Análisis de Datos de Pacientes y Consultantes con COVID-19

Pedro Remigio Pingarilho Fermin Gomez Miguel Di Luca
Tutor: Dra. María Juliana Gambini

29 de septiembre de 2021

Índice

1. Introducción **4**

2. Estado del Arte **5**

3. Descripción de los Métodos Utilizados en Este Trabajo **11**
3.1. K vecinos más cercanos con Clasificador Bayesiano Ingenuo (KNNNB) . . . 11
3.1.1. K vecinos más cercanos (KNN) . . . 11
3.1.2. Clasificador Bayesiano Ingenuo . . . 11
3.1.3. KNNNB - Instance Cloning Local Naive Bayes (ICLNB) . . . 11
3.2. Arboles de Decisión . . . 12
3.3. Random Forest . . . 14
3.4. Arboles de regresión . . . 15
3.5. Máquinas de vectores de soporte . . . 16
3.5.1. Clasificador de margen máximo . . . 16
3.5.2. Vectores de soporte para clasificación . . . 18
3.5.3. Máquinas de vectores de soporte . . . 19
3.6. C4.5 . . . 21
3.7. Regresión Lineal Múltiple . . . 21
3.8. Regresión Logística . . . 23
3.9. Redes Neuronales Artificiales . . . 23
3.9.1. Perceptrón Simple . . . 23
3.9.2. Perceptrón Multicapa . . . 24
3.10. Validación Cruzada en K Partes . . . 25
3.11. Selección de variables mediante el método de eliminación hacía atrás . . . 26

4. Análisis de tendencias en búsquedas de síntomas **28**
4.1. Conjunto de datos utilizados . . . 28
4.2. Aplicación del método Regresión Lineal Múltiple . . . 29
4.2.1. Resultados . . . 30
4.3. Aplicación del método Árboles de regresión . . . 37
4.3.1. Resultados . . . 39
4.4. Conclusiones . . . 44

5. Diagnosis COVID-19 alternativo basado en síntomas presentes **45**
5.1. Conjunto de datos utilizados . . . 45
5.2. Aplicación del método Arboles de Decisión . . . 47
5.2.1. Resultados . . . 48
5.3. Conclusiones . . . 50

6. Análisis de Casos Positivos en México en Base a las Pre-Condiciones de una Persona **51**
6.1. Conjunto de datos utilizados . . . 51
6.2. Aplicación del método C4.5 . . . 53
6.2.1. Resultados . . . 54
6.3. Aplicación del método Regresión Logística . . . 55
6.3.1. Resultados . . . 56
6.4. Aplicación del método Redes Neuronales - Perceptrón Multicapa . . . 58
6.4.1. Resultados . . . 58
6.5. Conclusiones . . . 60

7. Análisis de mortalidad en base a la dieta de los países **62**
7.1. Conjunto de datos utilizados 62
7.2. Aplicación del método Máquinas de Vectores de Soporte 64
7.2.1. Resultados 65
7.3. Aplicación del método K vecinos más cercanos con Clasificador bayesiano ingenuo (KNN-NB) 68
7.3.1. Resultados 69
7.4. Conclusiones 70

8. Conclusiones **72**

1. Introducción

El virus COVID-19 (SARS-CoV-2) fue identificado por primera vez en Diciembre del 2019 en la ciudad de Wuhan, provincia de Hubei, China.
A pesar de que hay muchas hipótesis sobre el origen del virus, el mismo todavía sigue indefinido. Una de las hipótesis [8] asevera que provino de la mutación de un virus de una especie de murciélago. El primer contagio no fue de tipo interpersonal, sino que el virus infecta a pangolines y estos transmiten la enfermedad.
Se conjetura que el virus se propaga por el aire, mediante micropartículas emitidas a través de la tos, el estornudo o el habla de personas que poseen el virus. El virus puede vivir en distintas superficies por horas o incluso por días, dependiendo del tipo de superficie.
Los primeros síntomas se manifiestan luego de aproximadamente seis días de haber contraído el virus, aunque este período puede variar entre uno y catorce días. Entre los síntomas más comunes se encuentran fiebre, tos y cansancio. Al agravarse la condición del paciente, puede causar dificultad al respirar, pérdida del olfato y gusto y dolores en el pecho. Sin embargo, muchos casos de la enfermedad son asintomáticos, lo que hace aún más difícil aislar a la población sana de la parte de la población que se encuentra infectada.

El 11 de marzo de 2020 la Organización Mundial de la Salud (OMS) reconoció esta enfermedad como pandemia. Debido a su rápido contagio, logró alcanzar 58,4 millones de casos acumulados al día 23 de noviembre de 2020 [16].
La velocidad de propagación del virus impacta principalmente a los hospitales, debido a que los recursos de los mismos son limitados, pueden llegar a colapsar. Lo cual implica que muchas personas puedan morir a causa de no poder recibir el tratamiento y la atención adecuada contra la enfermedad.
Las autoridades de los distintos países del mundo tomaron diversas medidas con el fin de contener la propagación del virus y evitar el colapso del sistema sanitario. Entre ellas, cerrar las fronteras, suspender servicios comunitarios, suspender la actividad escolar, suspender el transporte público, suspender actividades comerciales y confinamiento, entre otras. Sin embargo, las medidas tomadas no fueron suficientes en muchos países, sino que además trajeron problemas económicos, sociales, políticos y educativos.

El COVID-19 es un nuevo desafío para la humanidad, la gran cantidad de variables del virus que aún son desconocidas genera una incertidumbre que llama a explorar nuevas soluciones innovadoras que puedan aportar a reducir el impacto como la propagación de la enfermedad y la muerte de personas.
Este trabajo consiste en analizar los múltiples conjuntos de datos disponibles, que incluyen información clínica de pacientes, demográfica, crecimiento de casos, políticas gubernamentales, y redes sociales, entre otros, para luego combinar los datos útiles, y mediante métodos de aprendizaje automático poder encontrar nueva información útil sobre el virus.

2. Estado del Arte

En esta sección, se presentan algunos trabajos relacionados con la evaluación de políticas públicas y el avance del COVID-19 a través del análisis de datos.

Con el avance de la pandemia, se establecieron diferentes modelos de cuarentena en los distintos países del mundo. Como mencionamos anteriormente, la cuarentena afecta a la sociedad de diversas formas, por lo tanto, es muy importante encontrar el modelo de cuarentena que ofrece mejores resultados. El estudio [5] utiliza datos brindados por Google, que informan la cantidad de tiempo que las personas permanecen en un lugar determinado y lo compara con el tiempo que solía permanecer previo a la pandemia, se divide en actividades de tipo recreativas, compras, lugares de trabajo, entre otras. Al conjunto de datos se le agregó el estado de las escuelas, si estas permanecían abiertas o cerradas. Se utilizan algoritmos de aprendizaje automático, como Random Forest y K-Nearest Neighbor para encontrar que tan eficiente fue el cierre de una actividad en concreto, teniendo en cuenta la correlación entre el cambio de tiempo de estadía por cada actividad y la tasa de infección. Del mismo modo, evalúan la eficiencia de los distintos países teniendo en cuenta todas las actividades. Al finalizar, se llega a la conclusión de que cada actividad, individualmente, no tiene un impacto directo en la eficiencia de la cuarentena. En cambio, tiene un mayor impacto al tomar todos los factores en conjunto. De todas formas, el alto error obtenido indica que las actividades tomados en cuenta no son los únicos que modifican la eficiencia de la cuarentena. El estudio muestra una alta correlación entre una cuarentena estricta y la tasa de infección.

En algunos países de Europa, como en Alemania, se ha observado una menor tasa de mortalidad que en otros. Esto lleva a la analizar los distintos factores que pueden ser responsables de este fenómeno. En el estudio [10] se busca analizar si las dietas en los diferentes países pueden ser un factor determinante en la tasa de mortalidad. El estudio gira en torno a la enzima convertidora de angiotensina 2 (ECA2), la cual es el punto de entrada de los coronavirus, y a la enzima convertidora de angiotensina (ECA). La alimentación tiene un gran efecto en los niveles de ECA, una dieta rica en grasas saturadas la incrementa, mientras que otros alimentos funcionan como inhibidores de la ECA. A su vez, se supone que una modificación en la cantidad de ECA produciría un cambio en la expresión de ECA2, por lo cual, es importante analizar las dietas de los países para poder determinar cómo modifican la ECA, y por consiguiente, la ECA2. Se tiene en cuenta que, la evidencia actual parece indicar que los inhibidores de ECA no incrementen la cantidad de ECA2, por lo que los cambios en la ECA podrían no ser tan relevantes. Se observa que ciertas comidas que inhiben la ECA son consumidas en gran proporción en países con baja tasa de mortalidad como Corea y Taiwán. También se analizaron los casos de Bulgaria, Grecia, Rumania y Turquía, estos países poseen una baja tasa de mortalidad y sus habitantes suelen consumir alimentos que inhiben la ECA.

Debido a los diversos problemas causados por el SARS-CoV-2 el estudio [1] se especializa en analizar cinco problemas, usando el algoritmo Support Vector Regression, comparando con los resultados de aplicar regresión lineal y polinómica. Los trabajos realizados en el artículo son (I) predecir la proliferación del COVID por región, (II) analizar el crecimiento de la enfermedad con respecto a las mitigaciones de los países, (III) predecir cómo va a terminar la pandemia, (IV) analizar el contagio del virus y (V) relacionar la enfermedad con las condiciones climáticas.

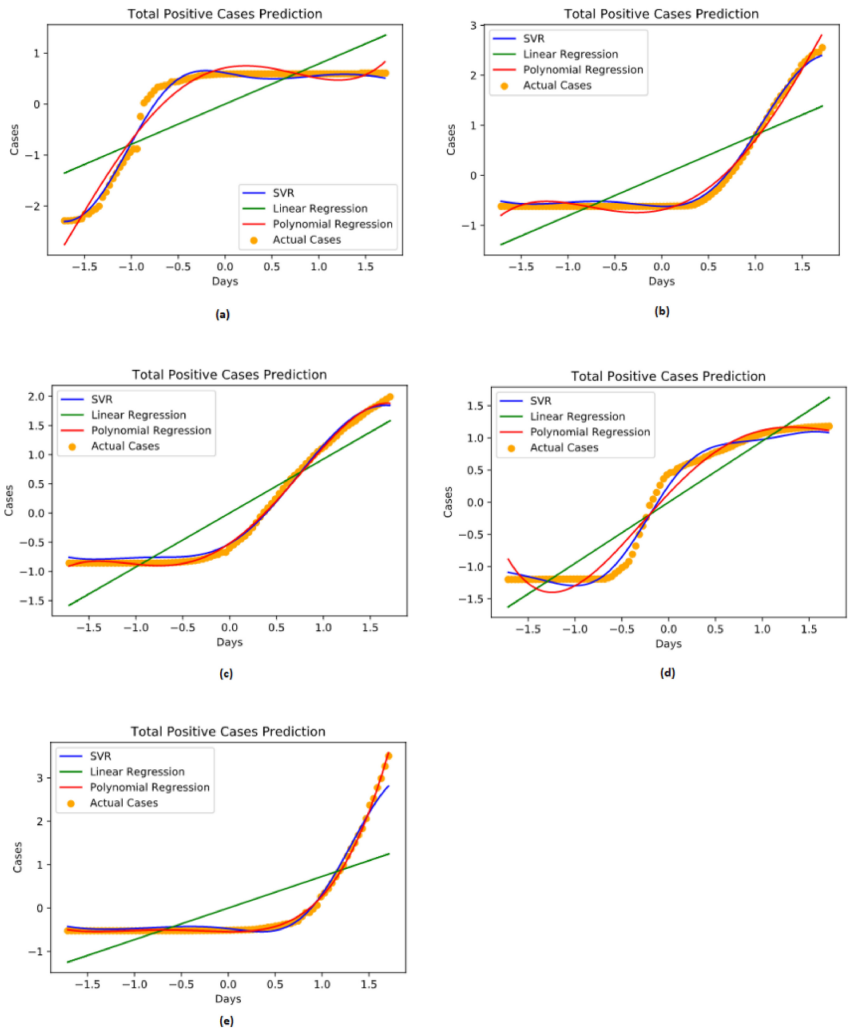
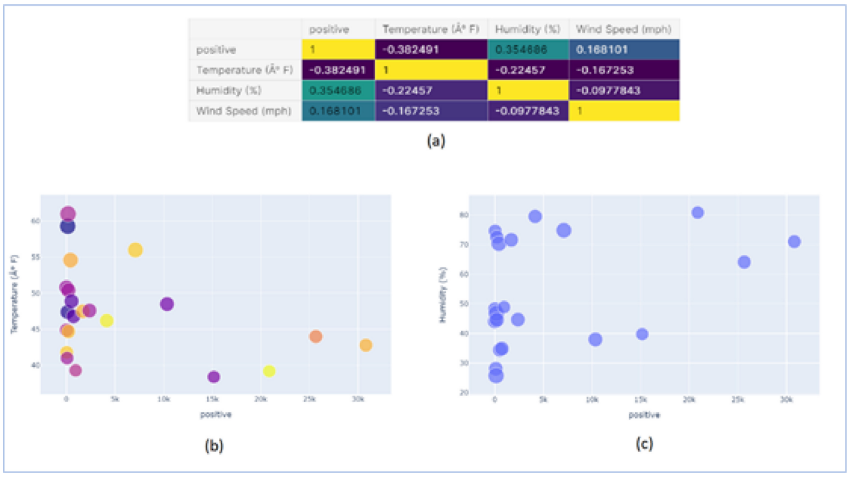


Figura 1: Exactitud de predecir el número de casos positivos en diferentes regiones (a) China, (b)EEUU, (c)Italia, (d) Corea, (e)India

La Figura 1 presenta las predicciones de casos positivos en diferentes regiones, usando los diferentes algoritmos para predecir las variaciones de acuerdo al tiempo. El estudio compara la exactitud de los casos reales con las predicciones hechas con los diferentes algoritmos. Para la mayor parte de las preguntas I, II y IV, el algoritmo SVR es más preciso que los otros algoritmos analizados. Sin embargo, para el estudio III tuvo mayor precisión la regresión polinómica por poca diferencia de SVR.



	positive	Temperature (Â° F)	Humidity (%)	Wind Speed (mph)
positive	1	-0.382491	0.354686	0.168101
Temperature (Â° F)	-0.382491	1	-0.22457	-0.167253
Humidity (%)	0.354686	-0.22457	1	-0.0977843
Wind Speed (mph)	0.168101	-0.167253	-0.0977843	1

Figura 2: Resultado del análisis de diferentes informaciones de correlaciones de Nueva York. (a) Tabla de correlaciones entre la temperatura, humedad, velocidad del viento y casos positivos, (b) correlación entre temperatura y casos positivos, (c) correlación entre casos positivos y humedad

Para el quinto estudio se usó la correlación de Pearson para ver la comparación. Se ve que el aumento en la humedad aumenta el número de infecciones. El aumento de temperatura también influye la contaminación, siendo una relación inversa (el aumento de la temperatura disminuye el número de casos). La Figura 2, presenta estos resultados en la ciudad de Nueva York.

El Artículo [13] realiza una comparación directa en las medidas tomadas entre Taiwán y Nueva Zelanda, dos países chicos que tuvieron un número de casos por millón de habitantes ínfimo si comparado con otros países en el mundo. El mismo resalta que Taiwán obtuvo un mejor rendimiento en número de casos y en la economía que Nueva Zelanda, dado que tuvo una respuesta mas rápida a las implementaciones de control, como el cierre de la frontera con China, el aislamiento de casos posibles de infección, la distribución de barbijos a la población (y concientización del uso por la misma, dado que no es la primera vez que tienen que combatir una enfermedades contagiosa respiratoria, como el brote de Sars en el 2003), control de lugares con alta densidad de personas, y una amplia comunicación sobre el tema de salud en la población. Dado eso, no fue necesario ningún tipo de confinamiento. Como punto positivo de Nueva Zelanda, aunque la comunicación fue tardía, tuvo mayor alcance a su población, dado que tradujeron todos los comunicados a diferentes idiomas nativos de la región. El artículo ofrece como recomendaciones para una próxima pandemia los siguientes puntos:

- establecer / fortalecer a una agencia nacional de salud
- formular planes genéricos para posibles pandemias
- Mayor investigación en recursos / infraestructura para acelerar la respuesta a posibles brotes en el futuro, como sistemas de vigilancia, políticas fronterizas de rápida implementación, reglas más robustas de cuarentena principalmente para visitantes, facilidades de monitoramento & tracing a las personas y la facilitación en la distribución de barbijos a la población.
- Hacer una revisión de los recursos humanos (fuerza laboral) para prevenir la falta de personal capacitado en situaciones de emergencia.
- Creación de sistemas para evaluar las respuestas a una pandemia.
- Intentar crear una cultura de aceptación a propuestas de combate a pandemias.

En el trabajo [6] se estudia probabilidades de que un paciente sea diagnosticado COVID-19 positivo utilizando distintos métodos de aprendizaje automático.
Los datos utilizados fueron extraídos de 235 pacientes del hospital Israelita Albert Einstein en São Paulo, Brasil de los cuales el 43 % fue diagnosticado COVID-19 positivo mediante el test RT-PCR. Se recolectaron las siguientes variables obtenidas de los resultados de recuento sanguíneo y PCR: edad, género, hemoglobina, plaquetas, glóbulos rojos, concentración de hemoglobina corpuscular media (MCHC), hemoglobina corpuscular media (MCH), ancho de distribución de glóbulos rojos (RDW), volumen corpuscular medio (MCV), leucocitos, linfocitos, monocitos, basófilos, eosinófilos y proteína c reactiva (PCR). Se probó el rendimiento de 5 métodos y luego se los compraron con el objetivo de probar la capacidad de los métodos de aprendizaje automático a la hora de predecir el riesgo de que un paciente tenga COVID.

Algoritmo	**AUC**	**Sensibilidad**	**Especificida**	**F1-score**	**Brier score**	**PPV**	**NPV**
Supp. Vec. Machines	0.847	0.667	0.850	0.724	0.160	0.778	0.773
Random Forests	0.847	0.667	0.850	0.724	0.161	0.778	0.773
Neural Networks	0.844	0.742	0.800	0.742	0.187	0.742	0.800
Logistic Regression	0.843	0.742	0.825	0.754	0.161	0.767	0.805
Grad. Boost. Trees	0.842	0.806	0.800	0.781	0.171	0.758	0.843

Tabla 1: Métricas de rendimiento para cada método usado con el conjunto de datos de prueba.

La Tabla 1 muestra para cada uno de los distintos métodos de aprendizaje automático utilizados los resultados obtenidos al evaluarlo con las siguientes métricas: AUC, Sensibilidad, Especificida, F1-Score, Brier score, PPV (Valor Predictivo Positivo o Precisión), NPV (Valor Predictivo Negativo). Los algoritmos de mejor rendimiento resultaron ser Support Vector Machines y Random Forests. Sin embargo, todos obtuvieron un AUC mayor a 8.4 lo cual es bastante prometedor y permitiría usar estos métodos para poder definir las probabilidades de que un paciente sea diagnosticado COVID-19 y usar los test de PCR de manera eficiente en caso de escasez.

El trabajo [3] consiste en el estudio de 336 casos de pacientes diagnosticados con COVID-19 mediante el test de PCR en Shanghai, China. El objeto del estudio fue identificar de 220 características clínicas la más significativas asociadas con casos graves/críticos, y predecir si un paciente evolucionará en un caso grave/crítico.
Se compararon las características clínicas dividiendo las muestras en dos grupos, según si son correspondientes a casos graves/críticos o no. Para ello se utilizó la prueba de t-Student y la prueba de rango de Wilcox, para las características con valores de $p < 0{,}01$ en ambos algoritmos se consideró que estaban significativamente asociadas con un síntoma grave/crítico. En total 36 características clínicas se asociaron con casos graves/críticos. Estas son: inhibidores de Colinesterasa, Prealbúmina, proteína de unión al Retinol, proteina total, Albúmina, Calcio total, Transferrina, valor absoluto de CD19, valor absoluto de CD4, porcentaje de Linfocitos, Triyodotironina libre, Sodio, Cloro, porcentage de CD3, porcentage de CD4, Tiroxina, Tiroxina libre, Endotoxina cuantitativa, Lactato Deshidrogenasa, Reductasa de GSH, Glicoproteína ácida, Proteína C-reactiva, PCR hipersensible, porcentaje de Neutrófilos, precursor del péptido B Batriurético, Procalcitonina, Mioglobina sérica, Tromboplastina activada, Nivel de ángulo de fibrinógeno, índice de coagulación, función plaquetaria MA, ALB/GLB, Edad, LDH, Cl.
En una segunda etapa se consideró la combinación de características para la predicción de casos graves/-críticos usando el método de Support Vector Machines. Para evitar el sobreajuste en el método utilizado se buscó la combinación óptima que use menos de 5 características. Para ello se utilizó el método de ataque exhaustivo combinando 2, 3 y 4 características. Para medir el rendimiento del modelo se utilizó el área por debajo de la curva ROC (AUC) tanto para el conjunto de prueba como para el de entrenamiento

del modelo. La Tabla 2 muestra los resultados para la métrica AUC de algunas de las combinación de características clínicas que obtuvieron mejor resultado.

Combinaciones	**AUC** (Conjunto de entrenamiento)	**AUC** (Conjunto de prueba)
Edad, GSH, porcentaje CD3, proteina total	0.999616858	0.975711
Porcentaje de Neutrófilos, Albúmina, GSH, porcentaje CD4	0.997318008	0.975711
HCRP, Mioglobina sérica, CL, Porcentaje de CD4	0.998357964	0.969466
Edad, Cl, Calcio, LDH	0.997318008	0.951748
Edad, Mioglobina sérica, Proteína de unión al Retinol, Glicoproteína ácida	0.990960452	0.951423
Porcentaje de Neutrófilos, Procalcitonina, Mioglobina sérica, Proteina total	0.977024482	0.958362

Tabla 2: Resultados de la predicciones más performantes

En el trabajo [2] se investiga la correlación entre variables climáticas y el brote de la enfermedad COVID-19. En una instancia inicial se demostró utilizando la regresión de mínimos cuadrados ordinarios (OLS) que los factores climáticos como temperatura y humedad influyen tanto en la cantidad como en la tasa de crecimiento de casos COVID-19 positivos.
Resultó importante destacar en la investigación que los factores climáticos no son los únicos que inciden en el brote de la enfermedad, sino que la densidad de la población, grado de urbanización y otros factores demográficos afectan fuertemente el grado de contagio. Como ejemplo, a mayor densidad de personas y horas de luz solar aumentan las probabilidades de que las personas salgan de sus respectivas casas y entren en contacto con el virus.
Por lo tanto se utilizó un conjunto de datos en el que para cada ciudad se tomaron los siguientes atributos: cantidad de casos confirmados, activos, muertes, recuperados, tamaño de la población, tasa de fertilidad, edad media, porcentaje de urbanización, camas ICU por 1000 habitantes, tasa de infección, temperatura, humedad, horas de luz solar y velocidad del viento. Se implementó un modelo de regresión para predecir el número de casos y muertes. Además, se utilizo el método de Random Forest para la selección de atributos mas significativos del conjunto de datos. Analizando los resultados se concluye que los efectos del clima tienen un mayor impacto en la cantidad de casos y muertes por COVID-19 que los demográficos. La Figura 3 muestra un gráfico de barras donde se puede visualizar la importancia de los atributos según su influencia en la cantidad de casos COVID-19 positivos.

