Modelo predictivo de mortalidad del COVID19 respecto a los datos en EE. UU.

Castro Mamani, Oscar
Bachiller en Computación Científica
UNMSM
oscar.castro@unmsm.edu.pe

Fernández Estela, Julio Cesar
Ingeniero de Petróleo y Gas Natural
UNI
jfernandeze@uni.pe

Dolci Flores, Juan Jose
Bachiller en Ingeniería de Sistemas
UNU
jdolcics@gmail.com

Mendez Rosales, Andres Alonso
Estudiante de Economía
andresaamr@gmail.com

***Abstract*—Realizaremos la exploración de los datos del dataset de COVID-19 case surveillance public use data de la CDC encontrado en Kaggle, para luego con sus respectivas variables significativas generar una comparativa entre los modelos de logistic regression, random forest classifier, decisión tree classifier, k neighbors classifier y ver cual es el más efectivo al predecir la muerte de las personas confirmadas con covid19.**

***Keywords*—Python, Covid19, Clasificador.**

I. INTRODUCCIÓN

El virus SARS-CoV-2 (COVID19) declarada pandemia el 22 Abril 2020 con 2.573.143 casos confirmados en 185 países [1] cuyo brote inicial fue en Wuhan, China en diciembre 2019 tiene hasta la fecha 106.902.907 casos confirmados y 2.341.004 muertes [2]. Como nos ha mostrado esta pandemia, su propagación es un proceso dinámico muy complejo y por ello hay una basta cantidad de factores que se ven englobados en su estudio como variables de susceptibilidad al patógeno (edad y condiciones de salud), comportamientos (cumplimiento con el distanciamiento social y el uso de máscaras), etc. Esto llevo a los gobiernos responder rápidamente con diferentes estrategias, por ejemplo, políticas de cierre de escuelas, lugares de trabajo, criterios para las pruebas y disponibilidad potencial de intervenciones farmacéuticas.

Las respuestas tomadas fueron gracias a los esfuerzos conjuntos de muchas áreas, pero principalmente a los que nos dieron un entendimiento mayor de esta enfermedad, para ello se está necesitando diferentes modelos y técnicas para comprender, pronosticar, planificar y responder las dinámicas muy cambiantes de esta pandemia.

Es así que en la literatura podemos encontrar métodos estadísticos y métodos de apoyo a la toma de decisiones que utilizan modelos de agentes múltiples, tales como:

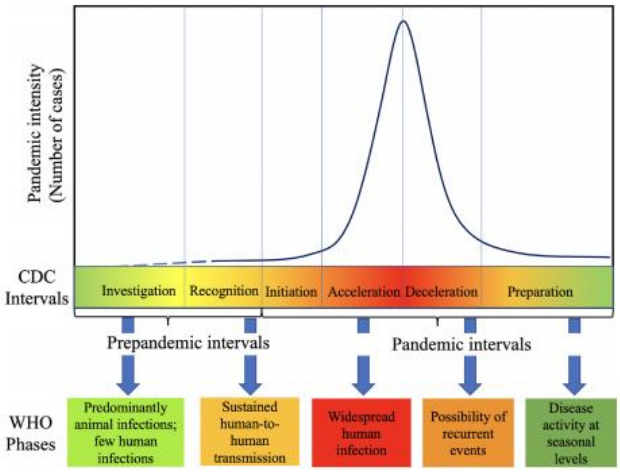
(i) pronosticar los resultados de la epidemia (por ejemplo, recuento de casos, mortalidad y demandas hospitalarias), utilizando un conjunto diverso de métodos basados en datos, por ejemplo, tipo ARIMA predicción de series temporales, técnicas bayesianas y aprendizaje profundo, por ejemplo [5] (ii) vigilancia de enfermedades [6] y (iii) análisis contrafáctico de epidemias utilizando modelos de múltiples agentes por ejemplo [7,8]; de hecho, estos dos últimos fueron muy influyentes en las decisiones tempranas de cierres en varios países.

