Proyecto Covid

Análisis de Casos Sars-Cov-2

Retirando Variables

Comenzamos analizando superficialmente la base de datos.
La base contenía muchos valores nulos, en particular había algunas variables que tenían más del 20% de valores faltantes, por lo que las eliminamos de la base ya que no podríamos asegurarnos de hacer una buena imputación.
Las variables que fueron retiradas de los modelos fueron las siguientes:

Variable	Descripción del Diccionario de Datos
FECDEF	Fecha de defunción
INTUBADO	Si el paciente está intubado
ANOSMIA	Si presentó anosmia
DISGEUSIA	Si presentó disgeusia
TXCROBIA	Si recibió tratamiento antimicrobial
ANTIVIRA	Tipo de antiviral que se administró
FECINITXANTIVI	Fecha que se inicio el tratamiento antiviral
CONOCASO	Si tuvo contacto con algun caso de COVID
CONANIMA	Con que animales tuvo contacto
FECVAEST	Fecha de la vacunación
TOMMUE*	Se tomó muestra del paciente
PUERPERIO	Si presenta puerperio
DIASPUERP	Cuántos días de puerperio tiene
UCI	Si se encuentra en una unidad de cuidados intesivos

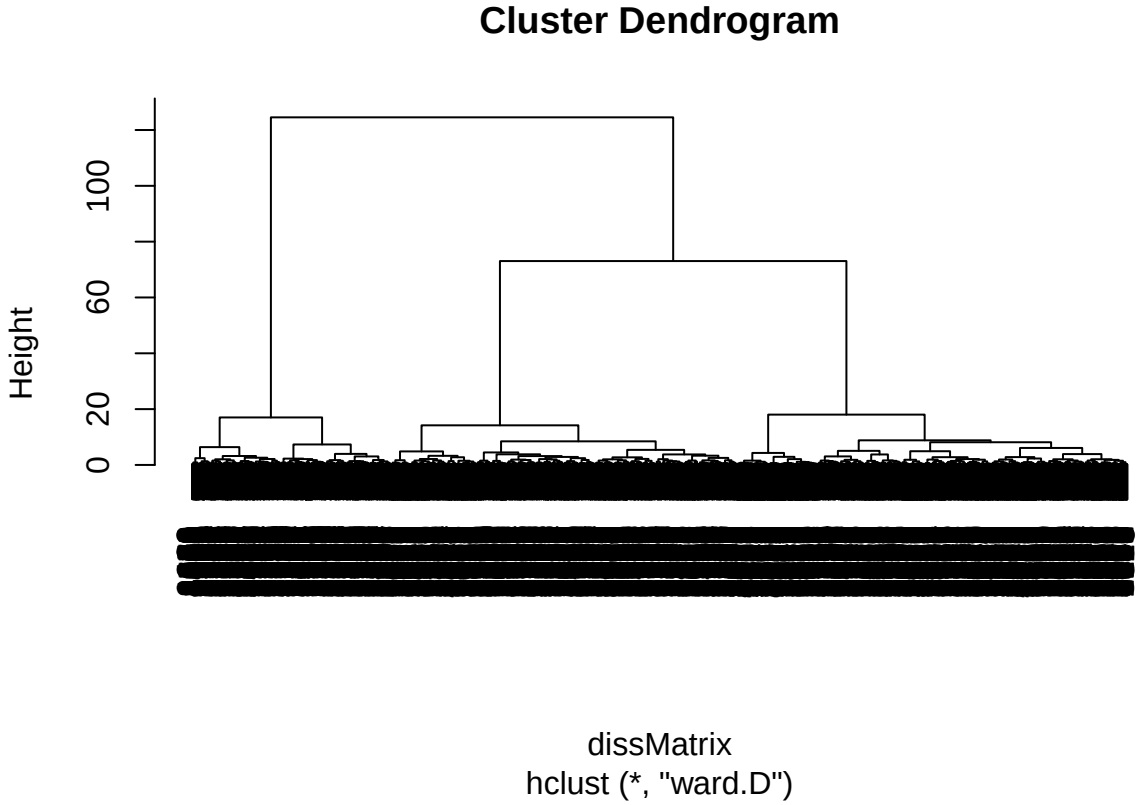
*TOMMUE se eliminó porque era constante “SI” en toda la base.

Imputación de los Valores Faltantes

Después de eliminar estas variables, imputamos los datos faltantes según correspondiera:
Con el método `norm` para las variables continuas,
El método de imputación `logreg` para las binarias,
Y el método `polyreg` para las categóricas.

Reduciendo las Categorías del Tipo de Evolución

Ya que imputamos en su totalidad la base, procedimos a intentar hacer grupos de las observaciones para ver si podíamos reducir a menos categorías la variable `EVOLUCI`.
Intentamos diversos métodos de análisis de conglomerados con diferentes métricas, pero el modelo que mejor conglomeró los datos fue haciendo aglomeramiento jerárquico, métrica `gower` para observaciones y métrica `ward.D` para clusters.
A continuación les mostramos esta conglomeración mencionada:



Aunque el cluster jerárquico nos parece indicar que podríamos cortar en 3 grupos, analizando los grupos, dos de ellos terminaron colapsando en uno mismo, ya que los tipos de evolución más prevalentes de covid en ambos cluster eran los mismos.

Es por esto que elegimos terminar reduciendo a dos niveles la variable evolución, y con esto logramos mejorar muchísimo la interpretación porque podemos pasar a preguntarnos únicamente si los nuevos casos serán casos donde no se complique el Sars-Cov-2, o si sí se complicará.