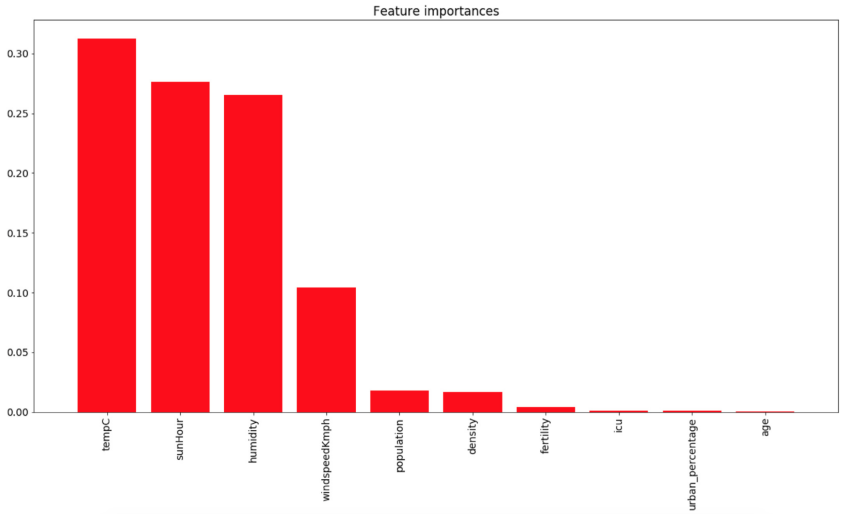


Figura 3: Importancia de atributos para casos COVID-19 positivos

La Figura 4 muestra un gráfico de barras donde se puede visualizar la importancia de los atributos según su influencia en la cantidad de muertes por COVID-19.

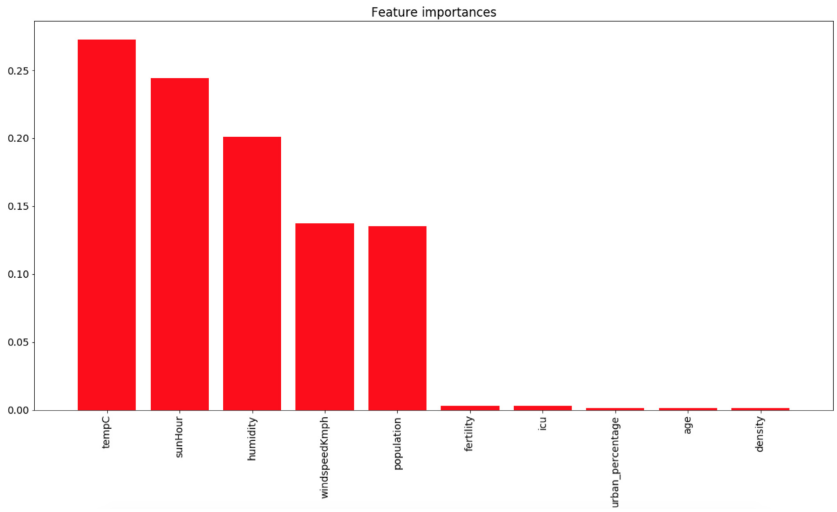


Figura 4: Importancia de atributos para muertes por COVID-19

Analizando ambas Figuras 3 y 4 se demuestra que los factores climáticas son más importantes cuando se comparan con otros factores como los obtenidos mediante un censo que incluyen tamaño de población, edad y porcentaje de urbanización.

3. Descripción de los Métodos Utilizados en Este Trabajo

3.1. K vecinos más cercanos con Clasificador Bayesiano Ingenuo (KNNNB)

El algoritmo K vecinos más cercanos con Clasificador Bayesiano ingenuo (KNNNB) es una variante del algoritmo K vecinos más cercanos (KNN) que hace uso de un Clasificador Bayesiano Ingenuo. Por eso, es necesario introducir ambos temas para poder desarrollar sobre el algoritmo utilizado en los análisis de esta tesis.

3.1.1. K vecinos más cercanos (KNN)

El método consiste en clasificar, de manera supervisada, datos en base a los de entrenamiento con valores cercanos a los suyos. Dado un número natural K y un dato para clasificar x_0, es posible decir que la probabilidad de que este dato pertenezca a la clase j es de

$$P(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j) \tag{1}$$

K es el número de datos cercanos al que se desea clasificar, o sea, el número de vecinos utilizado para definir la clasificación más probable. N_0 es el conjunto de los K puntos más cercanos, calculados usando la distancia Euclidea entre x_0 y cada punto en el conjunto de entrenamiento. I es una función que retorna 1 si el dato pertenece a la clase j y 0 sino. La clasificación predicha se da a la clase de mayor probabilidad.[7]

3.1.2. Clasificador Bayesiano Ingenuo

Es una técnica de clasificación simple creada del Teorema de Bayes (Ecuación 2). Es ingenuo pues presupone la independencia de las características descriptivas de un dato.

$$P(c|x) = P(c)\frac{P(x|c)}{P(x)} \tag{2}$$

El clasificador define la probabilidad que un dato x pertenezca a una clase c_j es de

$$P(Y = c_j|X = x) = \prod_{i=1}^{n} P(c_j|X_i = x_i) \tag{3}$$

siendo x_is las características descriptivas del dato x.

3.1.3. KNNNB - Instance Cloning Local Naive Bayes (ICLNB)

KNNNB es la alteración del clasificador KNN en el cual a los vecinos encontrados de cada punto se aplica el método de Bayes para encontrar la categorización adecuada. Para mejorar la predicción con Bayes el clasificador realiza duplicaciones de los datos en base al conjunto de vecinos. Dado la duplicación descrita, se le da al algoritmo el nombre de Instance Cloning Local Naive Bayes (Clon de Instancias Locales para el Clasificador Bayesiano Ingenuo).

Sea un vector de atributos (o características descriptivas) $< \alpha_0(x), \alpha_1(x), ..., \alpha_n(x) >$ con $\alpha_i(x)$ el i-ésimo atributo de la variable x. Para cada relación entre dos vecinos x_1 y x_2, se hace una copia en la lista de vecinos si $|\alpha_i(x_1) - \alpha_i(x_2)| < \lambda$ para variables numéricas o $\alpha_i(x_1) = \alpha_i(x_2)$ para otros tipos de variables. λ es un valor mínimo elegido para concluir que los valores son los mismos. Esto significa que valores repetidos de atributos tienen mayor importancia a la hora de la toma de decisiones con el método de Bayes. La función de la copia es el intento de darle más peso a los vecinos que tienen atributos idénticos.

El algoritmo funcionaria en el siguiente orden. Primero, se buscan los K vecinos más cercanos. Una vez encontrados, se realiza las comparaciones y los clones de los vecinos en base a sus atributos. A partir de este resultado, se le aplica el clasificador de Bayes Ingenuo con el conjunto de vecinos y sus clones como conjunto de entrenamiento. [11].

3.2. Arboles de Decisión

Los árboles de decisión son un modelo de aprendizaje supervisado, no paramétrico cuyos datos de entrada consiste en valores discretos. Se lo utiliza tanto para tareas de clasificación como de regresión.

En este método el proceso de aprendizaje se lo representa con un árbol. Los árboles consisten de dos elementos base: nodos y ramificaciones hacia otros nodos. Los nodos pueden ser raíz, intermedios u hoja también denominados nodos terminales. La Figura 5 muestra un ejemplo de árbol de decisión.

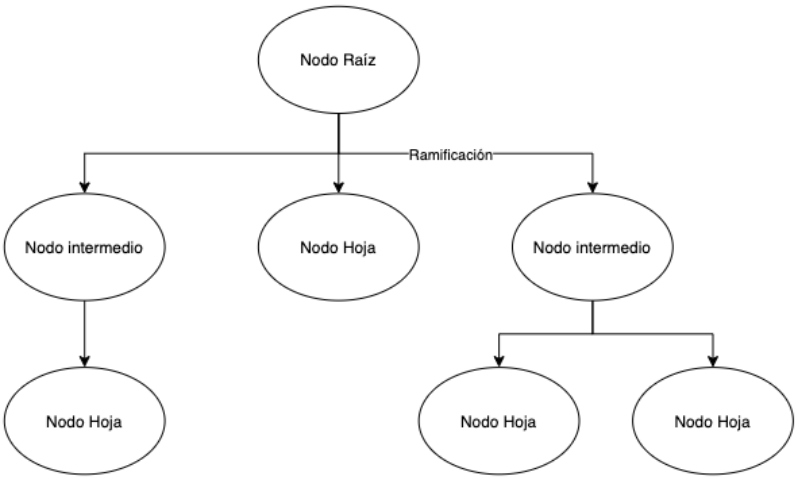


Figura 5: Estructura de un árbol.

En cada nodo se evalúa aquella característica descriptiva que contiene la mayor cantidad de "información" sobre la característica objetivo y mejor divide a los datos. Luego se realiza una ramificación a lo largo de los valores de esta característica de modo que los valores de las características objetivo para los sub-conjuntos resultantes contenga predominancia de observaciones de una sola clase. Este proceso de encontrar la característica "más informativa" se realiza hasta que logramos un criterio de detención en el que finalmente terminamos en los llamados nodos hoja.
Los nodos hoja contienen las predicciones que haremos para las nuevas instancias de consulta presentadas a nuestro modelo entrenado. Esto es posible porque el modelo ha aprendido la estructura subyacente de los datos de entrenamiento y, por lo tanto, puede, dadas algunas suposiciones, hacer predicciones sobre el valor de la entidad/clase objetivo de instancias de consulta desconocidas por el árbol.
Para de determinar cuál es la mejor característica para dividir el nodo utilizamos funciones de información de ganancia. Para esto necesitamos el cálculo de métricas como lo son el Índice de Gini o Entropía de Shannon. La Entropía de Shannon mide el grado de información que posee un conjunto de datos. Entonces:

Sea A es un atributo y $S_v = \{s \in S/A(s) = v\}$ el subconjunto de S para el cual el atributo A tiene valor v.

$$Entropía(S_v) = -\sum_{j=1}^{n} p_j^v * log_2(p_j^v) \tag{4}$$

donde p_j^v es la frecuencia relativa del valor v y clase j en S, $j = 1, \ldots, n$.
La función información de ganancia utilizando la Entropía de Shannon resulta:
Dado un atributo A, definimos:

$$Ganancia(S, A) = Entropía(S) - \sum_{v \in Valores(A)} \frac{|S_v|}{|S|} Entropía(S_v) \tag{5}$$

donde $S_v = \{s \in S | A(s) = v\}$ es el subconjunto de S para el cual el atributo A tiene valor v.

Algoritmo para construir el árbol:

Entrada:

- Ejemplos: Los ejemplos de entrenamiento.
- Atributo Objetivo: Es el atributo cuyos valores queremos predecir con el árbol.
- Atributos: Lista del resto de los atributos

Salida:

- Árbol de decisión que clasifica los ejemplos dados

Algoritmo:

$ID3(Ejemplos, AtributoObjetivo, Atributos)$

1. Crea un nodo raíz para el árbol
2. Si todos los ejemplos tienen el mismo valor en su atributo objetivo, devuelve la raíz del árbol de un solo nodo etiquetado con el valor del atributo de destino en los ejemplos.
3. Si el número de atributos de predicción está vacío, devolver la raíz del árbol de un solo nodo, con etiquetado con el valor más común del atributo objetivo en los ejemplos.
4. De lo contrario:
   4.1. $A \leftarrow$ El atributo que maximiza la Ganancia.
   4.2. $raíz = A$.
   4.3. Para cada valor v_i que pueda tomar el atributo A:
      4.3.1. $Ejemplos(v_i) \leftarrow$ subconjunto de ejemplos que tienen el valor v_i para A.
      4.3.2. Agregar la ramificación a *raíz* correspondiente a $A = v_i$ y subárbol $= ID3(Ejemplos - Ejemplos(v_i), AtributoObjetivo, Atributos - A)$

A modo de ejemplo la Figura 6 representa el árbol de decisión resultante creado con el algoritmo anterior utilizando como entrada un conjunto de 303 observaciones de distintos pacientes con enfermedades cardíacas.
La variable objetivo en cuestión es la presencia de enfermedad cardíaca. Por lo que un valor de salida del árbol de "Si" indica que el paciente posee una enfermedad cardíaca, mientras que un valor de "No" indica la falta. Por otro lado hay 4 variables predictoras: "Prueba de estrés con Talio: A", "Frecuencia cardíaca máxima", "Calcio" y "Dolor de pecho".

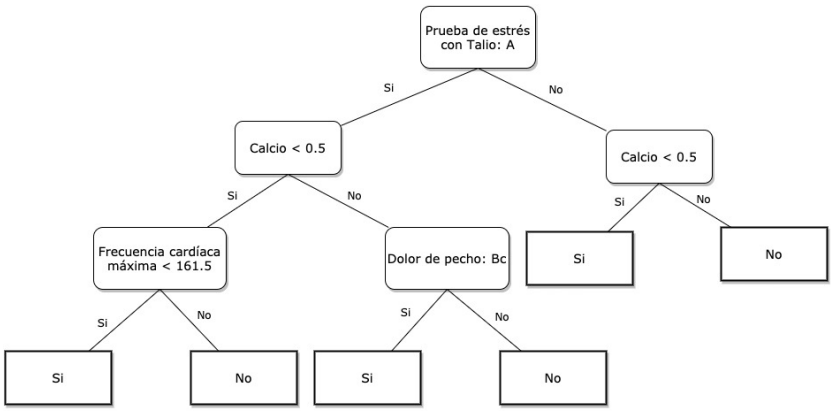


Figura 6: Árbol de decisión

Analizando el árbol resultante del método, vemos que el nodo raíz corresponde al atributo "Prueba de estrés con Talio: A", el cual divide el árbol por la rama izquierda a aquellas observaciones con un valor "normal" en esta prueba y en la rama derecha las demás observaciones.
El atributo "Calcio" divide el árbol en aquellos pacientes cuyo valor esta por encima de 0.5 o por debajo. Análogamente el atributo "Frecuencia cardíaca máxima" para aquellos con un valor por encima de 161.5 o por debajo.
Finalmente el atributo "Dolor de pecho" divide el árbol por la rama izquierda a aquellas observaciones con un valor "angina típica o angina atípica" en esta prueba y en la rama derecha con un valor "leve dolor y asintomático".

3.3. Random Forest

Los modelos de árboles de decisión son modelos de alta varianza y bajo sesgo. En consecuencia, son propensos a sobreajustar los datos de entrenamiento. Otro inconveniente de los modelos de árboles de decisión clásicos es que son relativamente inestables. Esta inestabilidad puede llevar a la situación de que un pequeño cambio en la composición del conjunto de datos lleve a un modelo de árbol completamente diferente. El enfoque de Random Forest se basa en dos conceptos, Bagging o bootstrap aggregatio y muestreo subespacial. La técnica de Bagging consiste en:

1. Divide el conjunto de Entrenamiento en subconjuntos de datos, obteniendo diferentes muestras aleatorias con las siguientes características:
   1.1. Muestras con la misma cantidad de individuos en cada conjunto.
   1.2. Muestras con reemplazo (los individuos pueden repetirse en el mismo conjunto de datos).
   1.3. El tamaño de la muestra es igual al tamaño del conjunto de entrenamiento, pero no contiene a todos los individuos ya que algunos pueden repetirse.
2. Se crea un modelo predictivo con cada conjunto, obteniendo modelos diferentes.
3. Se construye un único modelo predictivo, que es una síntesis de todos los modelos.

Como resultado se disminuye la varianza de un conjunto de datos al realizar remuestreo con reemplazo. También se reduce el sobreajuste. Esto ocurre porque los modelos no pueden sobreaprender ya que ninguno tiene todos los datos de entrenamiento. Como también reduce el ruido producido por los outliers, porque los outliers no pueden estar presenten en todos los modelos. Finalmente mejora la predicción, debido a que lo que no detecta un modelo lo detectan los otros.
Random forest aplica el método de Bagging con una muestra menor o igual al conjunto de entrenamiento

y para cada subconjunto de datos construye el árbol de decisión correspondiente. Luego, el resultado final es el resultado más votado por todos los árboles anteriores.

3.4. Arboles de regresión

El método de árboles de regresión es similar al de árboles de clasificación, pero en este caso, el resultado puede ser un número real y no necesariamente una categoría. Las variables en este método son de tipo continuo.

Veamos un ejemplo. Tenemos un conjunto de observaciones con tres atributos, el peso de la persona, su altura y la cantidad de puntos que obtiene en un partido de básquetbol. Nuestro objetivo es poder predecir la cantidad de puntos que una persona anotará en un partido, dada su altura y su peso. Por lo tanto, decidimos crear un árbol de regresión que ayude a solucionar nuestro problema.

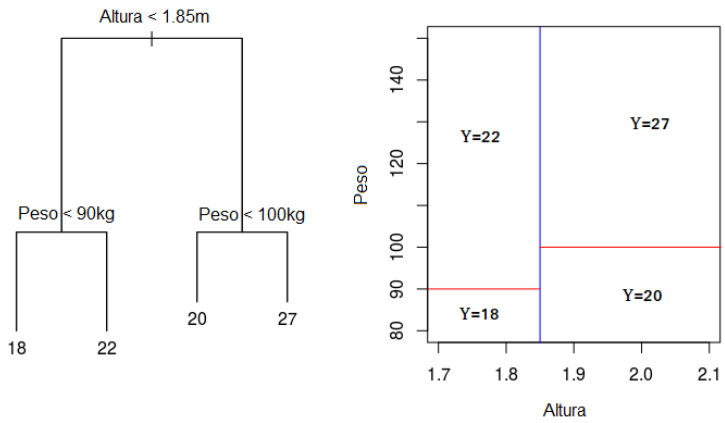


Figura 7: Ejemplo de un árbol de regresión.

En la Figura 7 se ve el resultado de nuestro modelo. Como podemos ver el método dividió al espacio en cuatro subespacios. Ahora cuando queramos clasificar una nueva observación, simplemente nos fijamos a que subespacio pertenece y de acuerdo a eso contestaremos con nuestra predicción. Ahora veamos como construimos nuestro árbol.

Para obtener el nodo raíz, para cada atributo X_j dividimos al conjunto en dos regiones, R_1 y R_2, estas regiones están dividas por un valor $s \in \Re$, todas las observaciones menores a s formarán parte de R_1 y todas las observaciones mayores o iguales a s se encontrarán en R_2.

Por lo indicado anteriormente, s podría ser cualquier valor, pero solo es de importante el valor de s que hace la suma de cuadrados residual (SCR) sea mínima.

$$R_1(j,s) = \{X|X_j < s\} \quad \text{y} \quad R_2(j,s) = \{X|X_j \geq s\} \tag{6}$$

$$SCR = \sum_{i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2 \tag{7}$$

donde y_i es el valor de la variable objetivo correspondiente a la observación i, $\hat{y}_{R_i}$ es el valor medio de la variable objetivo en el subconjunto R_i y como explicamos previamente, s es el valor en el cual se divide

al conjunto en R_1 y R_2.

Ahora tomamos el atributo con menor SCR y lo establecemos como raíz, para los nodos sucesivos realizamos el mismo procedimiento recursivamente. Para finalizar de expandir el árbol establecemos nuestro propio criterio, podemos dejar de expandirlo cuando la longitud llegue a un cierto valor, cuando un nodo no contenga más de cierta cantidad de observaciones, o podemos establecer que un nodo puede ser hoja únicamente si su cantidad de observaciones es igual o mayor a cierto valor, entre otros criterios posibles que podríamos elegir. A estos criterios los llamamos criterios de corte, que le indican al algoritmo cuando detener la expansión del árbol. Finalmente, el valor de los nodos hoja será el promedio de los valores de la variable objetivo de las observaciones que contenga el nodo.

La precisión del árbol se calcula con el coeficiente de determinación R^2.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \overline{y}_i)^2}{\sum_{i=1}^{n} (y_i - \hat{y})^2} \tag{8}$$

donde y_i es el valor real de la la variable objetivo en la observación i, $\overline{y}_i$ el valor predecido y $\hat{y}$ el valor medio de la variable objetivo.

3.5. Máquinas de vectores de soporte

Este algoritmo se utiliza para problemas de clasificación y regresión. Para entender las máquinas de vectores de soporte primero debemos repasar otros conceptos, para empezar comenzamos con el hiperplano.

3.5.1. Clasificador de margen máximo

Un hiperplano en un espacio de dimensión p es un subespacio de dimensión $p-1$. Su definición matemática esta dada por la siguiente ecuación:

$$A_0 + A_1X_1 + ... + A_pX_p = 0$$

Ahora supongamos que tenemos un conjunto de observaciones con p atributos donde cada observación pertenece a una de dos clases.

Cada observación esta representada por x e y de la siguiente manera:

$$x_i = (x_{i1} \; ... \; x_{ip})$$

$$y_i \in \{-1, 1\}$$

donde y_i indica la clase de la observación, es decir, -1 representa una clase y 1 la otra, y x_i representa los n atributos de la observación.

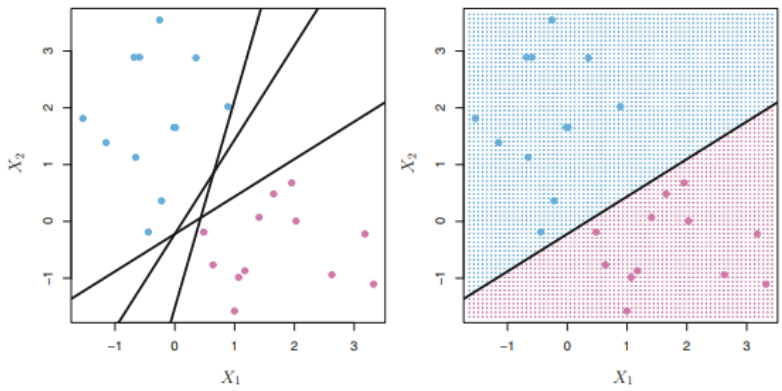


Figura 8: A la izquierda vemos múltiples hiperplanos que pueden dividir el conjunto, a la derecha el espacio separado por el hiperplano óptimo.

Como podemos ver en la Figura 8 en la parte izquierda, tenemos a todas las observaciones con el color de la clase a la cual pertenecen, y también se pueden observar los distintos hiperplanos que pueden dividir al conjunto.

En el gráfico derecho de la figura se muestra el hiperplano elegido y como el espacio queda dividido en dos subespacios, por lo tanto, para clasificar una observación nueva, bastaría con determinar a que subespacio pertenece para determinar su clase.

Siempre que exista un hiperplano capaz de dividir perfectamente a las observaciones según las clases, el hiperplano cumplirá

$$A_0 + A_1 x_{i1} + ... + A_p x_{ip} > 0 \text{ si } y_i = 1,$$
$$A_0 + A_1 x_{i1} + ... + A_p x_{ip} < 0 \text{ si } y_i = -1,$$

lo cual es equivalente a

$$y_i(A_0 + A_1 x_{i1} + ... + A_p x_{ip}) > 0.$$

Pero como se puede apreciar en el ejemplo, existen infinitos hiperplanos que dividan al conjunto, por lo tanto, el objetivo es encontrar el que divida al conjunto de la mejor manera.

Una opción para esto es usar el hiperplano de margen máximo. Para encontrarlo se calculan las distancias (perpendiculares al hiperplano) con todas las observaciones, la mínima distancia a una observación se denomina margen. Lo que queremos encontrar es el hiperplano cuyo margen es máximo.

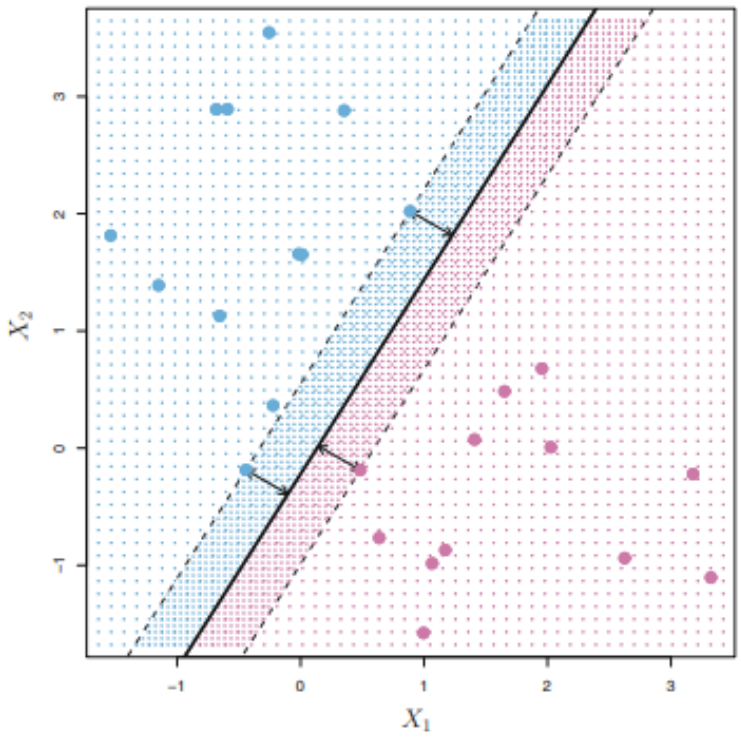


Figura 9: Hiperplano de margen máximo.

