UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

# MODELOS DE PREDICCIÓN PARA LA EVOLUCIÓN DE PACIENTES COVID-19

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INDUSTRIAL

JOAQUÍN ANDRÉS SEPÚLVEDA RODRÍGUEZ

PROFESORA GUÍA:
SUSANA MONDSCHEIN P.

MIEMBROS DE LA COMISIÓN:
RICHARD WEBER H.
CHARLES THRAVES C.

Este trabajo ha sido financiado por el proyecto COVID 0251 de la Agencia Nacional de Investigación y Desarrollo

SANTIAGO DE CHILE
2021

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL
POR: JOAQUÍN ANDRÉS SEPÚLVEDA RODRÍGUEZ
FECHA: 2021
PROF. GUÍA: SUSANA MONDSCHEIN P.

# MODELOS DE PREDICCIÓN PARA LA EVOLUCIÓN DE PACIENTES COVID-19

Este trabajo tiene como objetivo desarrollar modelos de Machine Learning para predecir el estado de salud de los pacientes COVID-19 en seguimiento domiciliario y hospitalizados para así anticiparse y apoyar en la gestión de pacientes, mejorando la calidad del servicio entregado y haciendo mas eficiente la utilización de los recursos.

El trabajo se enmarca dentro del proyecto COVID 0251, financiado por la Agencia Nacional de Investigación y Desarrollo. Este proyecto tiene como objetivo aumentar la efectividad del seguimiento de los pacientes COVID-19 confirmados, probables, sospechosos y contactos mediante una plataforma analítica escalable que integre diferentes fuentes de información, incluyendo la auto declarada por pacientes, y entregue analítica relevante para potenciar la toma de decisiones.

Para esto se proponen dos modelos: el primero consiste en predecir el desenlace de los pacientes hospitalizados confirmados de COVID-19, el cual utilice la información proveniente de la gestión hospitalaria y clasifique a los pacientes en alto riesgo de fallecer o bajo riesgo de fallecer (1). El segundo modelo consiste en predecir la severidad de la enfermedad en los pacientes COVID-19 clasificándolos en pacientes de alto y bajo riesgo, utilizando la información proveniente del seguimiento domiciliario (2).

La metodología utilizada en este trabajo corresponde a CRISP-DM, donde además de detallar el planteamiento y la evaluación de los modelos, también se detalla el procedimiento realizado para la limpieza de la base de datos, generación de atributos y el desarrollo de una aplicación web para desplegar los modelos.

Para el modelo de pacientes hospitalizados se utiliza el algoritmo XGBoost obteniendo un AUC de 0.77, accuracy de 0.74 y sensitivity de 0.84, y las variables mas relevantes son la edad, si el paciente es conectado a ventilación mecánica, el tramo de Fonasa, el tipo de establecimiento y si contaba con alguna enfermedad renal previa. El modelo de pacientes en seguimiento domiciliario se desarrolla utilizando el algoritmo Optimal Tree obteniendo un AUC de 0.87, un accuracy de 0.74 y una sensitivity de 0.87, las variables mas relevantes para este modelo son la presencia del síntoma disnea, el tipo de ingreso, la edad, la presencia del síntoma cefalea, enfermedades renales previas y la presencia del síntoma disgeusia.

Se concluye que los modelos obtienen un buen desempeño a la hora de clasificar a los pacientes según su nivel de riesgo tanto para pacientes en seguimiento como para pacientes hospitalizados. Las variables relevantes conversan con la experiencia internacional y se destaca el potencial que tienen la herramientas de analytics y datascience para apoyar en la toma de decisiones en la industria de la salud tanto como en cualquier otra.

*Para mi familia y amigos.*
***Muchas gracias por todo***

# Agradecimientos

Quiero partir agradeciendo a mi madre Pamela y a mis abuelos Momo y Pepe por estar siempre apoyándome y por todo lo que han hecho por mi durante toda mi vida.

También quiero agradecer a la profesora Susana quien me acompañó de manera directa durante todo el desarrollo de este trabajo, siendo un apoyo importante durante este camino y por tener la disposición de escucharme y ayudarme en todo momento. También agradecer al profesor Richard y a todo el equipo de analítica del proyecto COVID-0251, aprendí mucho y fue muy grato el grupo de trabajo que se formó.

Finalmente, quiero agradecer a mis amigos que han estado conmigo desde el inicio de la carrera y han hecho que esta sea lo mas entretenida posible, sin este grupo no me imagino como habrían sido estos años: Nacho, Diego, Benja, Mauri, Juanpa, Clara, Vale, Coni, Flo, Arica y Wladi. ¡Muchas gracias a todos!

# Tabla de Contenido

# Índice de Tablas

# Índice de Ilustraciones

# Capítulo 1

# Planteamiento del problema

## 1.1. Antecedentes generales

A continuación se desarrolla una presentación documentada de datos y antecedentes para entender el contexto en el cual se realiza el trabajo junto con una descripción de los actores relevantes dentro de este.

### 1.1.1. Contexto del trabajo

El seguimiento de los pacientes COVID-19 que realizan cuarentena domiciliaria es un proceso clave para enfrentar esta pandemia, ya que permite mantener actualizada la situación de salud de los pacientes y el desarrollo de la enfermedad. Este seguimiento debe ser diversificado en canal (distintas formas de llegar al paciente) y escalable de acuerdo con el avance de los contagios, permitiendo dar el tratamiento adecuado a cada paciente y generar información sobre posibles redes y focos de contagio.

La presente investigación se enmarca en el proyecto COVID 0251, financiado por la Agencia Nacional de Investigación y Desarrollo. El proyecto se titula "Sistema integrado de información para el seguimiento domiciliario de pacientes COVID-19 en servicios de salud".

El objetivo general de este proyecto es: "aumentar la efectividad del seguimiento a los pacientes COVID-19 confirmados, los casos sospechosos, los casos probables y sus contactos mediante una plataforma analítica escalable que integre diferentes fuentes de información, incluyendo la autodeclarada por pacientes, y entregue analítica relevante para potenciar la toma de decisiones". Es posible encontrar más información sobre el proyecto en el siguiente enlace: `https://www.sistemaspublicos.cl/gproyecto/covid0251/`

### 1.1.2. COVID-19: el virus, la enfermedad y la pandemia

El brote de la enfermedad por coronavirus (COVID-19), causado por el virus del síndrome respiratorio agudo severo tipo-2 (SARS-CoV-2), fue declarado como una pandemia en marzo de 2020. Las tasas de letalidad (proporción de personas que mueren por una enfermedad entre los afectados por la misma) se estiman entre 1 % y 3 %, afectando principalmente a los adultos mayores y a aquellos con comorbilidades, como hipertensión, diabetes, enfermedades cardiovasculares y cáncer [1]. El periodo de incubación promedio es de 5 días, pero puede ser hasta de 14 días. Muchos pacientes infectados son asintomáticos, sin embargo, debido a que liberan grandes cantidades de virus, son un desafío permanente para contener la propagación de la infección. La vigilancia intensa se torna vital para controlar la propagación del virus, y el aislamiento sigue siendo el medio conocido más efectivo para disminuir la transmisión [1].

A la fecha, abril de 2021, se han confirmado más de 150 millones de casos de COVID-19 a nivel mundial, con un estimado de 3 millones de muertes y más de 83 millones de pacientes recuperados, números que cambian día a día, y que pueden ser monitorizados en tiempo real en el sitio web de la Universidad Johns Hopkins[1].

### 1.1.3. COVID-19 en Chile

Según datos del sitio web de cifras oficiales del gobierno[2], a abril de 2021, más de 1 millón 138 mil casos acumulados y 25 mil muertes por COVID- 19 fueron registradas en Chile. El primer caso fue contactado el 4 de marzo del 2020 y el primer fallecido fue reportado el 23 de marzo de 2020. En las Figuras 1.1 y 1.2 se observa cómo ha ido evolucionando el número de contagios y fallecidos conforme avanza la pandemia respectivamente.


Figura 1.1: Número de nuevos casos confirmados por día, adaptado del sitio web de Cifras Oficiales COVID19.

[1]https://coronavirus.jhu.edu/map.html
[2]https://www.gob.cl/coronavirus/cifrasoficiales/

De las 25 mil muertes por COVID-19 en el país, gran parte se concentra en la Región Metropolitana, la cual supera las 13 mil muertes. Las regiones de Valparaíso y el Bío-Bío siguen a la capital en este registro. Los niveles más altos de mortalidad se alcanzaron durante los meses de mayo, junio y julio de 2020, luego de este periodo la cantidad de fallecidos bajó y se mantuvo estable hasta enero de 2021, desde entonces hasta la fecha la cantidad de fallecidos diarios ha ido aumentando de manera gradual sin aún superar los máximos alcanzados durante los primeros meses [2].

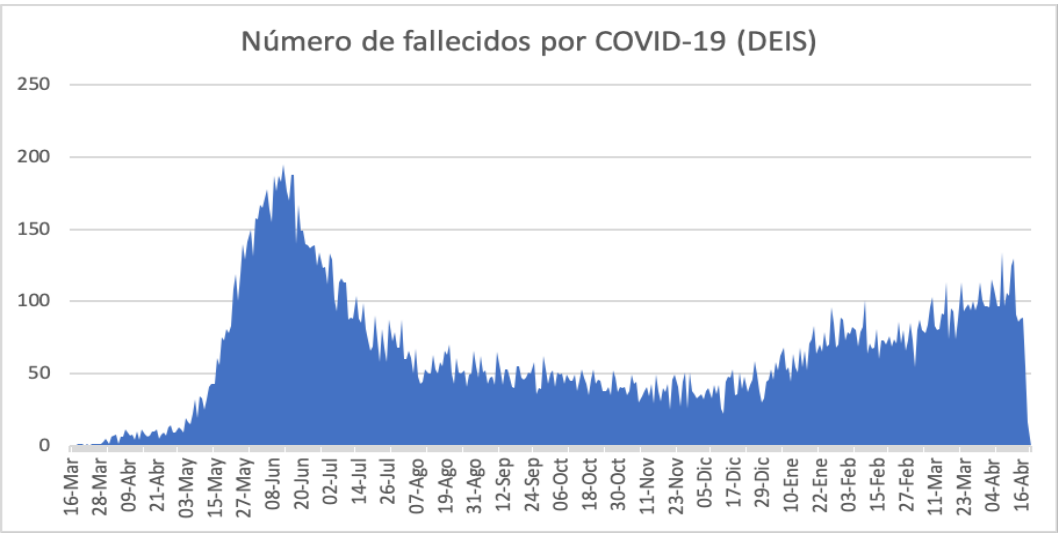



Figura 1.2: Número de nuevos fallecidos por día, adaptado del sitio web de Cifras Oficiales COVID19.

En la Figura 1.3 se puede observar la edad de los fallecidos en Chile, el 35 % de este grupo pertenece al rango etario de los 80 años y más, 28 % se encuentra en el rango de los 70 a 79 años, 21 % en el rango de los 60 a 69 años, 10 % en el rango de los 50 a 59 años, 5 % en el rango de los 30 a 49 y finalmente un 1 % en el rango que incluye a las personas con edades entre los 0 y 29 años de edad [3].

En las Figuras 1.4 y 1.5 se observan los síntomas y comorbilidades presentados en los casos confirmados de COVID-19 a nivel nacional, las cuales corresponden a las cifras oficiales reportadas por el Ministerio de Salud en sus Reportes Diarios de COVID-19 e Informe Epidemiológico COVID-19 [2].

De los síntomas se destaca la presencia de la cefalea (dolor de cabeza), tos (expulsión repentina y con fuerza del aire de los pulmones), mialgia (dolor de muscular) y fiebre (temperatura corporal promedio por sobre los 37 ºC). En el caso de las comorbilidades tanto en los pacientes hospitalizados como los no hospitalizados, se observa una mayor frecuencia en la hipertensión o HTA (presión de la sangre hacia las paredes de la arteria es demasiado alta), diabetes (niveles de glucosa de la sangre muy altos), obesidad (exceso de grasa en el cuerpo) y asma (enfermedad debida a la inflamación crónica de los bronquios) [2].

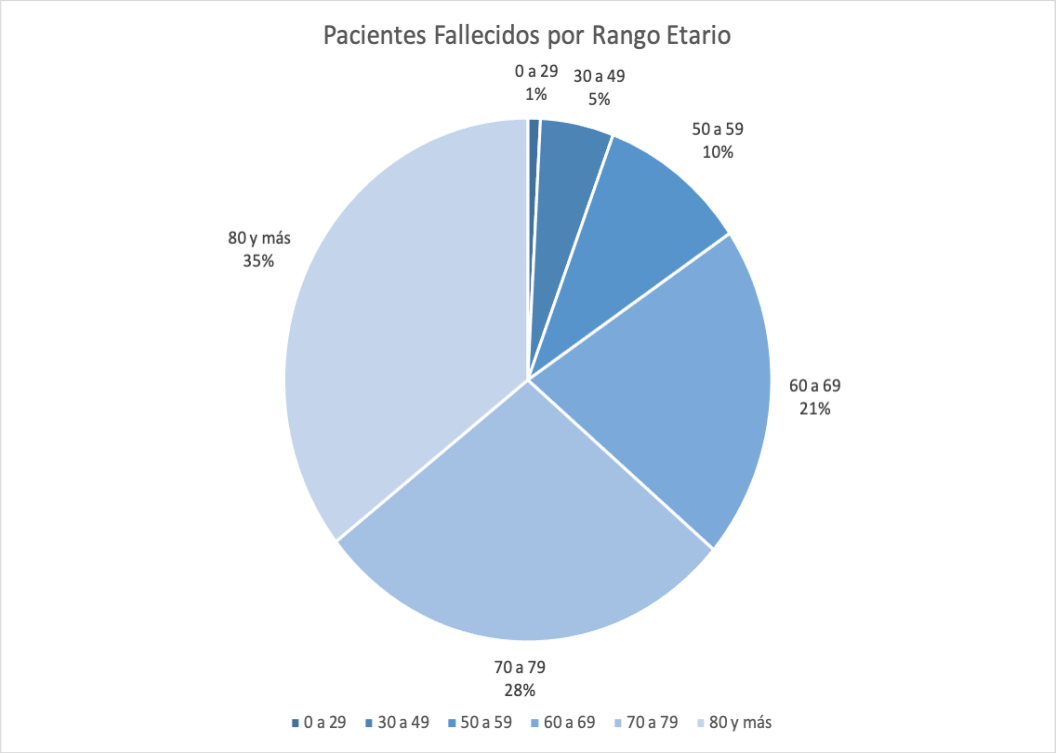



Figura 1.3: Pacientes Fallecidos por Rango Etario, adaptado del sitio web de Cifras Oficiales COVID-19.

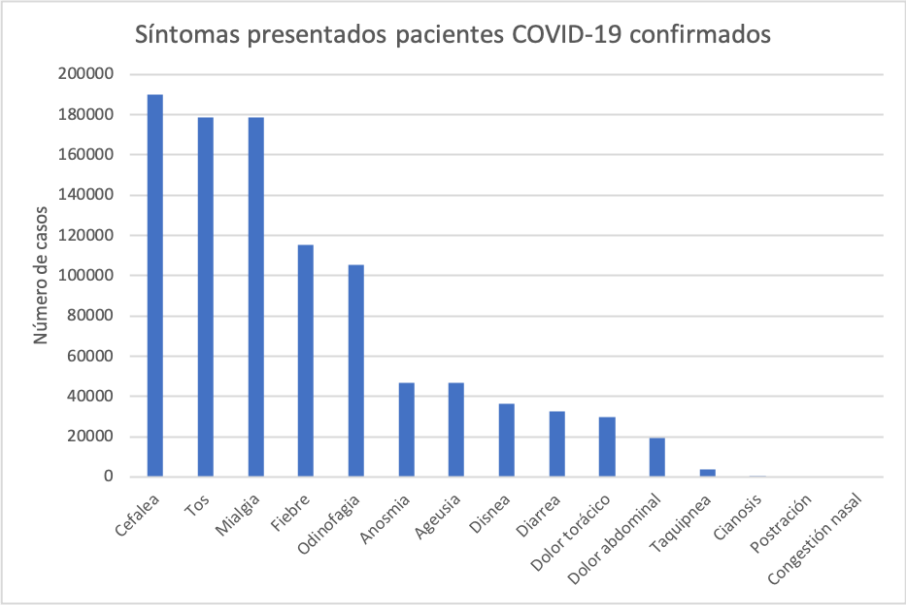
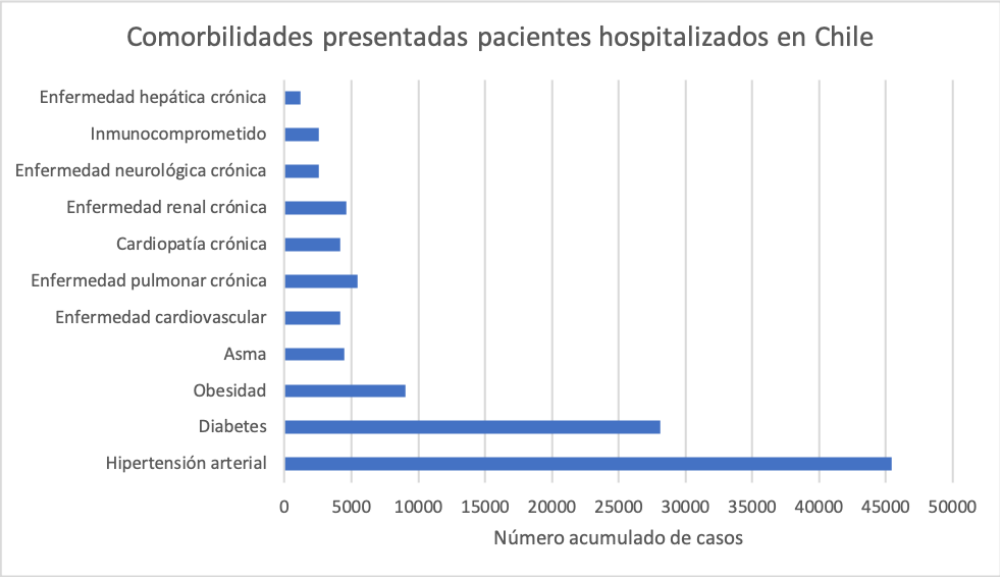
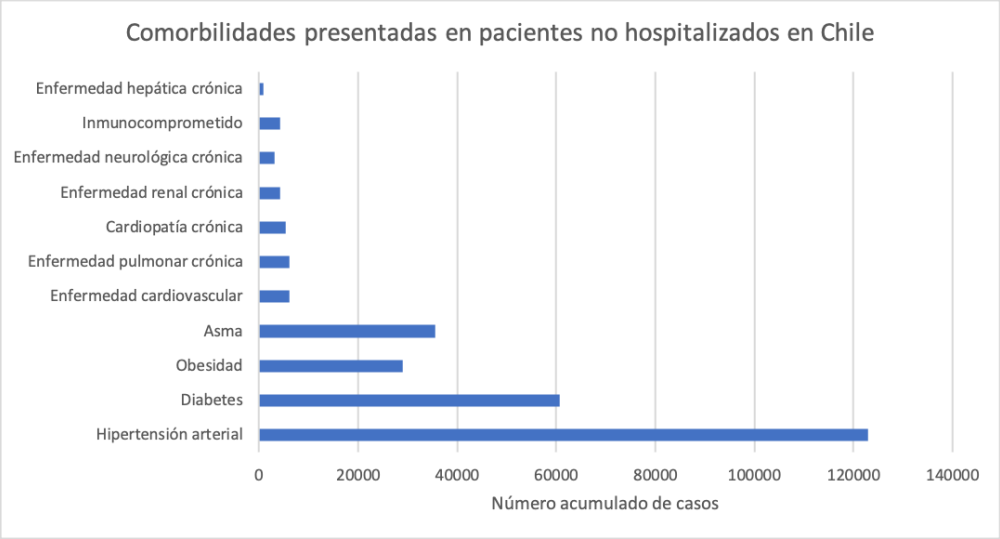



Figura 1.4: Síntomas presentados casos confirmados COVID-19 a nivel nacional, adaptado del sitio web de Cifras Oficiales COVID-19.




(a) Pacientes hospitalizados




(b) Pacientes no hospitalizados

Figura 1.5: Comorbilidades casos confirmados COVID-19 a nivel nacional, adaptado del sitio web de Cifras Oficiales COVID-19.

### 1.1.4. Actores relevantes

#### Servicio de Salud Metropolitano Sur Oriente (SSMSO)

El Servicio de Salud Metropolitano Oriente se inserta en el área Sur Oriente de Santiago, y comprende las comunas de Puente Alto, La Florida, San Ramón, La Granja, La Pintana, San José de Maipo y Pirque; dos de estas comunas, Puente Alto y La Florida, cuentan con las poblaciones más numerosas del país. El total de la población inscrita a la red es de 1.113.477 personas, equivalente a un 15 % del total de la Región Metropolitana.

La Red se divide en tres áreas funcionales o subredes, Cordillera, Santa Rosa y La Florida, las cuales se pueden apreciar en la siguiente ilustración:

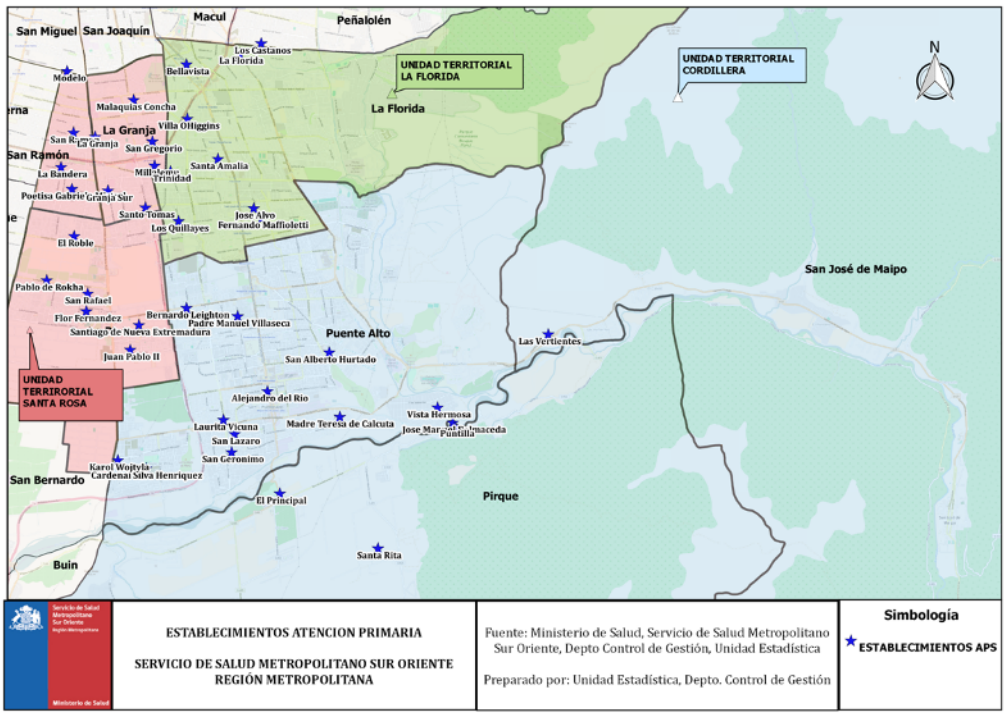



Figura 1.6: Red de establecimientos SSMSO, sitio web Servicio de Salud Sur Oriente.

