



[www.datascience.pe](http://www.datascience.pe)

# ***Predicción del grado de vulnerabilidad por COVID-19 en E.E.U.U.***

(Caso : Dataset Covid - 19)

# Contexto inicial

[www.datascience.pe](http://www.datascience.pe)

# Situación COVID-19

8 Enero

+++Coronavirus, minuto a minuto: EE. UU. supera las 4.000 muertes +++

Marzo

**BBC**

Coronavirus en EE.UU.: Donald Trump descarta declarar cuarentena en Nueva York y extiende medidas de distanciamiento hasta el 30 de abril

11 Abril

Coronavirus: Estados Unidos supera a Italia y se convierte en el país del mundo con más muertos y casos de covid-19

27 Junio

**BBC**

Coronavirus en Estados Unidos: 3 claves del "preocupante repunte de contagios"

3 Agosto

**BBC**

Coronavirus en Estados Unidos | La "nueva fase" de coronavirus que atraviesa el país y por qué es una amenaza mayor que cuando empezó la pandemia

Septiembre



Estados Unidos superó los 200.000 muertos, la cifra más alta del mundo

14 Noviembre

**BBC**

Coronavirus en Estados Unidos: 5 cifras que muestran cómo la pandemia de coronavirus está fuera de control en el país

30 Octubre

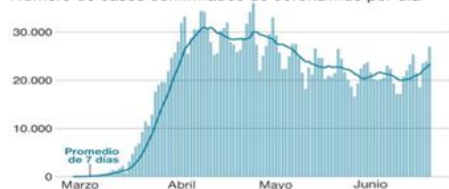
**BBC**

Coronavirus en EE.UU.: récord de contagios y más de 1.000 muertos en 24 horas a pocos días de las elecciones

2020

2021

Los casos de covid-19 crecen otra vez en EE.UU.  
Número de casos confirmados de coronavirus por día



Fuente: COVID Tracking Project


**BBC**

# Problemática

El gobierno de EEUU se ve desbordado por una pandemia **fuera de control.**



El COVID-19 representa la enfermedad #1 que causa más muertes en EEUU actualmente.

Ubicación	Total de casos ↓	Casos nuevos (1 día*)	Casos nuevos (últimos 60 días)	Casos por 1 millón de personas	Muertes
🌐 Todos los países	105,945,462	476,185		13,625	2,312,845
🇺🇸 Estados Unidos	26,957,001	105,027		81,798	462,037
🇮🇳 India	10,826,363	12,059		7,957	154,996
🇧🇷 Brasil	9,447,165	0		44,702	230,034
🇬🇧 Reino Unido	3,929,835	18,262		59,153	112,092
🇷🇺 Rusia	3,907,653	16,379		26,629	75,010

- No hay recursos (de toda índole) para apoyar en todos los estados a la vez.
- Penetración de la idea de pandemia en la sociedad. (ideología)

¿Por qué?



Soluciones



Prevención

Regulaciones

Información

# Hipótesis

Hipótesis
Crear un modelo de predicción que nos permita determinar qué factores conllevan un mayor grado de vulnerabilidad midiendo la tasa de causalidad de ciertos eventos (necesidad de internamiento en hospital, ingreso a UCI, fallecimiento).
Un modelo de forecasting, basándonos en factores temporales (estacionalidad del tiempo, medidas preventivas del gobierno) y sociales (edad, sexo, raza) disponibles nos puede permitir realizar proyecciones a corto plazo (semanal, 1M, 3M), lo cual permite tomar acciones de planificación por parte del sector sanitario.
Mediante técnicas de clusterización, encontrar patrones que inciden en la letalidad de la enfermedad (edad, sexo, raza, enfermedades preexistentes, ingreso a UCI) para la predicción del riesgo de muerte en el paciente.



# Antecedentes

Antecedente	País	Conclusión	Similitud
Adquisición de conocimiento sobre la letalidad de la COVID-19 mediante técnicas de inteligencia artificial.	Cuba (Estudio de pacientes de México)	De los pacientes de la muestra de 9000 casos positivos, los 700 pacientes fallecidos tenían rasgos comunes de letalidad, lo cual pudo ser analizado gracias a las técnicas de IA, como importante aporte a los médicos tratantes.	Se asemeja al estudio del caso de EEUU porque los rasgos como enfermedades preexistentes, también desencadenaron decesos en gran parte de los casos.
Un modelo basado en aprendizaje automático para la predicción de supervivencia en pacientes con infección grave por COVID- 19	China (Estudio con pacientes de Wuhan)	De los pacientes de la muestra de 404 pacientes infectados, se usaron herramientas de aprendizaje automático capturado a través del uso de biomarcadores predictivos de gravedad de enfermedad. De esa forma se distinguió de forma rápida los casos que requerían atención más inmediata y se planificó la capacidad logística de los hospitales.	Se asemeja al estudio del caso de EEUU porque los rasgos de letalidad nos pueden proporcionar información importante para la planificación y poder dimensionar la cantidad de camas y respiradores en los hospitales.