En la Figura 9 se visualizan tres observaciones que se encuentran a una misma distancia del hiperplano, esta distancia es el margen mencionado anteriormente. Estas tres observaciones son las únicas responsables en determinar el hiperplano y son denominados vectores de soporte. Si cualquier otra observación cambia de lugar, siempre y cuando no se ubique una distancia menor al margen del hiperplano, no afectara al hiperplano elegido.

Podemos decir entonces, que encontrar el hiperplano de margen máximo es equivalente a encontrar los $A_0, A_1, .., A_p$ que maximicen M.

$$y_i(A_0 + A_1 x_{i1} + ... + A_p x_{ip}) \geq M, \tag{9}$$

$$\sum_{j=1}^{p} A^2 = 1. \tag{10}$$

La primera ecuación indica que todas las observaciones deben esta a una distancia mayor o igual a M. Al aplicar la segunda ecuación se puede ver que la distancia de la observación i al hiperplano queda dada por

$$y_i(A_0 + A_1 x_{i1} + ... + A_p x_{ip}).$$

Por lo tanto, cada observación estará a una distancia mayor o igual a M y se ubicará del lado correcto del hiperplano.

3.5.2. Vectores de soporte para clasificación

En muchos casos, no existe un hiperplano que pueda separar el conjunto o el hiperplano que los separa no es el deseable, ya que como explicamos anteriormente, este dependerá únicamente de los vectores de soporte, y una sola observación aislada podría afectar drásticamente el hiperplano elegido.

Los vectores de soporte para clasificación permiten que ciertas observaciones puedan ubicarse en el lugar incorrecto del margen o clasificarse de manera incorrecta para poder obtener un clasificador menos sensible a observaciones aisladas.

El problema a resolver es similar al del clasificador de margen máximo. Nuevamente, se deben encontrar los $A_0, A_1, .., A_p$ que maximicen M.

$$y_i(A_0 + A_1 x_{i1} + ... + A_p x_{ip}) \geq M(1 - \epsilon_i), \tag{11}$$

$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C, \tag{12}$$

$$\sum_{j=1}^{p} A^2 = 1. \tag{13}$$

Donde C es un parámetro que limita la cantidad de observaciones que pueden clasificarse de manera incorrecta y ϵ_i indica en donde se encuentra la observación i, si $\epsilon_i = 0$, entonces se encuentra del lado correcto del margen, si $\epsilon_i > 0$ se encuentra en el lado incorrecto del margen y por ultimo si $\epsilon_i > 1$ entonces estará en el lado incorrecto del hiperplano.
Un valor alto de C implica un alto sesgo y una baja varianza, un valor bajo lo contrario.

Existen casos en los que el conjunto no será linealmente separable y el clasificador separara los conjuntos de una manera muy pobre, un ejemplo de esto se ilustra en la Figura 10.

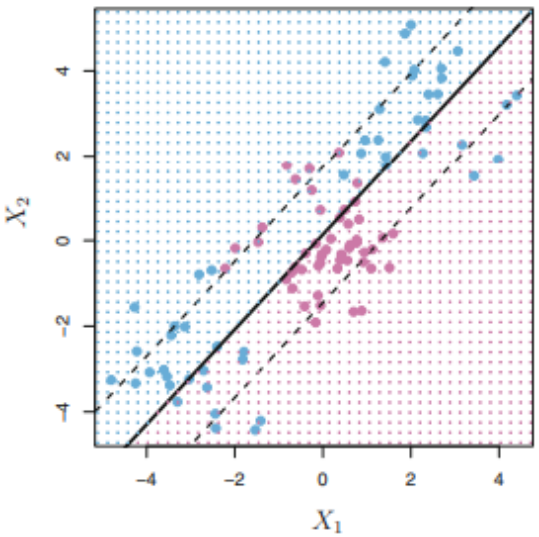


Figura 10: Clasificación de un conjunto no linealmente separable.

Para poder separar este conjunto de forma adecuada procederemos a estudiar las máquinas de vectores de soporte.

3.5.3. Máquinas de vectores de soporte

Para poder dividir un conjunto que no es linealmente separable utilizaremos las máquinas de vectores de soporte.

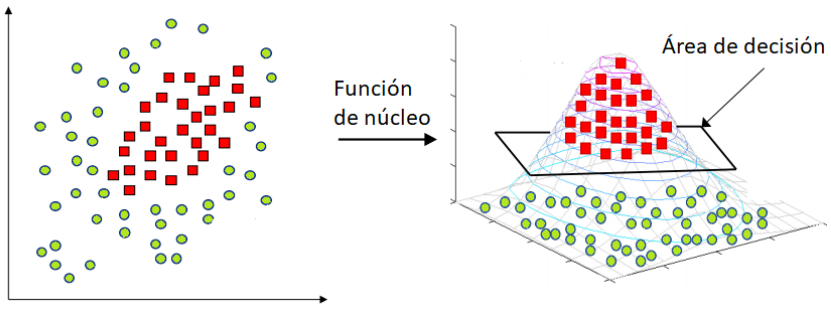


Figura 11: Separación de las clases en un espacio de dimensión superior al original.

Como ilustra la Figura 11, lo que haremos es llevar el conjunto a un espacio de mayor dimensión en donde las clases se puedan separar adecuadamente.

En primera medida podemos representar a un vector de soporte para clasificación como

$$f(x) = A_0 + \sum_{i \in S} \alpha_i \langle x, x_i \rangle, \tag{14}$$

donde

$$\langle x_i, x_{i'} \rangle = \sum_{j=1}^{p} x_{ij} x_{i'j}, \tag{15}$$

y S es el conjunto de las observaciones que funcionan como vectores de soporte, ya que todas las observaciones que no forman parte tendrán su $\alpha_i = 0$.

Para realizar la separación en un espacio de mayor dimensión haremos uso de la función

$$K(x_i, x_{i'}), \tag{16}$$

a la cual, de ahora en adelante, nos referiremos a ella como la función de núcleo.

Ahora, en lugar de utilizar el producto interno, la máquina de vectores de soporte se calculará de la siguiente forma

$$f(x) = A_0 + \sum_{i \in S} \alpha_i K(x, x_i). \tag{17}$$

Para empezar, podemos tomar la función de núcleo lineal

$$K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} x_{i'j}, \tag{18}$$

la cual seria equivalente a calcular simplemente el producto interno entre las observaciones, y por lo tanto, equivalente a los vectores de soporte para clasificación.

Podemos utilizar cualquier función de núcleo posible, entre ellas

$$K(x_i, x_{i'}) = (1 + \sum_{j=1}^{p} x_{ij} x_{i'j})^d, \tag{19}$$

$$K(x_i, x_{i'}) = e^{-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2}, \tag{20}$$

que representan la función polinomial y radial respectivamente, las cuales, junto con la función de núcleo lineal, son las más usadas a la hora de utilizar máquinas de vectores de soporte.

Hasta ahora solo hemos discutido la separación entre dos clases, pero es posible sutilizar las máquinas de vectores de soporte para separar una cantidad de clases mayor. A continuación detallaremos las dos estrategias mas conocidas.

One-Versus-One
Para poder dividir k clases, este método consiste en crear $\binom{k}{2}$ máquinas de vectores de soporte, esto corresponde a todas las formas posibles de comparar las clases de a pares. Para saber a que clase pertenece una observación nueva, simplemente la clasificamos en todos los clasificadores, por ultimo, la clase a la que se asignó más veces será a donde se asignará la observación.

One-Versus-All
Ésta alternativa consiste en crear K máquinas de vectores de soporte, cada una compara una clase en particular con el resto de las clases. A la hora de clasificar una observación nueva, elegimos a la que clasifique con una probabilidad mayor de pertenecer a la clase en concreto.

Todas las figuras a excepción de la Figura 11 fueron tomadas del libro [7].

3.6. C4.5

El algoritmo C4.5 es un algoritmo de clasificación estadística. Es una evolución del algoritmo de arboles ID3. Como mejoras del algoritmo original están:

1. Los datos de entrada pueden ser ambos continuo y discreto. Para aceptar los datos continuos, se crea umbrales para separar los datos.
2. Se aceptan datos faltantes, remplazando por el predeterminado ? y no es usado en el cálculo de entropía.
3. Se trabaja con datos de diferentes magnitudes
4. Al terminar de armar el árbol, se hace un segundo recorrido, podando caminos que no contribuyen al resultado y reemplazando con hojas.

Un dato que se puede calcular con este algoritmo es la importancia de cada característica para la toma de decisiones del algoritmo. Las características son las variables independientes. El valor para cada nodo del árbol se puede calcular con

$$FI = FM * n - LM * n_l - RM * n_r \tag{21}$$

siendo FI la importancia de la característica, FM la métrica de la característica, que seria la entropía calculada en cada nivel, n el número de instancias, LM la métrica del nodo hijo a la izquierda, RM a la derecha, n_l el numero de instancias para el nodo a la izquierda y n_r a la derecha. Para el valor total, se hace la suma total por característica.

3.7. Regresión Lineal Múltiple

Representa una extensión de la regresión lineal simple en la que podemos incluir más de una variable independiente a la vez, evaluando el efecto de cada variable en presencia del resto.

En una ecuación de regresión lineal múltiple, se asocia cada variable $(X_1, X_2, \ldots, X_p)$ un coeficiente $\beta_j,\ j = 1, \ldots, p$ que cuantifica la asociación entre las variables predictoras y la variable respuesta.

Variable de interés o variable dependiente:

- Se denota Y
- 1 variable aleatoria
- $\{y_1, ..., y_n\}$ son n observaciones de Y

Covariables o variables independientes:

- Se denotan $X = (X_1, ..., X_p)$
- p variables deterministas
- En este caso, hay $n.p$ observaciones de X

La idea es encontrar los coeficientes β_j tal que

$$Y_i = \sum_{j=1}^{p} \beta_j X_{ji} + \beta_0 + \epsilon_i,$$

- Con:
  - β_0 ordenada al origen
  - $\beta_j, j = 1...p$, el efecto de la covariable X_i
  - ϵ el error
- Hipótesis:
  - ϵ_i (variable aleatoria), $\epsilon_i \sim N(0, \sigma^2)$
  - $\forall i \neq k,\ \epsilon_i, \epsilon_k$ son independientes.

Para optimizar el modelo se utiliza el método de eliminación hacia atrás para la selección de las variables independientes, estadísticamente significativas para el modelo. Este método consiste en:

1. Entrenar el modelo de regresión lineal utilizando las variables independientes y dependientes.
2. Calcular los p-valores de las variables independientes, como indicador de qué relaciones son estadísticamente significativas.
3. Identificar la variable x con el p valor mas alto.
4. El p-valor para cada variable independiente, prueba la hipótesis nula de que la variable no tiene correlación con la variable dependiente. Si no hay correlación, no hay asociación entre los cambios en la variable independiente y los cambios en la variable dependiente. Entonces:
   Si el p-valor$(x) > 0.05$ aceptamos la hipótesis nula y descartamos x nuestro modelo. Volvemos al paso 2.
   Si p-valor$(x) < 0.05$ rechazamos la hipótesis nula y así validando el modelo.

3.8. Regresión Logística

La Regresión Logística es un método de clasificación en base a las variables independientes. Esta regresión busca la probabilidad de que un Y pertenezca a una categoría. En el caso de que las categorías sean 1 y 0, se busca $p(X) = Pr(Y = 1|X)$.

La probabilidad se calcula en base a la ecuación usada en la Regresión Lineal.

$$p_l(X) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p \tag{22}$$

pero con algunas alteraciones. Para encerrar a los resultados de la regresión adentro de los valores de clasificación, se usa la función logística, dejando entonces el cálculo de la probabilidad de la siguiente manera:

$$p(X) = \frac{\exp(p_l(X))}{1 + \exp(p_l(X))} \tag{23}$$

Los valores de β_i son incógnitas calculadas en base al set de entrenamiento. Para calcularlos, se usa la función de verosimilitud, buscando maximizar el resultado.

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})) \tag{24}$$

3.9. Redes Neuronales Artificiales

Las redes neuronales artificiales son un concepto inspirado en el cerebro humano y su funcionamiento. Intenta simular a las neuronas y las transmisiones de los impulsos nerviosos entre sí. En cada neurona, se recibe una señal, se hace modificaciones en la misma y se la envía a las próximas. El nombre dado a esta simulación de la neurona es Perceptrón.

3.9.1. Perceptrón Simple

El perceptrón simple es un clasificador binario lineal que fue creado como una representación computacional artificial de una neurona. La unidad básica del perceptrón consiste de un nodo central que recibe señales de entrada y mediante una operación, lo transforma en una señal de salida. El perceptrón corrige los errores en base a las diferentes entradas que va recibiendo.

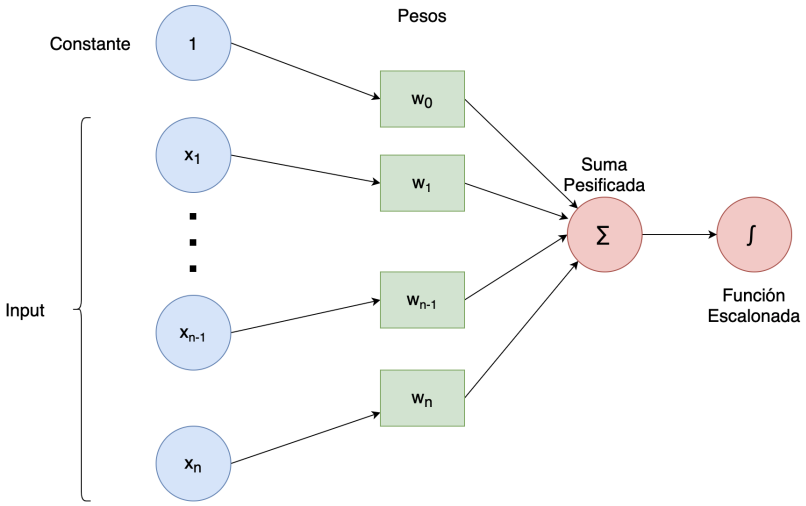


Figura 12: Representación de un perceptrón simple

La unidad de un perceptrón esta representado como lo muestra la Figura 12. Primero se definen los componentes:

- x_i denota a uno de los parámetros de entrada, correspondiente al índice i del vector de entrada.
- w_i denota al peso del parámetro de entrada en la posición i del vector de entrada
- y denota al parámetro de salida de la neurona

El cálculo realizado en un perceptrón simple consiste de transmitir el resultado del producto del vector de entrada y el vector de los pesos. $z = x \cdot w$ El resultado entra en una función escalonada de activación para transformar el resultado en un valor binario. Ejemplos de funciones escalonadas utilizadas son la tangente hiperbólica, la función logística y la rectificadora.
Si el vector de entrada viene en la etapa de entrenamiento, se balancea los pesos w usando la siguiente relación:

$$w = \eta * \epsilon * x^T$$

dónde η es la constante de aprendizaje y ϵ el error calculado por la resta del valor esperado y el valor calculado.
Para una predicción con menor error, se utiliza un valor constante además de los valores de entrada. A este valor se le nombra Bias. Este valor es tomado como si fuera uno de los componentes de x (entrada) y por eso tiene un peso propio calculado. [9]

3.9.2. Perceptrón Multicapa

El perceptrón multicapa es una red neuronal formada por diversas capas. Es una evolución del perceptrón simple. En vez de una sola neurona, el perceptrón multicapa posee diversas neuronas, distribuidas en las diferentes capas de la red. Este clasificador tiene n neuronas de salida y n de entrada. Entre estas dos capas, puede tener diferentes capas internas / ocultas.

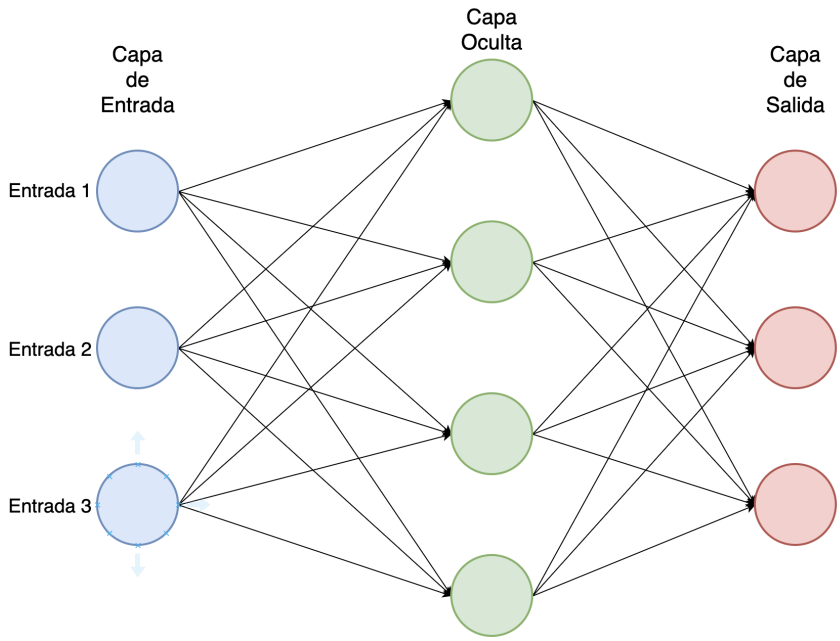


Figura 13: Representación de un perceptrón multicapa

La Figura 13 presenta un ejemplo de perceptrón multicapa con la capa de entrada con 3 nodos, una capa oculta con 4 nodos y la capa de salida con 3.

El aprendizaje se hace mediante propagación hacia atrás (retro-propagación). Los pasos para el aprendizaje son los siguientes[9]:

1. Se inicializa los pesos con valores con pequeños valores aleatorios
2. Se elije un patrón ξ_k^μ y se lo aplica a la capa de entrada $m = 0$:
   $V_k^0 = \xi_k^\mu$ para todo k
3. Se propaga la señal hacia adelante usando:
   $V_i^m = g(h_i^m) = g(\sum_j w_{ij}^m V_j^{m-1})$ con g como la función de activación
4. Se computa los deltas para la capa externa
   $\delta_i^M = g'(h_i^M)[\zeta_i^\mu - V_i^M])$ con ζ_i^μ como el valor esperado
5. Se computan los deltas para las capas anteriores
   $\delta_i^{m-1} = g'(h_i^{m-1}) \sum_j w_{ji}^m \delta_j^m$
6. Para entrenar la red, se actualiza el peso
   $\Delta w_{ij}^m = \eta \delta_i^m V_j^{m-1}$

Los pasos 2 en adelante se repiten por cada entrada recibida en el perceptrón.

3.10. Validación Cruzada en K Partes

Para asegurarse que un modelo realice predicciones lo mas precisas posible a partir de datos de entradas con el cual el modelo no fue entrenado, utilizamos el método de validación cruzada en k partes.

El algoritmo:

1. Mezclar el conjnuto de datos
2. Dividir el conjunto en k partes iguales
3. Por cada i-ésima parte
   3.1. Tomar la k_i parte como conjunto de testeo
   3.2. Tomar la k partes restantes como conjunto de entrenamiento
   3.3. Entrenar y testear el modelo con los respectivos conjuntos
   3.4. Anotar la evaluación del modelo
4. La evaluación final del modelo es el promedio de las evaluaciones hechas en el paso 3

3.11. Selección de variables mediante el método de eliminación hacía atrás

La selección de variables es una forma de reducir la cantidad de variables y, por lo tanto, reducir la complejidad computacional del modelo. Este método ayuda a determinar el conjunto más pequeño de variables que se necesitan para predecir la variable objetivo con alta precisión como también solucionar el problema de sobre ajuste.

La selección de variables mediante el método de eliminación hacía atrás cosiste en realizar el modelo de regresión lineal múltiple con todas las variables, y se eliminan las que tienen el t estadístico cercano a 1, o el p-valor mayor que un umbral, usualmente se toma 0.05. Las variables descartadas son aquellas que producen el deterioro más insignificante estadísticamente del ajuste del modelo. Al eliminar una variable se evalúa el modelo mediante el coeficiente de determinación adjunto Adj R^2 calculado como:

$$R^2_{adj} = 1 - (\frac{(1 - R^2)(n - 1)}{n - q})$$

donde n es el número de observaciones y q el número de coeficientes en el modelo. Al coeficiente de Determinación R^2 se lo obtiene mediante

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

El proceso se repite hasta que no se pueden eliminar más variables sin una pérdida de ajuste estadísticamente insignificante o al eliminar una variable el valor Adj R^2 disminuye respecto al calculado en el paso anterior, en ese caso la variable a descartar se la agrega de nuevo al conjunto.

El algoritmo:

1. Entrenar el modelo con el conjunto de datos utilizando todas las variables disponibles.
2. Calcular el valor Adj R^2 utilizando todas las variables disponibles.
3. Identificar la variable con el p-valor mas alto.
   3.1. Si el p-valor de la variable es menor al umbral ir al paso 4.
   3.2. Si el p-valor de la variable es mayor al umbral, calcular el valor Adj R^2 descratando la varibale en cuestion. Si el valor disminuye respecto al calculado en el paso 2 ir al paso 4.

3.3. Si el p-valor de la variable es mayor al umbral, calcular el valor Adj R^2 descratando la varibale en cuestion. Si el valor aumenta respecto al calculado en el paso 2 eliminamos la variable del conjunto. Finalmente ir al paso 1

4. Finaliza el algoritmo, para esta instancia obtuvimos el conjunto más pequeño de variables.

4. Análisis de tendencias en búsquedas de síntomas

La forma más sencilla de detectar la cantidad de casos de COVID-19 en una región es mediante la realización de pruebas de RT-PCR en un porcentaje de la población. Luego en base a los resultados obtenidos estimar el número para la totalidad de la población.
Este proceso es lento y limitado por costos. Requiere de un presupuesto para cubrir el traslado de las personas a ser diagnosticadas, especialistas y las pruebas de RT-PCR. Evaluar la muestra tomada de la población puede tomar más de un día y los resultados se obtienen luego de las 48hs.
El resto de la población que no tiene acceso a una prueba de RT-PCR, o no tiene el tiempo de someterse a esta prueba suele acudir a una búsqueda en internet de los síntomas que posee como método alternativo de auto-diagnóstico.
El análisis de las tendencias en los patrones de búsquedas en internet de síntomas relacionados al COVID-19 puede ayudar a realizar una estimación diaria y veloz de la cantidad de casos e inclusive detectar brotes.

4.1. Conjunto de datos utilizados

El conjunto [14] utilizado fue recopilado por Google y publicado a través de un repositorio en github, con el fin de a ayudar a los investigadores a comprender mejor el impacto de COVID-19. Estos datos reflejan el volumen de búsquedas de Google para un amplio conjunto de síntomas, signos y condiciones de salud.
Una sola consulta de búsqueda se puede asignar a más de un síntoma. Por ejemplo, la búsqueda de "reflujo ácido y tos con moco" se asigna a tres síntomas: "Tos", "Enfermedad por reflujo gastroesofágico" y "Acidez de estómago".
Los datos muestran la popularidad relativa de los síntomas en las búsquedas dentro de una región geográfica como una serie de tiempo diaria.
El proceso de normalización y escalamiento de la series de tiempo diaria en una región consiste en:

1. Se cuenta el número de búsquedas de cada síntoma en una región para ese día.
2. Se divide este recuento por la cantidad total de búsquedas en la región para ese día para calcular la popularidad relativa.
3. Se busca el valor máximo de la popularidad relativa en todo el rango de tiempo publicado para esa región, sobre todos los síntomas. Se escala este valor máximo a 100. Todos los demás valores se asignan a valores proporcionalmente más pequeños (escala lineal) en el rango 0-100.
4. Finalmente, se utiliza factor de escala para escalar valores (para la misma región) en versiones posteriores. En actualizaciones futuras, cuando la popularidad de un síntoma supera el valor máximo observado anteriormente (que se encuentra en el paso 3), los nuevos valores escalados pueden ser superiores a 100.