La misión del SSMSO es llegar a ser una Red de Salud Pública conformada por personas comprometidas con el propósito de satisfacer las necesidades de salud de la población usuaria preferentemente del territorio Metropolitano Sur Oriente, en un contexto de participación, brindando acciones sanitarias con oportunidad, calidad y equidad. La visión declarada por el SSMSO consiste en, al 2023, ser la mejor Red de Atención de Salud Pública del país.

El SSMSO cuenta con cuatro Hospitales (Hospital La Florida Dra. Eloísa Díaz Insunza, Hospital Padre Hurtado, Complejo Hospitalario San José de Maipo, Hospital Provincia Cordillera), el Centro Metropolitano de Sangre y Tejidos, un Centro de Diagnóstico y un Centro de Referencia de Salud. Además, cuenta con otros 83 establecimientos los cuales se dividen entre Centros de Salud Familiar (CESFAM), Consultorio de Salud Mental (COSAM) y Servicios de Atención Primaria de Urgencia (SAPU) [4].

### Unidad de Salud Digital (USD)

La Unidad de Salud Digital (USD) del SSMSO es la encargada de liderar los procesos relacionados a la implementación de la Estrategia del Sistemas de Información de la Red Asistencial (SIDRA) cuyo propósito es impulsar un plan de acción para digitalizar los establecimientos que conforman la red asistencial de salud.

La USD busca que los actores de los procesos asistenciales (usuarios, clínicos, informáticos y gestores) cuenten con información oportuna que les permita tomar mejores decisiones, como también ser un puente de comunicación entre ellos con el fin de fortalecer el trabajo en conjunto.

La misión de la USD consiste en conformar un equipo interdisciplinario con visión estratégica y colaborativa que facilite el uso, acceso y trazabilidad de la información a los distintos actores de la red a través de la implementación e integración de herramientas tecnológicas para el mejoramiento de la calidad de la atención en salud. Por otra parte, la visión de esta unidad es llegar a ser un referente nacional en el área informática clínica y destacar por contribuir al mejoramiento de la calidad de los servicios de salud entregados a los usuarios, a través de la incorporación de tecnologías de información y de los actores del sistema de salud [5].

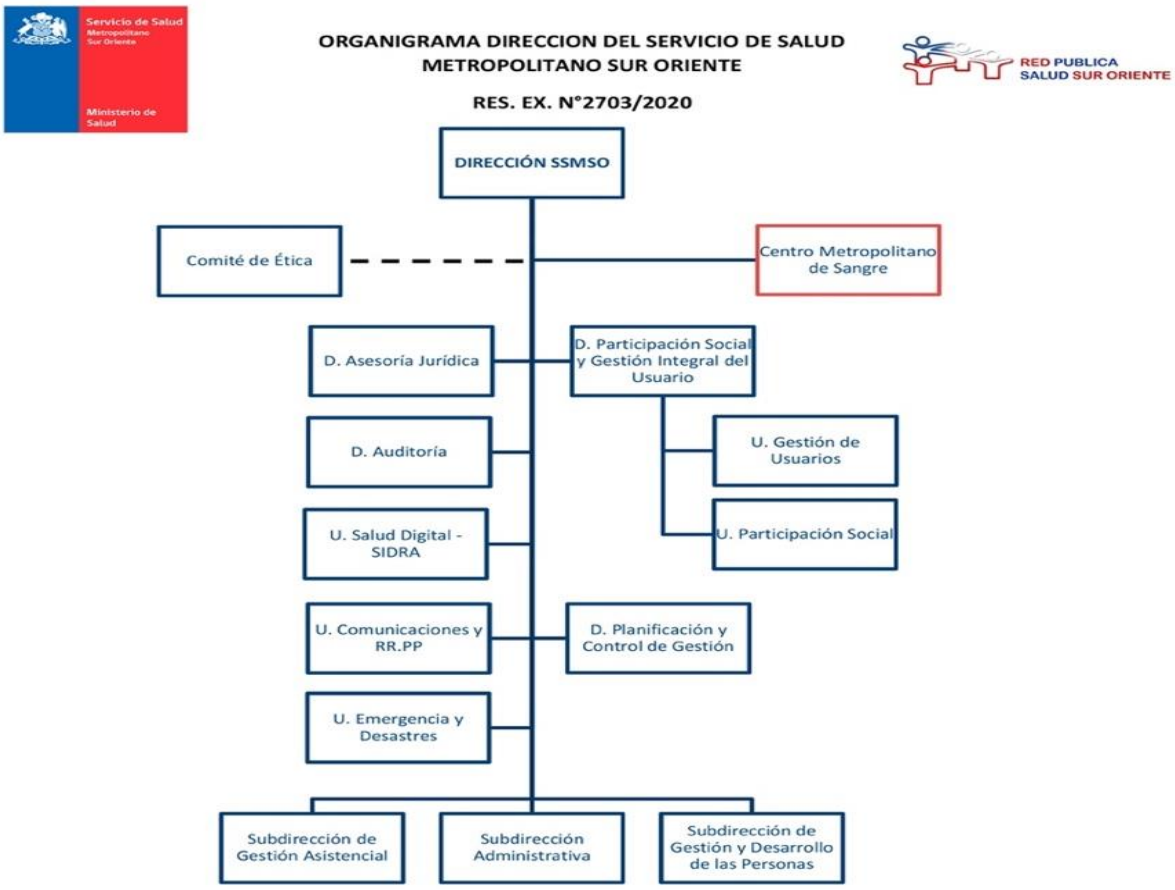



Figura 1.7: Organigrama SSMSO, sitio web Servicio de Salud Sur Oriente.

### Plataforma COVID

La plataforma de seguimiento del plan coronavirus tiene como objetivo ser una herramienta para documentar el seguimiento de las personas COVID-19 en la Red de establecimientos pertenecientes al SSMSO, además de entregar información significativa para la toma de decisiones en salud.

La plataforma fue desarrollada por la USD y la Unidad de Desarrollo de la red y comenzó a operar a fines de mayo del 2020. Mediante esta plataforma, los referentes de epidemiología y de seguimiento domiciliario pueden consultar resultados de exámenes, registrar casos confirmados y sus contactos, actualizar sus signos y síntomas y registrar el egreso o eventual hospitalización, además la plataforma se encuentra integrada con los sistemas de información de Fonasa y los sistemas de información del SSMSO, para así facilitar el ingreso de la información y obtener estadísticas diarias de la situación sanitaria por subred, comunas y establecimientos [6]. En el Apéndice B, se encuentra mas información acerca de como funciona la plataforma y que datos se obtienen a partir de esta.

## 1.2. Problema y justificación

A continuación se realiza una descripción del contexto de la oportunidad o problema que se quiere abordar junto con la justificación y valor agregado de llevarlo a cabo.

### 1.2.1. Área donde se desarrollará el trabajo

El trabajo será desarrollado en el equipo de analítica del proyecto COVID 0251, este equipo se compone de profesores, estudiantes y profesionales del departamento de Ingeniería Civil Industrial de la Universidad de Chile junto con trabajadores del SSMSO (tanto del personal médico, administrativos y analistas). En este equipo se trabajan temas relacionados con el análisis de datos provenientes de los registros de los seguimientos efectuados en la plataforma COVID19 y los registros de los seguimientos a pacientes dentro de los hospitales. Alguno de los temas que se trabajan en el equipo aparte del propuesto en esta tesis son la georreferenciación de pacientes, trayectorias sintomáticas de los pacientes, análisis de riesgo a nivel de manzanas y el desarrollo de una red social para identificar a los contactos estrechos de los pacientes confirmados de coronavirus.

### 1.2.2. Problema identificado y su relevancia

Como contexto, al momento de desarrollar el trabajo el equipo de analítica se encontraba realizando las primeras pruebas del sistema de IVR (Interactive Voice Response) o sistema de respuesta automático y complementar el seguimiento domiciliario de los pacientes contactos, sospechosos, probables y confirmados de Coronavirus. Con esto surge la interrogante de que pacientes son a los cuales se les debe de hacer un seguimiento más exhaustivo (más riesgosos) y que pacientes se pueden dejar en el sistema de respuesta automático, ya que son pacientes de menor riesgo.

Poder hacer esta distinción entre pacientes es relevante puesto que la capacidad del servicio para contactar a los pacientes es limitada (se cuenta con una cantidad fija de horas hombre para realizar los seguimientos) y además conociendo el riesgo de los pacientes con anticipación se podrían llegar a gestionar de mejor manera los recursos del servicio, tanto para el seguimiento de pacientes en casa como para los pacientes que se encuentran hospitalizados. Cabe mencionar que en cada uno de los seguimientos el encargado de llevarlos a cabo clasifica al paciente según los criterios de clasificación de riesgo (Apéndice B.5.), para definir si se trata de pacientes de alto, medio o bajo riesgo. Existen casos en donde los pacientes no cuentan con la información correspondiente a la clasificación de riesgo registrada por el personal del servicio, por lo que surge la necesidad de apoyar en esta tarea para los casos en donde no se tiene información y complementar la clasificación de riesgo provista por el personal del servicio con la de los modelos en los casos donde si se registra la clasificación de riesgo.

Gracias a la información que la USD entrega al equipo de analítica, surge la oportunidad de desarrollar modelos que permitan predecir el desenlace final de los pacientes en seguimiento

y hospitalizados usando las variables registradas una vez que son ingresados a la plataforma COVID19 en el caso de los pacientes en seguimiento y las variables registradas en la admisión hospitalaria para los pacientes hospitalizados. Entre las variables registradas se encuentran las características sociodemográficas de los pacientes (edad, género, comuna), los antecedentes médicos o comorbilidades previas, los síntomas registrados en cada uno de los seguimientos y variables relacionadas al ingreso hospitalario como el tipo de hospital, el tipo de cama de ingreso, etc. Estos modelos buscan clasificar a los pacientes en alto o bajo riesgo con el fin de poder complementar las predicciones con el IVR y poder automatizar gran parte de los seguimientos, liberando tiempo y recursos para los trabajadores del servicio.

### 1.2.3. Justificación del problema

El seguimiento de los pacientes COVID-19 (Casos confirmados, Probables, Sospechosos y Contactos) que realizan cuarentena domiciliaria ha demostrado ser un proceso clave para enfrentar esta pandemia. Pese a su importancia, los servicios de salud cuentan con limitada capacidad para integrar la información recopilada con los sistemas tecnológicos que poseen. Por ejemplo, según cifras del Ministerio de Salud al 4 de octubre del 2020, el 26 % de los fallecidos por COVID-19 murieron sin ser hospitalizados para tratar la enfermedad. En la Región Metropolitana, por ejemplo, de las 9.601 muertes causadas por el virus hasta la fecha, 2.772 (29 %) se produjeron en personas que nunca fueron hospitalizadas[3].

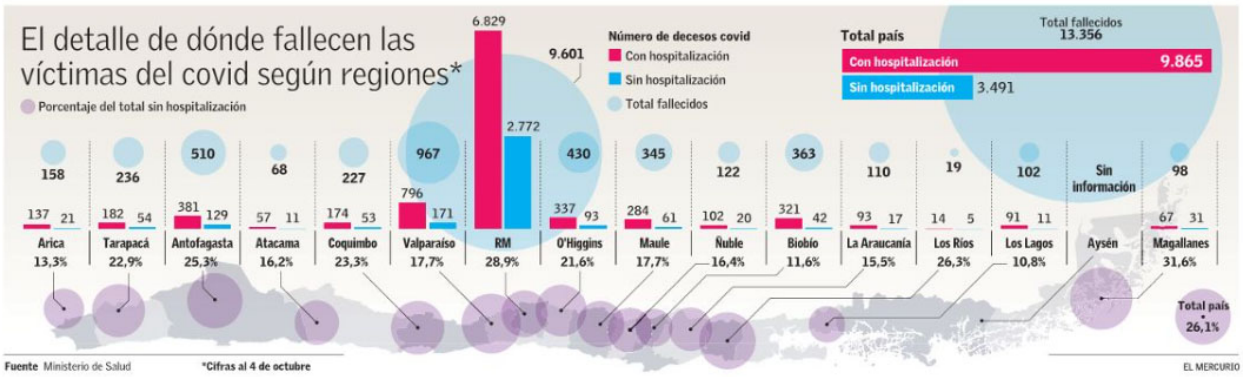



Figura 1.8: Detalle de donde fallecen las víctimas del COVID-19 [3].

Frente a esto, se identifica la oportunidad de poder aprovechar los datos provenientes de distintas fuentes del Servicio de Salud con el objetivo de construir modelos para predecir la evolución de la enfermedad y clasificar a los pacientes según su nivel de riesgo de enfermarse gravemente de coronavirus usando la información de sus comorbilidades, síntomas presentados, características demográficas y el resto de las variables que se recolectan en los sistemas de información del Servicio.

Con los resultados de la investigación se busca apoyar a las distintas unidades del Servicio de Salud del SSMSO en la toma de decisiones y en aumentar su capacidad de respuesta frente a la pandemia.

[3]https://www.clinicasdechile.cl/noticias/uno-de-cada-cuatro-fallecidos-por-covid-19-murio-sin-haber-sido-hospitalizado/

#### Apoyo a la toma de decisiones

En la medida que se implementen los modelos antes mencionados, los resultados aportarían el valor de promover la toma de decisiones con respecto al seguimiento activo de pacientes basados en la mejor evidencia posible (los datos).

Un ejemplo de como los modelos servirán para apoyar la gestión del seguimiento es en categorizar a aquellos pacientes que no cuentan con la categoría de riesgo entregada por el personal médico al momento de ser seguidos.

De los 140.868 ingresos al seguimiento domiciliario, un 45 % (63.391) no cuenta con la categoría de riesgo asignada por el personal. Automatizar esta categorización de riesgo conllevará a que todos los pacientes cuenten con una clasificación de riesgo, la cual puede ser complementada con la opinión del personal de salud posteriormente.

#### Aumentar capacidad de respuesta frente a la pandemia

El alertar a los tomadores de decisiones según la información proveniente de la plataforma ayudaría a identificar tempranamente a pacientes de alto riesgo sin siquiera la necesidad de que estos sean revisados por un médico o personal de la salud. Los resultados de los modelos servirían como una primera recomendación y un apoyo para conocer de antemano a que pacientes habría que seguir de manera más exhaustiva y a que pacientes se les podría asignar menos recursos del sistema puesto que presentan un menor riesgo de verse agravada su situación de salud.

Esto se puede evidenciar tomando los datos de una de las comunas del servicio, donde en promedio realizar un seguimiento para un caso activo demora 15 minutos y para un contacto estrecho 7 minutos. Considerando 10 hospitales de la comuna, en promedio se tienen 117 pacientes activos y 200 contactos estrechos a los que se le debe de realizar el seguimiento por hospital, lo que se traduce en 29 horas hombre para seguir a los casos activos y 25 horas hombres para el seguir a los contactos. En total se tiene la suma de 54 horas hombres diarias (7 trabajadores de la salud) dedicados únicamente a realizar el seguimiento.

### 1.2.4. Valor generado a través de la solución

#### Apoyo al plan de monitoreo del SSMSO

Utilizar los modelos para poder tener un plano general semana a semana del nivel de riesgo de los pacientes que se están siguiendo o que se encuentran en algún centro hospitalario. Este plan de monitoreo se presenta cada semana en el comité COVID del SSMSO, donde se juntan representantes de distintas áreas del servicio para analizar y tomar decisiones con respecto al avance de la pandemia.

### Liberación de recursos para el seguimiento

Como bien se menciono en la justificación del problema, los modelos en conjunto con el IVR liberarán recursos para que el personal médico se enfoque en los pacientes de mayor riesgo de enfermar gravemente producto del coronavirus.

### Identificar características propias de pacientes de mayor riesgo

Si bien el objetivo de este trabajo es predecir si el paciente sera de alto o de bajo riesgo, los resultados de los modelos son interpretables y se pueden conocer las características que hacen que un paciente presente mayor riesgo de enfermar gravemente que otro. También la interpretabilidad de los modelos servirá para verificar si las variables que determinan el riesgo en los pacientes de COVID del estudio conversan con la experiencia internacional y si existen características propias de la población chilena (como por ejemplo, el efecto de variables socioeconómicas).

# Capítulo 2

# Objetivos

A continuación se define el objetivo general, los objetivos específicos y los alcances del trabajo.

## 2.1. Objetivo general

El objeto de este trabajo consiste en estudiar la evolución de pacientes contagiados de coronavirus, específicamente, desarrollar modelos de aprendizaje automático que permitan predecir el estado de salud final de los pacientes en seguimiento y hospitalizados determinando que variables hacen que las personas sean más propensas a enfermarse gravemente producto de la enfermedad. Enfermarse gravemente se refiere a un paciente COVID-19 que necesita de hospitalización intensiva, asistencia mediante ventilación mecánica y a los pacientes que fallecen.

Los resultados de este estudio servirán de insumo para alertar y priorizar pacientes a los cuales habría que realizarle un seguimiento más exhaustivo, permitiendo enfocar los esfuerzos del personal de salud de la red.

## 2.2. Objetivos específicos

1. Consolidar una base de datos adecuada para el proceso de construcción de los modelos.
2. Desarrollar un modelo que permita clasificar pacientes según la probabilidad de enfermarse gravemente a partir de la información recolectada de los primeros seguimientos domiciliarios.
3. Desarrollar un modelo que permita clasificar a las personas según su riesgo de fallecer producto de la enfermedad al momento de ingresar a un establecimiento de salud.
4. Diseñar y desarrollar un prototipo de aplicación web que integre uno de los modelos y pueda ser utilizado por los funcionarios del SSMSO.

## 2.3. Alcances

### 2.3.1. Coronavirus, una historia en desarrollo

Debido a que la pandemia se encuentra en curso y aún no se tiene certeza cómo evolucionará con el tiempo la enfermedad, se hace necesario definir el espacio temporal utilizado para entrenar y validar los distintos modelos. El estudio contempla la información de los pacientes ingresados a la plataforma de seguimiento desde marzo de 2020 hasta enero de 2021, además se consideran los protocolos de salud y las definiciones de casos utilizados durante este periodo.

### 2.3.2. Modelos predictivos, sin capacidad prescriptiva

Los modelos a desarrollar no involucran líneas de acción para optimizar el proceso de seguimiento de pacientes, puesto que no se tienen los conocimientos clínicos ni los datos necesarios para definir que acciones tomar una vez se tengan las predicciones.

### 2.3.3. Integración con los sistemas de salud del SSMSO

El alcance de esta memoria es hasta la implementación de los modelos, entregando resultados teóricos y desarrollando un prototipo de aplicación web que permita que los modelos sean utilizables. No se considera dentro del alcance de la memoria la integración de la aplicación en la plataforma de seguimiento ni en ningún otro sistema de información del servicio.

# Capítulo 3

# Marco conceptual

A continuación, se describe el marco conceptual que tiene el propósito de dar a la investigación un sistema coordinado y coherente de conceptos y proposiciones que permitan abordar el problema.

## 3.1. Machine Learning

Machine Learning o aprendiza automático, es una de las ramas de la Inteligencia Artificial que permite a un sistema aprender de los datos en lugar de aprender mediante la programación explícita. El Machine Learning utiliza una variedad de algoritmos que aprenden iterativamente de los datos para mejorar la descripción y predicción de los resultados.

Un modelo de Machine Learning es la salida de información que se genera cuando se entrena un algoritmo con datos. Después del entrenamiento, al proporcionar un modelo con una entrada, se le dará una salida. Por ejemplo, un algoritmo predictivo, creará un modelo predictivo. Dependiendo de la naturaleza del problema, existen diferentes enfoques basados en el tipo y el volumen de datos. En este caso en particular, los modelos serán supervisados, esto consiste en hacer predicciones a futuro basadas en comportamientos o características que se han visto en los datos ya almacenados. El aprendizaje supervisado permite buscar patrones en datos históricos relacionando todos los campos con un campo especial, llamado campo objetivo. Por ejemplo, los correos electrónicos se etiquetan como “spam” o “legítimo” por parte de los usuarios [22].

## 3.2. Entrenamiento y validación de modelos de Machine Learning

Los modelos de Machine Learning aprenden de los datos con los que los entrenamos. A partir de ellos, intentan encontrar o inferir el patrón que les permita predecir el resultado para un nuevo caso. Pero, para poder calibrar si un modelo funciona, necesitaremos probarlo

con un conjunto de datos diferente. Por ello, en todo proceso de aprendizaje automático, los datos de trabajo se dividen en dos partes: datos de entrenamiento y datos de prueba o test [23].

- Conjunto de entrenamiento (Training): datos/observaciones con las que se entrena el modelo.
- Conjunto de test (Validation): Datos que el modelo no ha visto, sirve para ajustar parámetros y seleccionar el mejor algoritmo.
- Error de entrenamiento: Error que comete el modelo al predecir observaciones que pertenecen al conjunto de entrenamiento.
- Error de test: Error que comete el modelo al predecir observaciones del conjunto de test.

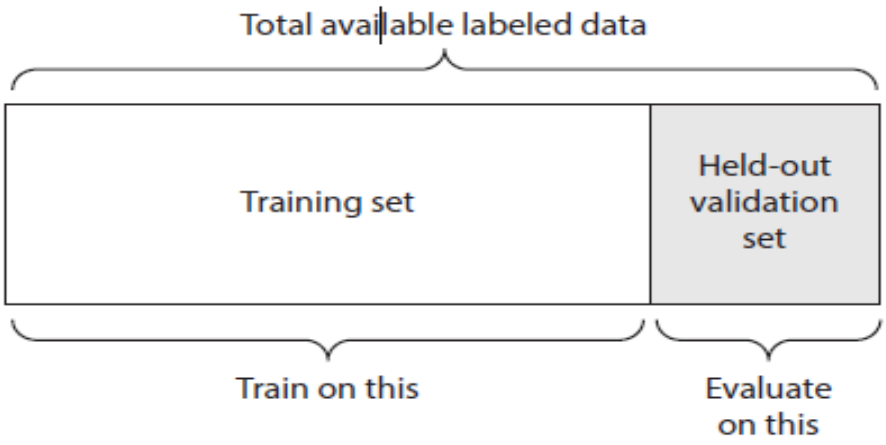



Figura 3.1: Separación de datos en set de entrenamiento y validación. Imagen recuperada del sitio web medium.com

De esta manera tendríamos, un set de datos de entrenamiento y un set de datos de validación. Si bien es la forma mas simple de entrenar y validar modelos, puede surgir un inconveniente cuando disponemos de una cantidad limitada de datos y es que la cantidad de los conjuntos de entrenamiento y validación pueden ser tan pequeños que no conseguiremos un modelo efectivo, para hacer frente a este problema, existen diferentes estrategias de validación, a continuación describimos las estrategias mas utilizadas [24]:

1. **Validación simple :** Repartir aleatoriamente las observaciones disponibles en dos grupos, uno para entrenar y otro para evaluar. Esta estrategia tiene dos problemas:
   - Estimación del error variable dependiendo de que observaciones se incluyan en cada grupo
   - Se pierde poder predictivo, ya que se dispone de menos información para entrenar el modelo (20 %-30 %)

2. **K-fold Cross-Validation :** Consiste en dividir los datos de forma aleatoria en k grupos de aproximadamente el mismo tamaño, k-1 grupos se emplean para entrenar el modelo y uno de los grupos se emplea como validación. Este proceso se repite k veces utilizando un grupo distinto como validación en cada iteración. El proceso genera k estimaciones del error cuyo promedio se emplea como estimación final.