# Exploración de datos

[www.datascience.pe](http://www.datascience.pe)



# Presentación del dataset

Variable	Significado
<i>cdc_report_dt</i>	Fecha en que se informó el CDC (Center for Disease Control and Prevention)
<i>pos_spec_dt</i>	Fecha de la primera recolección de muestras positivas
<i>onset_dt</i>	¿Cuál fue la fecha de inicio?
<i>current_status</i>	¿Cuál es el estado actual de esta persona?
<i>sex</i>	Género
<i>age_group</i>	Categorías de grupos por edad
<i>race and ethnicity (combined)</i>	Perfil demográfico
<i>hosp_yn</i>	¿Fue hospitalizado el paciente?
<i>icu_yn</i>	¿El paciente fue ingresado en una unidad de cuidados intensivos (UCI)?
<i>death_yn</i>	¿Murió el paciente como consecuencia de esta enfermedad?
<i>medcond_yn</i>	¿Tenían alguna condición médica subyacente y / o conductas de riesgo?

# Evolución del COVID en data

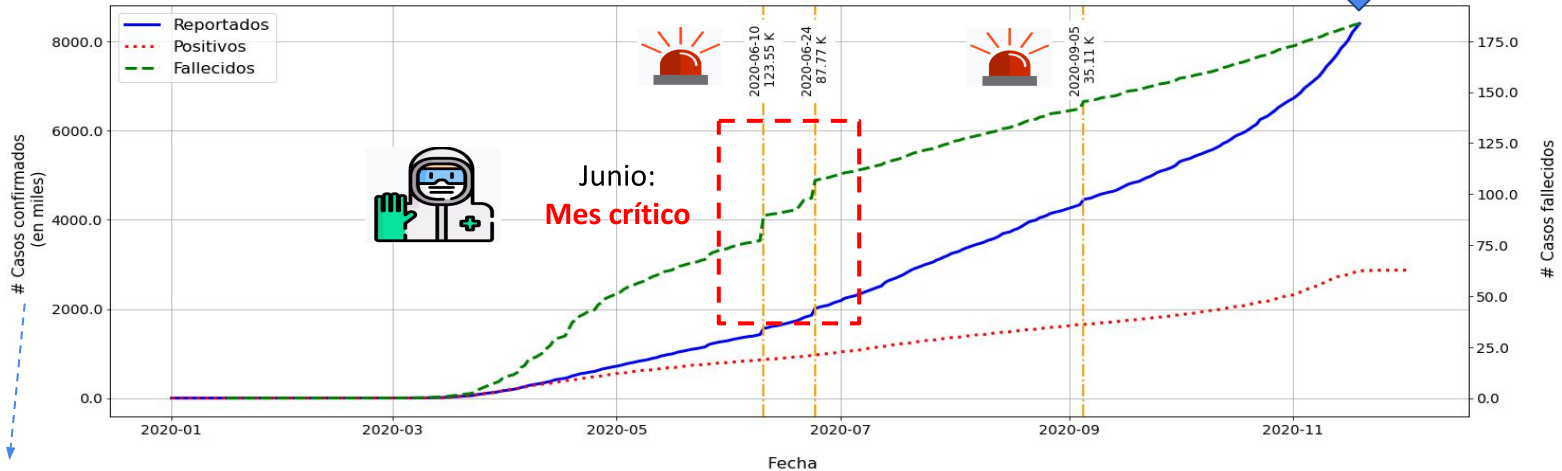
Se observa un crecimiento lineal del # de fallecidos (desde julio), mientras que los reportados crecen en curva. Además, la cantidad de positivos no sigue la tendencia (reducción de # exámenes realizados en últ. meses).