Modelos Predictivos

Habiendo optado por considerar únicamente dos categorías, si el paciente se complicará o no se complicará, entrenamos diferentes modelos de predicción y los calificamos usando el método `Repeated Holdout` con $B = 50$ (es decir, repetimos 50 veces el entrenamiento de los modelos tomando particiones aleatorias del 80%, y el promedio de estos 50 entrenamientos es la calificación del modelo final).
Mostramos en la siguiente tabla los resultados obtenidos, ordenando los modelos del mejor calificado en la precisión global al peor calificado.

Modelo	Global	Sensibilidad	Especificidad
Random Forest	0.9231431	0.9320920	0.8627907
LDA	0.9194022	0.9233452	0.8928094
Regresión: Probit	0.9178744	0.9482259	0.7131783
Regresión: Logit	0.9171780	0.9480760	0.7087968
Naive Classifier	0.9029029	0.9045977	0.8914729

Como podemos apreciar, el modelo ganador en precisión global es el Random Forest. Sin embargo es de muy buena información saber que tenemos otros modelos para respaldar futuras predicciones, en especial el LDA,

que es el mejor calificado en cuanto a especificidad. Es decir, que para reducir la cantidad de falsos negativos es un muy buen modelo y no está de más tenerlo en consideración.
*El nivel de referencia de los modelos es "No se complica".

Variables Más Influyentes

Afortunadamente tenemos modelos predictores bastante bastante buenos, con más del 90% de precisión global todos, sólo nos falta analizar qué variables son las más influyentes según los modelos.
Para esto mostraremos dos tablas. Primero una tabla listando qué variables fueron las más decisivas para saber si el paciente se complicará, y después una para las variables más decisivas para saber si el paciente no se complicará.

Modelo	Variables Más Influyentes (el paciente se complicará)
Random Forest	Tipo de Servicio de Hospital al que Ingresó, Tipo de Paciente (Hospitalizado/Ambulatorio), Diagnósito Probable, Diagnóstico Clínico, Entidad Federativa, Entidad Federativa de la Unidad Médica, Ocupación, Edad
LDA	Ingreso a UCI, Hospitalizado, Urgencias Adultos, Sector IMSS-Oportunidades, Diagnóstico Probable de IRAG, Diagnostico Clinico Afirmativo, Resultado Prueba Covid
Regresión: Logit	Tipo de Servicio de Hospital al que Ingreso, Sector Medico, Hospitalizado, Resultado Definitivo de la Prueba Covid, Clave Entidad Federativa en c(7, 11, 5, 27, 15, 26, 32, 21, 10, 28, 19), Ocupación, Diagnóstico Probable IRAG)

Es decir, parece que todos los modelos coinciden en la mayoría de cosas tales como:

- El tipo de servicio de hospital al que ingresó (urgencias, neumología, ... consultar el catálogo de SERINGRE para más información)
- Si fue hospitalizado
- Si fue diagnósticado probablemente con IRAG (Infección Respiratoria Aguda Grave)
- El sector médico donde se encuentra el paciente (consultar catálogo de SECTOR)
- Ocupación
- Resultado de la prueba covid

Incluso encontramos algunos detalles raros para el caso del random forest o la regresión logística, como que algunas entidades federativas también juegan un papel importante en el resultado final según estos modelos.

Modelo	Variables Más Influyentes (el paciente no se complicará)
LDA	Servicio de Hospital al que ingresó es UCIN, Resultado Definitivo de la Prueba Covid es Corona 229E o Influenza AH1N1 PMD, Ingresó a Urgencias-Cirugía
Regresión: Logit	Servicio de Hospital al que ingresó es UCIN, Ocupación Dentista, Ocupación Laboratorista, Resultado Definitivo de la Prueba Covid es Corona 229E o Influenza AH1N1 PMD, El Sector Municipal, Ingresó a Urgencias-Cirugía, Ocupación Enfermera

Nuevamente los modelos coinciden en las mismas variables, salvo la regresión logística que tiene más variables en comparación.

Entonces, las variables que nos pueden servir para saber si el paciente no se complicará, lo podemos reducir a:

*Si el paciente ingresó a UCIN

*Si su resultado definitivo de la prueba Covid es Corona 229E o Influenza

*Si el paciente ingresó a urgencias para cirugía

Las variables que no esperaríamos de primera entrada que fueran tan decisivas para saber si un paciente no se complicará, son las variables que no aparecen en el análisis de discriminante pero sí en la regresión logística. Como por ejemplo, enfermeras y dentistas tienen mayores probabilidades de no complicarse según los modelos, al igual que si el paciente es de un sector municipal.

Conclusiones Finales

Gracias a estos modelos, podemos hacer uso de ellos para futuros pacientes y saber con buena probabilidad cómo avanzará su caso.
Por otra parte, desafortunadamente las variables con mayor poder predictivo de estos modelos no son fáciles de cambiar si es que queremos mejorar las probabilidades de que los pacientes no se compliquen, sólo podemos predecir dado lo que ya tenga el paciente, tales como sus síntomas, su ocupación, el resultado de su prueba Sars-Cov-2, etc. mencionadas en el punto anterior.
Quizás lo único que podríamos hacer para evitar complicar a un paciente si es que se quiere cuidar a tiempo, sería intentar ingresar a una UCIN (si fuese posible) o intentar entrar a urgencias de cirugía.