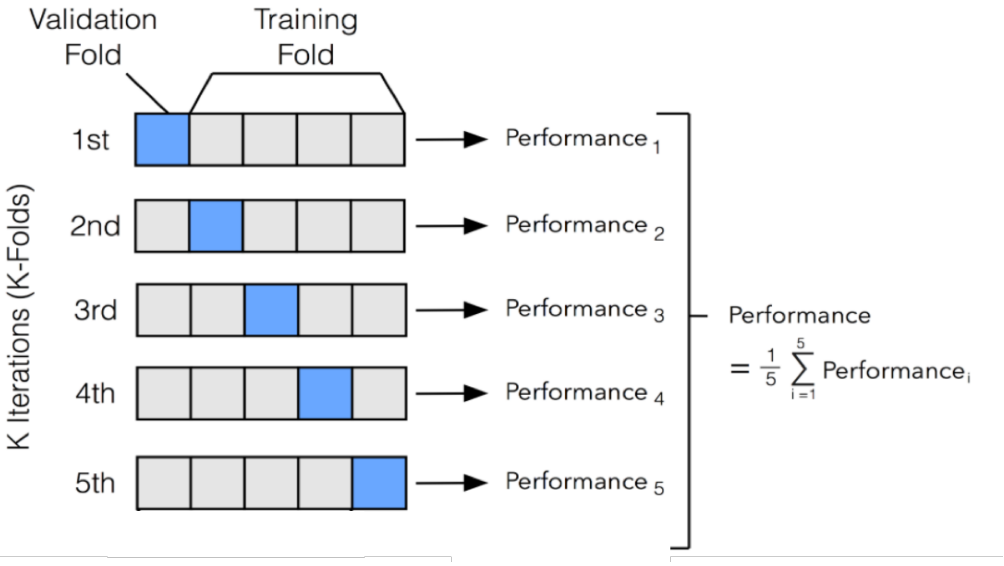



Figura 3.2: Ejemplo K-fold Cross- Validation (K=5). Imagen recuperada del sitio web stackoverflow.com

3. **Repeated K-fold-Cross-Validation:** Consiste en repetir el método K-Fold-Cross-Validation "n"veces.

## 3.3. Técnicas para manejo de datos desbalanceados

Los conjuntos de datos desbalanceados son muy comunes en todas las industrias, en este caso, los pacientes que terminan siendo asintomáticos o aquellos que tienen un ligero pasar de la enfermedad son mas que aquellos pacientes que enferman gravemente y mas aun que los pacientes que terminan falleciendo.

Debido a esto, la implementación de un modelo de clasificación en tales datos desequilibrados probablemente resultará en una precisión predictiva muy baja. Para solucionar esto existen diferentes métodos, el mas sencillo consiste en muestrear aleatoriamente el conjunto de datos de entrenamiento, ya sea, eliminando registros de la clase mayoritaria (random undersampling) o duplicando ejemplos de la clase minoritaria (random oversampling). Al usar undersampling existe el inconveniente de eliminar registros de la clase mayoritaria, lo que puede resultar en la pérdida de representatividad de la muestra y generar sesgo en el modelo. Al usar oversampling existe el riesgo de sobreajuste debido a que se duplica información existente. Una forma de reducir estos riesgos es utilizando ambos enfoques en menor proporción sobre la muestra [27].

### 3.3.1. Under-Sample

Consiste en balancear la distribución de los datos eliminando instancias de la clase mayoritaria. A pesar de su sencillez y de la reducción del tiempo de procesado de los datos, existe el riesgo de eliminar elementos de la muestra potencialmente importantes en el proceso de clasificación, por eso se han desarrollado métodos capaces de realizar una selección inteligente sobre los elementos del conjunto de datos de la clase mayoritaria.

### 3.3.2. Over-Sample

Contrariamente al Under-sample, esta técnica consiste en seleccionar y duplicar aleatoriamente datos de la clase con menos observaciones hasta que coincida con el número de la clase con más observaciones. Existe el riesgo de sobre ajustar el modelo.

### 3.3.3. SMOTE

SMOTE (Synthetic Minoriry Oversampling Technique) es una técnica estadística para aumentar el número de casos de la clase con menos observaciones. A diferencia de Over-Sample, las nuevas instancias se crean artificialmente a partir de un algoritmo que utiliza las características de los registros mas cercanos a la muestra. Este enfoque aumenta las características disponibles para cada clase y hace que las muestras sean más generales [7]. En esta técnica primero se selecciona aleatoriamente un registro de la clase minoritaria (a) y se encuentran los k vecinos de clase minoritaria más cercanos. Luego, el registro sintético se crea eligiendo aleatoriamente uno de los k vecinos más cercanos (b) y conectando a y b para formar un segmento de línea en el espacio de características. Las instancias sintéticas se generan como una combinación convexa de las dos instancias elegidas a y b.

## 3.4. Árboles de decisión (CART)

Son un tipo de algoritmo supervisado desarrollado por Breiman, Friedman, Olshen y Stone en 1984 [8]. Dependiendo del tipo de variable a predecir, se diferencian dos tipos de árboles, los de clasificación y los de regresión. Los de clasificación tienen una variable a predecir de tipo categórica (Ejemplo: Llueve o no llueve) y los de regresión una variable numérica.

**Terminología**

- **Nodo raíz:** Representa a toda la población o muestra y esto se divide en dos o más conjuntos homogéneos.
- **Ramificación o división:**Es un proceso de división de un nodo en dos o más subnodos.

- **Nodos de decisión:** Cuando un sub-nodo se divide en sub-nodos adicionales, se llama nodo de decisión.
- **Hojas o nodos terminales:** Los nodos sin hijos (sin división adicional) se llaman Hoja o nodo terminal.
- **Poda:** Cuando reducimos el tamaño de los árboles de decisión eliminando nodos (opuesto a la división), el proceso se llama poda.
- **Rama / Sub-arbol:** Una sub-sección del árbol de decisión se denomina rama o subárbol.
- **Nodo padre-hijo:** Un nodo, que se divide en sub-nodos se denomina nodo principal de sub-nodos, mientras que los sub-nodos son hijos de un nodo principal.

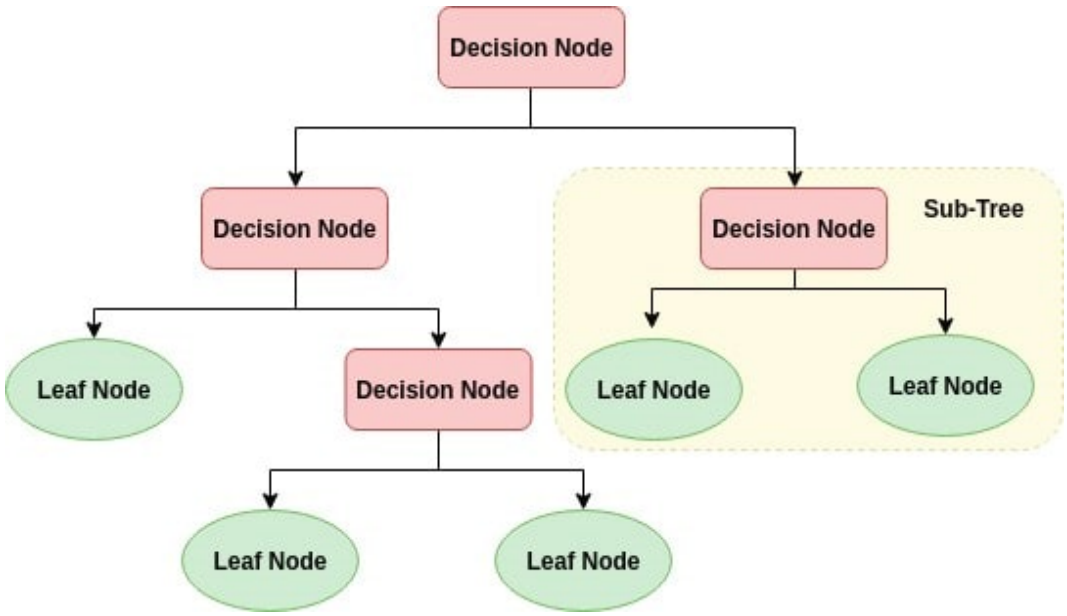



Figura 3.3: Diagrama árbol de decisión. Imagen recuperada del sitio web datacamp.com

**¿Cómo funciona el algoritmo del árbol de decisión?**

Los árboles de decisión clasifican las observaciones ordenándolos en el árbol desde la raíz hasta algún nodo hoja / terminal, y el nodo hoja / terminal proporciona la clasificación de la observación. Cada nodo del árbol actúa como un caso de prueba para algún atributo, y cada borde que desciende del nodo corresponde a las posibles respuestas al caso de prueba. Este proceso es de naturaleza recursiva y se repite para cada subárbol que desciende del nuevo nodo.

La técnica de árbol de clasificación sigue los siguientes pasos:

- En cada nodo, se toma la decisión de ramificar o parar
- Cuando se toma la decisión de parar, se elige la etiqueta del nuevo nodo hoja

- Cuando se toma la decisión de ramificar, se elige que variable se usara para la ramificación
- Cuando clasificamos los datos de entrenamiento acorde a la construcción del árbol, se elige a que nodo hoja será asignado cada uno de los datos de modo que se respecte la estructura del árbol.

**¿Cómo se construyen los árboles?**

Existen distintos criterios de selección de los atributos (variables) para la división óptima, todos tienen como objetivo encontrar los nodos mas puros o homogéneos posibles en cada una de las ramificaciones:

- **Error de clasificación:** Se define como la proporción de observaciones que no pertenecen a la clase más común en el nodo.

$$E_m = 1 - max_k(\hat{p}_{mk}) \tag{3.1}$$

- **Índice Gini:** Es una medida de la varianza total en el conjunto de las k clases del nodo m. Se considera una medida de pureza del nodo.

$$G_m = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{3.2}$$

  El algoritmo **CART (Classification and regression trees)** emplea Gini como criterio de división.

- **Chi Cuadrado:** Esta aproximación consiste en identificar si existe una diferencia significativa entre los nodos hijos y el nodo parental, es decir, si hay evidencias de que la división consigue una mejora.

$$X^2 = \sum_{k} \frac{(observado_k - esperado_k)^2}{esperado_k} \tag{3.3}$$

- **Entropía:** La entropía es otra forma de cuantificar el desorden de un sistema. En el caso de los nodos, el desorden se corresponde con la impureza. Si un nodo es puro, contiene únicamente observaciones de una clase, su entropía es cero. Por el contrario, si la frecuencia de cada clase es la misma, el valor de la entropía alcanza el valor máximo de 1.

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} log(\hat{p}_{mk}) \tag{3.4}$$

## 3.5. Algoritmos de ensamble

Los algoritmos de ensamble (o aditivos) son un método de combinar predictores débiles (con exactitud no muy superior a lanzar una moneda) para crear un predictor más robusto. Dos formas de algoritmos de ensamble son el Bagging y Boosting.

### 3.5.1. Bagging

Bagging (Anexo 6.1.) es un método para generar múltiples versiones de un predictor. La agregación promedia las predicciones de cada predictor generado en el caso de regresiones y toma el "voto mas común" en el caso de la clasificación. Estas múltiples versiones de algoritmos débiles se crean en base a un bootstrapping del dataset de entrenamiento [9].

**Random Forest**

Random Forests (Anexo 6.1.1.) es un algoritmo de ensamble que utiliza la técnica de Bagging añadiéndole el factor de aleatoriedad en la construcción de cada clasificador débil. Donde a diferencia de lo que sucede con el Bagging en cada split el algoritmo escoge sólo un subconjunto de los predictores, esto es, la aleatoriedad está presente tanto en las muestras de entrenamiento como en el conjunto de atributos.[10]

### 3.5.2. Boosting

Intuitivamente, la idea detrás de los métodos de Boosting (Anexo 6.2.) es que la combinación de varios predictores débiles produce un mejor modelo en términos de capacidad predictiva. Un predictor débil es un algoritmo cuya capacidad predictiva es levemente mejor que un resultado aleatorio, y con baja capacidad de generalización [11].

**Extreme Gradient Boosting**

Extreme Gradient Boosting (Anexo 6.2.1.) es un algoritmo predictivo supervisado que utiliza el principio de Boosting. La idea de este algoritmo es combinar el método de Boosting explicado anteriormente con un algoritmo de optimización con el fin de conseguir un modelo mas fuerte [12].

## 3.6. Optimal Classification Tree

Desarrollado por Dimitris Bertsimas y Jack Dunn en 2017 [13]. A diferencia de los algoritmos de ensamble con árboles de clasificación, los cuales utilizan heurísticas de forma recursiva para crear cada ramificación (lo que puede no capturar bien las características subyacentes del conjunto de datos), el árbol de decisiones óptimo intenta resolver esto creando la decisión completa del árbol (elegir todas las ramificaciones desde un principio) para lograr

la optimización global a través de lo que se conoce como MIO (Mixed Integer Optimization). Formulando el problema usando MIO se modelan todas las decisiones discretas que se realizan en un árbol de clasificación normal en un solo problema, de modo de encontrar el óptimo global, opuesto a lo que hacen el resto de los algoritmos de clasificación que buscan los mejores óptimos locales.

Este algoritmo tiene la ventaja de combinar el poder predictivo de los algoritmos de ensamble pero no perdiendo la interpretabilidad que tienen los algoritmos simples como los árboles de clasificación.

## 3.7. Métricas de desempeño de los algoritmos

Definimos ahora las métricas de desempeño para determinar la capacidad predictiva de los modelos desarrollados.

### 3.7.1. Matriz de confusión

Una matriz de confusión es una representación matricial de los resultados de las predicciones de cualquier prueba binaria que se utiliza a menudo para descubrir el rendimiento del modelo de clasificación sobre un conjunto de datos de prueba cuyos valores reales se conocen.

| | | **Predicción** | |
|---|---|---|---|
| **Observación** | | Positivos | Negativos |
| | Positivos | **VP** | **FN** |
| | Negativos | **FP** | **VN** |

Figura 3.4: Matriz de confusión.

Cada predicción puede ser uno de los cuatro resultados, basados en como coincide con el valor real:

- **Verdadero Positivo (VP):** Predicho Verdadero y Verdadero en la realidad
- **Verdadero Negativo (VN):** Predicho Falso y Falso en la realidad
- **Falso Positivo (FP):** Predicción de verdadero y Falso en la realidad
- **Falso Negativo (FN):** Predicción de falso y verdadero en la realidad

Ahora entendamos este concepto usando la prueba de hipótesis, una hipótesis es una especulación o teoría basada en pruebas insuficientes que se presta a mas pruebas y experimentación. Con mas pruebas una hipótesis puede ser probada como verdadera o falsa. Una Hipótesis Nula es una hipótesis que dice que no hay significancias estadísticas entre las dos variables de la hipótesis. En este caso es lo que se esta tratando de refutar. La hipótesis nula se rechaza cuando es falsa, y aceptamos la hipótesis nula cuando es realmente verdadera. Si bien las pruebas de hipótesis son de fiar, hay dos tipos de errores que pueden ocurrir:

- **Error tipo I:** equivalente a los Falsos Positivos (FP). Rechazar una hipótesis nula que es verdadera
- **Error tipo II:** equivalente a los Falsos Negativos (FN). Aceptar una hipótesis falsa nula

A partir de la matriz de confusión, se construyen las siguientes métricas capaces de medir el poder predictivo del modelo:

- **Precisión:** Indica el número de elementos clasificados correctamente en comparación con el número total de casos. Esta métrica no nos entrega suficiente información para cuando los datos que se utilizan para el modelo presentan clases desbalanceadas. Por ejemplo en un modelo de clasificación que separa a pacientes de alto riesgo con pacientes de bajo riesgo entrenado con datos de 990 pacientes que terminaron siendo bajo riesgo y 10 pacientes de alto riesgo, lo más probable es que el modelo se limite a responder "los pacientes son siempre de bajo riesgo"puesto que así tendría un acierto del 99 % en la validación.

$$\frac{(VP + VN)}{(VP + VN + FP + FN)} \tag{3.5}$$

- **Sensibilidad:** La métrica de sensibilidad muestra la cantidad de verdaderos positivos que el modelo ha clasificado en función del número total de valores. En términos médicos, indica la probabilidad de identificar correctamente a un paciente con la enfermedad.

$$\frac{VP}{(VP + FN)} \tag{3.6}$$

- **Especificidad:** Esta métrica muestra la cantidad de falsos negativos que el modelo ha clasificado en función del total de falsos. En términos médicos, indica la probabilidad de clasificar correctamente a un individuo no enfermo.

$$\frac{VN}{(VN + FP)} \tag{3.7}$$

- **Valor predictivo positivo:** También conocido como precisión, muestra el número de la clase positiva predicha correctamente como una proporción del total de predicciones de la clase positiva realizadas.

$$\frac{VP}{(VP + FP)} \tag{3.8}$$

- **Valor predictivo negativo:** Similar a la precisión de la clase negativa, muestra el número de clases negativas predichas correctamente como una proporción del total de predicciones de clases negativas realizadas.

$$\frac{FP}{(FN + FP)} \tag{3.9}$$

- **Prevalencia:** Muestra con qué frecuencia ocurre realmente la clase positiva en nuestra muestra, es decir, la proporción de pacientes de riesgo en la muestra.

$$\frac{VP + FN}{(Total)} \tag{3.10}$$

### 3.7.2. Curva ROC y AUC

Una curva ROC (curva de característica operativa del recepto) es un gráfico que muestra el rendimiento de un modelo de clasificación en todos los umbrales de clasificación. Esta curva representa dos parámetros:

- Tasa de verdaderos positivos (Sensibilidad)
- Tasa de falsos positivos (1-Especificidad)

El **AUC** es el área bajo la curva ROC, se interpreta como la probabilidad de que el modelo clasifique un ejemplo positivo aleatorio más alto que un ejemplo negativo aleatorio, su valor oscila entre 0 y 1 siendo 1 un modelo cuyas predicciones son un 100 % correctas. Como se puede apreciar en la Figura 11, cuando la curva ROC es la diagonal, el desempeño del modelo es el mismo que el que tuviese uno que clasifico de manera aleatoria (como flipar una moneda y predecir que saldrá cara).

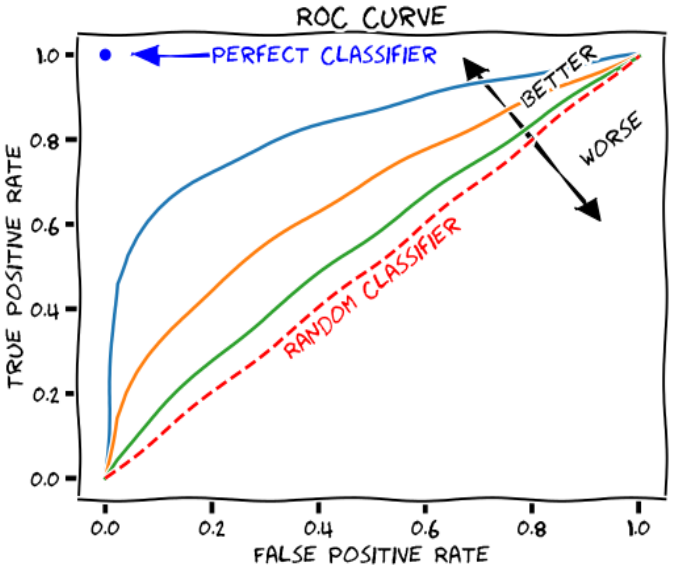



Figura 3.5: Curva ROC. Imagen recuperada del sitio web themachinelearners.com

# Capítulo 4

# Metodología

Dado el objetivo general de este trabajo, es necesario seguir una metodología que proporcione una descripción normalizada del ciclo de vida de un proyecto de análisis de datos. La metodología CRISP-DM(Cross Industry Standard Process for Data Mining) permite abordar las fases de un proyecto de Data Mining e identificar los objetivos de cada una de estas etapas, siguiendo el objetivo y contexto del trabajo [14]. Además, considera la etapa de implementación del modelo en los procesos del negocio y la iteración de las diferentes fases a medida que avanza el trabajo. Debido a estas razones, se va a considerar esta metodología para el desarrollo de este trabajo. Bajo esta metodología, el ciclo de vida del proyecto consiste en seis fases[1] :

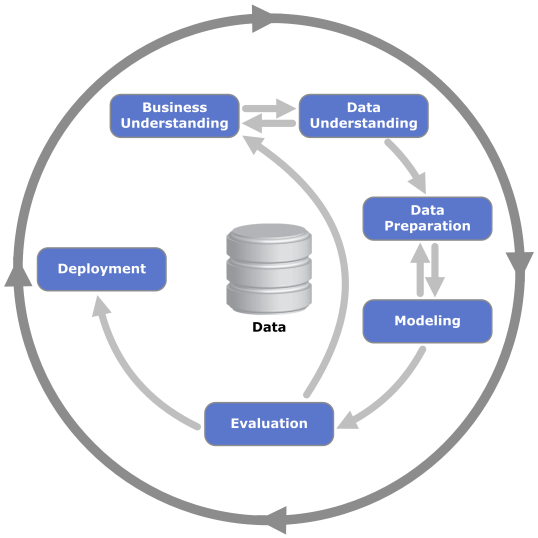



Figura 4.1: Diagrama metodología CRISP-DM

[1]Imagen recuperada del sitio web healthdataminer.com/data-mining/crisp-dm-una-metodologia-para-mineria-de-datos-en-salud/

Siguiendo esta metodología se propone desarrollar dos modelos, estos son:

1. Modelo para predecir riesgo de fallecer en pacientes hospitalizados por COVID-19
2. Modelo para predecir la severidad de la enfermedad en pacientes COVID-19 en seguimiento domiciliario

Debido a esto, a medida que se avanza en la explicación de la metodología, se van a aclarar las principales diferencias entre los modelos desarrollados.

## 4.1. Fase de comprensión del negocio

El objetivo de esta etapa es entender la organización, el área en que se va a trabajar, los procesos, objetivos estratégicos, políticas, entre otros aspectos relevantes. A partir de esto, se espera definir el problema que se va a abordar en el proyecto y los objetivo de este. Si bien gran parte de esta etapa es abordada en el capítulo 1, en este apartado se describirá de forma detallada el problema, algunas definiciones relevantes para el trabajo, el proceso de registro de datos en la plataforma COVID19 y distintas reglas de negocio utilizadas para el seguimiento y la gestión de pacientes dentro de los establecimientos de salud.

### 4.1.1. Entendiendo el problema u oportunidad

Para entender el problema se estudia el estado del arte considerando dos etapas, la primera consiste en conocer los factores de riesgo que influyen en la severidad de la enfermedad provocada por el COVID-19 basándonos en la literatura correspondiente y la segunda etapa consiste en recabar información sobre la experiencia internacional en cuanto a predecir el estado de salud de pacientes contagiados de coronavirus mediante modelos de Machine Learning.