En este contexto, nos valimos de un conjunto de datos COVID-19 Case Surveillance Public Use Data [3] publicada por Centers for Disease Control and Prevention (CDC) el cual presenta 12 variables que nos ayudara a dar una predicción de muerte asociado a estas variables como por ejemplo condiciones médicas, grupo de edad, estado de hospitalización, etc y como soporte un artículo de resumen de varios modelos [4] que dan un marco general de los diferentes modelos tomados en distintos escenarios.

II. TRABAJOS RELACIONADOS

MODELADO BASADO EN DATOS PARA DIFERENTES ETAPAS DE RESPUESTA PANDÉMICA (Aniruddha Adiga, Jiangzhuo Chen, Madhav Marathe, Henning Mortveit, Srinivasan Venkatramanan and Anil Vullikanti).

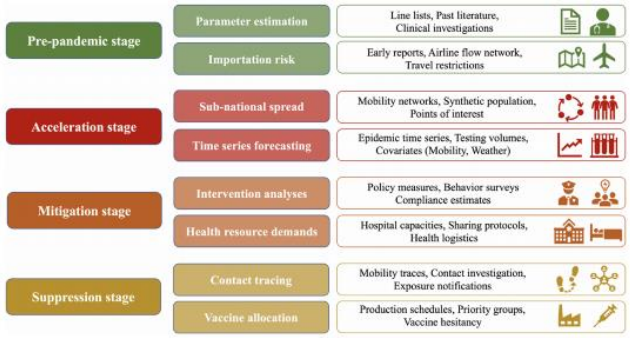
En este articulo se basan en un marco de trabajo ya establecido por la Worl Health Organization (WHO) y la CDC el cual describe que en el contexto de una pandemia de influenza hay 6 etapas que abarcan la investigación, ilustrada en la siguiente Figura.



Marco de intervalos pandémicos de los CDC y fases de la OMS

el reconocimiento y el inicio en la fase temprana, seguidas de la mayor parte de la propagación de la enfermedad que ocurre durante las etapas de aceleración y desaceleración. También proporcionan indicadores para identificar cuándo la pandemia ha pasado de una etapa a la siguiente.

Dentro de este marco, se consideran 4 etapas de respuesta pandémica, como se mostrará en la siguiente figura, en ella nosotros nos ubicaremos en la primera por el hecho del uso de ciertos modelos basados en datos con variables muy semejantes a las nuestras.



Etapas y data necesitada

III. METODOLOGÍA

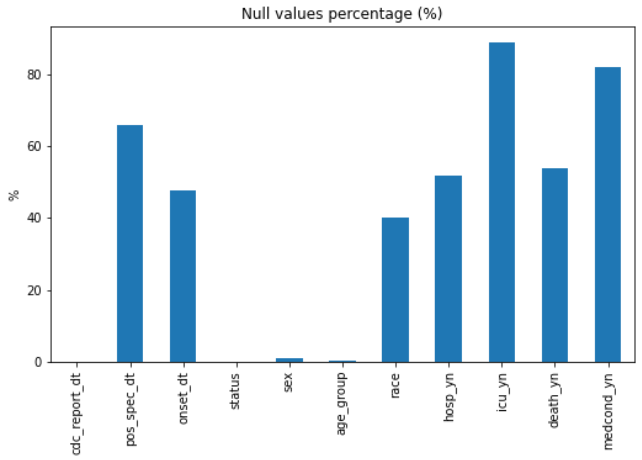
La metodología realizada es la siguiente, primero el análisis exploratorio de los datos, el cual nos sirvió para visualizar primero que es lo que nos muestra la data y sus relaciones, luego con ello nos planteamos nuestro dilema de si se podría predecir si alguien llegara a fallecer bajo ciertas características y nuestra hipótesis es la generación de un modelo predictivo de la mortalidad con un performance del 80 por ciento.