Para cada región y día, se escalan las popularidades relativas utilizando el mismo factor de escala.
Los datos abarcan 400 síntomas y distintas regiones dentro de los países Australia, Irlanda, Nueva Zelanda, Singapur, Reino Unido y Estados Unidos desde el 2017 y hasta el día de hoy se continua recopilando datos.
Para situar el conjunto en el contexto pandemia se seleccionaron los valores desde el día 01/01/2020 hasta el 18/03/2021. Además se redujo a 19 dentro de los síntomas más relevantes del COVID-19 según el reporte de la OMS[15].
Estos son: fiebre, tos, fatiga, esputo, dificultad para respirar, Taquipnea, dolor de garganta, dolor de cabeza, escalofríos, náuseas, vómitos, congestión nasal, diarrea, hemoptisis, conjuntivitis, erupción cutánea, dolor en el pecho, dolor ardoroso en el pecho, neumonía.

Para relacionar la popularidad relativa de los síntomas buscados en Google con la cantidad de casos para un día y región dadas se utilizo la API[4] creada por la organización The COVID Tracking Project.

4.2. Aplicación del método Regresión Lineal Múltiple

Se eligió la regresión Lineal Múltiple como modelo para predecir la cantidad de casos positivos según la popularidad relativa de los síntomas en las búsquedas en un día y región dadas.
Para optimizar el modelo se utilizó el método de eliminación hacia atrás para la selección de las variables independientes, en este caso la popularidad relativa de los síntomas del COVID-19.
A modo de ejemplo tomamos la región California y aplicamos el método.
Entonces de la totalidad del conjunto de datos solamente utilizaremos aquellos correspondientes al año 2020 para entrenar el modelo. Como en este caso particular se eligió la región California además se redujo el conjunto de datos de entrenamiento a los correspondientes de esta región.
La Tabla 3 muestra las iteraciones de la selección de variables mediante el método de eliminación hacía atrás para la región de California.
Las celdas en blanco representa aquellos síntomas que fueron eliminados del conjunto de variables.

Síntomas	**P-Valor**				
	Iteración 1	**Iteración 2**	**Iteracion 3**	**Iteracion 4**	**Iteracion 5**
Fiebre	0.205385	0.202779	0.214128	0.224920	0.214160
Tos	0.714772	0.719680			
Fatiga	0.000006	0.000006	0.000006	0.000003	0.000001
Esputo	0.740418	0.718768	0.639702		
Dificultad para respirar	0.003505	0.003143	0.003187	0.002805	0.002304
Taquipnea	0.326234	0.328348	0.260970	0.279822	0.235958
Dolor de garganta	0.000071	0.000050	0.000002	0.000001	0.000001
Dolor de cabeza	0.185427	0.191987	0.150822	0.134083	0.155761
Escalofríos	0.000410	0.000188	0.000195	0.000195	0.000201
Náuseas	0.000000	0.000000	0.000000	0.000000	0.000000
Vómitos	0.000023	0.000015	0.000007	0.000006	0.000006
Congestión nasal	0.000000	0.000000	0.000000	0.000000	0.000000
Diarrea	0.311167	0.276972	0.296074	0.303174	0.275451
Hemoptisis	0.505074	0.513813	0.555101	0.461641	
Conjuntivitis	0.000000	0.000000	0.000000	0.000000	0.000000
Erupción cutánea	0.810775				
Dolor en el pecho	0.192237	0.189664	0.184676	0.165414	0.160857
Dolor ardoroso en el pecho	0.003457	0.003398	0.003581	0.003636	0.002846
Neumonía	0.178040	0.174555	0.142837	0.151913	0.167560
Adj R^2	0.7406	0.7415	0.7423	0.743	0.7434

Tabla 3: Iteraciones del método de eliminación hacia atrás para la selección de síntomas mas significativos

En cada iteración del método se entrena un modelo de regresión lineal utilizando como variables independientes la popularidad relativa de los síntomas del COVID-19 en un día dado y como variable dependiente la cantidad de casos para ese día.
Una vez entrenado el modelo se calculan los p-valores de las variables independientes y el valor Adj R^2.
Paso siguiente se identifica la variable con el p-valor mas alto. Como podemos ver en la Tabla 3 en la primera iteración la variable correspondiente es "erupción cutánea" con un p-valor de 0.810775.
Paso siguiente utilizamos el p-valor como prueba de la hipótesis nula de que la variable no tiene corre-

lación con la variable dependiente. Como $0.810775 > 0.05$ aceptamos la hipótesis nula y descartamos la popularidad relativa del síntoma "erupción cutánea" de nuestro modelo siempre que al quitarla no disminuya el valor Adj R^2 calculado previamente.
Análogamente realizamos otra iteración del método hasta encontrar una variable independiente con p-valor menor a 0.05 que rechace la hipótesis nula o al quitar una variable el valor Adj R^2 disminuye.
Como resultado quedamos con las variables independientes estadísticamente más significativas para la región de California que validan el modelo de regresión lineal. Como podemos observar en la tabla 3 estas son: Fiebre, Fatiga, Dificultad para respirar, Respiración superficial, Dolor de garganta, Dolor de cabeza, Escalofríos, Náuseas, Vómitos, Congestión nasal, Diarrea, Conjuntivitis, Dolor en el pecho, Dolor ardoroso en el pecho, Neumonía.

Finalmente para evaluar el modelo de regresión lineal múltiple resultante para cada región del conjunto de datos se analiza el histograma de residuos.
El residuo se calcula mediante la diferencia en entre los valores observados y los valores del la predicción.

$$Residuo(\epsilon_i) = y_i - \widehat{y_i}$$

donde y_i denota al i-ésimo valor observado del conjunto de datos.
Si el modelo es válido se debe observar en el histograma una distribución normal de los residuos.

4.2.1. Resultados

Los siguientes resultados corresponden a predicciones diarias desde el día 01/01/2021 hasta el 18/03/2021 en Estados Unidos. De los 50 estados en Estados se tomaron los estados de: Alabama, Arizona, Arkansas, California, Colorado, Connecticut, Florida, Georgia, Idaho, Illinois, Indiana, Kentucky, Louisiana, Maryland, Massachusetts, Michigan, Minnesota, Mississippi, Missouri, Nevada, New Jersey, New Mexico, New York, North Carolina, Ohio, Oklahoma, Oregon, Pennsylvania, South Carolina, Tennessee, Texas, Utah, Virginia, Washington, Wisconsin.
En total son 31, ya que para los restantes no había suficiente cantidad de registros para el desarrollo del modelo de regresión lineal múltiple.
Para sintetizar esta sección presentaremos los resultados correspondientes a las regiones de California, Ohio y Indiana.
California

Los síntomas seleccionados por el método de eliminación hacia atrás para la región de California son: fiebre, fatiga, dificultad para respirar, respiración superficial, dolor de garganta, dolor de cabeza, escalofríos, Náuseas, vómitos, congestión nasal, diarrea, conjuntivitis, dolor en el pecho, dolor ardoroso en el pecho, neumonía.

La Figura 14 muestra el histograma de residuo entre los valores observados y los valores del la predicción para la región de California. Como esta muestra una distribución de forma gaussiana podemos validar el modelo de regresión lineal múltiple.

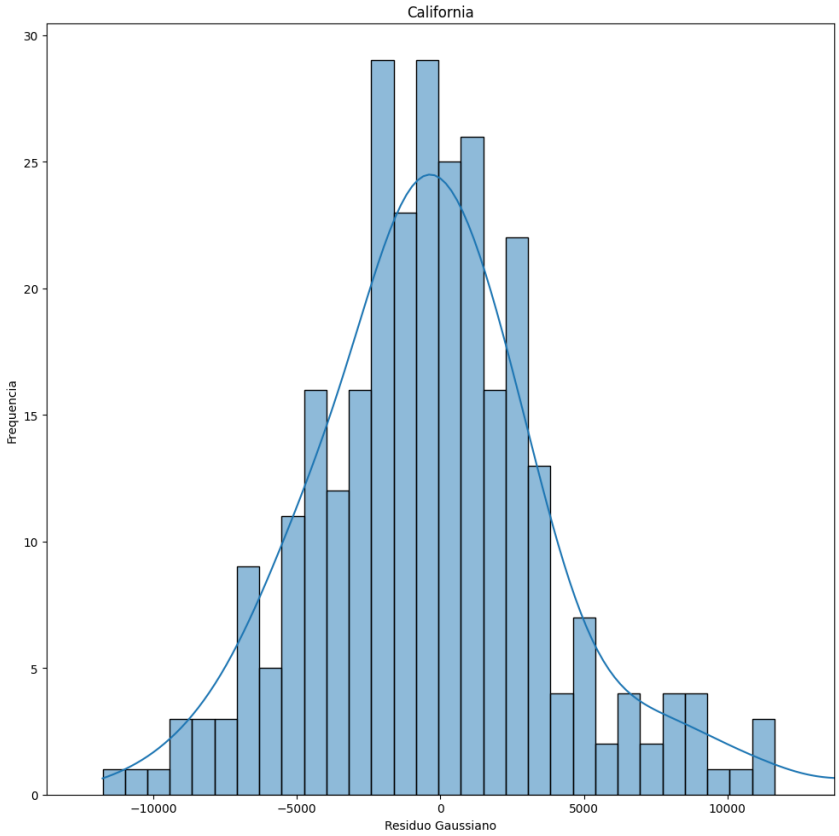


Figura 14: Histograma de residuo para la región de California

En la Figura 15 podemos visualizar un gráfico de dispersión de la cantidad de casos COVID-19 positivos registrados para un día dado en la región de California por la organización The COVID Tracking Project [4] en el eje x. Luego en el eje y tenemos la predicción de la cantidad de casos para el día utilizando el modelo de regresión lineal múltiple para la región de California. El área color verde delimita ancho de banda de error gaussiano, para este caso las predicciones se realizan con un error de ± 6103 casos.
Como medida de precisión tomamos el porcentaje de predicciones realizadas que se encuentran dentro del ancho de banda de error gaussiano. En la Tabla 4 podemos observar que para la región de California esta es de 52 %.

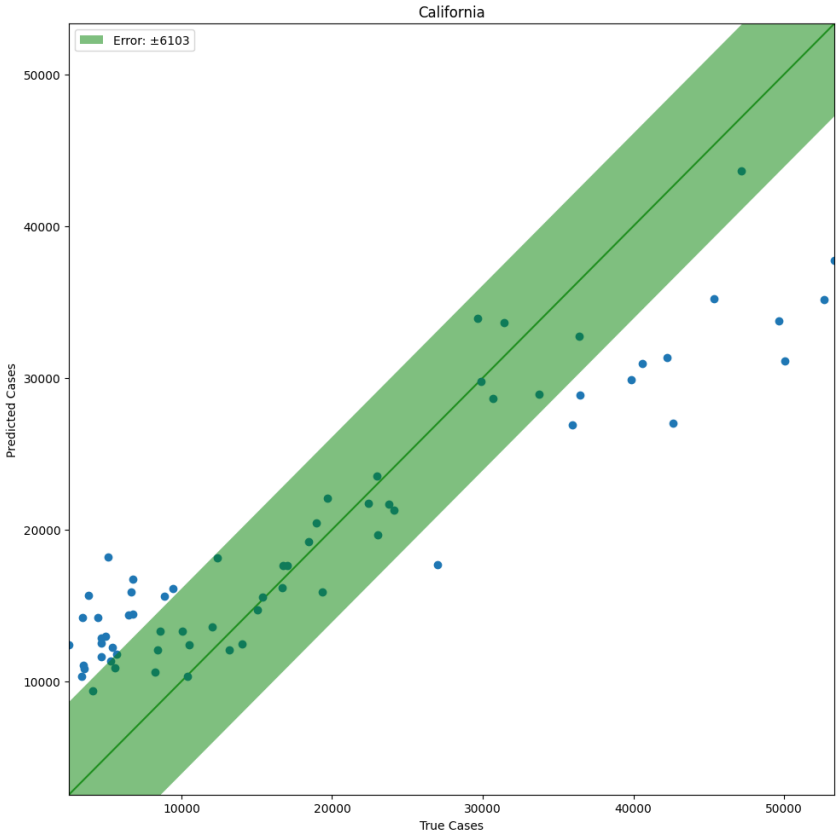


Figura 15: Diagrama de dispersión de la cantidad de casos reales vs predecidos junto con el ancho de banda de error gaussiano en la región de California y la racta x=y

Ohio

Los síntomas seleccionados por el método de eliminación hacia atrás para la región de Ohio son: fiebre, tos, fatiga, dificultad para respirar, dolor de garganta, dolor de cabeza, escalofríos, náuseas, vómitos, congestión nasal, diarrea, conjuntivitis, erupción cutánea, ardor en el pecho, neumonía.

La Figura 16 muestra el histograma de residuo entre los valores observados y los valores del la predicción para la región de Ohio. Como esta muestra una distribución de forma gaussiana podemos validar el modelo de regresión lineal múltiple.

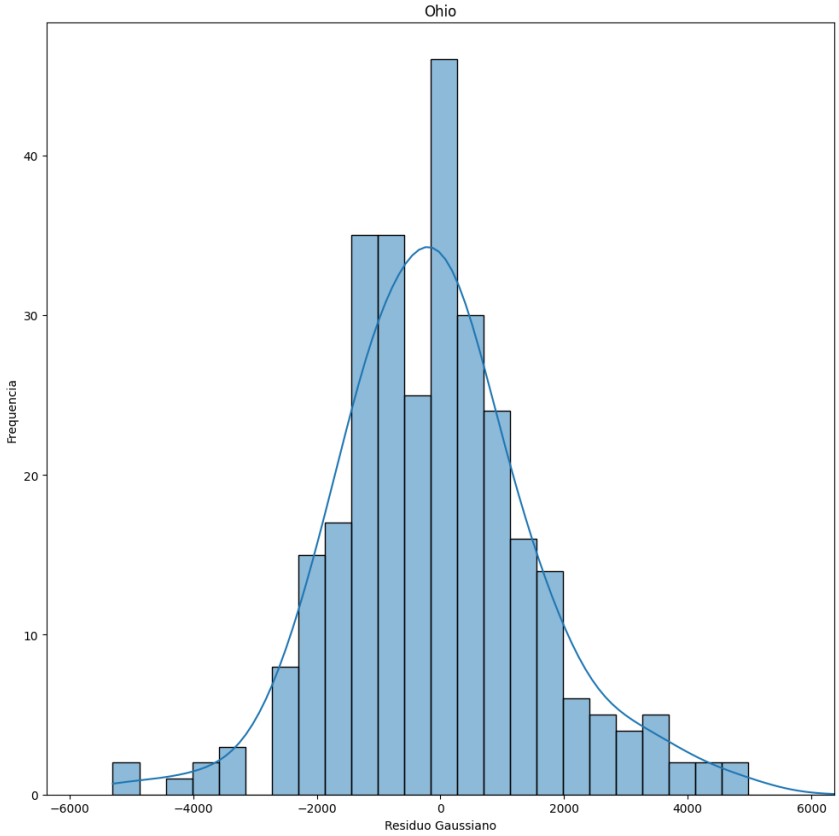


Figura 16: Histograma de residuo para la región de Ohio

En la Figura 17 podemos visualizar un gráfico de dispersión de la cantidad de casos COVID-19 positivos registrados para un día dado en la región de Ohio por la organización The COVID Tracking Project [4] en el eje x. Luego en el eje y tenemos la predicción de la cantidad de casos para el día utilizando el modelo de regresión lineal múltiple para la región de Ohio. El área color verde delimita ancho de banda de error gaussiano, para este caso las predicciones se realizan con un error de ± 1544.
En la Tabla 4 podemos observar que para la región de Ohio, el porcentaje de predicciones dentro del ancho de banda de error gaussiano es de 62 %.

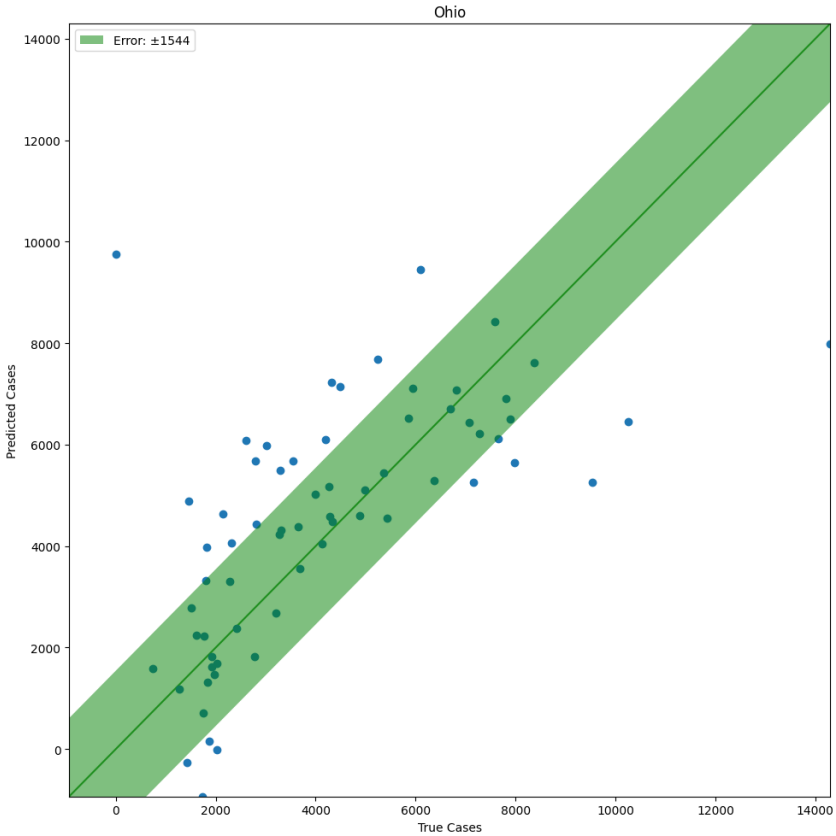


Figura 17: Diagrama de dispersión de la cantidad de casos reales vs predecidos junto con el ancho de banda de error gaussiano en la región de Ohio y la racta x=y

Indiana

Los síntomas seleccionados por el método de eliminación hacia atrás para la región de Indiana son: tos, fatiga, esputo, dificultad para respirar, respiración superficial, dolor de garganta, dolor de cabeza, escalofríos, náuseas, vómitos, congestión nasal, diarrea, conjuntivitis, erupción cutánea, neumonía.

La Figura 18 muestra el histograma de residuo entre los valores observados y los valores del la predicción para la región de Indiana. Como esta muestra una distribución de forma gaussiana podemos validar el modelo de regresión lineal múltiple.

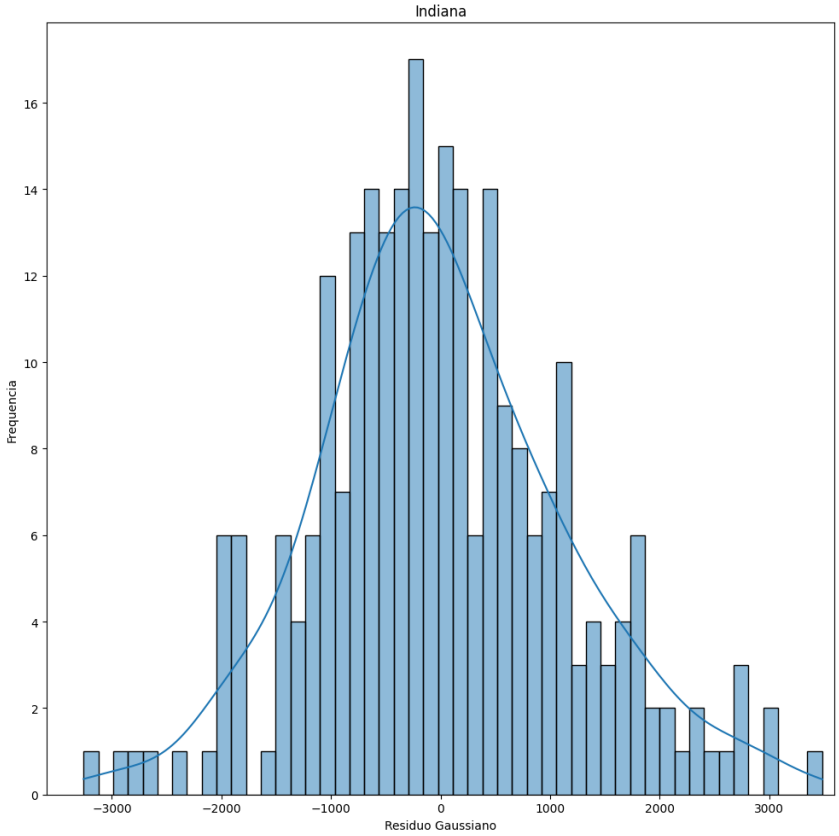


Figura 18: Histograma de residuo para la región de Indiana

En la Figura 19 podemos visualizar un gráfico de dispersión de la cantidad de casos COVID-19 positivos registrados para un día dado en la región de Indiana por la organización The COVID Tracking Project [4] en el eje x. Luego en el eje y tenemos la predicción de la cantidad de casos para el día utilizando el modelo de regresión lineal múltiple para la región de Indiana. El área color verde delimita ancho de banda de error gaussiano, para este caso las predicciones se realizan con un error de ± 926.
En la Tabla 4 podemos observar que para la región de Indiana, el porcentaje de predicciones dentro del ancho de banda de error gaussiano es de 61 %.

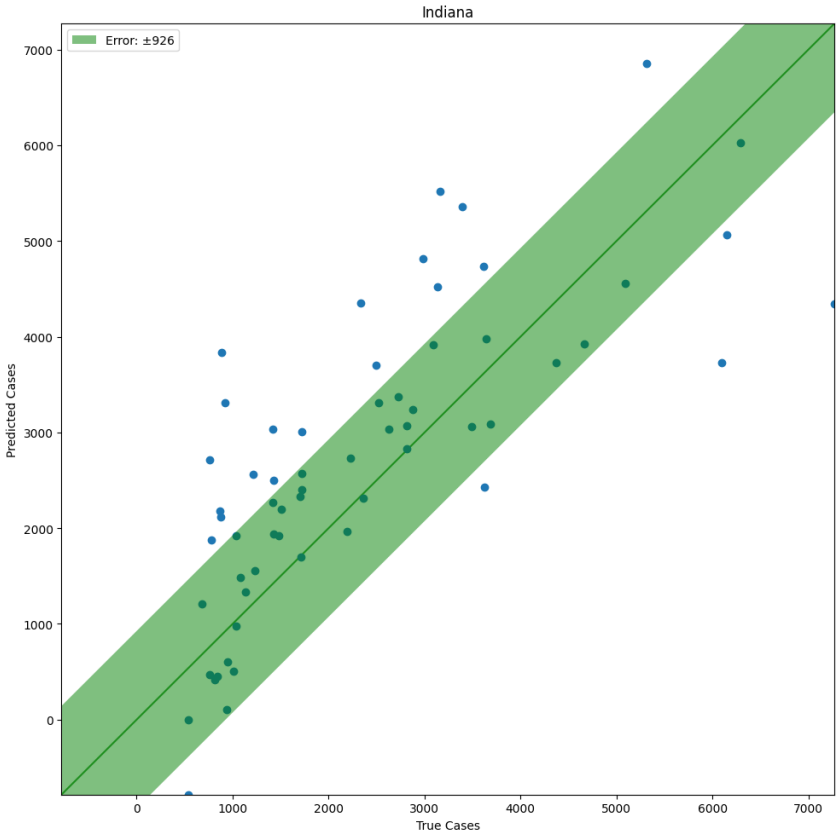


Figura 19: Diagrama de dispersión de la cantidad de casos reales vs predecidos junto con el ancho de banda de error gaussiano en la región de Indiana y la racta $x{=}y$

La Tabla 4 muestra la medida de precisión obtenida para todas los estados evaluados. Promediando el porcentaje de predicciones dentro del ancho de banda de error gaussiano y el error gaussiano para los distintos estados obtenemos como resultado un 57 % y un ± 1520 casos respectivamente.