La idea es poder conocer y tomar como referencia la experiencia internacional para luego poder validar los resultados y comparar la experiencia internacional con los resultados provenientes de los datos de Chile, específicamente los de la zona sur oriente de la Región Metropolitana.

#### I. Factores de riesgo

Con el objetivo de resumir los factores de riesgo que más influyen en la severidad de la enfermedad, se buscan estudios sistemáticos relacionados usando palabras clave propias del presente trabajo (e.g. “Covid”, “risk factors”, “comorbidities”, “symptoms”, “characteristics”, ”meta-analysis”, etc). Además se seleccionan estudios cuya fecha de realización es posterior a agosto del año 2020. Como resultado de la búsqueda seleccionamos 3 estudios

sistemáticos los cuales utilizamos para desarrollar una lista con los factores de riesgo que mas se repitieron, la cual se presenta a continuación:

**Comorbilidades**

Las comorbilidades qué más se repiten entre los estudios seleccionados son:

- **Obesidad:** Los pacientes hospitalizados con obesidad (IMC $\geq 30$) presentan el doble de riesgo de fallecer (1.2 veces) producto de la enfermedad en comparación con aquellos que no presentan esta comorbilidad [14]. Pacientes con obesidad (IMC $\geq 30$) son 1.6 veces más propensos a enfermarse gravemente producto de la enfermedad que los sin obesidad [17].
- **Hipertensión:** Pacientes hospitalizados con hipertensión son 2.1 veces propensos a fallecer producto de la enfermedad [15].
- **Diabetes:** Pacientes hospitalizados con diabetes son 1.9 veces más propensos a fallecer producto de la enfermedad [15] y 3.4 veces más propensos a enfermar gravemente [16].
- **Enfermedad cardiovascular:** Pacientes hospitalizados que presentan como antecedente alguna enfermedad cardiovascular tienen 2.5 veces más probabilidades de fallecer producto de la enfermedad [15].
- **Cáncer:** Pacientes hospitalizados que presentan como antecedente el Cáncer tienen 2.3 veces más probabilidades de fallecer producto de la enfermedad [15].
- Otras comorbilidades que se asocian con un riesgo mayor de fallecer son : enfermedad cerebro vascular, EPOC (enfermedad pulmonar obstructiva crónica), enfermedad coronaria, enfermedad renal crónica, enfermedad hepática crónica, enfermedad pulmonar crónica y enfermedad renal crónica [15].

**Manifestaciones Clínicas**

Para las manifestaciones clínicas se estudian las características de 344 mil pacientes aproximadamente, separándolos en dos grupos: los que se enfermaron gravemente y los que no [16]. Las manifestaciones clínicas en donde se presentaron diferencias entre los grupos son:

- La producción de flema, la cual fue significativamente mayor en el grupo que terminó enfermando gravemente.
- La incidencia significativa del síntoma Disnea (Ahogo o dificultad en la respiración) en el grupo que enfermó gravemente.
- La incidencia significativa de los síntomas de fatiga (sensación de cansancio), problemas de respiración y tos en el grupo que enfermó gravemente.
- El caso particular de presentar el síntoma hemoptisis y no morir producto del Covid. La hemoptisis se define como una tos severa con partículas de sangre.

- Otras variables en donde se presentaron diferencias pero las cuales no fueron estadísticamente significativas son en la incidencia de los síntomas Fiebre ($37{,}5^{\text{o}} \geq$), Faringitis (Dolor o irritación en la garganta), Náuseas y Vómitos en el grupo de pacientes que enfermó gravemente y la incidencia de la mialgias, diarrea y dolor de cabeza en el grupo de los pacientes que no presentaron mayores inconvenientes en el transcurso de la enfermedad.

### Características Sociodemográficas

En cuanto a las características sociodemográficas, estudios previos indican que existe una relación significativa en la severidad de la enfermedad con la edad de los pacientes. Se identifica que los pacientes hospitalizados con edad $\geq 65$ años tienen 3.6 más probabilidades de fallecer en comparación con los pacientes hospitalizados con edades $\leq 65$. También se identificó que el riesgo de mortalidad entre los pacientes hospitalizados con Covid-19 es 63 % mayor para los hombres que para las mujeres y que los pacientes ingresados en una unidad intensiva (UCI) son 3.7 veces más susceptibles de fallecer en comparación con los que se hospitalizaron en una cama básica [15].

También se evidencio que la proporción de hombres en el grupo que enferma gravemente (65.1 %) es mayor que en el grupo que no enferma gravemente (52.5 %) [16] y finalmente se identifica que a medida que aumentan los años, aumenta el riesgo de enfermar gravemente [17].

## II. Machine Learning en Pandemia: Experiencia Internacional

El estado de arte fue construido buscando trabajos relacionados usando palabras clave propias del presente trabajo (e.g. “Machine Learning”, “Risk Factors”, “disease severity”, “predict”, “mortality”, etc). Se consideraron 3 estudios, el primero un estudio internacional con pacientes de distintas partes del mundo, el segundo un estudio utilizando datos de pacientes de Estados Unidos, uno de los países mas afectados por el virus y finalmente un estudio con datos de la región de Wuhan en China, lugar de donde se tiene la hipótesis que se originó el virus.

1. COVID-19 mortality risk assessment: An international multi-center study [18].

   Este estudio consistió en desarrollar una herramienta que permite identificar pacientes con alto riesgo de fallecer una vez estando hospitalizados, usando modelos de Machine Learning como metodología. Se tiene el registro de pacientes de 33 hospitales de diferentes partes del mundo (Italia, USA, España y Grecia), y se utilizan como predictores para los modelos las características socio demográficas (edad, género y país de hospitalización), resultados de laboratorios, registros clínicos del ingreso hospitalario como por ejemplo la saturación de oxígeno y las comorbilidades previas de los pacientes. Se emplea el algoritmo XGBoost para entrenar y validar el modelo. Se obtuvo un AUC de 0.9 en set de validación. AUC de 0.92 set de prueba para pacientes en Sevilla, AUC de 0.87 set de prueba para pacientes en Grecia y AUC de 0.81 para pacientes en USA (Connecticut). Las variables que más inciden a la hora de clasificar a los pacientes según

su riesgo de fallecer son presentar una edad avanzada, disminución en la saturación de oxígeno, altos niveles de la proteína C-reactiva, alto niveles de nitrógeno ureico y altos niveles de creatinina en la sangre.

2. Machine learning based predictors for COVID-19 disease severity [19].

   En este trabajo se estudian los predictores para determinar la severidad de la enfermedad causada por el virus coronavirus en pacientes hospitalizados, para esto se desarrollan dos modelos, uno para predecir qué pacientes necesitarán de una hospitalización intensiva y otro para predecir qué pacientes necesitarán conexión a un ventilador mecánico. Se utilizan los registros de pacientes de dos hospitales del estado de California, Estados Unidos. Se entrena el modelo mediante el algoritmo Random Forest, utilizando como predictores las variables de la admisión hospitalaria, comorbilidades previas y el resultado de los exámenes de sangre realizados a los pacientes. Los modelos obtuvieron un AUC de 0.8 para predecir necesidad de hospitalización intensiva y AUC de 0.82 para la conexión a ventilador mecánico. Además de las comorbilidades y las características sociodemográficas, este estudio resalta la relevancia de la información proveniente del examen de sangre en identificar pacientes de alto riesgo de forma anticipada.

3. A machine learning-based model for survival prediction in patients with severe COVID-19 infection [20].

   En este trabajo se desarrolla un modelo de machine learning para identificar pacientes de alto riesgo utilizando la información presentada en los exámenes de sangre de los pacientes en cuestión. Se tiene la información de 404 pacientes ingresados al hospital de Tongji, Wuhan, China. De los 404 pacientes en observación, 191 fallecen producto de la enfermedad (47 %), esta alta tasa de letalidad se relaciona con el hecho de que este hospital se caracterizó por ingresar a los primeros casos severos de coronavirus. El modelo se desarrolla utilizando el algoritmo XGBoost, obteniendo un accuracy de 90 % en los datos de validación. Entre los principales factores de riesgo que remarca este estudio, se tiene la deshidrogenasa láctica (LDH) y altos niveles de la proteína C-reactiva.

### 4.1.2. Los datos

En esta subsección se incluyen las definiciones, protocolos y fuentes de información relevantes para el posterior desarrollo de los modelos.

#### I. Fuentes de información

Si bien gran parte de la información proviene de la plataforma de seguimiento, existen otros sistemas de información tanto del SSMSO como a nivel nacional que son de relevancia a la hora de entender con que datos se está modelando y con que otros datos se podrían complementar y mejorar los modelos desarrollados mas adelante. Las fuentes relevantes son:

1. **Plataforma COVID19 SSMSO :** Información del seguimiento de las personas COVID-19 en la red de establecimientos del SSMSO
2. **Epivigila:** Información de registros de las solicitudes que realizan los médicos frente a un caso sospechoso de COVID-19. Para acceder al examen PCR como paciente en cualquier laboratorio clínico del país y así poder descartar COVID-19
3. **Consolidado de exámenes:** Cada centro asistencial envía un archivo Excel con los datos de los resultados de exámenes Covid-19 a unidad de salud digital, con distintas frecuencias. Esta operación la realizan una vez a la semana.
4. **MIS (Más Inteligencia en Salud):** Repositorio de datos histórico de los pacientes. Atenciones de urgencia, consultas APS, fichas clínicas, histórica de crónicos.
5. **Departamento de estadísticas e información de salud (DEIS):** Organismo público dedicado a las estadísticas y datos en salud. De aquí se recolecta la información de las defunciones por COVID-19 de los pacientes del servicio
6. **Unidad de Gestión de camas (UGC):** Información del monitoreo de pacientes dentro de los hospitales.

## II. Definiciones relevantes

Para las definiciones de los casos de pacientes de coronavirus, el seguimiento se basa en el protocolo de coordinación para acciones de vigilancia epidemiológica del MINSAL [21]. En este documento se define la estrategia nacional de testeo, trazabilidad y aislamiento de pacientes.

1. **Caso Confirmado:** Persona que cumple los criterios de definición de caso sospechoso en que la prueba específica SARS-CoV-2 resultó positiva (RT-PCR)
2. **Confirmado Índice:** Persona confirmada de Covid-19 cuya detección da inicio a la investigación epidemiológica e investigación de contactos estrechos.
3. **Confirmado secundario:** Caso confirmados por COVID-19 con antecedentes de contacto con otro caso confirmado y sin antecedentes de viaje.
4. **Caso Sospechoso:**
   (a) Persona que presenta un cuadro agudo con al menos dos de los síntomas compatibles con Covid-19: fiebre (37,5°C o más), tos, disnea, dolor torácico, odinofagia, mialgias, calofríos, cefalea, diarrea, o pérdida o disminución brusca del olfato (anosmia o hiposmia) o del gusto (ageusia o disgeusia)
   (b) Cualquier persona con una infección respiratoria aguda grave que requiera hospitalización
5. **Caso Probable:** Los casos probables se deben manejar para todos los efectos como casos confirmados: Aislamiento por 11 días desde el inicio de síntomas, Identificación y cuarentena de contactos estrechos, Licencia médica. Se califican como probables por las siguientes razones.

(a) Caso Probable por Resultado de Laboratorio: Paciente que cumple con la definición de caso sospechoso en el cual el resultado del PCR es indeterminado o bien tiene una prueba antigénica para SARS-CoV-2 positiva.

(b) Caso Probable por Nexo Epidemiológico: personas que ha estado en CONTACTO ESTRECHO con un confirmado con Covid-19 y desarrolla fiebre (temperatura axilar 37,8 o más) y al menos dos síntomas compatibles con COVID-19 dentro de los 14 días posteriores al contacto. No requiere de PCR. Si por cualquier motivo, un caso probable se realizó un examen confirmatorio y este resulta positivo, se considerará como confirmado. Por el contrario si resulta negativo o indeterminado, se seguirá considerando como caso probable.

(c) Caso Probable por Imágenes: Caso sospechoso con resultado RT-PCR para ARS-CoV-2 negativo pero que cuenta con una tomografía computarizada de tórax con imágenes características de de COVID-19 según el informe radiológico.

(d) Caso Probable por Síntomas: Persona que presente pérdida brusca y completa del olfato (anosmia) o de sabor(ageusia) sin causa que lo explique.

6. **Contacto estrecho:** Contacto estrecho de un confirmado, se considera estrecho si cumple alguna de las siguientes condiciones.

   (a) Haber mantenido más de 15 minutos de contacto cara a cara, a menos de un metro, sin mascarilla.

   (b) Haber compartido un espacio cerrado por 2 horas o más, en lugares tales como oficinas, trabajos, reuniones, colegios, entre otros, sin mascarilla.

   (c) Vivir o pernoctar en el mismo hogar o lugares similares a hogar, tales como hostales, internados, instituciones cerradas, hogares de ancianos, hoteles, residencias, entre otros.

   (d) Haberse trasladado en cualquier medio de transporte cerrado a una proximidad menor de un metro con otro ocupante del medio de transporte que esté contagiado, sin mascarilla.

7. **Descartados:** Se descarta infección por Covid-19

### III. Criterios para finalizar el seguimiento activo

Los criterios establecidos por el MINSAL para definir el alta del paciente en seguimiento y el termino del periodo en el cual se consideran como "infectantes" son:

1. El paciente tuvo sintomatología leve con manejo domiciliario y se le considera "no infectante" 14 días después de que manifestó el virus o fue diagnosticado.
2. El paciente estuvo hospitalizado, fue dado de alta sin síntomas y se considera "no infectante" al día 14 tras manifestar el virus o ser diagnosticado.
3. El paciente fue hospitalizado con síntomas, pero no tuvo fiebre y, en este caso, a los 14 días luego de ser dado de alta se le estima "no infectante".

4. El paciente evidenció un compromiso previo en su sistema inmune y recién es considerado "no infectante" cuando se han cumplido 28 días desde que comenzó con la manifestación de síntomas.

El examen para definir la condición de no infectante corresponde a analizar la cantidad de Inmunoglobulina G (IGG) a través de una muestra de sangre.

### IV. Categorización de riesgo

Actualmente, el SSMSO utiliza el Protocolo de Seguimiento de Casos COVID-19 de la Red de centros Áncora de la Pontificia Universidad Católica de Chile. Este protocolo es de particular interés para el trabajo puesto a que en el se describen los factores epidemiológicos, clínicos y sociales para clasificar pacientes según su nivel de riesgo de enfermarse gravemente producto de la enfermedad [21].

Esto es precisamente lo que buscamos responder con los modelos, por lo que este protocolo servirá para validar los resultados de estos y además se espera que los resultados puedan complementar este protocolo con nueva información acerca de los factores de riesgo en pacientes confirmados de coronavirus.

Dado que existen múltiples factores de riesgo, la categorización queda a criterio del personal de salud que atiende al paciente, no obstante, el protocolo entrega una guía para objetivar la elección del riesgo.

En la Figura 4 del Apéndice B se encuentran los criterios de clasificación de riesgos de casos sospechosos y/o confirmados con COVID-19. A modo de resumen, el protocolo clasifica a pacientes en 3 niveles (alto, moderado y bajo) según el numero y tipo de factores epidemiológicos, clínicos y psicosociales presentados en los pacientes:

- Alto: 1 o más criterios mayores o 2 o más criterios intermedios.
- Moderado: 1 criterio intermedio o 2 o más criterios menores.
- Bajo: Ninguno de los factores de riesgo o sólo 1 criterio menor.

En síntesis, consideraremos las definiciones de casos del MINSAL para seleccionar a los pacientes que consideraremos más adelante como pacientes COVID-19 o pacientes contagiados de coronavirus, utilizaremos como fuentes de información la Plataforma Covid-19 del SSMSO, estadísticas acerca de las defunciones por COVID-19 del DEIS y la información de la UGC. También consideraremos las categorizaciones de riesgo de la Red de centros Áncora para poder comparar posteriormente con las categorizaciones de riesgo entregadas por los modelos desarrollados.

## 4.2. Fase de recolección y comprensión de los datos

En esta etapa se recolectan los datos que serán utilizados posteriormente. El objetivo es comprender las bases de datos y evaluar la calidad de estas. Específicamente en este trabajo se espera comprender las variables que existen en las bases de datos, especificar los datos con los que se va a trabajar y realizar el análisis exploratorio.

### 4.2.1. Base de datos

Como fue mencionado en el subcapítulo anterior, el desarrollo de este trabajo se basará en el set de datos que contiene los registros de todos los seguimientos de los pacientes COVID-19 provenientes de la plataforma COVID19 del SSMSO más la información de la unidad de gestión de camas para aquellos pacientes que fueron hospitalizados. Se consideran los registros realizados entre marzo del 2020 y abril del 2021 para entrenar y validar los modelos y luego se considera la información del mes de junio del 2021 para testear el desempeño de los modelos con nuevos datos.

Si bien un modelo es para pacientes en seguimiento domiciliario y el otro modelo es para pacientes ingresados en un establecimiento de salud, ambos usan las dos bases de datos (plataforma COVID19 y UGC). Para el modelo de pacientes en seguimiento es necesario utilizar la base de datos de la UGC para encontrar a aquellos pacientes que del seguimiento pasaron a estar hospitalizados y para el modelo de pacientes hospitalizados, utilizamos los datos de la plataforma COVID19 para conocer los antecedentes clínicos y comorbilidades previas de los pacientes hospitalizados.

### 4.2.2. Atributos de la base

A continuación se presentan los datos entregados por el SSMSO previo a la limpieza y selección de los datos. En los siguientes cuadros se pueden observar las tablas y variables de cada una de las fuentes de datos y el volumen de información que contienen cada una. A medida que se avanza en la metodología se detallara el proceso de limpieza y selección de variables junto con explicar el significado de cada una de estas.

Cabe mencionar que este volumen de datos corresponde a los ingresos realizados entre marzo de 2020 y abril de 2021, siendo este el primer filtro realizado a la base de datos.

| Datos | Volumen | Cantidad de variables |
|---|---|---|
| Ingresos seguimiento | 141,021 | 36 |
| Ingresos hospitalarios | 3,326 | 30 |

Tabla 4.1: Volumen y variables datos disponibles

| Tablas | Variables |
|---|---|
| Persona | Id del paciente |
| | Id ingreso |
| | Tipo ingreso |
| | Edad |
| | Genero |
| | Trabajador salud |
| | ¿Hospitalizado centro salud? |
| | Fecha inicio sintoma |
| | Tipo origen |
| Movimientos | Id movimiento |
| | Id egreso |
| | Id establecimiento |
| | Tipo movimiento |
| | Fecha registro |
| | ¿Hay contacto? |
| | ¿Presenta síntomas? |
| | Categorización |
| | ¿Cumple cuarentena? |
| | Dias proximo seguimiento |
| | Modalidad atención |
| Egresos | Id egreso |
| | Fecha registro |
| | Fecha egreso |
| | Resumen historial clinico |
| | Complicaciones |
| | Causal egreso |
| | Destino |
| | Criterio alta |
| Defunciones | Id persona |
| | Día defunción |
| | Fecha carga |
| | Lufar defunción |
| | Región |
| Antecedentes | Id ingreso |
| | Antedecente |
| Síntomas | Id movimiento |
| | Sintomas |
| Establecimientos | Id comuna |
| | Id establecimiento |

Tabla 4.2: Antecedentes base de datos plataforma COVID19

| Tabla | Variables |
|---|---|
| Egresos hospitalarios | Id per |
| | Id caso |
| | Tipo establecimiento |
| | Cesfam |
| | Tipo de atención |
| | Tipo de hospitalización |
| | Comuna |
| | ¿Es Fonasa? |
| | Previsión |
| | Sexo |
| | Edad |
| | Antecedentes morbidos |
| | Habitos nocivos |
| | Nocivos otras drogas |
| | Medicamentos |
| | Tramo Fonasa |
| | Tipo cama ingreso |
| | Tipo cama egreso |
| | Sindrome clinico asociado |
| | Fecha ingreso |
| | Cantidad de transiciones |
| | Fecha egreso |
| | Días de estada |
| | ECMO |
| | Fecha registro ECMO |
| | ECMO fecha desconexión |
| | VM conexión |
| | VM fecha conexión |
| | VM fecha desconexión |
| | Fecha ultima evolución |

Tabla 4.3: Atributos base de datos UGC

### 4.2.3. Modificación de atributos

Las primeras modificaciones realizadas a los atributos corresponden a transformar el formato de las tablas de antecedentes y síntomas. Estas tablas en su formato original contienen dos columnas un identificador para el ingreso o movimiento dependiendo si es un antecedente o un síntoma y otra columna con el síntoma o antecedente correspondiente. Los pacientes en un movimiento pueden tener varios síntomas al igual que en un ingreso presentar varios antecedentes clínicos registrados, por esto es que transformamos esta tabla en una en donde se presente un ID por movimiento o ingreso junto con los síntomas y antecedentes como variables dummies.

Posteriormente se modifican los atributos que indican el tipo de ingreso al seguimiento y

el tipo de cama al ingreso hospitalario. Ambos son atributos multicategóricos y tienen varias etiquetas, por lo tanto, se realiza una reducción en el numero de clases mediante la agrupación de clases similares. Con respecto al atributo del tipo de ingreso al seguimiento, se redefinen las siguientes categorías:

| **Tipo ingreso** | **Proporción** |
|---|---|
| Confirmado | 15 % |
| Contacto | 19 % |
| Descartado | 51 % |
| Probable/sospechoso | 15 % |
| **Total** | **100 %** |

Tabla 4.4: Distribución tipo de ingreso al seguimiento

Se agrupan las categorías Confirmado Índice (12,9 %) y Confirmado Secundario (1,7 %) en la categoría Confirmado, las categorías Descartado (8,2 %), Descartado Índice (43,2 %) y Descartado Secundario (0,1 %) en la categoría Descartado y finalmente las categorías Probable Índice (1,8 %), Probable Secundario (3,6 %) y Sospechoso (9,3 %) en la categoría Probable/-sospechoso.

Con respecto al tipo de cama al ingreso hospitalario, se agrupan las categorías cama básica (31 %) y cama media (12 %) en cama básica/media, cama UCI (21 %) y cama UTI (12 %) en cama UCI/UTI. Existen 821 registros Nulos (24 %) los cuales corresponden a hospitalizaciones ambulatorias las que por naturaleza no tienen una cama asignada.

| **Tipo cama ingreso** | **Proporción** |
|---|---|
| Básica/media | 43 % |
| UCI/UTI | 33 % |
| Nulo | 24 % |
| **Total** | **100 %** |

Tabla 4.5: Distribución tipo de cama ingreso hospitalario

### 4.2.4. Creación de atributos

El primer atributo creado corresponde a una variable multicategoría llamada *Rango etario*("grupos_edad"), este atributo se construye a partir de la variable "Edad" y la utilizamos para el análisis exploratorio, específicamente para encontrar diferencias en las características en los pacientes COVID19 a lo largo de los distintos grupos etarios.