IV. EXPERIMENTO

Teniendo en cuenta el csv obtenido de la pagina de kaggle relacionada al dataset dada por CDC [3]. Procedemos a su lectura y pronto análisis en la plataforma de Google colab y así comenzamos.

ANÁLISIS DE LOS DATOS

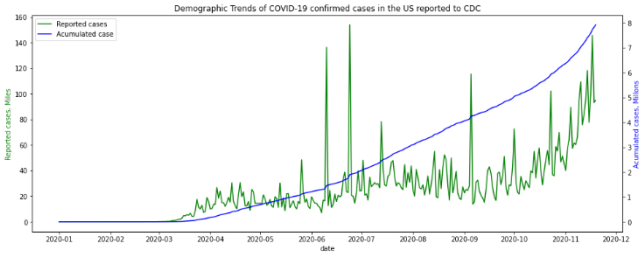
Para entender el conjunto de datos reemplazamos valores nulos por 'Missing' & 'Unknown'.



Se puede observar en el gráfico anterior que los datos iniciales contienen un gran porcentaje de datos nulos, los cuales fueron eliminados para lograr mejores modelos predictivos.

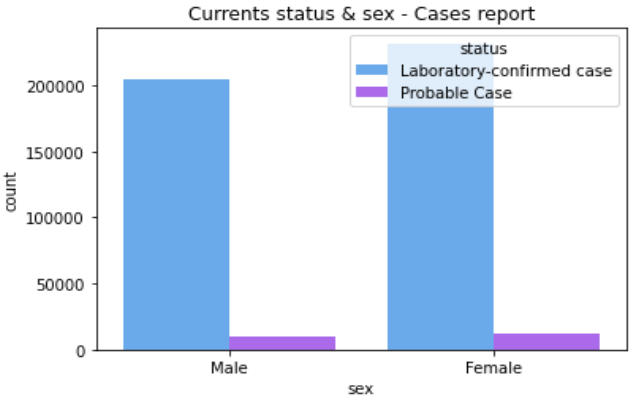
- EDA de cantidad de casos confirmados - recepcionados por CDC

El siguiente cuadro muestra en una serie de tiempo el número de casos confirmados, se observa que, desde marzo del 2020, el número de casos fue en aumento.

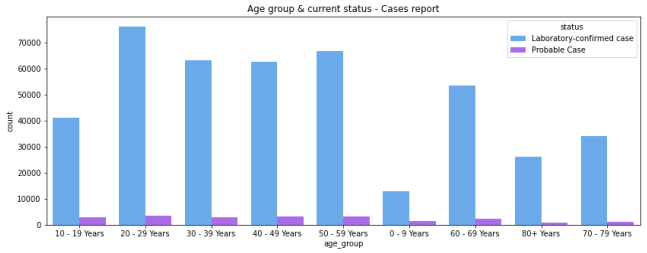


- ***Visualización del Data Set - EDA***

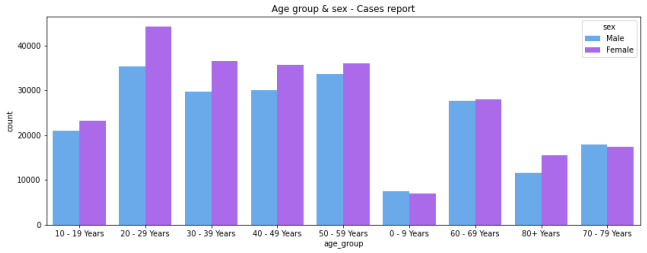
Se puede observar en el siguiente gráfico que hay una mayor cantidad de datos de casos de contagio confirmados por laboratorio que los probables casos de contagio, esto hará que los modelos sean más precisos, al reducir el margen de incertidumbre.