Estado	**Porcentaje de predicciones dentro del ancho de banda de error gaussiano**	**Error gaussiano**
Arizona	68 %	±1691
Arkansas	55 %	±848
California	52 %	±6103
Colorado	42 %	±1170
Connecticut	59 %	±1215
Florida	62 %	±2785
Georgia	58 %	±2209
Idaho	58 %	±469
Illinois	52 %	±2363
Indiana	61 %	±888
Iowa	47 %	±783
Kansas	58 %	±1287
Kentucky	68 %	±858
Louisiana	55 %	±1210
Maryland	56 %	±602
Massachusetts	59 %	±1340
Michigan	55 %	±1742
Minnesota	53 %	±1748
Mississippi	50 %	±608
Missouri	50 %	±1496
Nebraska	50 %	±475
Nevada	58 %	±553
New Jersey	59 %	±1042
New Mexico	53 %	±443
New York	53 %	±3981
North Carolina	64 %	±2156
Ohio	62 %	±1544
Oklahoma	58 %	±1020
Oregon	61 %	±282
Pennsylvania	58 %	±1356
South Carolina	61 %	±1362
Tennessee	59 %	±1683
Texas	59 %	±4915
Utah	58 %	±803
Virginia	53 %	±1443
Washington	65 %	±976
Wisconsin	50 %	±1412

Tabla 4: Tabla del porcentaje de predicciones acertadas para los distintos estados evaluados.

4.3. Aplicación del método Árboles de regresión

Se analizó con árboles de regresión el conjunto de distintas formas.

Primero, dado que la cantidad de casos puede variar según el tamaño de la población del estado que se está evaluando, no se encontró sentido alguno en mezclar todas las observaciones sin tener en cuenta lo

anterior. Por esto, se comenzó a analizar el conjunto separando las observaciones por regiones y analizando cada región de forma separada. De esta forma, los modelos obtenidos solo serían útiles en el estado para el cual fueron entrenados, lo cual no es óptimo.

Por este motivo, decidimos dividir la cantidad de casos diaria por el tamaño de la población de cada estado y de esta forma nuestra variable objetivo será la proporción de contagiados. Sin embargo, ya que la popularidad de los síntomas fue multiplicada por un factor distinto y que desconocemos en cada región, no podemos hacer comparaciones entre regiones de la forma que los presenta el conjunto, una forma de poder compararlos es dividir cada síntoma por su valor medio en su región, con esto quitamos de la ecuación el factor por el cual fue multiplicado previamente. De esta forma estaremos prediciendo la proporción de casos con la variación de la popularidad de los síntomas con respecto a su valor medio. Con estos cambios ahora podemos mezclar todas las observaciones y construir un modelo más genérico.

Se realizaron pruebas con árboles de regresión y random forest regressor, pero dado que la precisión es más alta con random forest y que el método es una mejora sobre utilizar un único árbol de regresión, se decidió mostrar solo los resultados obtenidos con este método.
Para cada prueba se realizaron tres iteraciones y se promediaron los resultados obtenidos. Se utilizó un 75 % del conjunto para entrenamiento y el 25 % restante para pruebas.

A continuación se listan los parámetros utilizados, los parámetros que no se encuentran listados no fueron modificados de sus valores por defecto.

n_estimator	min_samples_leaf	max_depth
400	3	10

- n_features: Cantidad de árboles a utilizar
- min_samples_leaf: Mínima cantidad de observaciones necesarias para que un nodo hoja sea considerado
- max_depth: Máxima profundidad del árbol

Si bien al incrementar la máxima profundidad del árbol podremos mejorar levemente la precisión, esto produce sobreajuste, por lo tanto se la limitó a diez.

Más información sobre los parámetros disponible aquí.

En el conjunto observamos algunos estados en los cuales los datos no eran correctos.

Fecha	Casos diarios nuevos
2020-10-08	384
2020-10-09	290
2020-10-10	0
2020-10-11	0
2020-10-12	1339
2020-10-13	320

Tabla 5: Casos por día en Connecticut

Fecha	Casos diarios nuevos
2020-05-30	0
2020-05-31	0
2020-06-01	292
2020-06-02	0
2020-06-03	159
2020-06-04	0

Tabla 6: Casos por día en Kansas

En la Tabla 6 podemos observar como no se reportan los casos en ciertos días, sin encontrar un patrón claro. Por otro lado, en la Tabla 5 podemos ver que para las fechas 2020-10-10 y 2020-10-11 no se encuentran casos, estos días son sábado y domingo respectivamente. También vemos que el valor del día lunes (2020-10-12) es considerablemente superior al viernes y martes, ya que los casos no contados durante el fin de semana se acumulan los lunes.

Dado que no es factible utilizar el conjunto con estos días en los que no se reportan casos, decidimos filtrar estos datos, para esto aplicamos tres formas distintas de filtrado:

- Quitar todos: Elimina todos los días con cero casos que se encuentran luego del primer caso detectado en la región
- Quitar regiones: Elimina la región si tiene una cierta cantidad de días con cero casos luego del primer caso detectado
- Quitar próximo día: Igual al método "Quitar todos", pero además elimina el próximo día si tiene una cantidad nueva de casos mayor a diez, esto ayuda a mantener los datos si la cantidad de casos oscila entre cero y diez.
- Ninguno: No se realiza ningún filtrado

4.3.1. Resultados

Primero se ilustra en las Figuras 20, 21, 22, 23, 24 y 25, 26 los resultados por regiones para los distintos tipos de filtrado.

La precisión la calculamos con el coeficiente de determinación R^2, definido en la ecuación 8.

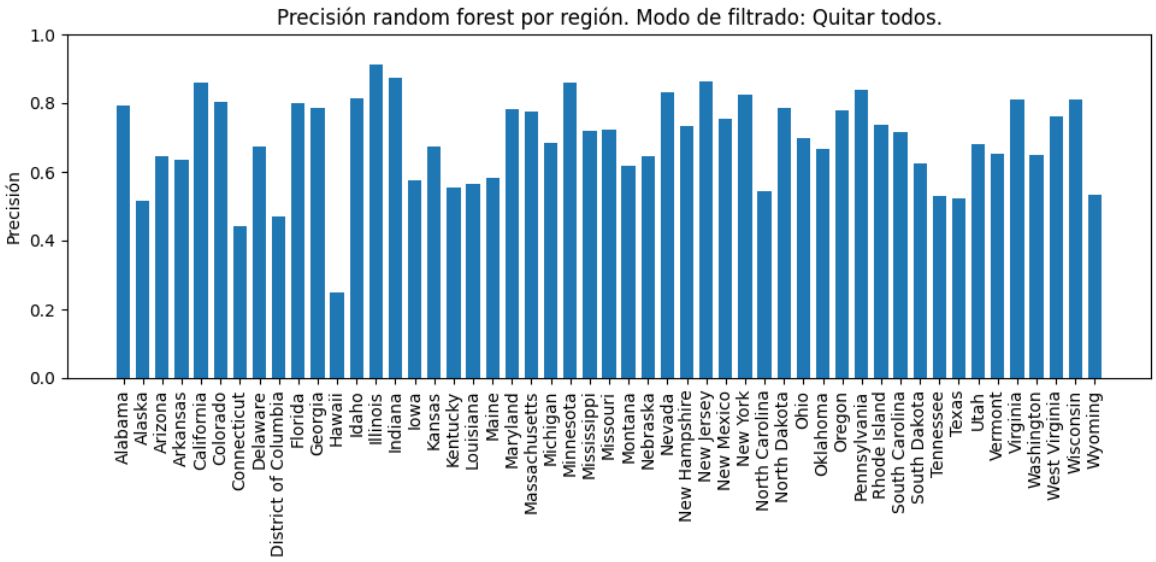


Figura 20: Regresión con random forest. Modo de filtrado: Quitar todos. Precisión media: 0.694

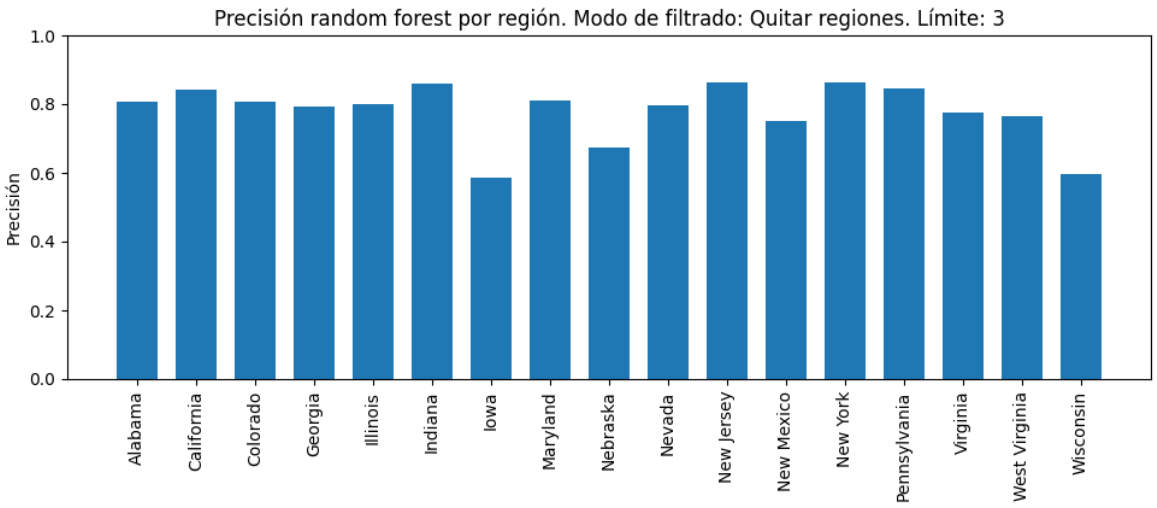


Figura 21: Regresión con random forest. Modo de filtrado: Quitar regiones con límite 3. Precisión media: 0.779

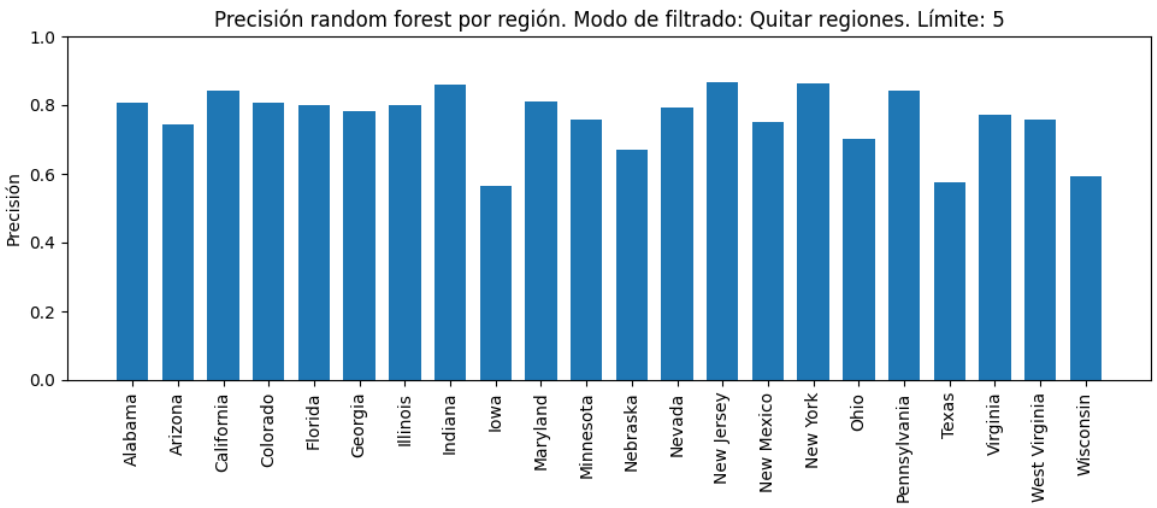


Figura 22: Regresión con random forest. Modo de filtrado: Quitar regiones con límite 5. Precisión media: 0.762

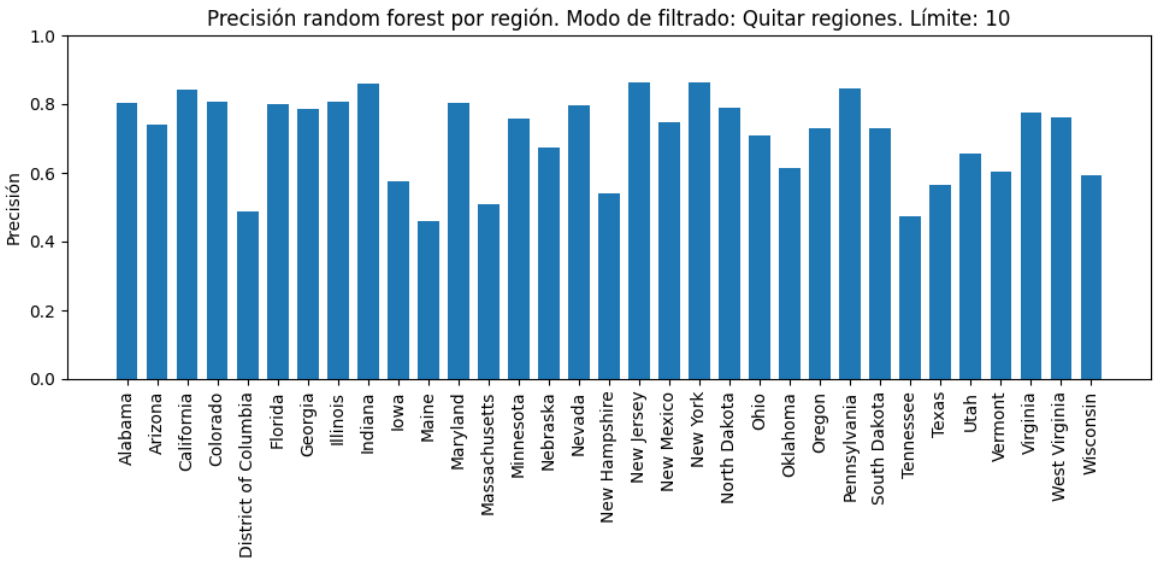


Figura 23: Regresión con random forest. Modo de filtrado: Quitar regiones con límite 10. Precisión media: 0.708

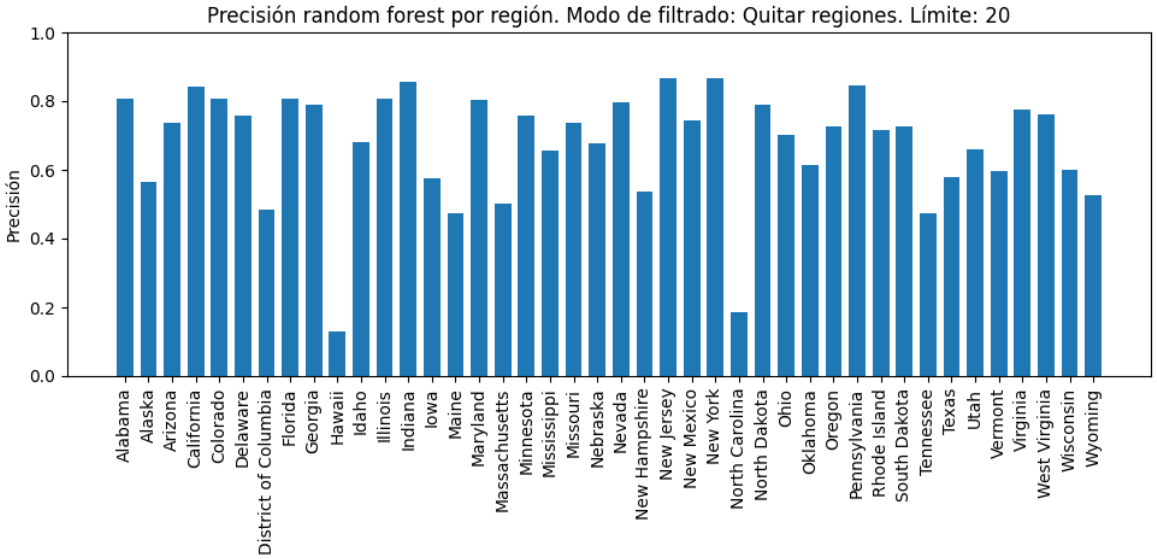


Figura 24: Regresión con random forest. Modo de filtrado: Quitar regiones con límite 20. Precisión media: 0.675

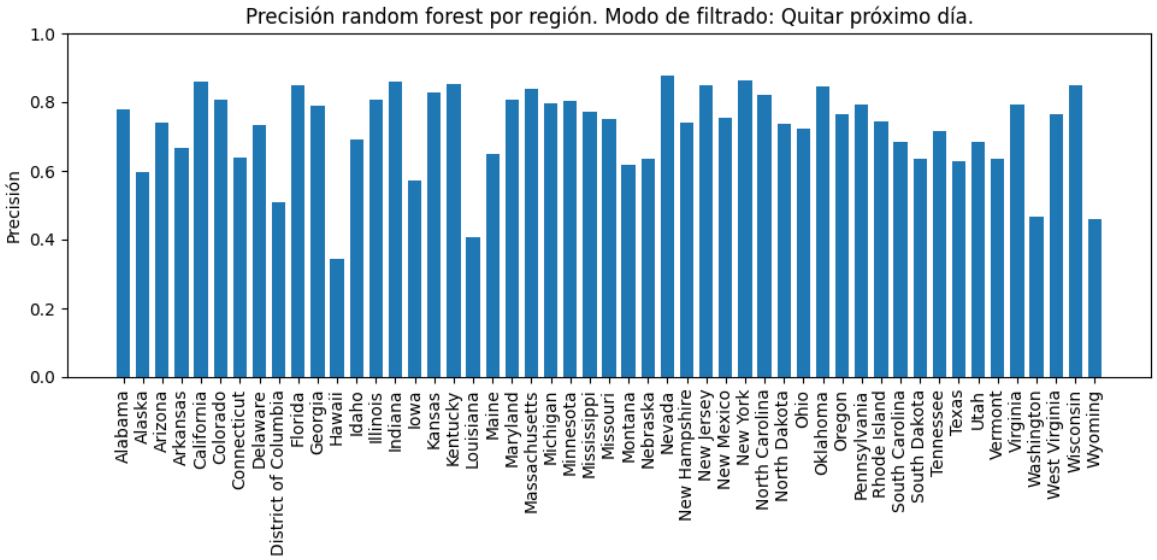


Figura 25: Regresión con random forest. Modo de filtrado: Quitar próximos días. Precisión media: 0.722

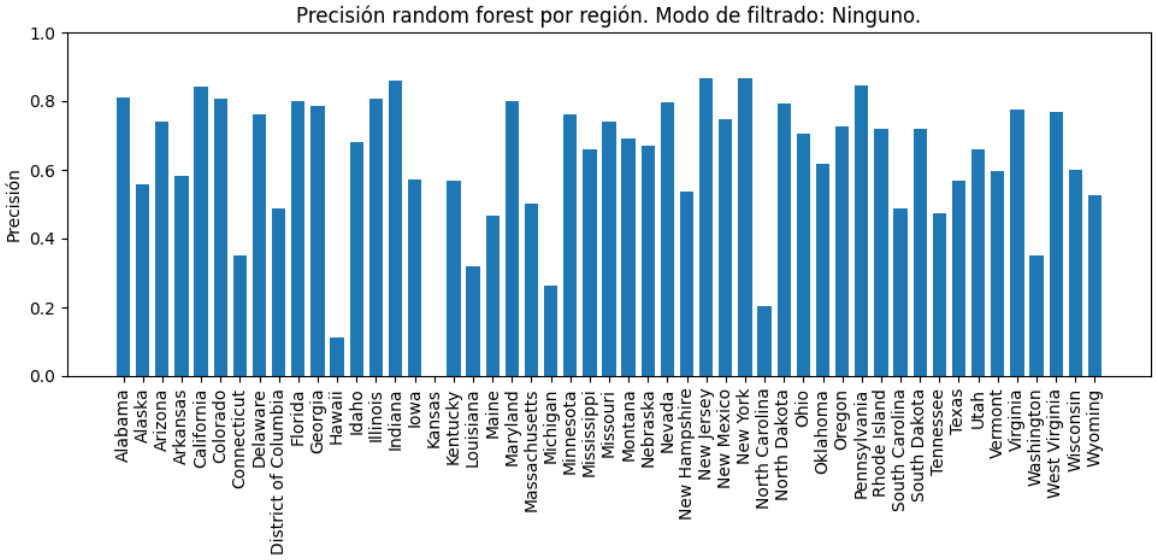


Figura 26: Regresión con random forest. Modo de filtrado: Ninguno. Precisión media: 0.624

Para los próximos resultados se realizaron los cambios descriptos en la sección anterior para poder juntar todas las observaciones.

Modo de filtrado	Precisión
Ninguno	0.634
Quitar todos	0.654
Quitar regiones Límite = 3 Regiones = 17	0.760
Quitar regiones Límite = 5 Regiones = 22	0.764
Quitar regiones Límite = 10 Regiones = 33	0.697
Quitar regiones Límite = 20 Regiones 42	0.682
Quitar próximo día	0.714

Tabla 7: Precisión del árbol para los distintos tipos de filtrado

En la Tabla 7 "Límite" indica la cantidad máxima de días con cero casos que puede tener una región para ser tenida en cuenta. "Regiones" indica la cantidad de regiones que cumplieron la condición para ese límite en concreto.

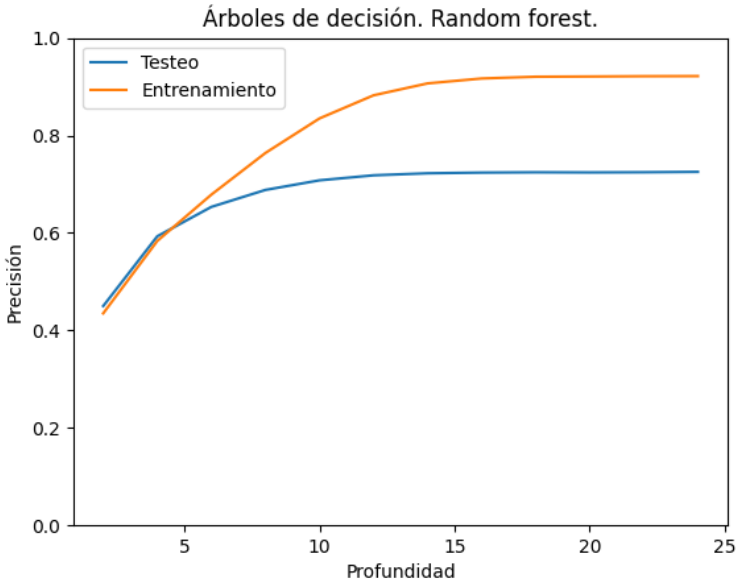


Figura 27: Prueba de sobreajuste. Modo de filtrado: Quitar próximos días.

En la Figura 27 realizamos la prueba de sobreajuste, decidimos tomar una profundidad máxima de diez nodos, ya que a partir de ese punto se produce sobreajuste, creciendo considerablemente la precisión del conjunto de entrenamiento pero muy levemente la del conjunto de testeo.

4.4. Conclusiones

A modo de conclusión, predecir el número de casos COVID-19 positivos en un día no es una tarea sencilla. Más aún cuando la variable objetivo no siempre es certera. Suelen suceder situaciones en las que se acumulan datos durante un periodo corto de tiempo y luego se registran los casos todos juntos en forma de lote para como si hubiesen ocurrido el mismo día.
Sin embargo con el modelo de regresión lineal múltiple en el mejor de los escenarios se obtuvo un 68 % de predicciones correctas y considerando todas la regiones evaluadas un 57 %.
Si bien este porcentaje es bajo, sirve como motivación para continuar analizando el conjunto de datos con otros métodos de aprendizaje automático para encontrar el modelo de predicción óptimo.
Por otro lado demuestra la utilidad de recolectar las tendencias diarias de búsquedas de síntomas en internet. Una de las desventajas del modelo es que solo es aplicable para los 6 países para los cuales se registraron datos. Queda demostrada la importancia de recopilar datos de todos los países, para así extender el modelo o bien realizar futuros análisis distintos a este.