De las distribuciones por edad, notamos que las proporciones entre pacientes en seguimiento y ingresados al hospital difieren notoriamente. Específicamente notamos que en porcentaje, los pacientes en seguimiento se concentran en el rango etario de los 0 a los 29 años (38 %) seguido por los pacientes en el rango de los 30 a 49 años (30 %). También notamos que a medida que se avanza en la edad, la proporción de pacientes disminuye.

| **Rango etario** | **Proporción** |
|---|---|
| 0 a 29 | 38 % |
| 30 a 49 | 30 % |
| 50 a 59 | 13 % |
| 60 a 69 | 11 % |
| 70 a 79 | 6 % |
| 80 y más | 2 % |
| **Total** | **100 %** |

Tabla 4.6: Distribución por edad pacientes en seguimiento

Para aquellos pacientes que hicieron su ingreso en algún establecimiento de salud, el mayor número de pacientes se concentra en el rango etario de los 30 a los 49 (26 %) años seguido por el rango de los 60 a 69 años (20 %). En este caso no se observa el comportamiento de la distribución de disminuir la proporción de pacientes a medida que se avanza en la edad.

| **Rango etario** | **Proporción** |
|---|---|
| 0 a 29 | 15 % |
| 30 a 49 | 26 % |
| 50 a 59 | 18 % |
| 60 a 69 | 20 % |
| 70 a 79 | 14 % |
| 80 y más | 7 % |
| **Total** | **100 %** |

Tabla 4.7: Distribución por edad pacientes gestión hospitalaria

En lo que procede, se explica creación de tres nuevas variables, las cuales serán utilizadas como variables explicativas o predictoras para el modelo de desenlace de pacientes hospitalizados. La primera de ellas es la variable *Hospitalización previa* ("hosp_prev"), esta indica si el paciente presentaba o no un ingreso hospitalario previo producto del COVID-19.

$$hosp\_prev = \begin{cases} 1 & \text{si } f_i(ID\_PER) > 1 \\ 0 & \text{sino} \end{cases} \tag{4.1}$$

La segunda variable corresponde a *Ventilación Mecánica al ingreso*("vm_ingreso"), esta indica si el paciente fue conectado a ventilación mecánica al momento de ingresar al establecimiento de salud. Se considera como al "momento de ingresar" si el paciente fue conectado durante las primeras 24 hrs desde su ingreso.

$$vm\_ingreso = \begin{cases} 1 & \text{si fecha conexión ventilador} - \text{fecha ingreso hospital} \leq 1 \\ 0 & \text{sino} \end{cases} \tag{4.2}$$

Finalmente identificamos la fecha correspondiente al inicio del seguimiento y al ingreso al hospital para construir la variable *Días en seguimiento previa hospitalización*("seg_prev")

$$seg\_prev = fechaingresohospital - fechaingresoseguimiento \tag{4.3}$$

### 4.2.5. Variable a predecir

Luego de la creación de las variables mencionadas, se construyen las variables que se buscan predecir en cada uno de los modelos.

#### I. Machine learning para predecir el desenlace de los pacientes COVID hospitalizados

Para el modelo de pacientes COVID hospitalizados se busca predecir el desenlace una vez termina su periodo de hospitalización. Esta información se encuentra en la variable "Tipo de egreso" de la tabla "Egresos hospitalarios", la cual en su formato original considera 4 posibles estados finales:

| **Tipo de egreso** | **Proporción** |
|---|---|
| Alta domicilio | 77,6 % |
| Fallecimiento | 8,1 % |
| Centro privado | 1,7 % |
| Centro público | 2,9 % |
| Nulo | 9,7 % |
| **Total** | **100,0 %** |

Tabla 4.8: Distribución estado final pacientes hospitalizados

Es necesario redefinir la variable del estado final de los pacientes por dos razones:

- Para el objetivo del modelo, es necesario considerar dos posibles estados finales: Alta domicilio y fallecimiento. Puesto a que no conocemos el desenlace final del paciente cuando su egreso corresponde a trasladarse a otro centro (público o privado), no consideramos estos registros para el modelo.
- Los pacientes con tipo de egreso nulo son aquellos que, al momento de ser entregados los datos por la contraparte, estos pacientes permanecían hospitalizados por lo que no se conoce su estado final y por ende no son considerados para la construcción del modelo.

#### II. Machine learning para predecir la severidad de la enfermedad en pacientes COVID

Para el modelo de predicción de la severidad de la enfermedad para pacientes en seguimiento, creamos la variable riesgo la cual consiste de las categorías Alto riesgo y Bajo riesgo. Definimos a una persona de alto riesgo según su probabilidad de enfermar gravemente

producto de la enfermedad, para esto consideramos como "enfermar gravemente" a aquellos pacientes que pasan por alguno de los siguientes estados:

| **Causal riesgo** | **Casos** |
|---|---|
| Cama ingreso UCI/UTI | 1,086 |
| Cama egreso UCI/UTI | 328 |
| Requiere ventilación mecanica | 504 |
| Fallecimiento hospital | 270 |
| Fallecimiento seguimiento | 247 |
| **Total** | **2,435** |

Tabla 4.9: Causal de alto riesgo pacientes COVID-19

También es relevante mencionar que pacientes que fallecen en el hospital en la mayoría de los casos son los mismos pacientes que fallecen provenientes del seguimiento, puesto que empezaron del seguimiento domiciliario y luego fueron trasladados a un establecimiento hospitalario. Posterior a la limpieza de los datos, la cual será explicada en el próximo capitulo, la distribución de pacientes de alto y bajo riesgo es la siguiente:

| **Riesgo** | **Proporción** |
|---|---|
| Alto | 2 % |
| Bajo | 98 % |
| **Total** | **100 %** |

Tabla 4.10: Distribución riesgo pacientes COVID en seguimiento

Se puede observar que las clases en las variables a predecir están desbalanceadas, dada la naturaleza de la enfermedad la cual posee una tasa de letalidad estimada entre un 1 % y un 3 % (ver en el subcapítulo 1.1). Esto hace que sea necesario balancear los datos con alguna de las técnicas descritas en el marco conceptual.

### 4.2.6. Análisis descriptivo de los datos

En primer lugar, se realiza un análisis descriptivo de la información proveniente de la gestión hospitalaria. En la Figura 4.2 se puede observar la evolución de los ingresos hospitalarios en el tiempo, diferenciando por el tipo de cama que requirieron los pacientes al ingresar (ambulatorio en caso de no necesitar cama).

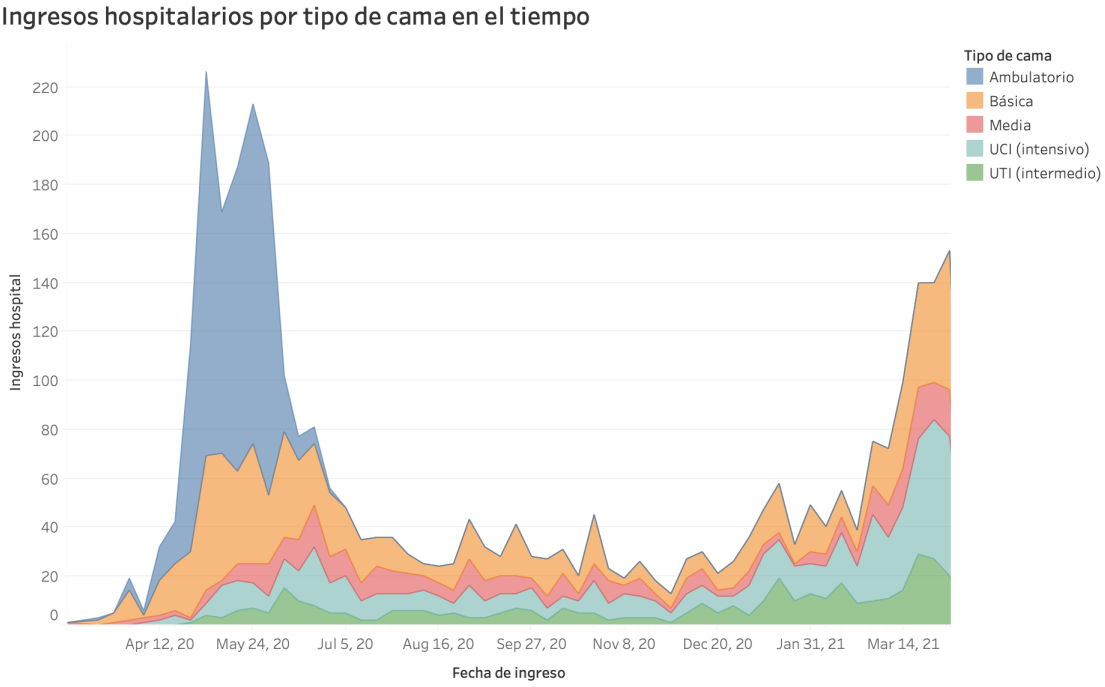



Figura 4.2: Ingresos en el tiempo según tipo de cama

Notamos que los peaks de ingresos hospitalarios ocurren durante la primera semana de mayo de 2020 y durante la última semana de marzo del 2021, en esta última semana se requirieron de 57 camas UCI y 27 camas UTI, siendo esta la semana con más demanda para el SSMSO. Se observa también que en el primer peak existe un gran número de ingresos ambulatorios, a diferencia del segundo donde las personas ingresados si requirieron de algún tipo de cama.

Con relación a la evolución de los egresos hospitalarios en el tiempo, notamos de la Figura 4.3 que el peak de fallecidos ocurre durante la semana del 28 de marzo al 4 de abril de 2021 (18 padecimientos), coincidiendo con la semana con más demanda por camas críticas del periodo de estudio.

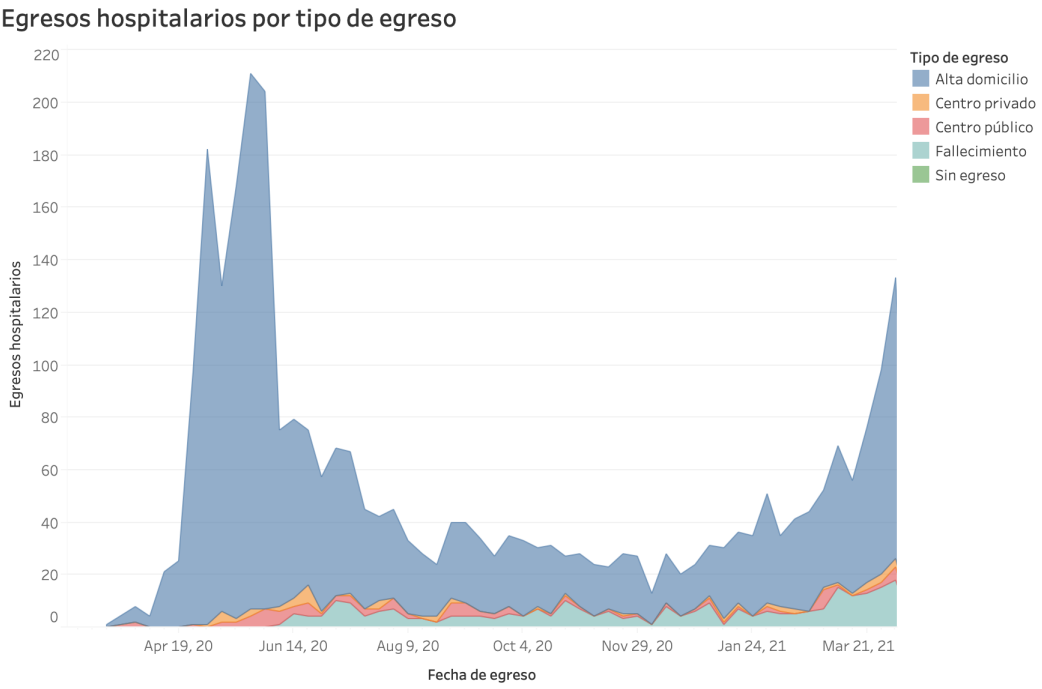



Figura 4.3: Egresos en el tiempo según tipo de egreso

Luego, se realiza un análisis de las diferencias en el tipo de cama al ingreso y el tipo de egreso según la edad de los pacientes. Notamos de la Figura 4.4 que la demanda por tipo de cama varia porcentualmente en cada uno de los rangos etarios, donde destacamos que el 30 % de los pacientes que requieren al ingreso cama UCI pertenecen al rango de los 56 a los 69 años, en el caso de las camas UTI también este rango etario es el con mayor demanda con un 27 % de las camas de este tipo. Para las camas básica y hospitalizaciones ambulatorias el rango etario con mayor porcentaje es el de 29 a 49 años de edad.

**Tipo de cama al ingreso según rango etario de los pacientes**

| | Rango etario | | | | | |
|---|---|---|---|---|---|---|
| Tipo de cama | [0,29] | (29,49] | (49,59] | (59,69] | (69,79] | (79,120] |
| Ambulatorio | 34% | 41% | 16% | 6% | 2% | 1% |
| Básica | 15% | 22% | 16% | 21% | 16% | 10% |
| Media | 4% | 19% | 21% | 22% | 23% | 12% |
| UCI (intensivo) | 3% | 21% | 22% | 30% | 21% | 3% |
| UTI (intermedio) | 4% | 20% | 21% | 27% | 17% | 11% |

Figura 4.4: Tipo de cama al ingreso según rango etario de los paciente

De la Figura 4.5 observamos que del grupo de pacientes que egresaron de alta al domicilio, un 29 % corresponde al grupo de los 29 a los 49 años, siendo este el rango etario con mayor

cantidad de pacientes que pasaron por el seguimiento. Del grupo de pacientes que egresan producto de padecer, notamos que a mayor edad, mayor es el proporción de fallecidos, siendo el grupo de los 69 a los 79 años el mas afectado. Del total de pacientes, la edad promedio es de 52 años con una desviación estándar de 19 (33-71).

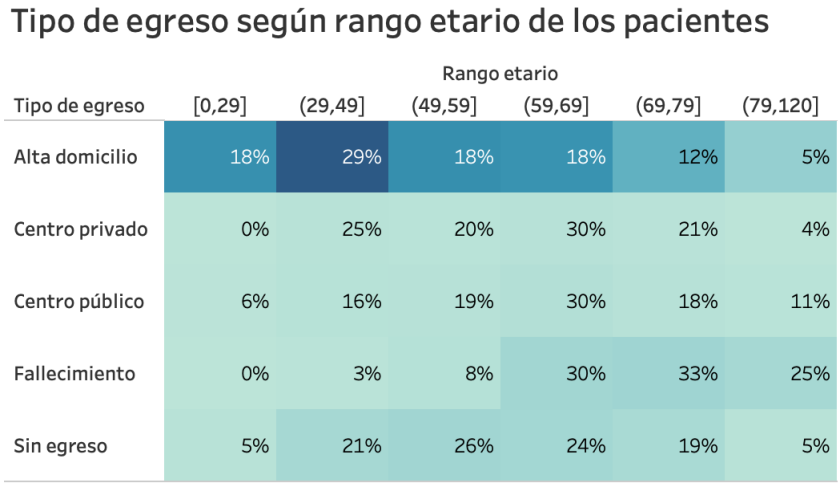



Figura 4.5: Tipo de egreso según rango etario de los paciente

De la Tabla 4.11 notamos que los pacientes que ingresan en cama UCI en promedio necesitan aproximadamente 10 días mas de hospitalización.

| **Tipo de cama** | **Tiempo promedio en el hospital (SD)** |
|---|---|
| Básica | 15.3 (15.9) |
| Media | 16.2 (20.9) |
| UCI | 25.5 (19.8) |
| UTI | 16.4 (17.3) |

Tabla 4.11: Tiempo promedio hospitalización por tipo de cama al ingreso

Y con relación al rango etario notamos en la Figura 4.12 que a medida que los pacientes son de mayor edad, mayor es el tiempo de hospitalización requerido, siendo el grupo de 69 a 79 años quienes en promedio requieren mas tiempo con un periodo de hospitalización promedio de 22 días aproximadamente.

| Rango etario | Tiempo en el hospital (SD) |
|---|---|
| 0 - 29 | 6.1 (12.0) |
| 29 - 49 | 8.4 (13.4) |
| 49 - 59 | 13.4 (19.3) |
| 59 - 69 | 18.1 (18.8) |
| 69 - 79 | 21.6 (21.7) |
| 79 o más | 16.1 (15.3) |

Tabla 4.12: Tiempo promedio hospitalización por rango etario

La proporción hombres y mujeres es equivalente (50 % cada grupo), del total de ingresos hospitalarios un 78 % (2,580) provienen del seguimiento y de los 3,326 ingresos, 3,129 corresponden a pacientes con un ingreso hospitalario y 192 pacientes ingresan a los registros hospitalarios en mas de una ocasión.

En cuanto a los seguimientos domiciliarios, se tienen 141,021 ingresos de los cuales 120,907 corresponden a pacientes con un único ingreso y 20,114 son pacientes que ingresaron más de una vez a la plataforma de seguimiento. En la Figura 4.6 se observan los ingresos al seguimiento según el tipo de caso COVID que se ingresa.

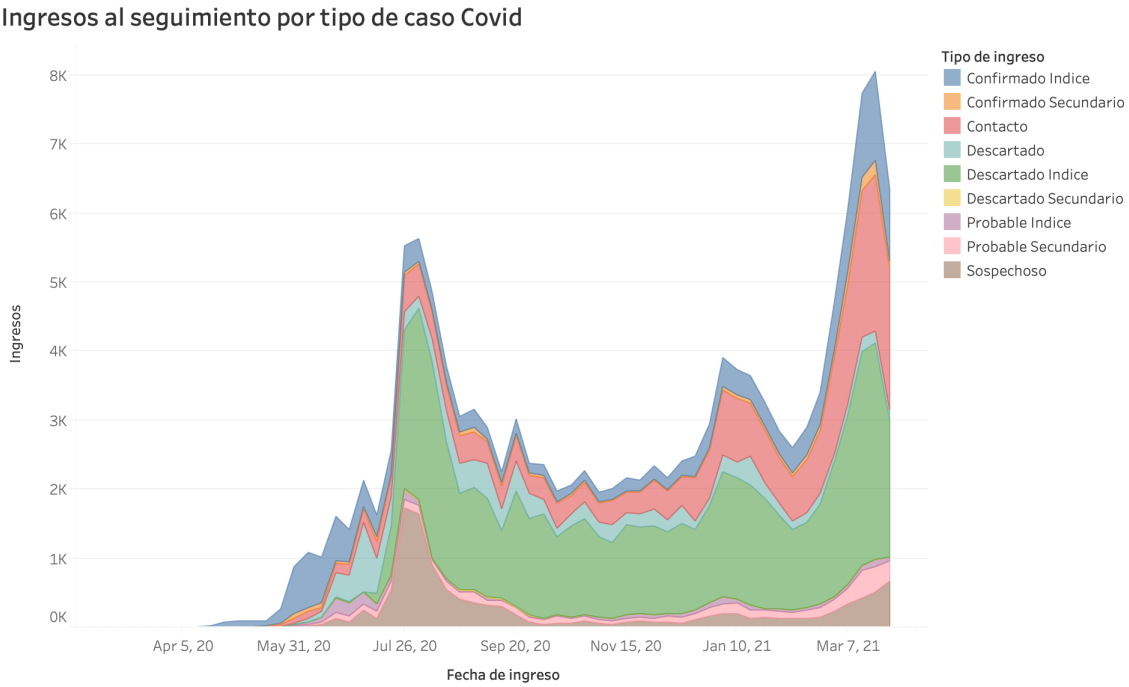



Figura 4.6: Ingresos al seguimiento por tipo de caso COVID19

Notamos que los peaks de ingresos al seguimiento difieren de los peaks de ingresos hospitalarios, específicamente el primer peak de ingresos ocurre en la primera semana de agosto de 2020 y el segundo peak en la penúltima semana de marzo de 2021. En este último, ingresaron 1,491 pacientes confirmados de COVID, 2,260 nuevos contactos de pacientes confirmados y aproximadamente 800 contactos entre sospechosos y probables.

Con relación a los antecedentes clínicos de los pacientes ingresados al seguimiento, la

hipertensión, la diabetes y el asma son los antecedentes que más se repiten. En la Figura 4.7 se puede apreciar la proporción de pacientes con determinado antecedente por rango etario, aquí notamos que el asma se presenta en mayor proporción en pacientes jóvenes (0 a los 29 años), que la diabetes se presenta en mayor proporción en el rango de los 59 a los 69 años al igual que la hipertensión y por último destacamos que el grupo etario con mayor proporción de consumidores de tabaco es el de los 0 a los 29 años.

**Antecedentes clínicos según rango etario**

| | Rango etario | | | | | |
|---|---|---|---|---|---|---|
| Antecedente | [0,29] | (29,49] | (49,59] | (59,69] | (69,79] | (79,120] |
| Alcohol | 29% | 45% | 13% | 10% | 3% | 1% |
| Alergias | 35% | 33% | 12% | 11% | 6% | 3% |
| Asma | 42% | 20% | 13% | 13% | 9% | 4% |
| Cáncer | 4% | 19% | 17% | 28% | 19% | 12% |
| Cardiopatías | 6% | 7% | 10% | 24% | 31% | 20% |
| Diabetes | 3% | 15% | 22% | 30% | 22% | 8% |
| Drogas | 48% | 44% | 6% | 1% | 0% | 0% |
| Embarazada | 63% | 37% | 0% | 0% | 0% | 0% |
| Enfermedad Renal | 6% | 13% | 13% | 25% | 25% | 19% |
| Epoc | 1% | 4% | 12% | 32% | 32% | 20% |
| Fibrosis Pulmonar | 4% | 5% | 7% | 29% | 33% | 23% |
| HTA | 2% | 14% | 21% | 29% | 23% | 10% |
| Obesidad | 24% | 36% | 17% | 14% | 7% | 2% |
| Paciente en Diálisis | 3% | 23% | 16% | 32% | 18% | 8% |
| Púerpera | 46% | 54% | 0% | 0% | 0% | 0% |
| Tabaco | 27% | 43% | 15% | 12% | 3% | 1% |
| TACO | 4% | 10% | 11% | 18% | 28% | 29% |
| Vacuna Influenza | 25% | 21% | 15% | 19% | 14% | 6% |
| Vacuna Pneumococo | 12% | 3% | 2% | 25% | 38% | 20% |

Figura 4.7: Antecedentes clínicos según rango etario

Luego analizamos las clasificaciones de riesgo de los pacientes al ingreso del seguimiento. En la Tabla 4.13 se encuentra la distribución de las clasificaciones al ingreso, donde recalcamos el hecho de que gran parte de los pacientes no cuenta con su categoría de riesgo en un principio lo que justifica el hecho de complementar esta categorización con los resultados de los modelos a desarrollar.

| **Clasificación** | **Proporción** |
|---|---|
| 1 | 5 % |
| 2 | 7 % |
| 3 | 42 % |
| NA | 45 % |
| **Total** | **100 %** |

Tabla 4.13: Clasificación de riesgo al ingreso del seguimiento

En cuanto al seguimiento, en promedio se realizan 4 seguimientos por paciente con una desviación estándar de 3. En la tabla 4.14 separamos el número de seguimientos promedio según la clasificación de riesgo al ingreso, como se espera, se realizan más seguimiento en promedio a pacientes con una clasificación de riesgo más alta.

| Clasificación | Seguimientos |
|---|---|
| 1 | 5.6 |
| 2 | 5.2 |
| 3 | 4.8 |

Tabla 4.14: Numero de seguimientos promedio por clasificación de riesgo

Finalmente separamos a los pacientes en las comunas del SSMSO que al momento de realizar el trabajo se encontraban utilizando la plataforma COVID19. En proporción la comuna con mas pacientes es La Florida seguida de La Pintana y en menor medida La Granja.

| Comuna (ID) | Proporción | Seguimientos |
|---|---|---|
| La Florida (298) | 35 % | 5.3 |
| La Granja (300) | 25 % | 3.9 |
| La Pintana (301) | 31 % | 4 |
| San Ramón (302) | 7 % | 2.5 |
| San José de Maipo (321) | 1 % | 5.6 |
| **Total** | **100 %** | **4.3** |

Tabla 4.15: Número de seguimientos promedio por comuna

También, con el propósito de conocer si existen diferencias en cuanto al seguimiento entre las comunas, se calcula el número de seguimientos promedio realizado en cada una. Se observa que la comuna que en promedio realiza mas seguimientos es San José de Maipo seguida de La Florida y la comuna con menos seguimientos promedio es San Ramón. Si bien esta información sirve de indicio para identificar diferencias entre las comunas, estas son diferentes entre si tanto en población como geográficamente, por lo que estas diferencias se podrían explicar por variados factores.