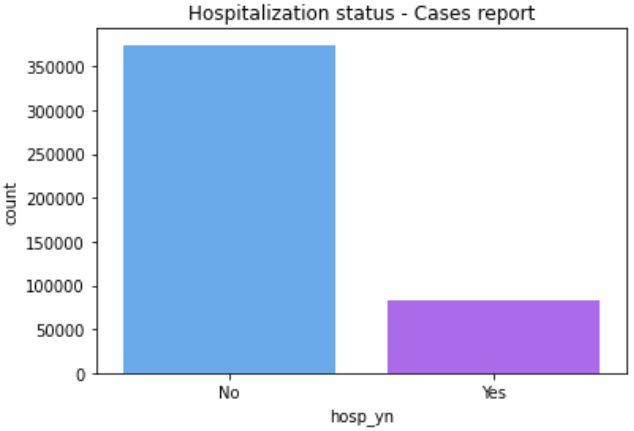
El siguiente gráfico muestra la distribución de los datos por grupo de edades.



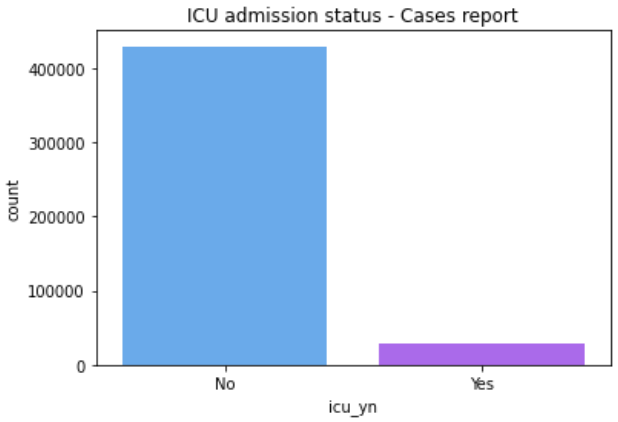
También se observa que la cantidad de datos correspondiente a casos en género masculino y femenino están equilibrados.



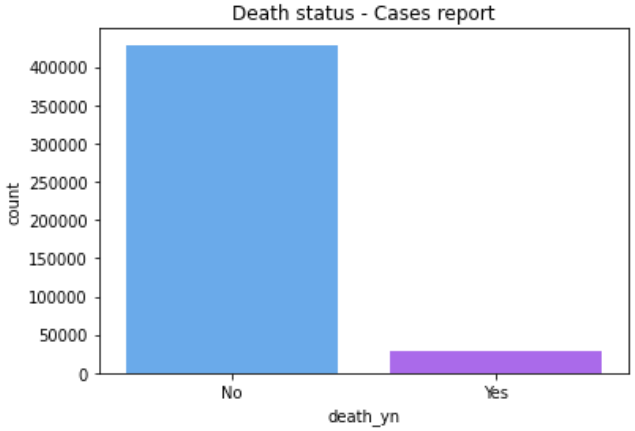
La cantidad de casos que tuvieron hospitalización vs los que no lo tuvieron.



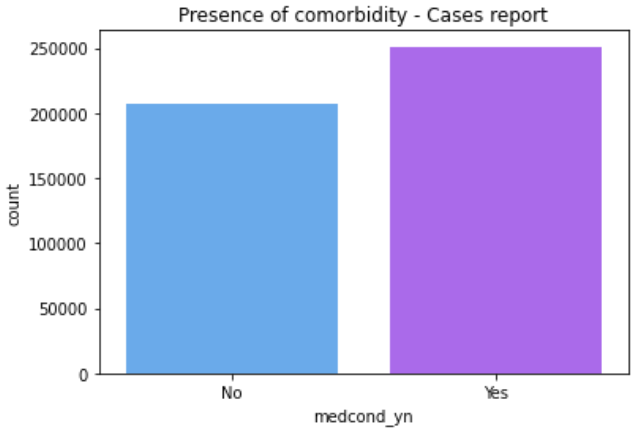
Y se muestra también la cantidad de personas que fueron atendidos en camas UCI.