Tasa de mortalidad

(2.25 vs.  
Act: 2.23%)

Se cierra nov. con aprox. 180K  
fallecidos en poco más de 8MM  
de infectados

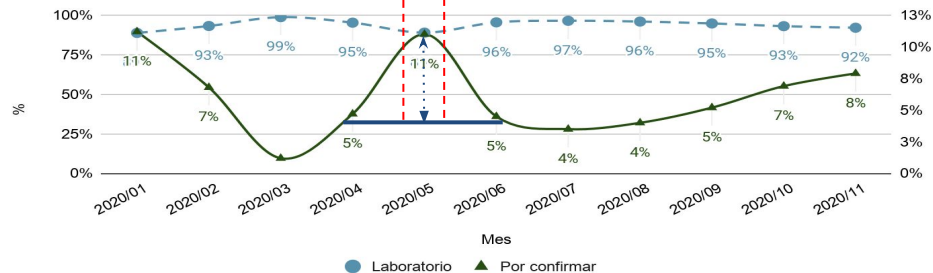
Curva acumulada de casos reportados  
Enero - Diciembre



Referencia a  
reportados y  
positivos

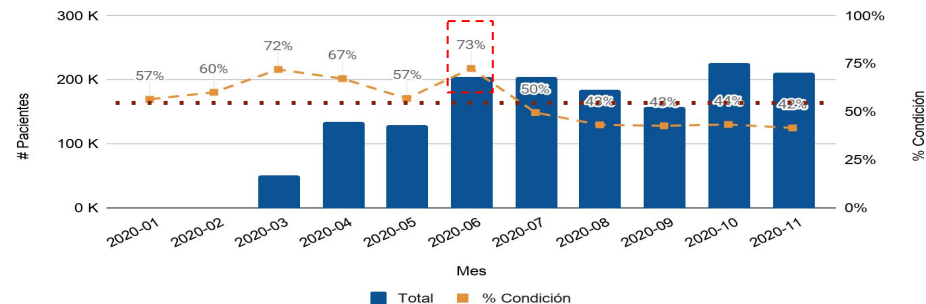
# ¿Que pasó en junio?

Evolución de casos confirmados



Posible evolución del pico de casos por confirmar de laboratorio de mayo a junio.

Evolutivo del % de pacientes con condición

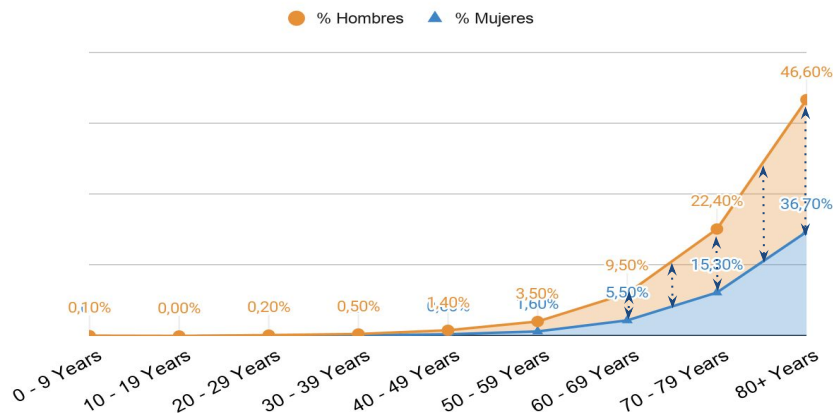


Seguido de marzo (a diferente volumen), es el mes con mayor % de pacientes con comorbilidades.