En cuanto al modelo de árboles de regresión inicialmente conseguimos una precisión media de 62.4 % al analizar el conjunto por regiones y 63.4 % al independizar los datos de la regiones. Se realizaron distintos tipos de filtrado que ayudaron a mejorar hasta cerca de un 9 %, con el tipo de filtrado "Quitar próximo día", se llegó a alcanzar una precisión media del 72.2 % analizando por regiones y 71.4 % en los datos independientes a las regiones.
El incremento en la precisión muestra la importancia de contar con datos precisos que no contengan errores, esto queda aún más en evidencia cuando se filtró estableciendo un límite de días en los que no se pueden no haber reportado casos, aumentando la precisión hasta un 76.4 %, pero la cantidad de datos que eliminamos al realizar este tipo de filtrado hacen este modelo menos genérico.
Para finalizar, creemos que al mezclar los datos de todas las regiones fuimos capaces de construir un modelo más genérico, que podría llegar a ser utilizado en otros países, y con una precisión suficientemente alta para ser de utilidad a la hora de predecir los casos.

5. Diagnosis COVID-19 alternativo basado en síntomas presentes

Desde el inicio de la pandemia el alto número de contagios continúa desafiando al sistema de salud de países subdesarrollados como también desarrollados. La alta demanda de camas de hospital, la escasez de equipamiento médico y falta de test RT-PCR son algunas de las dificultades que se tienen.
Es por esto que se deben tomar decisiones eficaces del uso de recursos médicos. Además tienen que ser rápidas ya que la vida de muchos individuos están en juego.

Una detección rápida y temprana del virus ayudaría al personal médico de todo el mundo a clasificar a los pacientes y tomar decisiones en un contexto de recursos limitados.
Si lográsemos un modelo de predicción que puede ser utilizado por cualquier usuario en cualquier lugar del mundo a través de una pagina web o un celular. Esto podría ayudar a mitigar gran parte de la carga sobre el sistema de salud, como también reducir la tasa de infecciones.

En esta sección analizaremos los síntomas presentes en pacientes que realizaron un test de RT-PCR. El objetivo es desarrollar un modelo para predecir si un individuo esta infectado o no con la enfermedad a partir de una serie de preguntas sobre los síntomas experimentados

5.1. Conjunto de datos utilizados

Con la llegada del COVID-19 en el mes de Febrero 2020 comenzaron los testeos en Israel. Con el fin de realizar un post-análisis, a cada individuo se le realizó una serie de preguntas de los síntomas presentados antes de tomar el test RT-PCR.
Estas preguntas son:

1. Experimentó tos antes de realizar el test?
2. Experimentó fiebre antes de realizar el test?
3. Experimentó dolores de garganta antes de realizar el test?
4. Experimentó falta de aire antes de realizar el test?
5. Experimentó dolores de cabeza antes de realizar el test?
6. Por qué motivo realiza el test?

La Tabla 8 muestra el conjunto de variables categóricas obtenidas de los datos recopilados.

Fecha test RT-PCR	dd/mm/yy	
Resultado test RT-PCR	0: Negativo 1: Positivo	
Edad	0: 60 > edad 1: 60 ≤ edad	
Genero	0: Hombre 1: Mujer	
Motivo	0: Contacto con contagiado 1: Otro 2: Llegada del extranjero	
Síntomas	Tos	0: Síntoma no presente 1: Síntoma presente
	Fiebre	0: Síntoma no presente 1: Síntoma presente
	Dolor de garganta	0: Síntoma no presente 1: Síntoma presente
	Falta de aire	0: Síntoma no presente 1: Síntoma presente
	Dolores de cabeza	0: Síntoma no presente 1: Síntoma presente

Tabla 8: Variables categoricas

Durante el período de 22-03-2020 a 14-04-2020 se recolectaron un total de 136294 datos de distintos individuos, numero suficiente para realizar el desarrollo del modelo.

		Porcentaje Respecto del Total	Porcentaje de Positivos	Porcentaje de Negativos
Tos	0: Síntoma no presente	81.77 %	4.98 %	95.02 %
	1: Síntoma presente	18.23 %	20.43 %	79.57 %
Fiebre	0: Síntoma no presente	90.71 %	5.00 %	95.00 %
	1: Síntoma presente	9.29 %	35.14 %	64.86 %
Dolor de garganta	0: Síntoma no presente	98.92 %	6.88 %	93.12 %
	1: Síntoma presente	1.08 %	92.06 %	7.94 %
Falta de aire	0: Síntoma no presente	99.22 %	7.14 %	92.86 %
	1: Síntoma presente	0.78 %	91.89 %	8.11 %
Dolores de cabeza	0: Síntoma no presente	98.48 %	6.43 %	93.57 %
	1: Síntoma presente	1.52 %	96.10 %	3.90 %
Edad	0: Síntoma no presente	82.61 %	7.61 %	92.39 %
	1: Síntoma presente	17.39 %	8.66 %	91.34 %
Genero	0: Síntoma no presente	50.74 %	8.70 %	91.30 %
	1: Síntoma presente	49.26 %	6.87 %	93.13 %
motivo	0: Contacto con contagiado	5.43 %	69.81 %	30.19 %
	1: Otro	83.91 %	3.59 %	96.41 %
	2: Llegada del extranjero	10.66 %	9.28 %	90.72 %

Tabla 9: Características de los datos del gobierno de Israel

En la Tabla 9 se pueden observar como se distribuyen las variables sobre el total de individuos, total con COVID-19 positivo y negativo.
Se puede observar una una mayor proporción de casos negativo sobre casos positivo. Analizando los datos registrados se noto que la mayoría de los individuos cuyo test dio negativo no presento ninguno de los síntomas en cuestión. Las entradas del conjunto de datos que presentan esta característica no aportan información suficiente para el análisis que se quiere llevar acabo. La tabla 10 muestra el conjunto resultante, removiendo las entradas en cuestión.

		Porcentaje Respecto del Total	Porcentaje de Positivos	Porcentaje de Negativos
Tos	0: Síntoma no presente	23.18 %	74.00 %	26.00 %
	1: Síntoma presente	76.82 %	20.43 %	79.57 %
Fiebre	0: Síntoma no presente	60.86 %	31.37 %	68.63 %
	1: Síntoma presente	39.14 %	35.14 %	64.86 %
Dolor de garganta	0: Síntoma no presente	95.45 %	30.02 %	69.98 %
	1: Síntoma presente	4.55 %	92.06 %	7.94 %
Falta de aire	0: Síntoma no presente	96.72 %	30.84 %	69.16 %
	1: Síntoma presente	3.28 %	91.89 %	8.11 %
Dolores de cabeza	0: Síntoma no presente	93.59 %	28.51 %	71.49 %
	1: Síntoma presente	6.41 %	96.10 %	3.90 %
Edad	0: Menos de 60 años	85.03 %	31.17 %	68.83 %
	1: 60 años o más	14.97 %	42.38 %	57.62 %
Genero	0: Hombre	53.24 %	34.91 %	65.09 %
	1: Mujer	46.76 %	30.50 %	69.50 %
motivo	0: Contacto con contagiado	21.17 %	75.45 %	24.55 %
	1: Otro	40.00 %	31.76 %	68.24 %
	2: Llegada del extranjero	38.83 %	10.74 %	89.26 %

Tabla 10: Características de los datos del gobierno de Israel, después de remover entradas no informativas

5.2. Aplicación del método Arboles de Decisión

Como el conjunto de datos esta compuesto de variables categóricas y con distribución discreta se eligió el método de random forest para predecir si un consultante tiene o no COVID-19 a partir de la 8 preguntas mencionadas anteriormente. Del total de los datos se utilizo un 70 % para el entrenamiento del modelo y el 30 % restante para testearlo.
Los parámetros que configuran el modelo de random forest son:

- Total de árboles: 100
- Mínimo número de observaciones para dividir un nodo: 2
- Mínimo número de observaciones para un nodo terminal: 1
- Máxima profundidad del árbol: Los nodos se expanden hasta que todas las hojas sean puras o hasta que todas las hojas contengan menos que el mínimo de observaciones para dividir un nodo
- Máximo número de nodos hoja: ilimitado
- Máximo número de atributos a considerar para la ramificación: 8

Se utilizó el método de validación cruzada en k partes para optimizar los parámetros seleccionados. Para ello se lo emplea en el conjunto de entrenamiento con k = 10, se eligió el valor-f como métrica para comparar los distintos modelos entrenados.
Al valor-f se lo calcula de la siguiente forma:

$$F_1 = 2 \cdot \frac{\text{Precisión} \cdot \text{Exhaustividad}}{\text{Precisión} + \text{Exhaustividad}} \tag{25}$$

donde,

$$\text{Precisión} = \frac{Tp}{Tp + Fp} \tag{26}$$

La precisión mide los positivos verdaderos sobre todos los que dieron positivo.

$$\text{Exhaustividad} = \frac{Tp}{Tp + Fn} \tag{27}$$

La exhaustividad mide los positivos verdaderos sobre todos los que son positivos.

$$Tp = \text{Verdaderos positivos}$$
$$Tn = \text{Verdaderos negativos}$$
$$Fp = \text{Falsos positivos}$$
$$Fn = \text{Falsos negativos}$$

Nuestra principal preocupación es evitar que el modelo prediga falsamente que un individuo no posee el virus COVID-19, son estos los que luego podrían contagiar al resto de la población esparciendo el virus. Por ello elegimos el valor-f como medida ya que consiste en un valor ponderado de la precisión y la exhaustividad que expresa ambas propiedades. Esto nos permite minimizar tanto los falsos positivos como los falsos negativos.

5.2.1. Resultados

En esta sección se expondrán los resultados obtenidos a partir de las predicciones realizadas con el conjunto de testeo y los resultados reales esperados.

Exactitud	0.886
Precisión	0.906
Exhaustividad	0.731
Valor-F	0.809
AUC (Área debajo de la curva ROC)	0.926

Tabla 11: Métricas de Evaluación

La Tabla 11 muestra las distintas métricas de evaluación del modelo obtenidas a partir de la tasa de verdaderos positivos, verdaderos negativos, falsos positivos, falsos negativos representadas por la matriz

de confusión.
La métrica AUC se obtuvo mediante el calculo del área debajo de la curva ROC de la Figura 29.
La Figura 28 define la matriz resultante, donde cada columna representa el número de predicciones COVID-19 negativas y positivas, mientras que cada fila representa el número real de casos COVID-19 negativos y positivos.
En la Figura 30 podemos visualizar los gráficos de validación cruzada en 10 partes para el conjunto de entrenamiento y testeo. Los parámetros variados son: total de árboles, máxima profundidad del árbol, mínimo número de observaciones para dividir un nodo y mínimo número de observaciones para un nodo terminal.

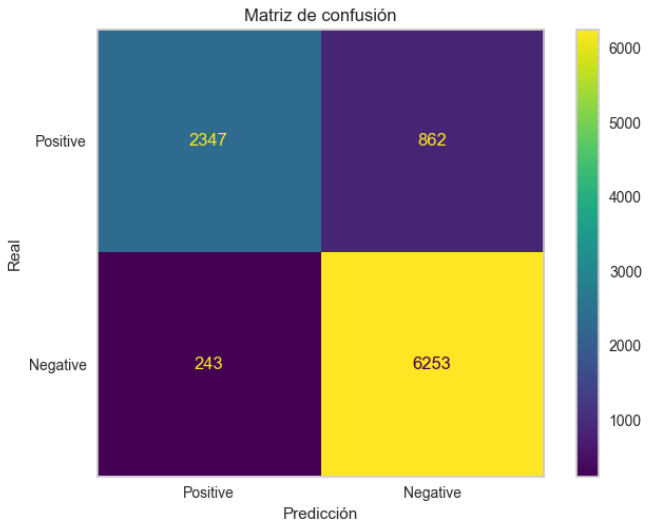


Figura 28: Matriz de Confusión

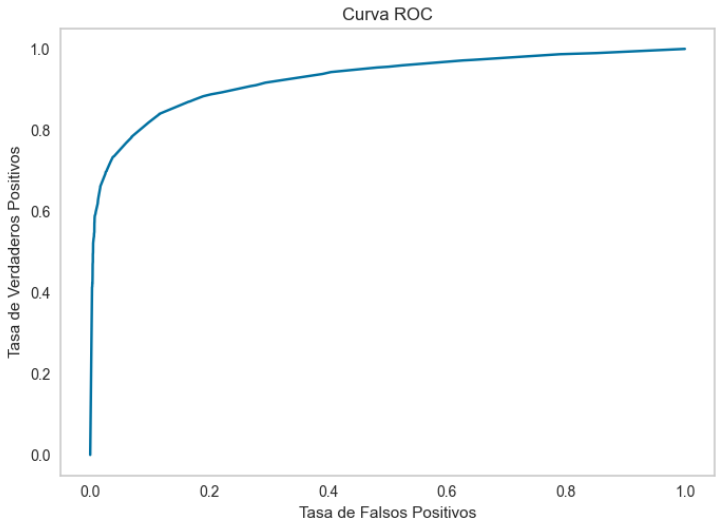


Figura 29: Curva ROC

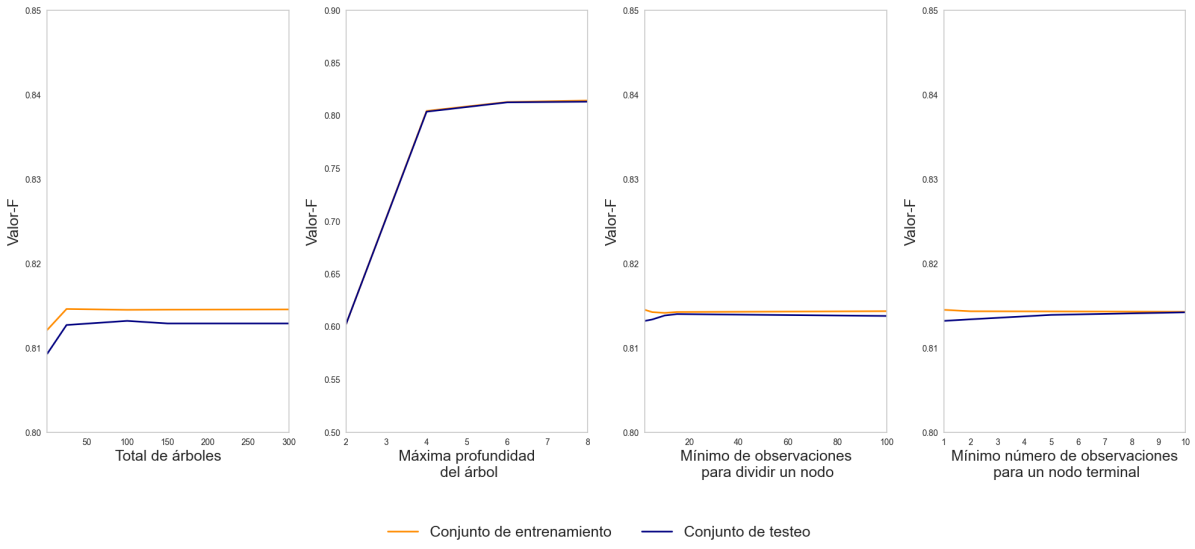


Figura 30: Validación cruzada en 10 partes utilizando el valor-f como métrica

5.3. Conclusiones

El análisis demostró que a partir de métodos de aprendizaje automático, la posibilidad de predecir con un alto grado de efectividad el riesgo de que un paciente obtenga un resultado COVID-19 positivo en un test RT-PCR. Creemos que es evidencia suficiente del potencial que posee el aprendizaje automático sobre la toma de decisiones clínicas.

La ventaja del modelo desarrollado es que realiza predicciones según los 8 síntomas básicos y sencillos de ser detectados. Esto es lo que lo hace altamente portátil y utilizable tanto por médicos expertos como por individuos sin conocimiento clínico.

Se proponen como futuras extensiones al trabajo utilizar otros métodos a modo de comparación como XGBoost, SVM, entre otros. Por otro lado se debe realizar un estudio en otra ubicación distinta de Israel para evaluar la capacidad del modelo para generalizar.

Finalmente incentivamos el desarrollo de una aplicación para celulares inteligentes, tabletas digitales como también una aplicación web para alcanzar un amplio grupo de usuarios. Esta constaría en un formulario con las 8 preguntas a responder y utilizando el método de random forest lograr una predicción lo mas temprana posible de casos COVID-19 positivos.

6. Análisis de Casos Positivos en México en Base a las Pre-Condiciones de una Persona

En México se dispuso por decreto poner a disposición de la población todos los datos de interés público

a la población de manera pro-activa. Con la pandemia global empezada en el año de 2020, la Dirección General de Epidemiología Mexicana siguió el decreto, creado en 2015, y empezó a compartir todos los datos definidos como importantes acerca del virus, manteniendo siempre la anonimidad de los pacientes / población de la cual los datos eran compartidos.

Una de las primeras y grandes dudas que vinieron sobre el virus SARS-CoV-2 es la mortalidad del patógeno. La población mundial se asustó con su alta velocidad de propagación, aumentando la visibilidad de la situación calamitosa y el número de muertes consecuentes de la enfermedad.

Encontrar una relación entre las comorbilidades / pre-condiciones y la letalidad del virus puede desencadenar diversas políticas públicas para controlar instalaciones y equipamientos fundamentales al tratamiento del virus, como respiradores y camas hospitalarias. Esto es fundamental en momentos de pandemia en el cual los países no están preparados para soportar en los hospitales las consecuencias de una enfermedad con alto grado de contagio. Los países en vías de desarrollo o subdesarrollados no tienen ni como fabricar o equipar los hospitales en el tiempo necesario, ni como conseguir el personal necesario para trabajar en la salud pública.

6.1. Conjunto de datos utilizados

En los datos proporcionados por el gobierno Mexicano, el que se usó en esta sección del trabajo posee datos personales de pacientes que hicieron el testeo de la enfermedad en el país. Además de las informaciones personales de cada paciente como edad y género, se identificó diferentes pre-condiciones de cada paciente. Finalmente, de acuerdo con el cuadro de cada paciente se identificó si el mismo se infectó con la enfermedad, si tuvo que internarse y si en caso positivo, la fecha de muerte. Los datos analizados de cada persona fueron descriptos en las Tablas 12 y 13.

Para este análisis, se decidió analizar las personas que ya estaban infectadas con el virus. Por esto, se usó los datos con resultados positivos de los exámenes de COVID. Así mismo, para alcanzar el objetivo principal del análisis, se borraron algunas columnas que no tenían sentido analizar. Estas son: El resultado del examen de COVID, por ya haber filtrado los casos positivos, el contacto con otros infectados, si el paciente estuvo en ambulatorio o hospitalizado, si el mismo estuvo intubado o si estuvo en la Unidad de Cuidados Intensivos.

Como hay muchos datos faltantes por filas de datos de pacientes de COVID, se usó imputación de datos con KNN, buscando los valores vecinos mas próximos para cada columna sin datos. De esta manera, intenta simular los valores inexistentes con opciones mas próximas en base a los propios datos.

Observando los datos, se puede ver que hay una mayor proporción de personas que sobrevivieron a la infección comparado a los que fallecieron. Para que los resultados de los entrenamientos de los modelos sean mas cercanos a la realidad, se uso la técnica de undersampling para balancear a los datos. La técnica consiste en bajar el número de datos marcados que están en exceso para que la relación entre categorías quede proporcional. En el caso de este análisis, se equiparó el dato de sobrevivientes con el de pacientes muertos.

Género	1: Masculino 2: Femenino
Tipo de paciente	1: Ambulatorio 2: Hospitalizado
Fecha de entrada en el hospital	dd-mm-yyyy
Fecha de los primeros síntomas	dd-mm-yyyy
Fecha de muerte	dd-mm-yyyy
Intubado	1: Si 2: No 97,98,99: Datos faltantes
Neumonía	1: Si 2: No 97,98,99: Datos faltantes
Edad	Años
Embarazo	1: Si 2: No 97,98,99: Datos faltantes
Diabetes	1: Si 2: No 97,98,99: Datos faltantes
Enfermedad pulmonar obstructiva crónica (EPOC)	1: Si 2: No 97,98,99: Datos faltantes
Asma	1: Si 2: No 97,98,99: Datos faltantes
Inmunosupresión	1: Si 2: No 97,98,99: Datos faltantes
Hipertensión	1: Si 2: No 97,98,99: Datos faltantes
Otras enfermedades	1: Si 2: No 97,98,99: Datos faltantes
Cardiovascular	1: Si 2: No 97,98,99: Datos faltantes
Insuficiencia Renal Crónica	1: Si 2: No 97,98,99: Datos faltantes

Tabla 12: Variables categóricas de pacientes Mexicanos

Tabaquismo	1: Si 2: No 97,98,99: Datos faltantes
Contacto con otros infectados de SARS-CoV-2	1: Si 2: No 97,98,99: Datos faltantes
Resultado del examen de COVID	1: Si 2: No 3: Esperando resultados
Unidad de Cuidados Intensivos (UCI)	1: Si 2: No 97,98,99: Datos faltantes

Tabla 13: Variables categóricas de pacientes Mexicanos, continuación.

Para los diferentes métodos, se decidió enfocar en encontrar si es que hay una relación entre las muertes y las pre-condiciones de un paciente, con el objetivo de encontrar la relación entre la mortalidad y las comorbilidades de una persona.

Los algoritmos usados para analizar esa relación son C4.5, Regresión Logística y Perceptrón Multicapa. Los tres algoritmos fueron elegidos porque todos hacen predicciones categóricas, o sea, hacen predicciones en base a una característica de las variables de un paciente. En este caso, la variable categórica es si el paciente falleció o no.

6.2. Aplicación del método C4.5

Al entrenar el algoritmo con el 80 % de los datos, se encontró la precisión, la exactitud, la exhaustividad y la relación entre la precisión y la exhaustividad (F1) para el conjunto de entrenamiento y testeo. La 14 muestra los valores obtenidos para las distintas metricas.

La Figura 31 presenta una matriz de confusión para los valores testeados. En el caso de estar marcado como 1, entonces el valor que el algoritmo predijo es que el paciente falleció. En el caso contrario - marcado como 0 - el paciente sobrevivió a la enfermedad.

El Gráfico 32 en compensación, presenta las pre-condiciones que más impactan al modelo a la hora de clasificar si paciente falleció, calculadas sumando los valores calculados con la ecuación 21, por cada característica presente. El impacto se da por la cantidad de nodos en el árbol que usan a la variable en la toma de decisiones y la entropía de la variable en el nivel correspondiente.

(a) Valores para el conjunto de entrenamiento

Precisión	85.71 %
Exactitud	83.21 %
Exhaustividad	79.37 %
Valor-F	82.42 %

(b) Valores para el conjunto de testeo

Precisión	79.50 %
Exactitud	82.15 %
Exhaustividad	85.60 %
Valor-F	82.43 %

Tabla 14: Tablas con valores de precisión, exactitud, exhaustividad y F1

6.2.1. Resultados

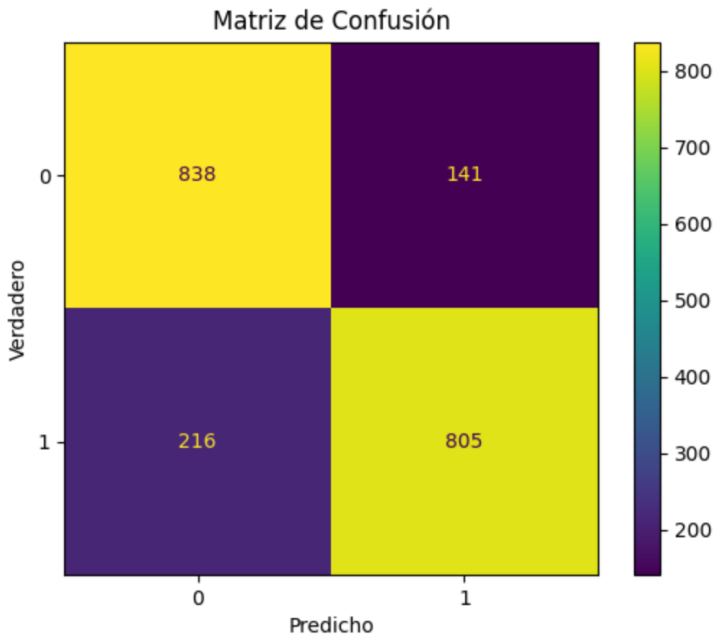


Figura 31: Matriz de Confusión

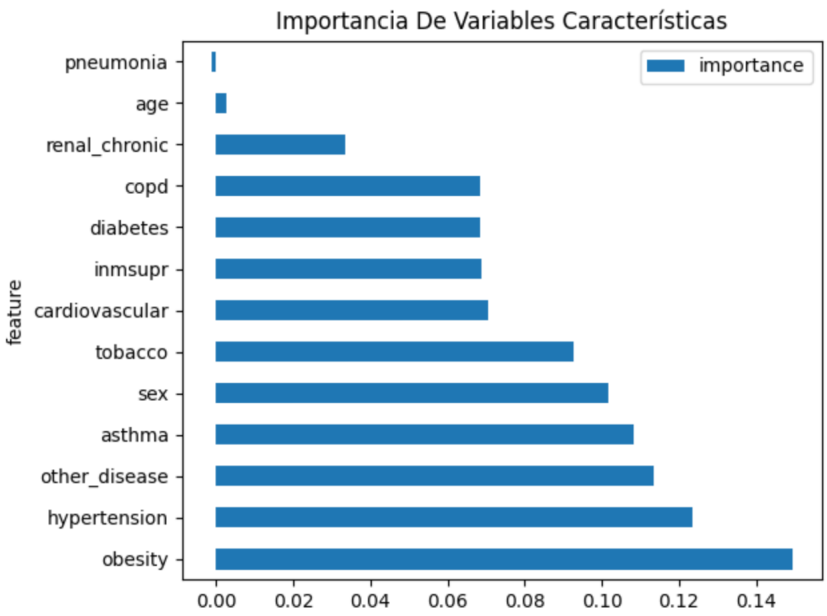


Figura 32: Importancia de las Pre-Condiciones

6.3. Aplicación del método Regresión Logística

Para la Regresión Logística, se usó el método de eliminación hacia atrás en la selección de variables. Se analizaron los p-valores y por iteraciones se eliminaron el de valor más alto. Se repitió este algoritmo hasta tener todos los valores con el p-valor $<= 0.05$.