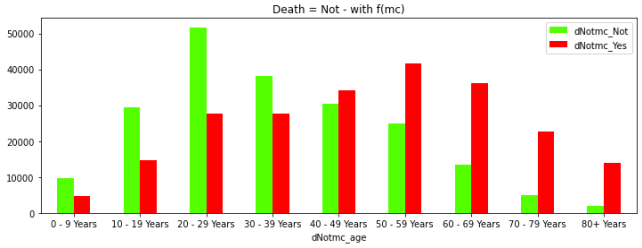
El siguiente grafico muestra la cantidad de casos de personas que han fallecido.



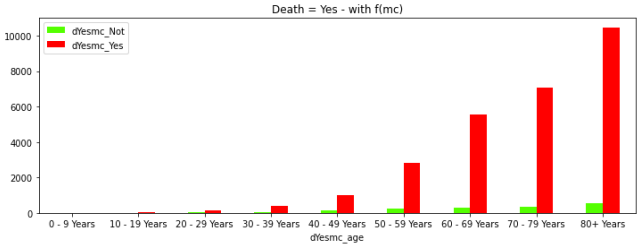
Para el dataset también se tuvo en cuenta si la persona contagiada tenia condiciones médicas preexistentes, la cantidad de casos positivos no están muy lejanos de los negativos, y eso es una ventaja al momento de hacer el análisis con los algoritmos de machine learning.



En el siguiente cuadro se muestra la distribución de las personas contagiadas que no fallecieron, separando los que tienen condiciones médicas preexistentes de las que no y agrupadas por edad.



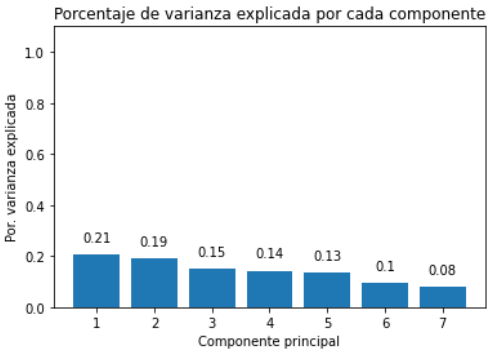
El siguiente cuadro nos muestra una relación entre los casos de personas que tenían condiciones médicas preexistentes y el número de fallecidos, en cada grupo de edad la cantidad de personas fallecidas con condiciones medias es mucho más grande que las que no tenían condiciones médicas.



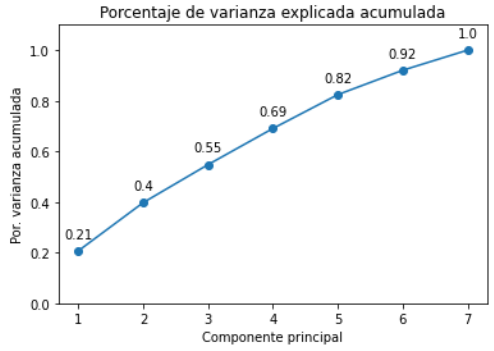
V. COMPARACIÓN DE LOS MODELOS UTILIZADOS

Para iniciar a modelar los algoritmos de Machine Learnig (*ML*), se inicia con la optimización de componentes principales. Típicamente utilizamos PCA (*Principal-Component-Analysis*) para reducir dimensiones del espacio de características original. Se utilizará el método para calcular la "proporción de variación explicada "de cada característica e ir tomando dimensiones hasta alcanzar un mínimo que nos propongamos, hasta alcanzar a explicar el 82% de la variabilidad total.

La gráfica siguiente nos muestra la distribución de cada uno de los componentes principales máximos para el conjunto de datos.

