintos valores, las predicciones obtenidas no eran las esperadas por lo ya mencionado al final de la discusión. Esto da a entender lo distintos que son los países estudiados, que a pesar de que algunos tengan características estructurales parecidas, sus comportamientos con las variables contingentes son muy distintas. Lo anterior se puede deber a que existen variables que no se consideran en este estudio y que podrían ser relevantes, como lo son las políticas de control adoptadas por los gobiernos, la cultura que tengan los ciudadanos de cada país o bien su idiosincrasia. Además se debe tener en consideración que no todos los países están igual de preparados para que sus habitantes puedan hacer una cuarentena total, el no respeto de esta medida, puede deberse a necesidades económicas que tengan los sectores más pobres, quienes viven del dinero que pueden generar en el día a día. En base a lo anterior tampoco se toman en cuenta el impacto que pueden tener las políticas de subvención por parte de los gobiernos hacia sectores más vulnerables.

Un punto de vista que puede ser de utilidad, teniendo en cuenta el presente estudio y el desarrollo de la pandemia en Chile, es considerar las regiones o distritos del país como si fueran naciones, las cuales cuentan con características consideradas en este estudio, pero adaptadas al contexto local, como lo es por ejemplo el Fondo Común Municipal de los diferentes municipios que componen una región (o distrito) representando un "PIB regional". Los municipios pueden administrar estos fondos públicos para combatir la crisis sanitaria actual, concentrando esfuerzos en el control de los primeros contagios en cada región, la capacidad asistencial de los hospitales regionales, junto con la cantidad de especialistas que tenga el país en la zona, y otras variables que sean potencialmente explicativas en futuros estudios.

Finalmente, se destaca la relevancia obtenida de las variables relativas al porcentaje del PIB gastado en salud además del porcentaje de personas de 15 a 64 años del país en todos los modelos, las cuales pueden ser tomadas como base para estudios posteriores sobre las predicciones de los distintos números para cada país.