Los p-valores en la Tabla 16 que están representados con 0.0000, no dicen necesariamente que el p-valor quedo igual a 0, pero su primer decimal distinto a 0 no es lo suficiente grande para aparecer. Como se puede ver, apenas las pre-condiciones de enfermedades cardiovasculares y asma fueron eliminadas. Para mejor análisis, se buscaron en la Tabla 17 los valores de los β calculados en el entrenamiento del modelo.

Otro análisis realizado es la curva ROC, representado en la Figura 34, que mide la relación entre el porcentaje de falsos y verdaderos positivos. Para finalizar, se buscó la curva Precisión-Exhaustividad (Figura 35), que busca la relación entre la precisión y exhaustividad en el modelo. Con este valor se calcula también la precisión promedio del modelo, usando como peso la variación de la exhaustividad. Este valor se calcula con la ecuación:

$$AP = \sum_n (R_n - R_{n-1})P_n \tag{28}$$

siendo n un valor que va variando la exhaustividad (R) de acuerdo con los resultados de la curva. Las matriz de confusión de la predicción está representada en la Figura 33.

6.3.1. Resultados

Síntomas	**P-Valor**		
	Iteración 1	**Iteración 2**	**Iteracion 3**
Sexo	0.0000	0.0000	0.0000
Neumonía	0.0000	0.0000	0.0000
Edad	0.0000	0.0000	0.0000
Diabetes	0.0000	0.0000	0.0000
EPOC	0.0001	0.0001	0.0002
Asma	0.0868	0.1139	-
Inmunosupresión	0.0572	0.0011	0.0007
Hipertensión	0.0000	0.0000	0.0000
Otras Enfermedades	0.0000	0.0000	0.0000
Cardiovascular	0.6910	-	-
Obesidad	0.0000	0.0000	0.0000
Insuficiencia Renal Crónica	0.0000	0.0000	0.0000
Tabaquismo	0.0205	0.0073	0.0126

Tabla 16: Test de hipótesis para las pre-condiciones

Síntomas	**Coeficiente β**
Sexo	-0.260488
Neumonía	1.218859
Edad	1.010289
Diabetes	0.197355
EPOC	0.053202
Inmunosupresión	0.047503
Hipertensión	0.118006
Otras Enfermedades	0.120418
Obesidad	0.139300
Insuficiencia Renal Crónica	0.197084
Tabaquismo	-0.034024

Tabla 17: Coeficientes para las pre-condiciones

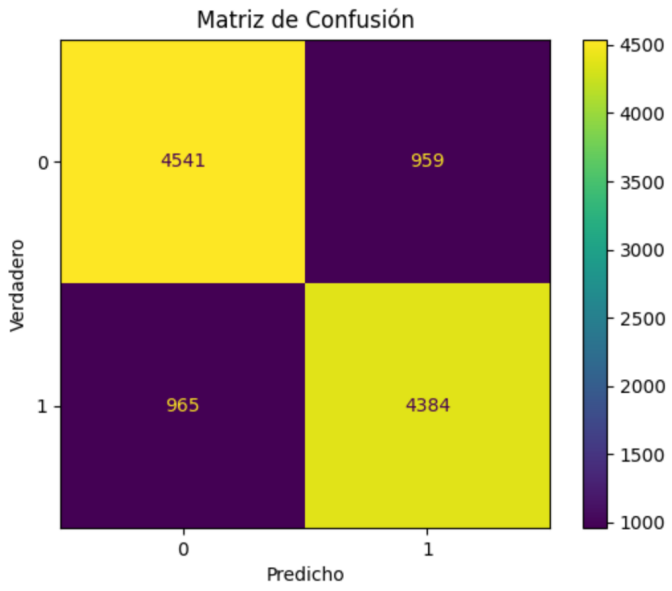


Figura 33: Matriz de Confusión para la Regresión Logística

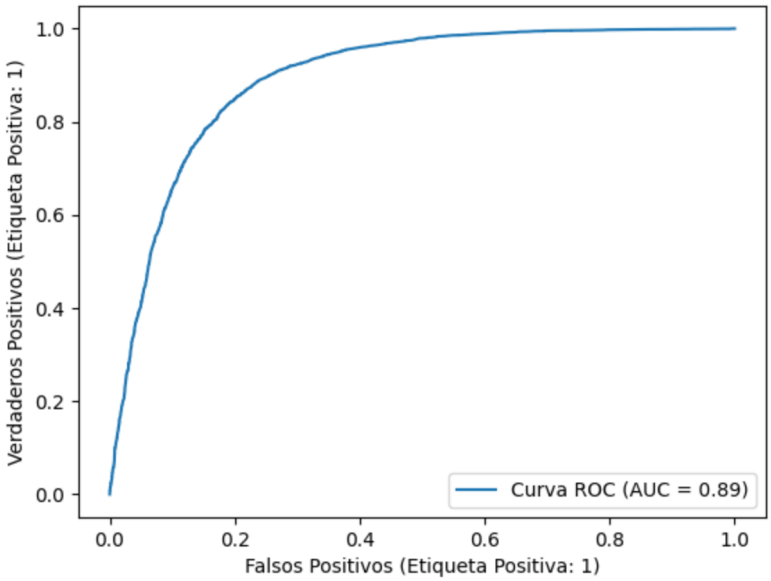


Figura 34: Curva ROC para la Regresión Logística

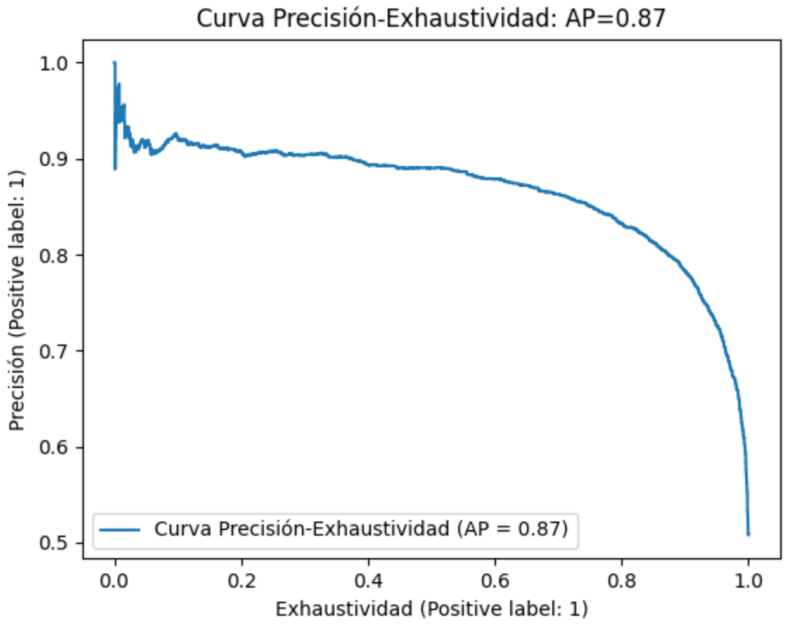


Figura 35: Curva Precisión-Exhaustividad para la Regresión Logística

6.4. Aplicación del método Redes Neuronales - Perceptrón Multicapa

Para los Perceptrones Multicapa, se buscó en el análisis la relación entre el Valor-F del modelo mientras se hacia cambios en características en el mismo. Lo primero analizado es el tamaño del perceptrón, con el número de capas internas y la cantidad de nodos en cada una. Después, se analizó la variación de las funciones escalonadas, variando entre la tangente hiperbólica, la función escalón simple y la función logística. Una vez analizado esto, se varió los valores de β, que representan la pendiente de la función escalonada, o sea, el cuan sensible es la función a la hora de calcular la salida de la clasificación. Finalmente, se varió la tasa de aprendizaje. En las Tablas 19 y 21, están los resultados de cada análisis con mayor Valor-F.

Se repitió el mismo proceso, pero la segunda vez se tomaron en cuenta además los datos del paciente: Si fue intubado, si se internó o si fue a la Unidad de Cuidados Intensivos.

6.4.1. Resultados

La Figura 36 presenta los gráficos encontrados para los casos sin usar los datos de internación, mientras que la Figura 37 presenta los resultados para los casos teniendo en cuenta la información de internación de los pacientes. Las Tablas 18 y 20 presentan los resultados de precisión, exactitud y exhaustividad de los modelos, mientras las Tablas 19 y 21 los valores finales usados en el perceptrón para los modelos.

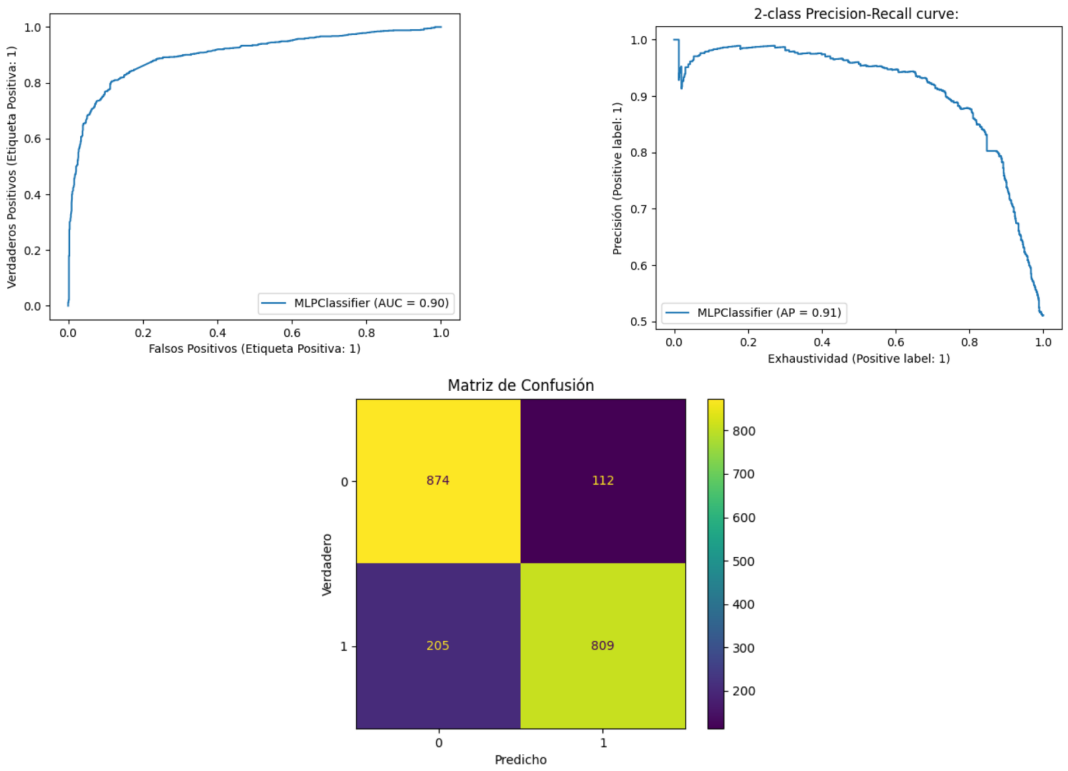


Figura 36: Matriz de confusión y curvas ROC y AUC para las pre-condiciones

Precisión	81.00 %
Exactitud	84.15 %
Exhaustividad	88.64 %
Valor-F	84.65 %

Tabla 18: Resultado Final para las pre-condiciones usando el perceptrón

Capas Ocultas	(4,)
Función de Activación	ReLu
Tasa de Aprendizaje	0.235
β	0.317

Tabla 19: Valores Finales del Perceptrón Multicapa para las pre-condiciones

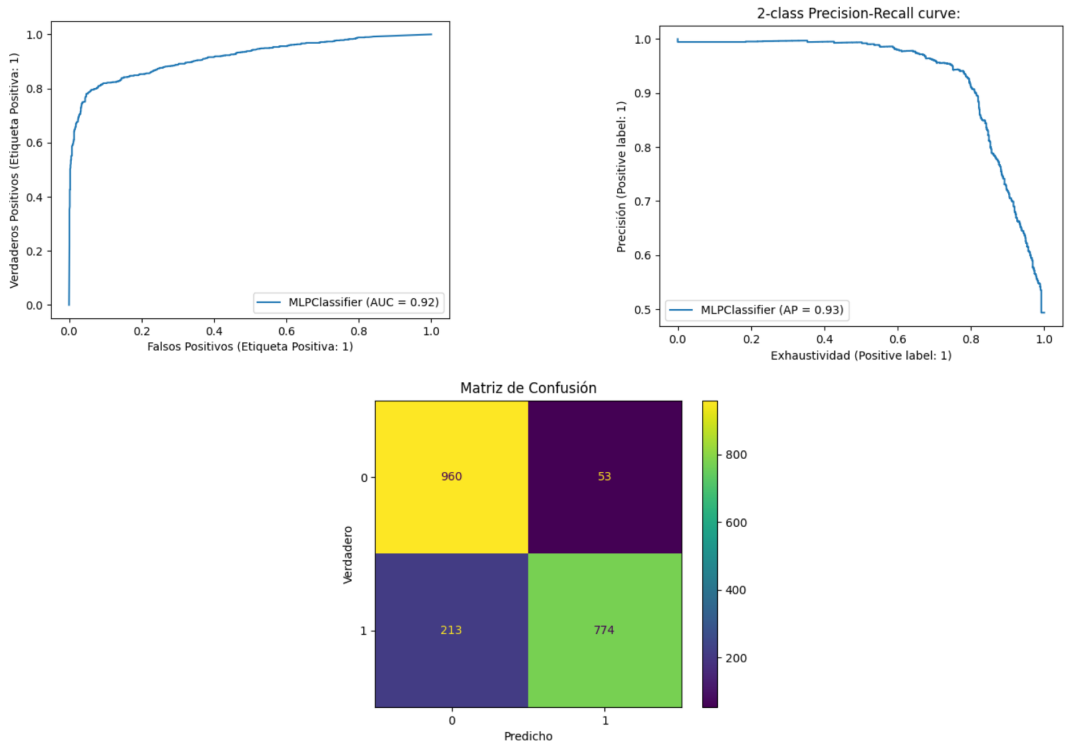


Figura 37: Matriz de confusión y curvas ROC y AUC para las pre-condiciones y condiciones del paciente

Precisión	81.84 %
Exactitud	86.7 %
Exhaustividad	94.77 %
Valor-F	87.83 %

Tabla 20: Resultado Final para las pre-condiciones usando el perceptrón

Capas Ocultas	(2,6)
Función de Activación	Logística
Tasa de Aprendizaje	0.2025
β	1.979

Tabla 21: Valores Finales del Perceptrón Multicapa para las pre-condiciones y datos de internación

6.5. Conclusiones

Se pudo concluir del análisis, la posibilidad de predecir con un grado de efectividad de arriba del 80 % la relación entre las pre-condiciones de una persona infectada y la letalidad de la enfermedad. Tiene suficiente evidencia de la posibilidad de encontrar relaciones entre ambos factores.

Se pudo observar en las Tablas 18 y 20 de los resultados del perceptrón, que la exhaustividad, precisión y exactitud aumentan si se toma en cuenta las condiciones médicas del paciente, o sea, si el mismo fue intubado, si estuvo en el hospital internado o si llego a estar en la Unidad de Cuidados Intensivos. Para los resultados de la regresión logística, se puede ver la baja significaría estadística para las variables de asma y enfermedades cardiovasculares.

Se proponen como futuros complementos al trabajo utilizar otros métodos a modo de comparación como

XGBoost. Es preferible usar datos mas globales, para no enfocar apenas en pacientes de México, sino también de otras nacionalidades. Además, se podría complementar el estudio analizando mas pre-condiciones o especificar un poco mas ciertas pre-condiciones, como tipo de diabetes, tipos de enfermedades de inmunosupresión entre otros.
Finalmente incentivamos el desarrollo de una aplicación para celulares inteligentes, tabletas digitales como también una aplicación web para utilizar estos algoritmos para predecir las consecuencias de la infección. De esta manera, para casos futuros, se podría enfocar en los pacientes con menores probabilidades de sobrevivir a la hora de decidir tratamientos, camas hospitalarias o intubaciones.

7. Análisis de mortalidad en base a la dieta de los países

A la hora de analizar los efectos del COVID-19 en el mundo, lo primero en lo que nos enfocamos suele ser el modelo de cuarentena aplicado por los distintos países. Esto nos hizo pensar si existirán otros factores que afecten la transmisión o letalidad del virus.

Dado que la dieta de una persona puede ayudar a la hora de combatir una enfermedad o de fortalecer su sistema inmunológico, nos hizo pensar si la misma podría ser un factor de importancia a la hora de combatir el COVID-19. Por lo tanto, decidimos analizar los hábitos alimenticios de los distintos países, y otros factores que parcialmente dependen de dichos hábitos, como lo es la obesidad o desnutrición, y contrastarlos con la tasa de mortalidad, en búsqueda de encontrar una relación entre ellos.

7.1. Conjunto de datos utilizados

El conjunto [12] incluye el porcentaje consumido de los distintos tipos de comida por cada país, a esto también se le añadió información sobre los casos de COVID-19 en cada país, con el fin de poder obtener una relación entre la enfermedad y la dieta del país. Además, se incluye el porcentaje de obesidad y desnutrición por cada país.

Para comenzar, veremos las columnas del conjunto de datos relacionadas con la dieta de cada país en la Tabla 22, el mismo conjunto ofrece una explicación sobre que indica cada una.

Categoría	Elementos
Bebidas alcohólicas	Alcohol, no alimentos; Cerveza; Bebidas Alcohólicas; Bebidas Fermentadas; Vino
Grasas animales	Mantequilla, Ghee; Crema; Grasas animales crudas; Pescado, aceite corporal; Pescado, aceite de hígado
Productos animales	Animales acuáticos, otros; Plantas acuáticas; Carne de vacuno; Mantequilla, Ghee; Cefalópodos; Crema; Crustáceos; Pescado demersal; Huevos; Grasas animales crudas; Pescado, aceite corporal; Pescado, aceite de hígado; Pescado de agua dulce; Peces marinos, otros; Carne, mamíferos acuáticos; Carne, Otros; Leche - Excluida la mantequilla; Moluscos, otros; Carne de cordero y cabra; Despojos comestibles; Peces pelágicos; Carne de cerdo; Carne de ave
Productos acuáticos	Animales acuáticos, otros; Plantas acuáticas; Carne, Mamíferos Acuáticos
Cereales	Cebada y productos; Cereales, otros; Maíz y productos; Mijo y productos; Avena; Arroz (Equivalente Molido); Centeno y productos; Sorgo y productos; Trigo y productos
Huevos	Huevos
Pescado y mariscos	Cefalópodos; Crustáceos; Pescado demersal; Pescado de agua dulce; Peces marinos, otros; Moluscos, otros; Peces pelágicos
Frutas	Manzanas y productos; Plátanos; Cítricos, otros; Fechas; Frutas, otras; Pomelo y productos; Uvas y productos (excepto vino); Limones, limas y productos; Naranjas, mandarinas; Piñas y productos; Plátanos
Carne	Carne de vacuno; Carne, Otros; Carne de Cordero y Cabra; Carne de cerdo; Carne de ave
Leche	Leche
Otros	Comida infantil; Otros
Despojos	Despojos comestibles

Cultivos oleaginosos	Cocos - Incl Copra; Semilla de algodón Cacahuetes (sin cáscara Eq); Otros cultivos oleaginosos; Aceitunas (incluidas las conservas); Semillas de palma; Colza y semillas de mostaza; Semilla de sésamo; Habas de soja; Semilla de girasol
Legumbres	Frijoles; Guisantes; Legumbres, Otros y productos
Especias	Clavos de olor; Pimienta; Pimiento morrón; Especias, Otros
Raíces almidonadas	Yuca y productos; Patatas y productos; Raíces, Otras; Patatas dulces; Batatas
Estimulantes	Granos y productos de cacao; Café y productos; Té (incluido mate)
Azúcar y edulcorantes	Miel; Azúcar (equivalente crudo); Azúcar no centrífugo; Edulcorantes, Otros
Cultivos de azúcar	Remolacha azucarera; Caña de azúcar
Nueces de árbol	Nueces y productos
Aceites vegetales	Aceite de coco; Aceite de algodón; Aceite de cacahuete; Aceite de germen de maíz; Otros aceites de cultivos oleaginosos; Aceite de oliva; Aceite de palma; Aceite de semilla de palma; Aceite de colza y mostaza; Aceite de salvado de arroz; Aceite de semilla de sésamo; Aceite de soja; Aceite de semilla de girasol
Vegetales	Cebollas; Tomates y productos; Verduras, Otro
Productos vegetales	Alcohol, no alimentario; Manzanas y productos; Plátanos; Cebada y productos; Frijoles; Cerveza; Bebidas Alcohólicas; Bebidas Fermentadas; Yuca y productos; Cereales, otros; Cítricos, otros; Clavos de olor; Granos y productos de cacao; Aceite de coco; Cocos - Incl Copra; Café y productos; Semilla de algodón; Aceite de algodón; Fechas; Frutas, otras; Pomelo y productos; Uvas y productos (excepto vino); Aceite de cacahuete; Cacahuetes (sin cáscara Eq); Miel; Comida infantil; Limones, limas y productos; Maíz y productos; Aceite de germen de maíz; Mijo y productos; Diverso; Frutos secos y productos; Avena; Otros aceites de cultivos oleaginosos; Otros cultivos oleaginosos; Aceite de oliva; Aceitunas (incluidas las conservas); Cebollas; Naranjas, mandarinas; Semillas de palma; Aceite de palma; Aceite de semilla de palma; Guisantes; Pimienta; Pimiento morrón; Piñas y productos; Plátanos; Patatas y productos; Legumbres, Otros y productos; Aceite de colza y mostaza; Colza y semillas de mostaza; Arroz (Equivalente Molido); Aceite de salvado de arroz; Raíces, Otras; Centeno y productos; Semilla de sésamo; Aceite de semilla de sésamo; Sorgo y productos; Aceite de soja; Habas de soja; Especias, Otras; Azúcar (equivalente crudo); Remolacha azucarera; Caña de azúcar; Azúcar no centrífugo; Semilla de girasol; Aceite de semilla de girasol; Patatas dulces; Otros edulcorantes; Té (incluido el mate); Tomates y productos; Verduras, otras; Trigo y productos; Vino; Batatas

Tabla 22: Columnas relacionadas a las comidas

Recordemos que cada valor las columnas en la Tabla 22 corresponde al porcentaje de comida que se consume en el país.

Además de las columnas mencionadas anteriormente, el conjunto de datos nos proporciona las siguientes columnas.