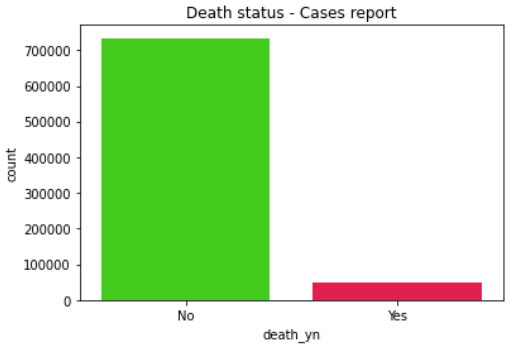


Posteriormente a normalizar los datos de entrada, aplicamos PCA y veremos que con 5 de las nuevas dimensiones (y descartando 2) obtendremos hasta un 82% de variación explicada y buenas predicciones, como se muestra en la figura siguiente.

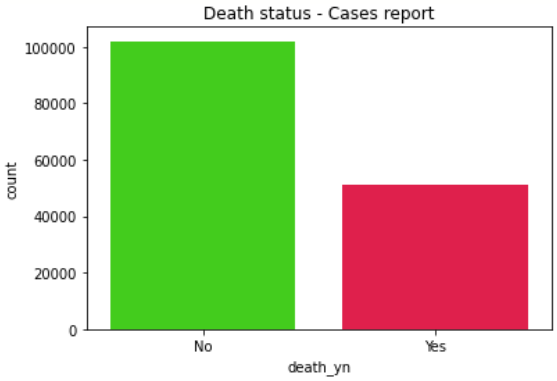


	PC1	PC2	PC3	PC4	PC5	death_yn
0	0.985521	-0.448567	0.192294	0.917744	-0.356422	0
1	1.015843	0.089883	-1.286757	1.072357	1.035685	0
2	0.708189	-1.175002	-1.564682	-1.040017	1.159033	0
3	-1.356901	0.920063	0.317409	0.934843	-0.734222	0
4	1.203327	1.692441	1.075080	0.942398	-0.916435	0

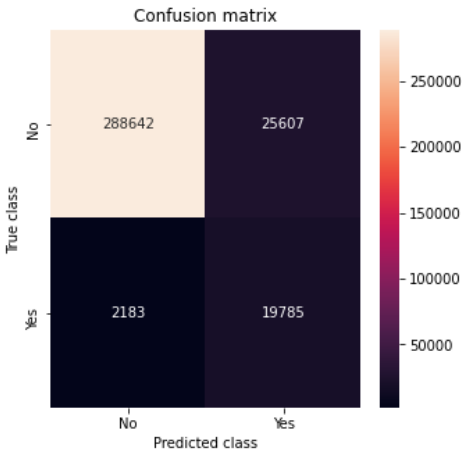
Luego de optimizar los componentes principales del conjunto de datos, realizaremos una distribución de la variable objetivo.



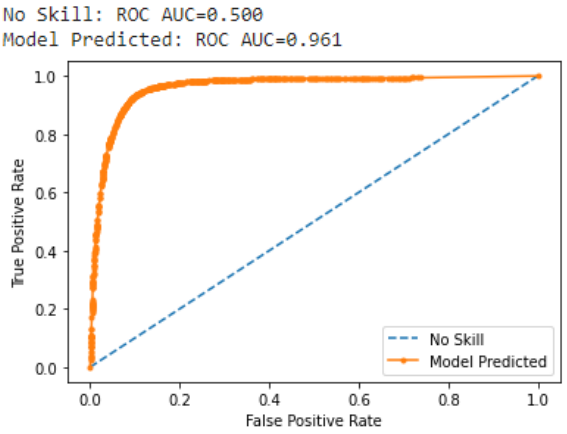
Podemos observar, que la clase "No" equivalentes a persona no fallecidas son mayoritarios, existiendo así un desbalance de clases, como se muestra en la figura anterior. Sin embargo, ello puede ser balanceado utilizando la librería *imblearn – under-sampling*, el cual resulta en la siguiente figura.