## 4.3. Fase de análisis y selección de datos

En esta etapa se llevan a cabo los procedimientos necesarios para obtener el conjunto final de datos que será utilizado para el modelamiento. Para este trabajo se llevará a cabo la limpieza de la base, se definirán los criterios para seleccionar los datos y variables para los modelos y finalmente se describirán las características de la población de estudio para cada uno de los modelos.

### 4.3.1. Población de estudio pacientes ingresados al hospital

#### Consolidación de la base de datos

Para desarrollar el modelo predictivo es necesario primero construir una base con todos los datos pertinentes de forma consolidada. Para esto, utilizamos los datos de la UGC y la información proveniente del seguimiento domiciliario, específicamente usamos la tabla de personas para extraer las características sociodemográficas, la tabla de antecedentes de donde recogemos las comorbilidades previas y variables relacionadas al cuidado de salud de las personas, la tabla de establecimientos para conocer la comuna de las personas y finalmente la tabla de ingresos hospitalarios, con la cual podemos saber que pacientes ingresaron al hospital junto con recoger las variables registradas al momento de la admisión al establecimiento de salud (4.2.2. Atributos de la base).

Los datos de la gestión hospitalaria contienen información acerca de los pacientes ingresados y/o atendidos en algún establecimiento de salud del servicio entre los meses de marzo del 2020 y abril 2021. En la Tabla 4.16 se detallan las variables de esta fuente que se incluirán en el modelo.

Además de la información proveniente de la gestión hospitalaria, utilizamos los datos provenientes de la plataforma de seguimiento, los cuales incluyen las características socio demográficas de las personas, las comorbilidades y variables relacionadas con el cuidado de la salud, como por ejemplo el hecho de fumar o de beber alcohol. En la Tabla 4.17 se describen las variables consideradas para incluir en el modelo.

| Variable | Descripción |
|---|---|
| ID Caso | Número único de ingreso |
| ID Persona | Identificador de persona, existen pacientes con más de un ingreso |
| Tipo establecimiento | Indica si la persona se atendió en un establecimiento público o privado |
| Tipo de atención | Indica si la atención fue hospitalaria o ambulatoria |
| Tipo de hospitalización | Domicilio o hospital |
| Síndrome clínico asociado | Patología del paciente al momento de ingresar al hospital |
| VM inicio | Si se conecto a un ventilador al momento de ingresar |
| Tramo Fonasa | Indica el tramo de Fonasa (A,B,C,D) o Isapre |
| Tipo de egreso | Fallecer, alta domicilio o traslado a otro centro |
| Tipo de cama ingreso | Indica el tipo de cama al ingreso hospitalario (básica,media, UCI/UTI) |

Tabla 4.16: Variables de la gestión hospitalaria UGC

| Variable | Descripción |
|---|---|
| ID Persona | Número único de persona |
| ID Ingreso | Número que identifica un seguimiento, se tienen personas con más de un ingreso |
| Edad | Edad del paciente |
| Género | Género del paciente |
| Comuna | Comuna del paciente |
| Antecedentes | Indica las comorbilidades y variables relacionadas al cuidado de la salud |

Tabla 4.17: Variables del seguimiento

La base de datos consolidada que se utilizará más adelante para entrenar el modelo, se construye a partir del Diagrama Entidad-Relación presentado en la Figura 4.8.

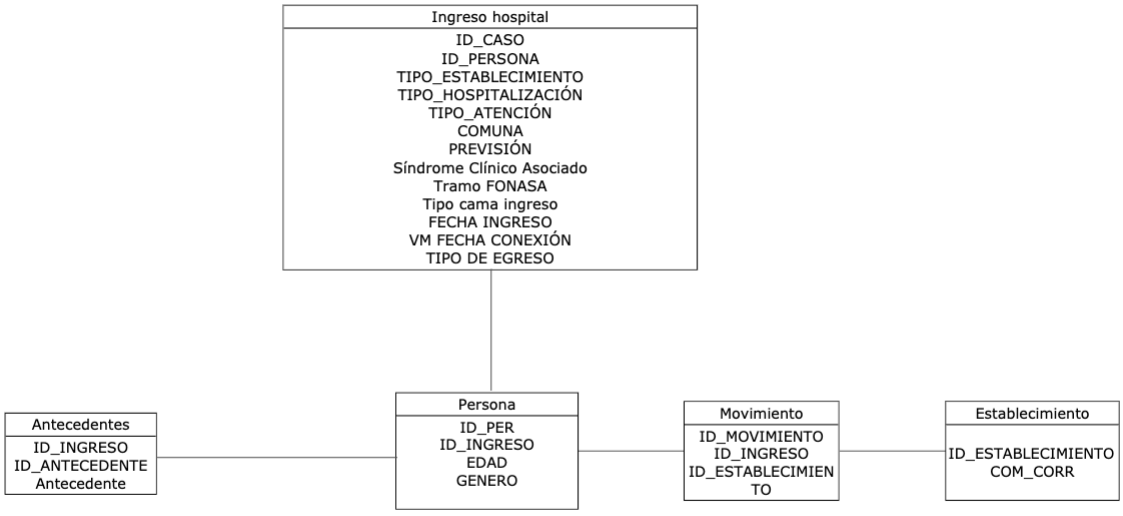



Figura 4.8: Diagrama Entidad-Relación datos consolidados

### Criterios de inclusión y exclusión

En este apartado se indican los criterios utilizados para seleccionar a la población de estudio. De los 3,226 ingresos hospitalarios de los datos de la UGC:

- Excluimos atenciones ambulatorias y hospitalizaciones en el domicilio, lo que nos deja 2,260 casos de ingresos y hospitalizaciones en un establecimiento de salud.
- Utilizamos el último ingreso hospitalario de cada persona, quedan 2,149 ingresos.
- Excluimos a 62 pacientes que se trasladaron a otro centro de salud (privado o público) puesto que no conocemos su desenlace final. Quedan 2,087 pacientes.
- Excluimos a 321 pacientes que siguen hospitalizados. Quedan 1,766 pacientes.
- Finalmente, seleccionamos pacientes que se encuentran en la base de datos provenientes de la plataforma de seguimiento del SSMSO, lo que nos deja una muestra de 1,555 pacientes.

### Población de estudio

El estudio comprende 27 establecimientos de salud a lo largo de 5 comunas de la zona sur oriente de la Región Metropolitana. Se consideran establecimientos públicos y privados y se consideran los ingresos hospitalarios desde marzo de 2020 hasta abril de 2021.

Luego de seleccionar a la población según los criterios de inclusión y exclusión, nos queda una muestra de 1,555 pacientes cuya tasa de mortalidad es de un 16 %.

| **Comuna** | **Número de hospitales** | **Pacientes** | **Tasa de mortalidad** |
|---|---|---|---|
| 298 | 10 | 561 | 16.6 % (93) |
| 300 | 4 | 367 | 13.4 % (50) |
| 301 | 6 | 444 | 18.7 % (83) |
| 302 | 5 | 147 | 12.2 % (18) |
| 321 | 2 | 36 | 14.3 % (5) |
| **SSMSO** | **27** | **1,555** | **16.1 % (249)** |

Tabla 4.18: Población de estudio pacientes hospitalizados

La tasa de mortalidad para pacientes hospitalizados a nivel nacional no es tan simple de establecer. Según un estudio del CIPER (Centro de investigación periodística)[2] las diferencias entre la tasa de mortalidad de establecimiento público v/s privado es muy alta, tomando de ejemplo al Hospital Padre Hurtado perteneciente al SSMSO cuya tasa de mortalidad es de un 25 % mientras que en el otro extremo la Clínica Las Condes presenta un 5 % de letalidad entre los pacientes hospitalizados. Por otro lado, el MINSAL desmintió las cifras, declarando que la diferencia entre las tasas de mortalidad entre los tipos de establecimientos es solamente de un 2 % (18 % privado y 20 % público). Puesto a que el SSMSO incluye establecimientos de los dos tipos, consideramos que una tasa de mortalidad de 16 % para pacientes hospitalizados se aproxima al menos a la realidad nacional.

### Características sociodemográficas

De las características sociodemográficas de la muestra, notamos que:

- La edad promedio del grupo sobreviviente es de 49 años y la edad del grupo no sobreviviente es de 63, presentando una diferencia en años entre los grupos de 14 años.
- En el género no se presentan mayores diferencias en el total de pacientes ni en el grupo sobreviviente, en el grupo de los no sobrevivientes se puede observar una leve diferencia a favor del genero Femenino.
- Del grupo no sobrevivente un 67 % pertenece al tramo B de FONASA, las personas de este tramo tienen la característica de percibir un ingreso imponible mensual menor o igual a $326.500, lo que nos indica a priori que el grupo social económicamente mas desfavorable es mas susceptible a fallecer producto de la enfermedad una vez estando hospitalizado.

También se realiza un Test de dos colas para la variable Edad y un Test de exactitud de Fisher para el resto de las variables cualitativas para concluir si las diferencias porcentuales

[2]https://www.elperiodista.cl/2020/06/covid-19-tasa-de-mortalidad-en-hospitales-publicos-metropolitanos-supera-con-creces-a-la-de-clinicas-privadas/

presentadas entre los grupos son estadísticamente significativas o no. Con el parámetro P-Valor podemos concluir que variables con P-valores menores a 0.05 son aquellas variables donde la proporción entre los grupos es significativa con un 95 % de confianza. En este caso se presentarían diferencias entre los grupos en la edad, en la composición de los rangos etarios exceptuando el rango de edad 0-15 años y diferencias en el Tramo Fonasa de los pacientes en cada uno de los grupos, donde la significancia existe para los tramos A, B, D y para pacientes con Isapre.

| **Variable** | **Total (N=1,555)** | **Sobrevivientes (N=1,306)** | **Fallece (N=249)** | **P-Valor** |
|---|---|---|---|---|
| **Edad, promedio (SD)** | 50.1 (32.6-67.6) | 48.6 (31.6-65.6) | 63.3 (51.1-75.5) | 2.2e-16 |
| **Rango etario** | | | | |
| **[0-15]** | 27 (2 %) | 26 (2 %) | 1(0.4 %) | 0.1078 |
| **(15-45]** | 303 (19 %) | 297 (23 %) | 6 (2.4) | 2.2e-16 |
| **(45-65]** | 619 (40 %) | 548 (42 %) | 71 (29.6 %) | 7.179e-05 |
| **(65-120]** | 606 (39&) | 435 (33 %) | 171 (68.6 %) | 2.2e-16 |
| **Femenino** | 758 (49 %) | 641 (49 %) | 117 (47 %) | 0.5801 |
| **Tramo Fonasa A** | 185 (12 %) | 168 (13 %) | 17 (7 %) | 0.0054 |
| **B** | 737 (47 %) | 571 (44 %) | 166 (67 %) | 3.059e-11 |
| **C** | 178 (12 %) | 157 (12 %) | 21 (9 %) | 0.1276 |
| **D** | 317 (20 %) | 283 (22 %) | 34(13 %) | 0.003467 |
| **ISAPRE** | 138 (9 %) | 127 (9 %) | 11 (4 %) | 0.00505 |

Tabla 4.19: Características sociodemográficas pacientes hospitalizados

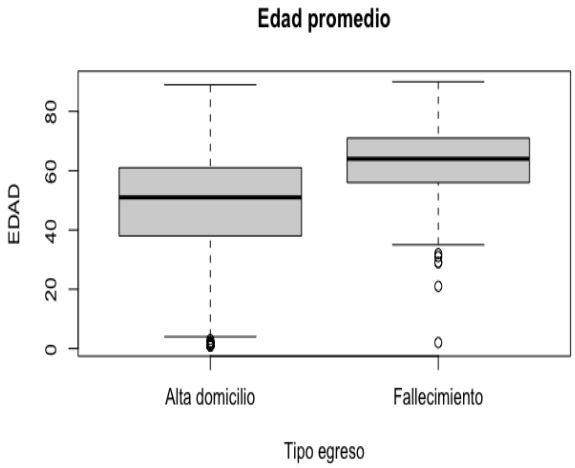



Figura 4.9: Diferencias de edad entre los grupos sobreviviente y no sobreviviente. Elaboración propia

De la Figura 4.9, los datos adentro de la caja representan desde el cuartil 1(25 %) al cuartil 3 (75 %), esto nos dice que el 50 % de los pacientes en el grupo de alta domicilio se encuentran en el rango de edad entre los 40-60 y en el grupo de fallecimiento el rango de edad esta entre los 50 y 70 años aproximadamente.

### Comorbilidades y cuidados de salud

En relación a las comorbilidades y a los cuidados de salud de los pacientes hospitalizados de la muestra se observa que las comorbilidades que más se presentan en los pacientes son hipertensión (34.7 %), diabetes (20.7 %) y obesidad (3.4 %), también notamos que una gran parte de la población se ha vacunado previamente con la vacuna de la influenza (22.4 %).

Podemos notar diferencias entre los grupos sobrevivientes y no sobrevivientes en la hipertensión (49.8 % vs 31.8 %), cáncer (1.5 % vs 3.2 %), Enfermedad pulmonar obstructiva crónica o EPOC (5.3 % vs 2.1 %),enfermedades renales(6.5 % vs 2.2 %) y Diabetes (19.6 % vs 26.3 %) . Se realiza un Test de exactitud de Fisher para determinar si las diferencias son significativas, y se encuentra que con un 95 % de confianza, los grupos presentan diferencias significativas en la proporción de pacientes con las comorbilidades Cáncer, Diabetes, EPOC, Enfermedad Renal y HTA.

| **Comorbilidad** | **Total pacientes (N=1,555)** | **Grupo sobreviviente (N=1,306)** | **Grupo no sobreviviente (N=249)** | **P-Valor** |
|---|---|---|---|---|
| **Alcohol** | 13 (0.8 %) | 12 (0.9 %) | 1(0.4 %) | 0.7055 |
| **Alergias** | 19 (1.2 %) | 17 (1.3 %) | 2 (0.8 %) | 0.7547 |
| **Asma** | 72 (4.5 %) | 58 (4.3 %) | 14 (5.7 %) | 0.4112 |
| **Cáncer** | 27 (1.7 %) | 19 (1.5 %) | 8 (3.2 %) | 0.06257 |
| **Cardiopatías** | 60 (3.8 %) | 48 (3.6 %) | 12 (4.9 %) | 0.3716 |
| **Diabetes** | 323 (20.7 %) | 256 (19.6 %) | 67 (26.3 %) | 0.01058 |
| **Drogas** | 0 (0 %) | 0 (0 %) | 0 (0 %) | 0.1388 |
| **Embarazada** | 14 (0.8 %) | 14 (0.9 %) | 0 (0 %) | 0.1447 |
| **EPOC** | 40 (2.6 %) | 27 (2.1 %) | 13 (5.3 %) | 0.00769 |
| **Fibrosis Pulmonar** | 11 (0.7 %) | 9 (0.7 %) | 2 (0.8 %) | 0.6921 |
| **Enfermedad renal** | 45 (2.9 %) | 29 (2.2 %) | 16 (6.5 %) | 0.001273 |
| **Hipertensión (HTA)** | 538 (34.7 %) | 416 (31.8 %) | 122 (49.4 %) | 3.188e-07 |
| **Obesidad** | 53 (3.4 %) | 42 (3.2 %) | 11 (4.5 %) | 0.3405 |
| **Diálisis** | 13 (0.8 %) | 10 (0.8 %) | 3 (1.2 %) | 0.4491 |
| **Puérpera** | 6 (0.4 %) | 6 (0.5 %) | 0 (0 %) | 0.5977 |
| **Tabaco** | 27 (1.7 %) | 23 (1.8 %) | 4 (1.6 %) | 1 |
| **TACO** | 22 (1.4 %) | 17 (1.3 %) | 5 (2.0 %) | 0.3792 |
| **Vacuna influenza** | 352 (22.4 %) | 283 (21.5 %) | 69( 27.1 %) | 0.03908 |
| **Vacuna Neumococo** | 44 (2.8 %) | 37 (2.8 %) | 7 (2.8 %) | 1 |

Tabla 4.20: Comorbilidades y antecedentes pacientes hospitalizados

### Características relacionadas al ingreso hospitalario

De las características relacionadas al ingreso hospitalario, notamos diferencias entre los grupos en el tipo de cama en la cual hicieron ingreso al hospital, en el tipo de establecimiento que se hospitalizaron, en si se conectaron en un principio a un ventilador mecánico, en

presentar neumonía grave o neumonía leve (síntoma que se presento en mayor porcentaje en el grupo sobreviviente) y en SDRA (síndrome de diestras respiratorio agudo). Aplicando un Test de exactitud de Fisher, encontramos que existen diferencias significativas entre los grupos en las variables Cama ingreso, el tipo de establecimiento, si fueron o no conectados a ventilación mecánica y en la presencia de los síndromes clínicos Neumonía grave, Neumonía leve y SDRA[3].

| **Variable** | **Total pacientes (N=1,555)** | **Grupo sobreviviente (N=1,306)** | **Grupo no sobreviviente (N=249)** | **P-Valor** |
|---|---|---|---|---|
| **Cama ingreso UCI/UTI** | 750 (48.2 %) | 593 (45.5 %) | 157 (63 %) | 3.779e-07 |
| **Tipo establecimiento Público** | 1,125 (72.4 %) | 919 (70.4 %) | 206 (83 %) | 3.956e-05 |
| **Ventilador mecánico** | 191 (12.3 %) | 127 (9.7 %) | 64 (26 %) | 1.367e-10 |
| **Síndrome clínico** | | | | |
| **Sin complicaciones** | 138 (9 %) | 123 (9.5 %) | 15 (6.1 %) | 0.08919 |
| **Neumonía grave** | 673 (43.4 %) | 546 (42 %) | 129 (51.4 %) | 0.004177 |
| **Neumonía leve** | 592 (38.2 %) | 531 (40.5 %) | 65 (26.3 %) | 1.314e-05 |
| **SDRA** | 127 (8.2 %) | 93 (7 %) | 35 (14.2 %) | 0.0006047 |
| **Hospitalización previa** | 91 (5.8 %) | 71 (5.4 %) | 20 (8 %) | 0.1388 |

Tabla 4.21: Variables del ingreso hospitalario

Del total de pacientes, un 5.8 % ingresa por segunda vez al hospital debido a la enfermedad del coronavirus, si separamos por grupo, un 8 % del grupo no sobreviviente se estaba hospitalizando por segunda vez vs un 5.4 % del grupo sobreviviente (lo que podría dar ciertos indicios de la gravedad de la enfermedad en estos pacientes).

Considerando a los pacientes que se hospitalizaron en un establecimiento de salud Público tenemos que de los 1,125 pacientes, un 18 % fallece v/s los 430 pacientes que se hospitalizaron en un establecimiento privado en donde fallece un 10 %.

### 4.3.2. Población de estudio pacientes en seguimiento domiciliario

#### Consolidación de la base de datos

Análogo al primer modelo, es necesario construir una base de datos consolidando las bases de la gestión hospitalaria y la plataforma COVID19 para el seguimiento domiciliario. Para este modelo específicamente utilizamos las variables del seguimiento de la Tabla 4.17 agregándole la variable "Síntomas" la cual indica los síntomas que se registran en cada uno de los movimientos y la variable "Tipo egreso" que indica como egreso el paciente del seguimiento domiciliario (Traslado, Alta, Fallece). De las variables de la gestión hospitalaria de la Tabla 4.16, utilizamos la información de las variables Ventilador mecánico, Tipo de cama al ingreso y al egreso y el Tipo de egreso hospitalario para construir la variable dependiente referente al riesgo de los pacientes.

[3]Síndrome de Diestrés Respiratorio Agudo

### Criterios de inclusión y exclusión

De los 141,021 ingresos al seguimiento domiciliario:

- Excluimos a pacientes con edades negativas y con sobre los 120 años, lo que nos deja 140,968 ingresos.
- Seleccionamos ingresos de pacientes Confirmados, probables, contactos y sospechosos, quedan 68,402 ingresos.
- Excluimos a aquellos pacientes que aun no egresan, lo que nos deja un total de 54,891 ingresos.
- Finalmente se excluyen a aquellos pacientes que abandonaron el seguimiento, lo que nos deja un total de 54,447 pacientes.

### Población de estudio

Análogo a la muestra del primer modelo, se consideran a los pacientes de 5 comunas de la zona sur oriente de la Región Metropolitana. Luego de seleccionar a la población según los criterios de inclusión y exclusión, nos queda una muestra de 54,445 pacientes donde el porcentaje de pacientes de alto riesgo es de un 1.8 %.

| **Comuna** | **Numero de hospitales** | **Pacientes** | **Tasa alto riesgo** |
|---|---|---|---|
| 298 | 10 | 20,932 | 1.5 % (324) |
| 300 | 4 | 10,853 | 2.1 % (233) |
| 301 | 6 | 18,615 | 1.9 % (367) |
| 302 | 5 | 2,446 | 3.5 % (87) |
| 321 | 2 | 1,601 | 1.0 % (16) |
| **SSMSO** | **27** | **54,447** | **1.8 % (1,027)** |

Tabla 4.22: Población de estudio pacientes en seguimiento

### Características sociodemográficas

De las características sociodemográficas de la muestra, notamos que las diferencias entre los grupos son estadísticamente significativas en las variables de edad, grupo etario, género y el tipo de ingreso. La edad promedio de la muestra es de 37.4 años, donde la del grupo de bajo riesgo es de 36.7 y la del grupo de alto riesgo es de 62.3, teniendo una diferencia de edad en promedio de 25 años aproximadamente.

| Variable | Total pacientes (N=54,447) | Grupo bajo riesgo (N=53,420 ) | Grupo alto riesgo (N=1,027) | P-Valor |
|---|---|---|---|---|
| **Edad, mean, (+-SD)** | 37.4 (15.7-59.1) | 36.7 (15.2-58.2) | 62.3 (45.7-78.9) | 2.2e-16 |
| **Grupo etario** | | | | |
| **[0-15]** | 9,669 (18 %) | 9,658 (18 %) | 11(1 %) | 2.2e-16 |
| **(15-45]** | 24,767 (45 %) | 24,627 (46 %) | 140 (14 %) | 2.1e-16 |
| **(45-65]** | 13,896 (26 %) | 13,495 (25 %) | 401 (39 %) | 2.2e-16 |
| **(65-120]** | 6,115 (11 %) | 5,640 (11 %) | 475 (46 %) | 2.2e-16 |
| **Género F** | 28,843 (53 %) | 28,407 (53 %) | 436 (42 %) | 1.005e-11 |
| **Confirmado** | 18,071 (33 %) | 17,287 (32 %) | 784 (76 %) | 2.2e-16 |
| **Contacto** | 21,802 (40 %) | 21,719 (41 %) | 83 (8 %) | 2.4e-16 |
| **Probable** | 6,521 (12 %) | 6,442 (12 %) | 79 (8 %) | 2.706e-07 |
| **Sospechoso** | 8,053 (15 %) | 7,972 (15 %) | 81 (8 %) | 5.49e-16 |

Tabla 4.23: Características sociodemográficas pacientes en seguimiento

### Comorbilidades y cuidados de salud

En relación a las comorbilidades y a los cuidados de salud, del total de pacientes las comorbilidades que más se presentaron son la hipertensión (12.5 %), diabetes (7 %) y asma (3 %). Además un 3.4 % de la muestra declara consumir tabaco y un 13.2 % del total de pacientes se vacuno previamente con la vacuna de la influenza.