País	Nombre del país
Obesidad	Porcentaje de obesidad
Desnutrición	Porcentaje de desnutrición
Confirmados	Porcentaje de la población con casos confirmados
Muertes	Porcentaje de la población que murió a causa del COVID-19
Recuperados	Porcentaje de la población que se infectó y se recuperó
Activos	Porcentaje de la población con casos activos de COVID-19

Tabla 23: Columnas no relacionadas a las comidas

En la Tabla 23 podemos ver columnas como "Obesidad" y "Desnutrición" que no necesariamente tienen que ver con la dieta de cada país, pero de igual forma utilizamos estos datos para intentar encontrar una relación entre ellos y la tasa de mortalidad. Por último, tenemos la cantidad de casos confirmados y la cantidad de muertos, lo cual utilizaremos para calcular la tasa de mortalidad del país.

7.2. Aplicación del método Máquinas de Vectores de Soporte

La idea al utilizar este método fue encontrar una relación entre el tipo de dieta y la tasa de mortalidad, para esto, creamos distintas clases de mortalidad.

Primero, calculamos la tasa de mortalidad de cada país a partir de los datos que ya nos ofrece el conjunto. A continuación, clasificamos a los países en tres grupos según el porcentual en que se encuentra su tasa de mortalidad. Tomamos tres percentiles para clasificar las clases, por lo tanto, nos quedamos con tres clases, baja, media y alta. Recordemos para entender los resultados, que un número mas alto de clase indica una mayor tasa de mortalidad.

Animal	Tipos de alimento que contienen o son derivados de animales.
Vegetal	Solo vegetales o derivados.
Salud	Consideramos desnutrición y obesidad.
Todas	Se tienen en cuenta todas las columnas descriptas en la tabla 22 junto con obesidad y desnutrición.

Tabla 24: Categorías definidas para entrenar los distintos modelos.

Se entrenó a múltiples máquinas de vectores de soporte, con distintas variables cada una, los tipos de variables que usamos en cada máquina se pueden observar en la Tabla 24. Además, por cada categoría, se entrenó a una máquina con función de núcleo lineal, otra utilizando la función polinómica y otra con la función radial.

Los valores nulos se reemplazaron por la media del conjunto. Para el caso de desnutrición, cuando es menor que 2.5 %, no indica la cantidad exacta, solo indica que es menor o igual a dicha cantidad, por eso, en esos casos, reemplazamos el valor por 2.5 %.

Se utilizó un 70 % del conjunto para entrenamiento y un 30 % para testeo.

7.2.1. Resultados

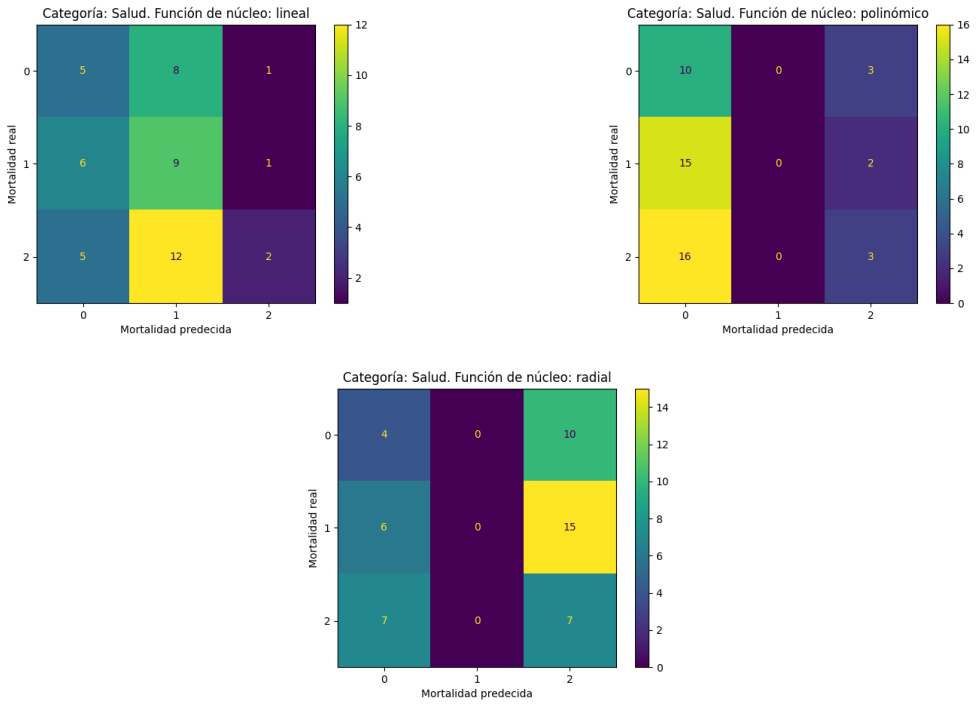


Figura 38: Matriz de confusión para la categoría "Salud"

Función de núcleo	Precisión
Lineal	32.65 %
Polinómica	23.53 %
Radial	22.45 %

Tabla 25: Resultados para la categoría "Salud"

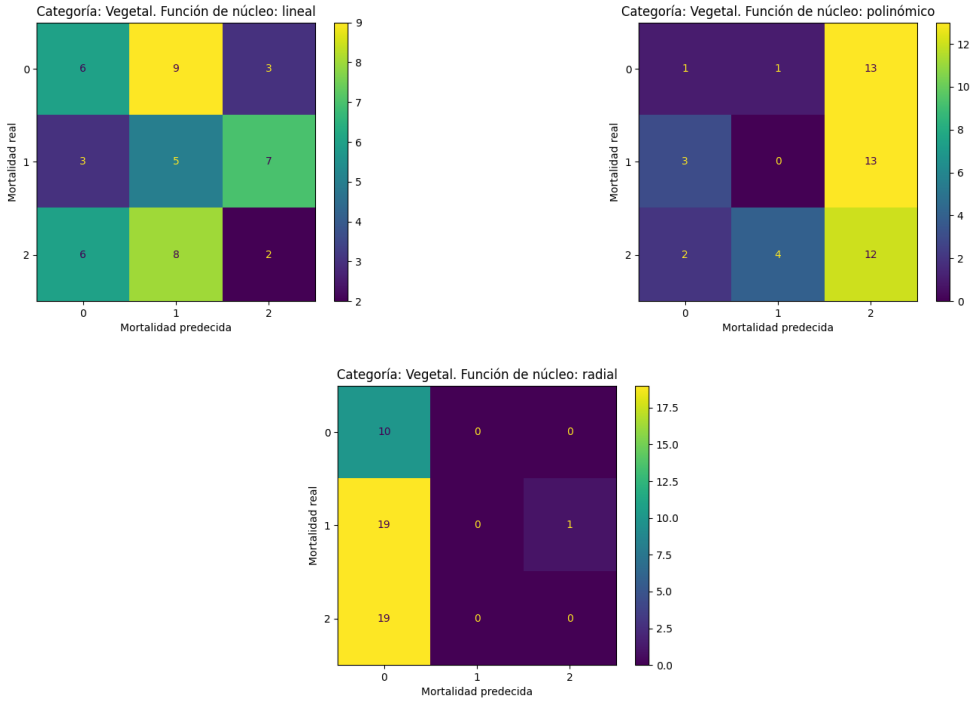


Figura 39: Matriz de confusión para la categoría “Vegetal”

Función de núcleo	Precisión
Lineal	26.53 %
Polinómica	26.55 %
Radial	20.41 %

Tabla 26: Resultados para la categoría “Vegetal”

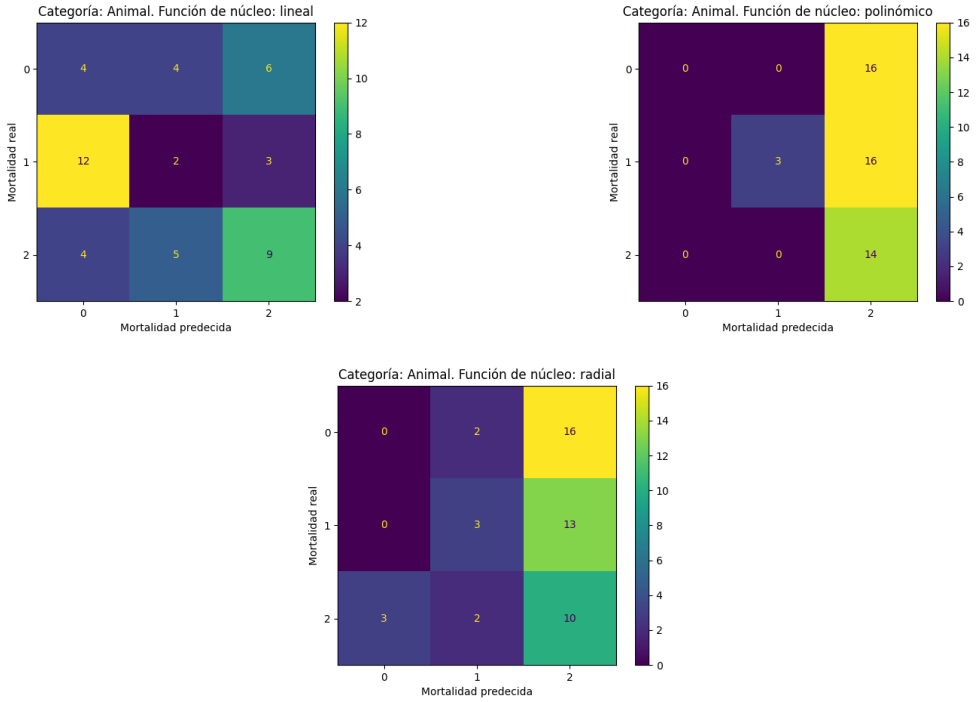


Figura 40: Matriz de confusión para la categoría “Animal”

Función de núcleo	Precisión
Lineal	30.61 %
Polinómica	34.69 %
Radial	26.53 %

Tabla 27: Resultados para la categoría “Animal”

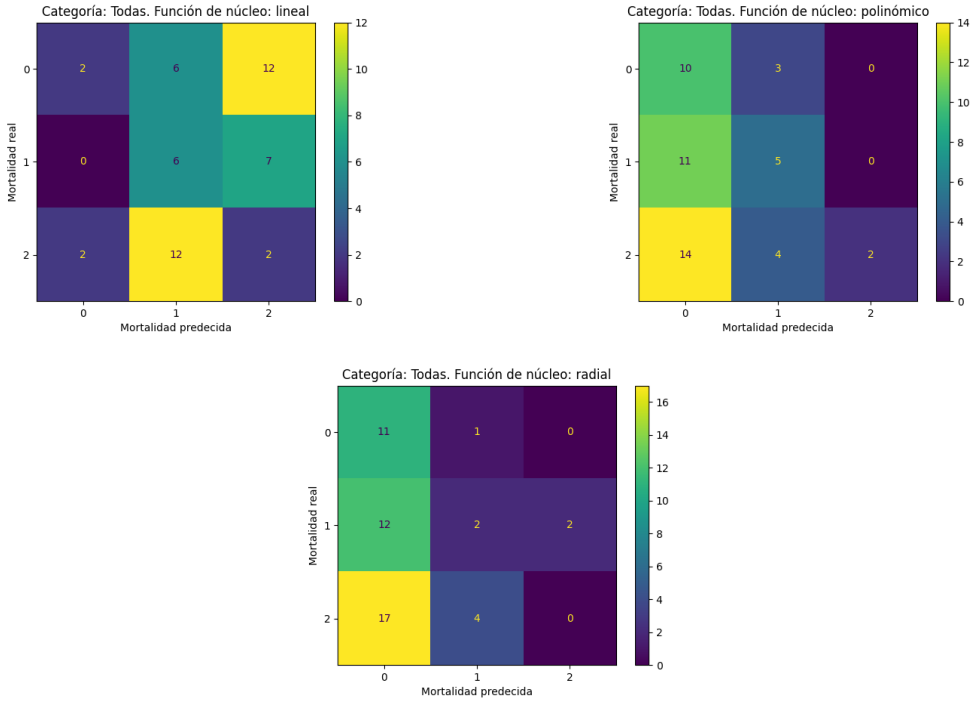


Figura 41: Matriz de confusión para la categoría "Todas"

Función de núcleo	Precisión
Lineal	20.41 %
Polinómica	34.69 %
Radial	26.53 %

Tabla 28: Resultados para la categoría "Todas"

En las Figuras 38, 39, 40 y 41 se visualizan las matrices de confusión de los modelos entrenados según las distintas categorías. La tasa de mortalidad esta dividida en tres percentiles, 0, 1 y 2, donde un mayor número indica una mayor tasa de mortalidad.

En las tablas 25, 26, 27 y 28 podemos ver la precisión de los modelos para los distintos tipos de funciones de núcleo y su categoría especifica.
Las categorías están definidas en la Tabla 24.

7.3. Aplicación del método K vecinos más cercanos con Clasificador bayesiano ingenuo (KNN-NB)

Para este método, se analizó la letalidad de los países en base a dieta, de la misma manera que fue analizado en los vectores de soporte. Otro análisis que se realizó es el efecto de a obesidad y desnutrición en cada país y su relación con la letalidad.

Para ambos análisis se buscaron para cada país los 10 vecinos más próximos. Además, se usó un λ fijo de 0.01. Se dividió la letalidad en 4 grupos, uno por cuartil, así era posible hacer la clasificación. Se utilizó así como en el algoritmo anterior, un 70 % del conjunto para entrenamiento y un 30 % para testeo.

7.3.1. Resultados

Las Figuras 42 y 43 muestran la matriz de confusión encontrada para las predicciones en base a la dieta y sobre el efecto de problemas relacionados a la alimentación. Las Tablas 29 y 32 muestran los resultados de precisión, exactitud y exhaustividad por cuartil de letalidad. Las Tablas 31 y 34 presentan los resultados finales para todos los cuartiles.

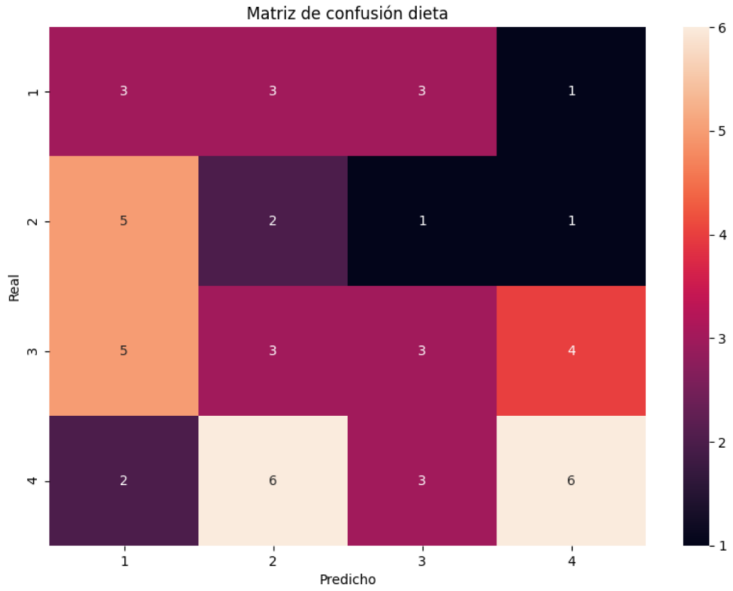


Figura 42: Matriz de Confusión para la dieta de los países

Precisión	20.00 %
Exactitud	62.75 %
Exhaustividad	30.00 %
Valor-F	24.00 %

(a) Primer Cuartil

Precisión	14.29 %
Exactitud	62.75 %
Exhaustividad	22.22 %
Valor-F	17.39 %

(b) Segundo Cuartil

Precisión	20.00 %
Exactitud	62.75 %
Exhaustividad	30.00 %
Valor-F	24.00 %

(c) Tercer Cuartil

Precisión	50.00 %
Exactitud	66.67 %
Exhaustividad	35.29 %
Valor-F	41.38 %

(d) Cuarto Cuartil

Tabla 29: Resultados para las Dietas

Precisión	27.45 %
Exactitud	63.73 %
Exhaustividad	27.45 %
Valor-F	27.45 %

Tabla 31: Resultado Final para la Dieta

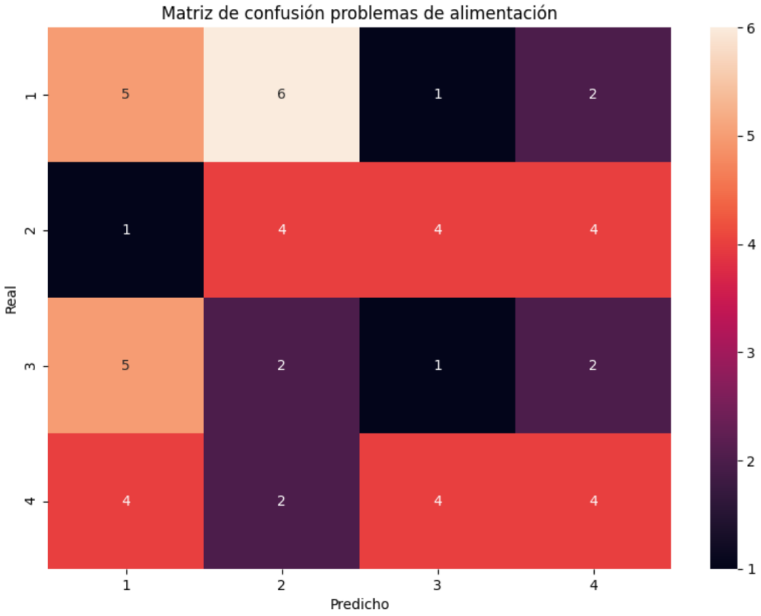


Figura 43: Matriz de Confusión para los problemas de alimentación de los países

Precisión	33.33 %
Exactitud	62.75 %
Exhaustividad	35.71 %
Valor-F	34.49 %

(a) Primer Cuartil

Precisión	28.57 %
Exactitud	62.75 %
Exhaustividad	30.77 %
Valor-F	29.63 %

(b) Segundo Cuartil

Precisión	10.00 %
Exactitud	64.71 %
Exhaustividad	10.00 %
Valor-F	10.00 %

(c) Tercer Cuartil

Precisión	33.33 %
Exactitud	64.71 %
Exhaustividad	28.57 %
Valor-F	30.77 %

(d) Cuarto Cuartil

Tabla 32: Resultados para los Problemas de alimentación

Precisión	27.45 %
Exactitud	63.73 %
Exhaustividad	27.45 %
Valor-F	27.45 %

Tabla 34: Resultado Final para los Problemas de alimentación

7.4. Conclusiones

Para finalizar, como se puede observar en los resultados, todos los métodos tienen una precisión muy baja, incluso menor a categorizar a los países en cada percentil de forma aleatoria.

Tomando los resultados de KNN-NB por ejemplo, se puede ver que para ambos análisis, tanto en la tabla 31 como en la tabla 34, la precisión no pasa de 27.45 %, que es un valor muy cercano a elegir de manera aleatoria. Observando ambas las matrices de confusión (Figuras 42 y 43), no hay un padrón con la mayoría de los resultados en la diagonal principal, mostrando la baja tasa de aciertos del modelo.

Esto no quiere decir que la dieta no pueda ser un factor que influya en la tasa de mortalidad, pero quedan afuera del análisis otros factores, como la cantidad de contagiados, cantidad y calidad de hospitales, acceso a la salud, el modelo de cuarentena impuesto, entre otros. Es muy probable que éstos tengan una mayor influencia en la tasa de mortalidad que la dieta, y por lo tanto, para poder analizar la influencia de la dieta deberíamos comparar entre países que tengan factores externos muy similares, lo cual es, cuanto menos, muy difícil de realizar.

8. Conclusiones

Al inicio del proyecto se planteó como objetivo hallar nueva información del virus COVID-19 utilizando conjuntos de datos relacionados a la enfermedad y métodos de aprendizaje automático.

Al finalizar el análisis se encontraron diferentes resultados en base a los datos analizados. Por un lado no se pudo encontrar una relación entre la dieta y la letalidad del virus, pero debido a todas las variables que hay en juego al momento de una pandemia no significa que no haya una relación.

Sin embargo, se obtuvieron buenos resultados en algunos métodos, como a la hora de decidir si una persona tiene el virus en base a sus síntomas. Esto es muy importante para poder detectar el COVID-19, pero a la vez, también es útil en la detección de futuras enfermedades, ya que con un método similar se podría saber si las personas se encuentran enfermas.
Además fue favorable el desempeño de los algoritmos para pronosticar brotes o número de casos con COVID-19. Para esto los datos recolectados por Google, sobre las tendencias en los síntomas buscados sirvieron para proporcionar un estimado de la cantidad de casos en una región. Esto resulta muy importante como alarma preventiva.

Resulta interesante destacar la importancia de la calidad de los datos para poder llegar al objetivo deseado. Se encontraron dificultades como que ciertos conjuntos de datos tenían observaciones erróneas, lo cual resultaba en una peor precisión de los métodos. En otros casos, se necesitó combinar conjuntos de datos para enriquecer el número de variables del conjunto de datos y así desarrollar un modelo acorde a al problema en cuestión.

Para futuros estudios sería interesante estudiar la relación entre el avance del virus y la cuarentena aplicada en los distintos países, con el fin de poder encontrar que modelo funcionó mejor para futuras situaciones similares.

Referencias

[1] Analysis on novel coronavirus (COVID-19) using machine learning methods. *Chaos, Solitons & Fractals*, 139:110050, October 2020.

[2] Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches. *Chaos, Solitons & Fractals*, 138:110137, September 2020.

[3] Combination of four clinical indicators predicts the severe/critical symptom of patients infected COVID-19. *Journal of Clinical Virology*, 128:104431, July 2020.

[4] The covid tracking project data api. `https://covidtracking.com/data/api`, 2021.

[5] Mohammed Al Zobbi, Belal Alsinglawi, Omar Mubin, and Fady Alnajjar. Measurement method for evaluating the lockdown policies during the covid-19 pandemic. *MDPI*, Aug 2020.

[6] Andre Filipe de Moraes Batista, Joao Luiz Miraglia, Thiago Henrique Rizzi Donato, and Alexandre Dias Porto Chiavegatto Filho. COVID-19 diagnosis prediction in emergency care patients: a machine learning approach. *medRxiv*, page 2020.04.04.20052092, April 2020.

[7] Trevor Hastie Gareth James, Daniela Witten and Robert Tibshirani. *An Introduction to Statistical Learning*. Springer, 2013.

[8] National Geographic. La hipótesis del origen del actual Coronavirus. `https://www.nationalgeographicla.com/ciencia/hipotesis-del-origen-del-actual-coronavirus`, October 2020.

[9] John A. Hertz, Anders Krogh, and Richard G. Palmer. *Introduction to the theory of neural computation*. CRC Press, Taylor & Francis Group, 2018.

[10] C. Akdis J. Bousquet, N. Stafford, D. Jarvis J. Sunyer, C. Tedijanto SM. Kissler, M. Chetboun A. Simonnet, F. Cosentino TJ. Guzik, M. Maziarz MD. Miedema, AO. Rinaldi U. Radzikowska, HS. Buttar S. Jain, M. Javadi E. Razmpoosh, and et al. Is diet partly responsible for differences in covid-19 death rates between and within countries? *Clinical and Translational Allergy*, Jan 1970.

[11] Liangxiao Jiang, Harry Zhang, and Jiang Su. Instance cloning local naive bayes. In Balázs Kégl and Guy Lapalme, editors, *Advances in Artificial Intelligence*, pages 280–291, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[12] Maria Ren. Covid-19 healthy diet dataset. 2021.

[13] Dr Jennifer Summers, Dr Hao-Yuan Cheng, Professor Hsien-Ho Lin, Dr Lucy Telfar Barnard, Dr Amanda Kvalsvig, Professor Nick Wilson, and Professor Michael G Baker. Potential lessons from the Taiwan and New Zealand health responses to the COVID-19 pandemic. *The Lancet Regional Health - Western Pacific*, page 100044, October 2020.

[14] O. Wahltinez et al. Covid-19 open-data: curating a fine-grained, global-scale data repository for sars-cov-2. 2020. Work in progress.

[15] WHO. Coronavirus. `https://www.who.int/health-topics/coronavirus`.

[16] WHO. Coronavirus disease (COVID-19) – World Health Organization. `https://www.who.int/emergencies/diseases/novel-coronavirus-2019`.