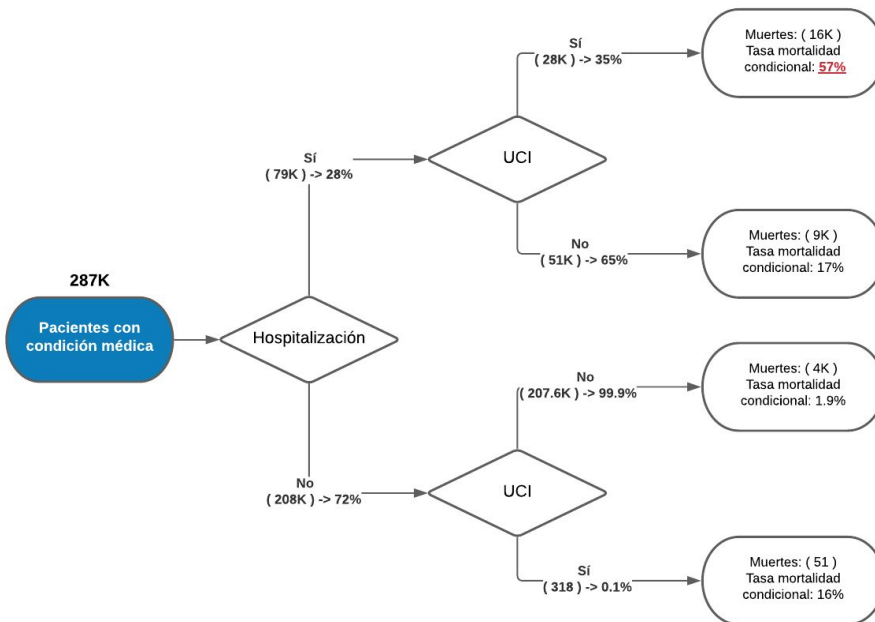


# Condiciones de vulnerabilidad

Influencia de la edad y sexo en el % fallecidos



Se observa un aumento en la brecha de las curvas por sexo del % de fallecidos a medida que pasamos los 50 años de edad, llegando a una diferencia máxima de 10% (80 a más).



# Modelamiento

[www.datascience.pe](http://www.datascience.pe)

Tratamiento de las variables categóricas, para que puedan participar en el entrenamiento.

```
from sklearn.preprocessing import LabelEncoder
```

+ Code

+ Markdown

```
LB_Encode = LabelEncoder()
```

```
data['current_status']=LB_Encode.fit_transform(data['current_status'])
data['sex']=LB_Encode.fit_transform(data['sex'])
data['age_group']=LB_Encode.fit_transform(data['age_group'])
data['Race']=LB_Encode.fit_transform(data['Race'])
data['hosp_yn']=LB_Encode.fit_transform(data['hosp_yn'])
data['icu_yn']=LB_Encode.fit_transform(data['icu_yn'])
data['death_yn']=LB_Encode.fit_transform(data['death_yn'])
data['medcond_yn']=LB_Encode.fit_transform(data['medcond_yn'])
```

Eliminar las variables que no aportan al modelo como las fechas.

```
data.drop(['cdc_report_dt', 'Periodo'], axis=1, inplace=True)
```

[+ Code](#)[+ Markdown](#)

```
data.head()
```

:

	current_status	sex	age_group	hosp_yn	icu_yn	death_yn	medcond_yn	Race
0	0	1	1	1	2	1	1	2
1	0	1	1	1	1	1	1	2
2	0	1	1	1	1	1	1	2
3	0	1	1	0	0	1	0	2
4	0	1	1	1	1	1	3	2

Se separaron los datos para 'X' e 'y'

Se realizó un análisis de componente principal

```
y = data_.pop('death_yn')  
X = data_
```

+ Code

+ Markdown

```
from sklearn.preprocessing import scale  
  
#Scaling the values  
X.iloc[:, :] = scale(X.iloc[:, :])
```

```
import numpy as np  
from sklearn.decomposition import PCA
```

```
hpc = PCA(n_components=7).fit(X.iloc[:, :])  
hpc
```

```
] PCA(n_components=7)
```



Se puede apreciar que los 5 primeros componentes en conjunto representan el 86%

Consideraremos esas variables para el modelo.

```
var1=np.cumsum(np.round(hpc.explained_variance_ratio_, decimals=4)*100)
var1|
```

```
array([ 29.68,  45.39,  59.72,  73.59,  86.21,  95.22, 100.  ])
```

[+ Code](#)[+ Markdown](#)

```
hpc = PCA(n_components=5).fit(X.iloc[:, :])
hpc
```

```
PCA(n_components=5)
```

Se entrena el modelo con el hpc

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(hpc,y,test_size=0.33, random_state = 0)
```

```
from sklearn.ensemble import RandomForestClassifier
rf = RandomForestClassifier()
rf.fit(x_train, y_train)
prediction=rf.predict(x_test)
from sklearn.metrics import confusion_matrix,classification_report,accuracy_score
print(confusion_matrix(y_test,prediction))
print(accuracy_score(y_test,prediction))
print(classification_report(y_test,prediction))
```

```
[[167342  2997]
 [ 5452  5344]]
0.953355232285312
```

	precision	recall	f1-score	support
0.0	0.97	0.98	0.98	170339
1.0	0.64	0.49	0.56	10796
accuracy			0.95	181135
macro avg	0.80	0.74	0.77	181135
weighted avg	0.95	0.95	0.95	181135