Podemos notar diferencias entre los grupos de alto y bajo riesgo en la comorbilidad hipertensión (39.8 % vs 12 %), diabetes (24.2 % vs 6.3 %), enfermedad pulmonar obstructiva crónica o EPOC (3.3 % vs 0.6 %) y enfermedad renal (0.4 % vs 2.8 %).

Se realiza un Test de exactitud de Fisher para determinar si las diferencias son significativas, y se encuentra que con un 95 % de confianza, los grupos presentan diferencias significativas en la proporción de pacientes con las comorbilidades asma, cáncer, cardiopatías, diabetes, consumo de drogas, EPOC, fibrosis pulmonar, enfermedad renal, hipertensión, obesidad, diálisis, tabaco, TACO y en las vacunas para la influenza y neumococo.

| Comorbilidad | Total pacientes (N=54,447) | Grupo bajo riesgo (N=53,420) | Grupo alto riesgo (N=1,027) | P-Valor |
|---|---|---|---|---|
| **Alcohol** | 628 (1 %) | 620 (1.2 %) | 8 (0.8 %) | 0.1706 |
| **Alergias** | 543 (1 %) | 533 (1 %) | 10 (1 %) | 0.9381 |
| **Asma** | 1,623 (3 %) | 1,574 (2.9 %) | 49 (4.8 %) | 0.006525 |
| **Cáncer** | 241 (0.4 %) | 230 (0.4 %) | 11 (1.1 %) | 0.04736 |
| **Cardiopatías** | 474 (0.8 %) | 433 (0.8 %) | 41 (4 %) | 2.464e-07 |
| **Diabetes** | 3,623 (7 %) | 3,374(6.3 %) | 249 (24.2 %) | 2.2e-16 |
| **Drogas** | 202 (0.4 %) | 201 (0.4 %) | 1 (0.1 %) | 0.005802 |
| **Embarazada** | 225 (0.4 %) | 222 (0.4 %) | 3 (0.3 %) | 0.4699 |
| **EPOC** | 341 (0.6 %) | 307 (0.6 %) | 34 (3.3 %) | 1.17e-06 |
| **Fibrosis Pulmonar** | 52 (0.1 %) | 45 (0.1 %) | 7 (0.7 %) | 0.02038 |
| **Enfermedad renal** | 226 (0.4 %) | 197 (0.4 %) | 29 (2.8 %) | 2.425e-06 |
| **Hipertensión (HTA)** | 6,828 (12.5 %) | 6,419 (12 %) | 409 (39.8 %) | 2.2e-16 |
| **Obesidad** | 937 (1.7 %) | 884 (1.7 %) | 53 (5.2 %) | 4.959e-07 |
| **Diálisis** | 61 (0.1 %) | 55 (0.1 %) | 6 (0.6 %) | 0.04371 |
| **Puérpera** | 23 (0.0 %) | 23 (0.0 %) | 0 (0.0 %) | 1.616e-06 |
| **Tabaco** | 1,870 (3.4 %) | 1,848 (3.5 %) | 22 (2.1 %) | 0.004177 |
| **TACO** | 142 (0.3 %) | 124 (0.2 %) | 18 (1.8 %) | 0.000221 |
| **Vacuna influenza** | 7,197 (13.2 %) | 6,961 (13 %) | 236 ( 23 %) | 1.1e-13 |
| **Vacuna Neumococo** | 498 (0.9 %) | 451 (0.8 %) | 47 (4.6 %) | 1.475e-08 |

Tabla 4.24: Comorbilidades y cuidados de salud pacientes en seguimiento

### Síntomas al ingreso

En el caso de los síntomas presentados por los pacientes, se consideran los síntomas de los dos primeros seguimientos (al ingreso y primer seguimiento) esto debido a que en el ingreso gran parte de los pacientes no tiene síntomas registrados sino que se registran en el primer seguimiento inmediatamente después. En el Apéndice C.1. se pueden encontrar unas definiciones coloquiales de los síntomas entregadas por un médico general del Hospital Metropolitano de Santiago.

Del total de pacientes, los síntomas que más se presentaron son la cefalea (16.2 %), mialgias (12.4 %), odinofagia (6.9 %), anosmia (6.8 %) y disgeusia (6.0 %). Se realiza un Test de exactitud de Fisher para determinar si las diferencias son significativas, y se encuentra que con un 95 % de confianza, los grupos presentan diferencias significativas en la proporción de pacientes que presentaron los síntomas anorexia, anosmia, cianosis, compromiso de conciencia, compromiso de estado general, congestión nasal, decaimiento, disgeusia, disnea, dolor abdominal, dolor torácico, fatiga, mialgias, postración, retracción costal, taquipnea, vómitos y fiebre (mayor a 37.5 grados celsius).

| **Síntoma** | **Total pacientes (N=54,447)** | **Grupo bajo riesgo (N=53,420)** | **Grupo alto riesgo (N=1,027)** | **P-Value** |
|---|---|---|---|---|
| **Anorexia** | 303 (0.6 %) | 286 (0.5 %) | 17 (1.7 %) | 0.005161 |
| **Anosmia** | 3,727 (6.8 %) | 3,697 (6.9 %) | 30 (2.9 %) | 1.907e-13 |
| **Calofríos** | 248 (0.5 %) | 240 (0.4 %) | 8 (0.8 %) | 0.2325 |
| **Cefalea** | 8,800 (16.2 %) | 8,612 (16.1 %) | 188 (18.3 %) | 0.07312 |
| **Cianosis** | 11 (0 %) | 6 (0 %) | 5 (0.5 %) | 0.02887 |
| **Comp de Conciencia** | 20 (0 %) | 8 (0 %) | 12 (1.2 %) | 0.0006101 |
| **Comp Estado General** | 376 (0.7 %) | 288 (0.5 %) | 88 (8.6 %) | 2.2e-16 |
| **Congestión nasal** | 2,608 (4.8 %) | 2,577 (4.8 %) | 31 (3 %) | 0.0008963 |
| **Decaimiento** | 2,007 (3.7 %) | 1,903 (3.6 %) | 104 (10.1 %) | 6.677e-12 |
| **Diarrea** | 1,780 (3.3 %) | 1,743 (3.3 %) | 37 (3.6 %) | 0.5626 |
| **Disgeusia/Ageusia** | 3,246 (6.0 %) | 3,210 (6 %) | 36 (3.5 %) | 1.927e-05 |
| **Disnea** | 2,086 (3.8 %) | 1,799 (3.4 %) | 287 (27.9 %) | 2.2e-16 |
| **Dolor abdominal** | 928 (1.7 %) | 896 (1.7 %) | 32 (3.1 %) | 0.008455 |
| **Dolor Torácico** | 1,214 (2.2 %) | 1,161 (2.2 %) | 53 (5.2 %) | 1.809e-05 |
| **Fatiga** | 728 (1.3 %) | 687 (1.3 %) | 41 (4 %) | 1.123e-05 |
| **Mialgias** | 6,766 (12.4 %) | 6,574 (12.3 %) | 192 (18.7 %) | 2.228e-07 |
| **Náuseas** | 655 (1.2 %) | 634 (1.2 %) | 21 (2 %) | 0.05375 |
| **Odinofagia** | 3,747 (6.9) | 3,670 (6.9 %) | 77 (7.5 %) | 0.4495 |
| **Postración** | 23 (0 %) | 13 (0 %) | 10 (1 %) | 0.002014 |
| **Retracción costal** | 24 (0 %) | 18 (0 %) | 6 (0.6 %) | 0.02094 |
| **Rinorrea** | 1,566 (2.9 %) | 1,540 (2.9 %) | 26 (2.5 %) | 0.4789 |
| **Taquipnea** | 34 (0.1 %) | 29 (0.1 %) | 5 (0.5 %) | 0.04702 |
| **Vómitos** | 461 (0.8 %) | 441 (0.8 %) | 20 (1.9 %) | 0.009734 |
| **Fiebre** | 2,504 (4.6 %) | 2,366 (4.4 %) | 138 (13.4 %) | 2.2e-16 |

Tabla 4.25: Síntomas al ingreso del seguimiento

### 4.3.3. Perspectiva general del estudio

A modo de síntesis, en la Figura 4.10 se presenta un diagrama que explica de forma resumida los modelos desarrollados, indicando que datos usa cada uno y que predicciones entregan.

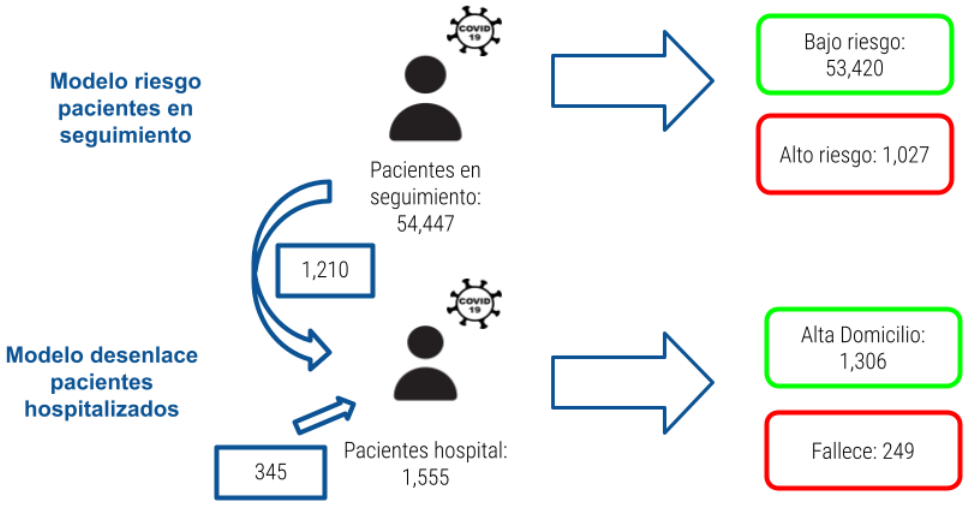



Figura 4.10: Perspectiva general del estudio

El primer modelo para pacientes en seguimiento se entrena y valida con los datos de la plataforma COVID19 y clasifica a los pacientes al momento de ingresar en alto o bajo riesgo de enfermar gravemente. El segundo modelo utiliza los datos de la gestión hospitalaria y clasifica a los pacientes en fallece y alta domicilio (o no fallece).

## 4.4. Fase de modelado

De acuerdo a lo mencionado anteriormente, se plantean 2 modelos, los cuales se entrenan a partir de distintos enfoques o técnicas de machine learning. El primer modelo (riesgo hospital) se desarrolla mediante algoritmos de machine learning clásicos usando el paquete CARET del software estadístico R y para el segundo modelo utilizamos el algoritmo "Optimal Tree" [].

### 4.4.1. Machine Learning para predecir riesgo de fallecer en pacientes hospitalizados por COVID-19

#### Introducción

Los modelos de Machine Learning proveen un apoyo sólido basado en datos para los profesionales de la salud, sobretodo cuando la demanda hospitalaria crece y la capacidad de respuesta de los servicios de salud se mantiene. Este modelo busca apoyar a los profesionales de salud en el manejo de pacientes una vez ingresan a un establecimiento hospitalario, entregando una clasificación de riesgo según la probabilidad de fallecer de cada paciente apenas estos hacen su ingreso al hospital. El modelo busca predecir el riesgo de fallecer de pacientes contagiados o probables de COVID-19 al momento que ingresan al hospital, usando como variables las características sociodemográficas, comorbilidades y la información recolectada al momento de la admisión hospitalaria.

#### Entrenamiento y selección del algoritmo

Dada la naturaleza del problema la variable de interés *tipo_egreso* presenta un 16 % de casos en donde el paciente fallece, por lo que existe un desbalance en la variable dependiente del modelo a entrenar. Esto implica que un predictor que clasifique a todos los pacientes como Alta domicilio tendrá en promedio un Accuracy de 84 %, lo que hace de esta medida una mala aproximación del desempeño del modelo predictivo. Por esto, para medir el desempeño del modelo, se proponen las métricas AUC, Specificity y PPV, cuyas definiciones se encuentran en el capítulo 2.7 del Marco Conceptual. También se verifica la correlación entre las variables explicativas a partir de una matriz de correlación (Apendíce C).

Los modelos se entrenan usando un 85 % de los datos y posteriormente se validan los resultados en el 15 % restante. Para el problema de desbalance de datos se utiliza la técnica Under-Sample. Además el entrenamiento se realiza ocupando la técnica de validación cruzada "Repetead K-Folds Cross Validation" con k=10 y n=10. Es decir, se realiza una partición aleatoria de los datos de entrenamiento y se deja el resto de los datos para testeo, proceso iterativo que se realiza 10 veces con distintas particiones y luego este proceso se repite 10 veces por completo para finalmente entregar un promedio de las métricas de desempeño de los modelos.

Cabe destacar que dado que los modelos basados en árboles incorporan en el algoritmo

la selección de variables, se consideran todas las variables disponibles en el entrenamiento de los modelos. Los resultados de los algoritmos probados se presentan en la Tabla 4.26.

| Algoritmo | AUC | Accuracy | Sensitivity | Specificity | Precision (PPV) |
|---|---|---|---|---|---|
| CART | 0.74 | 0.69 | 0.81 | 0.66 | 0.95 |
| RF | 0.72 | 0.70 | 0.75 | 0.69 | 0.94 |
| GMB | 0.75 | 0.71 | 0.81 | 0.69 | 0.95 |
| XGBoost | 0.77 | 0.74 | 0.84 | 0.72 | 0.96 |
| Optimal Tree | 0.77 | 0.61 | 0.95 | 0.54 | 0.67 |

Tabla 4.26: Desempeño algoritmos modelo hospital

**Resultados**

Para construir el modelo se utiliza el algoritmo XGBoost, el cual se basa en la técnica de Boosting. Si bien este algoritmo es el con mejor desempeño, al ser un algoritmo de ensamble (compuesto de varios árboles) pierde interpretabilidad. En la Figura 4.11 se presenta la curva ROC del modelo, la cual tiene un AUC de 0.77, esto nos dice que con una probabilidad de 77 %, el modelo podrá diferenciar a un paciente que fallecerá de uno que no.

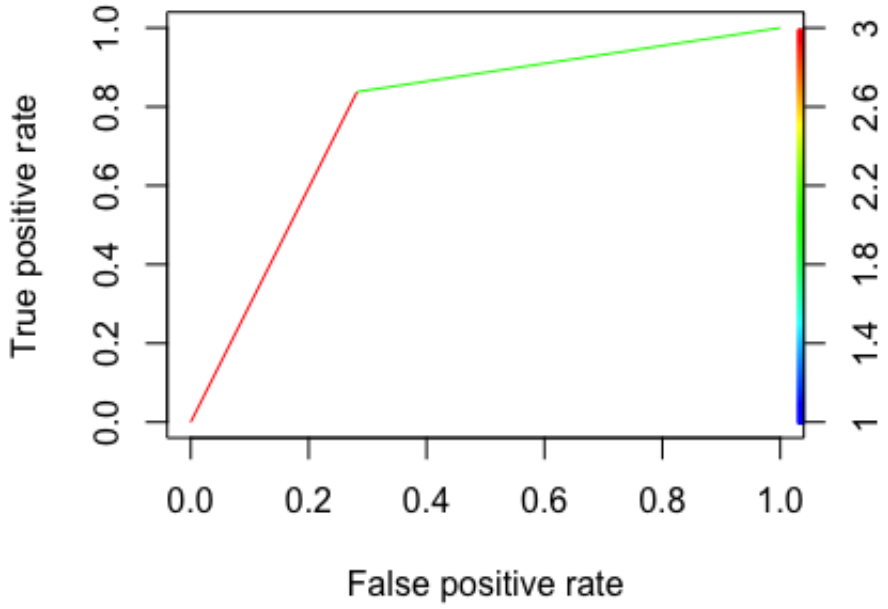



Figura 4.11: Curva ROC Modelo XGBoost

El modelo XGBoost nos indica las variables mas relevantes a la hora de clasificar pacientes según su riesgo de fallecer son la edad, el hecho de ingresar o no conectado a un ventilador

mecánico (vm_ingreso), pertenecer al tramo FONASA B, hospitalizarse en un establecimiento público y presentar antecedentes renales.

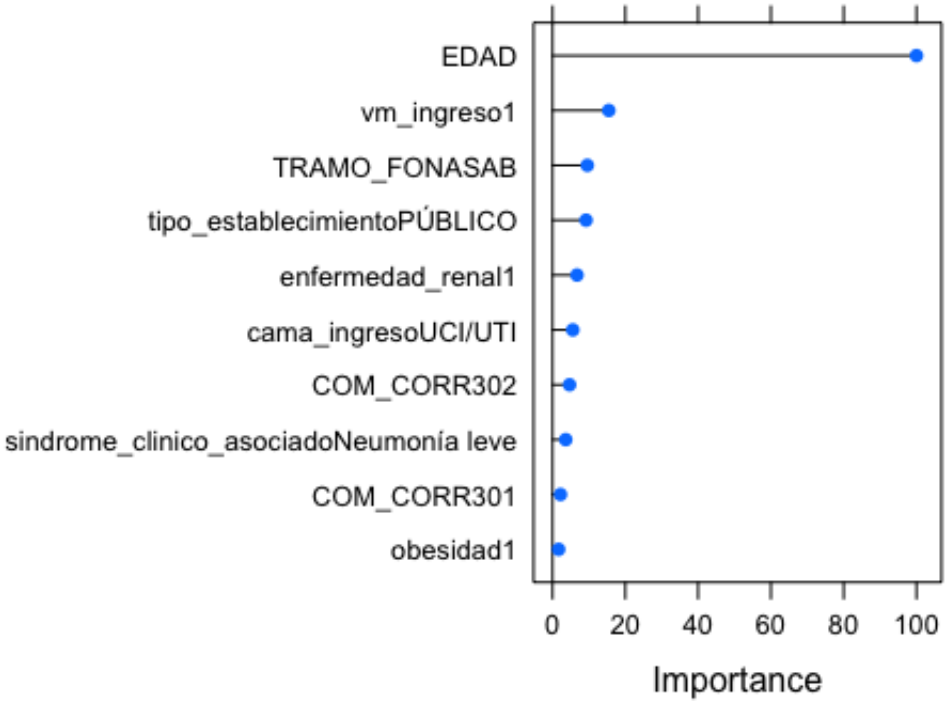



Figura 4.12: Importancia variables modelo XGBoost

Se puede notar que el modelo CART es el que presenta peor performance en cuanto a accuracy, sin embargo, la métrica sensitivity es de 0.81, a diferencia de los otros modelos el árbol es el único interpretable visualmente y esto da luces de algunos criterios de clasificación.

Podemos interpretar el árbol de la Figura 4.13 de la siguiente manera:

- Del primer nodo, pacientes menores a 48 años de edad presentan un 84 % de probabilidad de egresar con alta al domicilio
- Luego del nodo 3, pacientes con edad mayor a 48 presentan un 64 % de probabilidad de fallecer
- Pacientes mayores a 48 que ingresan a cama UCI/UTI presentan una probabilidad de un 74 % de fallecer (nodo 7), en cambio si ingresan en una cama básica presentan un 50 % (nodo 6)
- Para los pacientes mayores a 48 que ingresaron a cama UCI/UTI y que además tienen mas de 60 años presentan una probabilidad de fallecer de un 82 % (nodo 15), vs aquellos pacientes que presentan una edad entre 48 y 58 los cuales fallecen con una probabilidad de un 72 % (nodo 29). Se puede observar que los pacientes con edad entre 58 y 60 años presentan una probabilidad de fallecer de un 37 % (nodo 14), lo que probablemente se deba a la composición de nuestra muestra en especifico, puesto que en términos

médicos no hace mucho sentido que pacientes de estas edades en particular tengan un mejor desenlace de la enfermedad.

- Para aquellos pacientes que ingresaron a cama básica, si tienen edad menor a 61 años presentan una probabilidad de un 27 % de fallecer (nodo 12), en cambio si la edad es mayor a este numero la probabilidad aumenta a un 62 % (nodo 13).
- Finalmente notamos que pacientes que ingresan en cama básica, tienen una edad mayor a 61 años y además se hospitalizan en un establecimiento privado, tienen una probabilidad fallecer de un 22 % (nodo 26) vs los mismos pacientes que se hospitalizan en un establecimiento público, donde la probabilidad de fallecer aumenta a un 65 % (nodo 27)

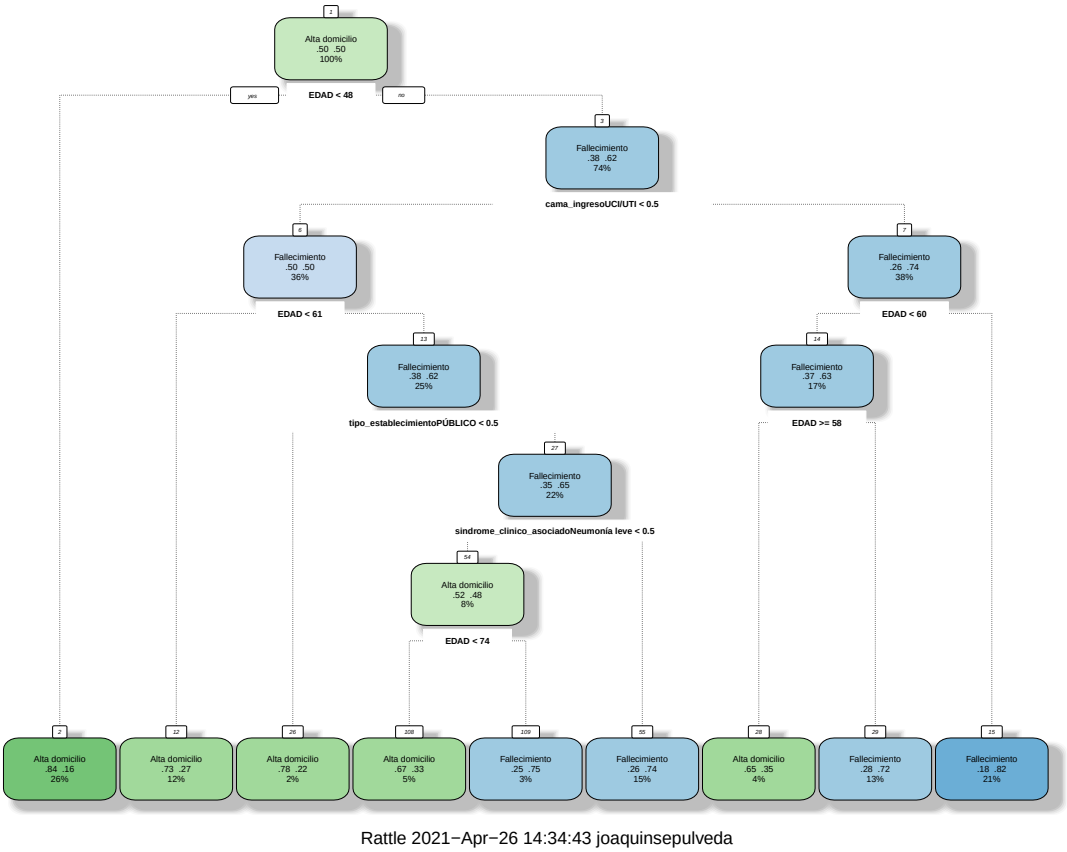



Figura 4.13: Árbol de clasificación (CART)

### 4.4.2. Machine Learning para predecir la severidad de la enfermedad en pacientes COVID-19

#### Introducción

Este modelo tiene por objetivo clasificar a los pacientes que ingresan al seguimiento en alto o bajo riesgo según su probabilidad de enfermar gravemente. Enfermar gravemente corresponde a fallecer, necesitar de una cama critica UCI o UTI y/o requerir de ventilación mecánica. Este modelo se entrena con las variables sociodemograficas de los pacientes, sus comorbilidades y los síntomas presentados en sus dos primeros seguimientos.