Posterior al balance de datos, se procede a realizar el modelamiento ML y comparación de algoritmos. En la gráfica siguiente una matriz de confusión para el modelo de *Decision Tree Classifier Model (DTC)*. El resultado del modelo nos da una aproximación de predicción del 92%, el cual es significativo para poder predecir la mortalidad de un determinado paciente afectado por el Covid-19.



	precision	recall	f1-score	support
0	0.99	0.92	0.95	314249
1	0.44	0.90	0.59	21968
accuracy			0.92	336217
macro avg	0.71	0.91	0.77	336217
weighted avg	0.96	0.92	0.93	336217



También podemos observar, que el modelo ML de DTC tiene un AUC = 0.961 el cual es significativo para explicar el comportamiento de mortalidad para el conjunto de datos.

En la tabla siguiente se muestran modelos desarrollados, para diferentes algoritmos *Logistic Regression(LRM), Random Forest Classifier (RFC), Decision Tree Classifier (DTC) y KNeighbors Classifier (KNC).*

Model	LRM	RFC	DTC	KNC
precisionNot	0.990	0.990	0.990	0.990
precisionYes	0.350	0.440	0.440	0.400
recallNot	0.890	0.920	0.920	0.910
recallYes	0.840	0.900	0.900	0.900
f1-scoreNot	0.940	0.950	0.950	0.950
f1-scoreYes	0.490	0.590	0.590	0.560
accuracy	0.890	0.920	0.920	0.910
ROC_AUC	0.936	0.963	0.961	0.941
run_time	3.115	23.810	2.972	21.560

Como podemos en la tabla anterior los valores de aproximación para explicar el modelo de mortalidad superan los 90%, siendo el modelo óptimo en tiempo y aproximación, el modelo ML DTC.

CONCLUSIONES

- El modelo tiene una aproximación superior al 90% para la predicción de mortalidad de una persona afectado por COVID-19.
- El modelo DTC tiene un desempeño óptimo para el modelamiento del conjunto de datos.
- Para una evaluación del efecto de enfermedades pre-existentes y su relación con la mortalidad es necesario tener la clasificación de enfermedades.
- El grupo de personas que superan los 60 años a mas son más susceptibles a fallecer siempre en cuando presente una enfermedad pre-existente.

REFERENCIAS

[1] Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis. 2020 Feb 19. PMID: 32087114 .

[2] Dashboard WHO COVID19. https://covid19.who.int/

[3] CDC. COVID-19 case surveillance public use data | data | centers for disease control and prevention. https://data. cdc.gov/Case-Surveillance/COVID-19-Case-Surveillan ce-Public-Use-Data/vbim-akqf. Accessed 24 Aug 2020

[4] Aniruddha Adiga, Jiangzhuo Chen, Data-Driven Modeling for Different Stages of Pandemic Response. J. Indian Inst. Sci. |VOL 100:4|901–915 October 2020|journal.iisc.ernet.in

[5] Perone G (2020) An arima model to forecast the spread and the fnal size of covid-2019 epidemic in Italy (frst version on SSRN 31 March). SSRN Electron J

[6] Healthmap. https://healthmap.org/en/. Accessed 28 Oct 2020

[7] Ferguson N, Laydon D, Nedjati Gilani G, Imai N, Ainslie K, Baguelin M, Bhatia S, Boonyasiri A, Cucunuba Perez Z, Cuomo-Dannenburg G et al (2020). Report 9: impact of non-pharmaceutical interventions (npis) to reduce covid19 mortality and healthcare demand. Imperial College Technical Report,2020.https://www.imperial.ac.uk/media/imperial-college/medicine/mrc-gida/2020-03-16-COVID19-Report-9.pdf

[8] IHME COVID, Murray CJL et al (2020) Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months. MedRxiv. https://www.medrxiv.org/content/10.1101/2020.03.27.20043752v1