#### Entrenamiento y selección del algoritmo

Dada la naturaleza del problema la variable de interés *riesgocovid* presenta un 1.8 % de casos en donde el paciente resulta ser de alto riesgo, por lo que existe un desbalance en la variable dependiente del modelo a entrenar. Esto implica que un predictor que clasifique a todos los pacientes como bajo riesgo tendrá en promedio un accuracy de 98 %, lo que hace de esta medida una mala aproximación del desempeño del modelo predictivo. Por esto, al igual que en el primer modelo, se proponen las métricas AUC, specificity y PPV.

Los modelos se entrenan usando un 70 % de los datos y posteriormente se validan los resultados en el 30 % restante. Para el problema de desbalance de datos se utiliza la técnica "auto-balance" incluida en el algoritmo de optimal tree, este ajuste le da mas peso a la clase con menos registros de tal forma de poder clasificar sin perder ni generar registros extras.

| Algoritmo | AUC | Accuracy | Sensitivity | Specificity | Precision (PPV) |
|---|---|---|---|---|---|
| CART | 0.80 | 0.80 | 0.81 | 0.80 | 0.07 |
| RF | 0.80 | 0.80 | 0.81 | 0.79 | 0.07 |
| GMB | 0.82 | 0.80 | 0.85 | 0.79 | 0.07 |
| XGBoost | 0.81 | 0.84 | 0.84 | 0.81 | 0.08 |
| Optimal Tree | 0.87 | 0.74 | 0.87 | 0.74 | 0.12 |

Tabla 4.27: Desempeño algoritmos modelo seguimiento

El algoritmo con mejor desempeño en este caso es el Optimal Tree con un AUC de 0.87 y una sensitivity de 0.87, recordemos que, este algoritmo a diferencia de las otras técnicas empleadas selecciona el óptima global de todas las combinaciones de arboles posibles, en cambio otras heurísticas como Random Forest o Gradient Boosting, buscan los mejores óptimos locales y toman la mejor decisión ramificación por ramificación.

Destacamos el hecho de que este modelo al tener clases tan desbalanceadas al momento de testear con el set de validación se tiene un PPV (Positive Predictive Value) bajo. Esto debido a que en el entrenamiento las clases tienen el mismo peso, pero en la validación no, por lo que se clasifican como clase de alto riesgo a muchos pacientes que en realidad no son de alto riesgo.

**Resultados**

En la Figura 4.14 se presenta la curva ROC del modelo, la cual tiene un AUC de 0.87, esto nos dice que con una probabilidad de 87 %, el modelo podrá diferenciar a un paciente de alto riesgo de uno de bajo riesgo.

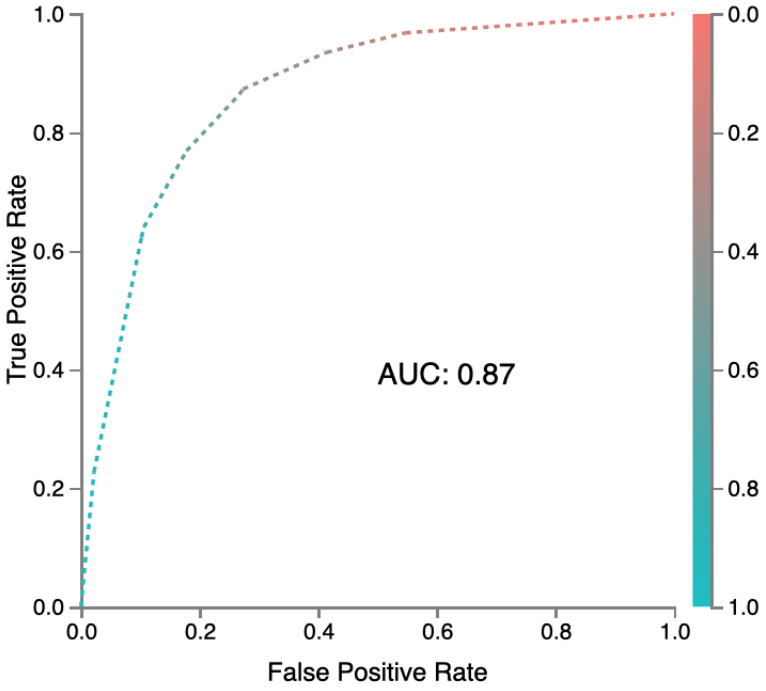



Figura 4.14: AUC árbol óptimo

En la Figura 4.15 se encuentra el árbol óptimo resultado de el entrenamiento de los datos, las ramificaciones en cada nodo se definen bajo el Índice de Gini:

$$gini = \sum_{k} p_k(1 - p_k) \tag{4.4}$$

Es decir, se suma, para todas las clases, el producto de su proporción por 1 menos su proporción y a mayor valor de este índice, mayor es la homogeneidad de la población en un determinado nodo.

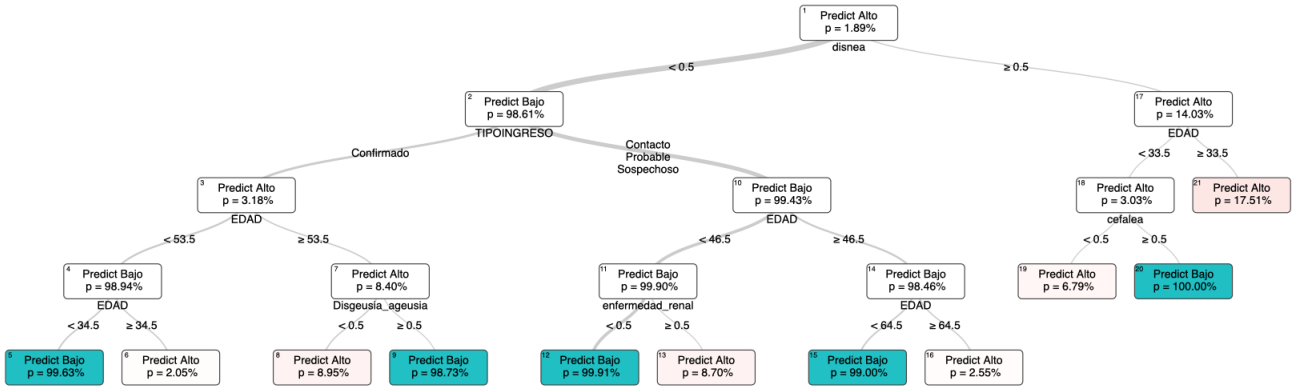



Figura 4.15: Optimal Tree

Podemos interpretar el árbol de la Figura 4.15 de la siguiente manera:

- De la primera ramificación de la derecha, pacientes que presentan síntoma disnea se clasifican con alto riesgo, aumentando la probabilidad cuando estos tienen una edad mayor a los 33 años.
- En caso de presentar disnea, ser menor a 33 años y presentar cefalea, el paciente se predice como bajo riesgo lo que nos indica que el síntoma cefalea sirve para identificar a pacientes de bajo riesgo, puesto que en el caso de no presentar este síntoma se sigue clasificando como de riesgo alto
- De la primera ramificación a la izquierda, pacientes se separan según el tipo de ingreso entre los confirmados y los probables, sospechosos y contactos.
- Pacientes confirmados con edad menor a 34 se clasifican como bajo riesgo, entre 34 y 54 se clasifica como alto riesgo, mayores a 54 que no presentan el síntoma disgeusia se clasifican como alto riesgo y lo que presentan este síntoma como bajo riesgo, lo que indica que este síntoma también lo usa el modelo para clasificar a los pacientes como bajo riesgo.
- Luego del resto de pacientes (probables, sospechosos y contactos), pacientes mayores a 64 se clasifican como alto riesgo y pacientes con edades entre 46 y 64 se clasifican como bajo riesgo. De los pacientes menores a 46 años, aquellos sin enfermedades renales se clasifican como bajo riesgo y aquellos que si presentan esta comorbilidad se clasifican como alto riesgo.

Finalmente comparamos los resultados de las predicciones con las clasificaciones entregadas por el personal médico encargado de realizar el seguimiento, para esto debido a que se utiliza la información de los dos primeros seguimientos y la clasificación entregada comprende de 3 niveles (alto, medio y bajo), promediamos las clasificaciones y separamos a los pacientes según las divisiones de la siguiente matriz de confusión, donde en vez de comparar con las predicciones del modelo, se compara la clasificación promedio con el riesgo real del paciente.

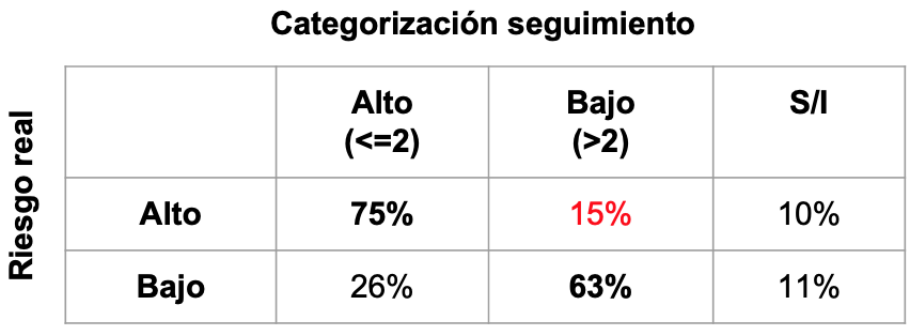



Figura 4.16: Categorización personal médico versus riesgo real

Notamos que de los pacientes que terminaron siendo de alto riesgo, la categorización de riesgo que entrega el personal médico identifica a un 75 % de los pacientes de alto riesgo versus el 87 % que predice de forma correcta el modelo. En el caso de clasificar a los pacientes como bajo riesgo, el modelo tiene una precisión de un 74 % versus el 63 % de los pacientes que se clasificaron como bajo riesgo según la categoria entregada por el personal de salud.

## 4.5. Fase de evaluación y despliegue

Finalmente, en esta fase se especifican los requerimientos y preparación de datos con el objetivo de poder integrar los modelos dentro de los procesos de toma de decisiones de la organización. Además se evalúan las predicciones de los modelos con nuevos datos para verificar que estos tienen un buen desempeño en datos que no se han usado en la construcción del modelo.

Para esta etapa solo se considera la evaluación y despliegue del primer modelo, esto debido a que otro integrante del equipo de analítica desarrollo un modelo de categorización de riesgo en seguimiento el cual en términos de usabilidad cumple con el mismo objetivo que el modelo de categorización desarrollado en este trabajo.

### 4.5.1. Evaluación con nuevos datos

Para validar el desempeño del modelo utilizamos la información de los pacientes ingresados a un establecimiento hospitalario del SSMSO entre el 24 de mayo y el 21 de junio del 2021. El desempeño del modelo con estos nuevos datos es el siguiente:

- Accuracy: 79 %
- Sensitivity: 83 %
- Specificity: 79 %

En la Figura 4.17 se observa el desempeño del modelo con los nuevos datos:

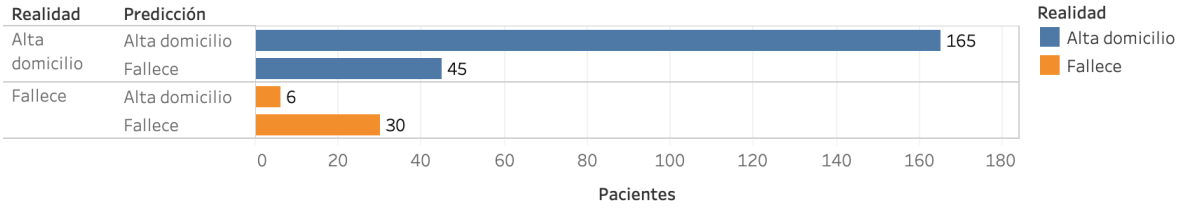



Figura 4.17: Evaluación modelo con nuevos datos

Uno de los objetivos del trabajo es poder apoyar con la gestión de los pacientes de la gestión hospitalaria. Para esto se propone que los modelos sean ejecutados una vez cada dos semanas y los resultados sean incluidos en el plan de monitoreo que realiza el servicio cada 15 días. Se propone que los resultados sean desagregados a nivel comuna o a nivel establecimiento, tal como se puede apreciar en la figura 4.18.

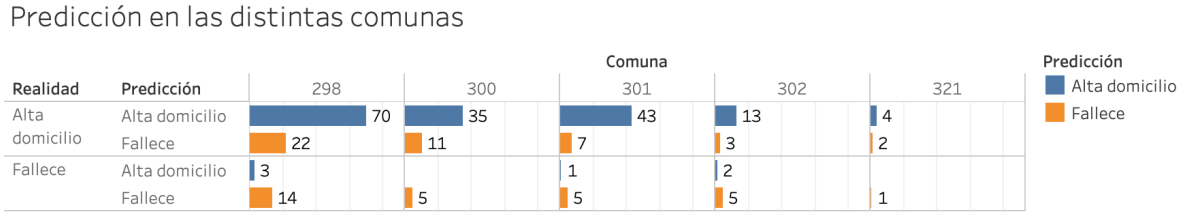



Figura 4.18: Predicción modelo por comuna

### 4.5.2. Diseño de la aplicación

Finalmente se desarrolla una aplicación web en donde se despliega el modelo para que usuarios encargados de ingresar la información de la admisión hospitalaria puedan obtener la clasificación de los pacientes al momento que estos ingresan. El diseño de la aplicación se puede observar en la Figura C.2 del Apéndice. Para que la aplicación fuera simple de usar, se seleccionan las 10 variables mas relevantes del modelo (Figura 4.12), ya que esto disminuye las métricas de desempeño en: 2 % en accuracy (0.74 a 0.72), 3 % en sensitivity (0.84 a 0.81) y un 1 % en specificity (0.72 a 0.71).

# Capítulo 5

# Conclusiones

## 5.1. Resumen del trabajo realizado

En el desarrollo del trabajo realizado fueron abordados dos problemáticas que surgieron con la llegada del coronavirus: el seguimiento domiciliario de los pacientes confirmados, probables, sospechosos y contactos de contagiados y la gestión de pacientes ingresados al hospital producto de la enfermedad. Se propone apoyar con estas problemáticas con modelos de predicción basados en machine learning, con el objetivo de identificar anticipadamente a los pacientes en riesgo para así apoyar con la gestión de pacientes, liberando tiempo y recursos del servicio. Tras finalizar el trabajo, se propone utilizar los modelos integrándolos en los distintos sistemas del servicio. Específicamente el modelo de predicción de riesgo de fallecer para pacientes hospitalizados obtuvo un AUC de 0.77 y Sensitivity de 0.81, y el modelo de predicción de riesgo de enfermar gravemente obtuvo un AUC de 0.87 y Sensitivity de 0.87. Nos interesan estas dos métricas puesto que la primera se interpreta como qué tan bien el modelo está prediciendo en general y con qué probabilidad logra diferenciar a un paciente en riesgo de uno que no y la Sensitivity cuantos de los pacientes que realmente estuvieron en riesgo fueron clasificados como pacientes en riesgo por el modelo.

Además del desarrollo de los modelos se realiza un análisis descriptivo y junto con el uso del parámetro P-Valor se logran determinar diferencias significativas con relación a las variables registradas en ambos modelos entre los pacientes de alto y bajo riesgo. También se desarrolla una aplicación web y se propone una forma de evaluar y utilizar los modelos para llevar a cabo el monitoreo de los pacientes.

Finalmente se prueba el modelo de pacientes hospitalizados con nuevos datos aparte de los usados para el entrenamiento y validación para conocer el error real del modelo. En esta experiencia, los resultados fueron positivos con un Accuracy de 0.79, Sensitivity de 0.83 y Specificity de 0.79. El personal de salud del servicio evalúa el desempeño del modelo de forma positiva y se propone entregar los resultados semana a semana separando por las distintas comunas del SSMSO.

## 5.2. Principales conclusiones

A partir de los resultados obtenidos se puede concluir que los modelos obtienen un buen desempeño a la hora de clasificar pacientes según su nivel de riesgo de enfermar gravemente para el caso de los pacientes que ingresan al seguimiento y en el riesgo de fallecer a la hora de que los pacientes ingresan al hospital.

Las variables relevantes en los modelos conversan con la experiencia internacional, presentar disnea en los primeros seguimientos, tener de comorbilidad previa alguna enfermedad renal y presentar una edad avanzada sirven para diferenciar a los pacientes de alto riesgo de los de bajo riesgo, además la presencia de síntomas como la cefalea y la disgeusia servirían de indicio para pensar que se podría tratar de un paciente de bajo riesgo. Para el caso de los pacientes hospitalizados, variables socioeconómicas como el plan de salud y el tipo de establecimiento (hospital publico o privado) son las que según el modelo mas pesan a la hora de clasificar a los pacientes según el riesgo de fallecer producto de la enfermedad.

Para el caso de la gestión de pacientes hospitalizados se proponen dos formas de utilizar los modelos. La primera es, a partir de la aplicación web, predecir el riesgo de los pacientes al momento que estos ingresan al hospital, para esto se necesita integrar la aplicación a los sistemas tecnológicos de la gestión hospitalaria. La segunda es evaluar el riesgo de los pacientes de manera periódica (semanal, quincenal o mensual) para añadir los resultados en el plan de monitoreo del servicio, tal de tener una visión general a nivel comuna o por establecimiento del riesgo de los pacientes que se encuentran en observación. Para el modelo de seguimiento domiciliario, se propone que estos sean usados para complementar las categorizaciones de riesgo entregadas por los funcionarios de salud encargados de realizar esta labor.

Finalmente destacar que la información correspondiente a los síntomas y a los antecedentes de esta base en su mayoría es autodeclarada por los mismos pacientes, esto quiere decir que los síntomas pueden estar sobre reaccionados o puede que no se declaren en su totalidad. Con los antecedentes relacionados al cuidado de salud y comorbilidades sucede el mismo fenómeno, por ejemplo cuando una persona se declara obesa, este no se mide con el IMC sino que es la percepción que tiene la persona que se declara como obesa. Similar pasa cuando se le pregunta por alcohol, tabaco y drogas.

## 5.3. Trabajo futuro

### 5.3.1. Integrar los modelos con los sistemas del SSMSO

Tal como se menciono en los alcances del trabajo, no se considera la integración de los modelos con los sistemas de salud. Integrar los modelos con la plataforma de seguimiento y específicamente con el sistema de respuesta de voz interactiva (IVR) para liberar recursos y automatizar el seguimiento de pacientes de bajo riesgo.

### 5.3.2. Vacunas

También se propone añadir el factor vacunas como variable explicativa de los modelos. El servicio hace entrega de datos de pacientes vacunados pero debido a que gran parte de esta muestra consideraba a pacientes que se habían inoculado en un periodo posterior a su paso en el seguimiento domiciliario o en su periodo de hospitalización, se decide no incluirla en este momento al modelo. Sin embargo, se añade un análisis de los pacientes vacunados en el Apéndice D.

### 5.3.3. Modelo tiempo en estadía

Finalmente además de las clasificaciones de riesgo se propone implementar modelos de regresión para estimar el tiempo de estadía en el hospital de los pacientes al momento de ingresar, con el fin de poder anticiparse a la demanda de camas y poder preparar la disposición de esta con semanas de anticipación. Se desarrolla un modelo de regresión para llevar a cabo esta idea con las variables de la admisión hospitalaria sin embargo probando diferentes algoritmos los modelos obtienen un error en la predicción del orden de 6-8 días por lo que se propone que sean desarrollados una vez se tengan mas registros para entrenar los diferentes algoritmos.

# Bibliografía


[1] Francisco Javier Díaz-Castrillón and Ana Isabel Toro-Montoya. Sars-cov-2/covid-19: The virus, the disease and the pandemic. *Medicina & laboratorio*, 24(3):183–205, 2021.

[2] Gobierno digital: Cifras oficiales covid-19. `https://www.gob.cl/coronavirus/cifrasoficiales/`, Marzo 2020. Accedido en Abril de 2021.

[3] Revista pauta: El perfil de los fallecidos por covid-19. `https://www.pauta.cl/nacional/el-perfil-de-los-fallecidos-por-covid-19-en-chile-actualizado`, Marzo 2020. Accedido en Abril de 2021.

[4] Sitio ssmso: Población ssmso. `http://estadistica.ssmso.cl/poblacion-ssmso/`, Marzo 2020. Accedido en Diciembre de 2020.

[5] Sitio ssmso: Usd. `https://saluddigital.ssmso.cl/`, Marzo 2020. Accedido en Diciembre de 2020.

[6] Sitio SSMSO. Covid ssmso. `https://covid.ssmso.cl/`, Marzo 2020.

[7] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[8] Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. *Classification and regression trees*. Routledge, 2017.

[9] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[10] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[11] Jerome H Friedman. Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378, 2002.

[12] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4, 2015.

[13] Dimitris Bertsimas and Jack Dunn. Optimal classification trees. *Machine Learning*, 106(7):1039–1082, 2017.

[14] Health data miner: Crisp-dm. https://healthdataminer.com/data-mining/crisp-dm-una-metodologia-para-mineria-de-datos-en-salud/.

[15] Farha Musharrat Noor and Md Momin Islam. Prevalence and associated risk factors of mortality among covid-19 patients: a meta-analysis. *Journal of community health*, 45(6):1270–1282, 2020.

[16] Li Zhang, Jie Hou, Fu-Zhe Ma, Jia Li, Shuai Xue, and Zhong-Gao Xu. The common risk factors for progression and mortality in covid-19 patients: a meta-analysis. *Archives of virology*, pages 1–17, 2021.

[17] You Li, Thulani Ashcroft, Alexandria Chung, Izzie Dighero, Marshall Dozier, Margaret Horne, Emilie McSwiggan, Azwa Shamsuddin, and Harish Nair. Risk factors for poor outcomes in hospitalised covid-19 patients: A systematic review and meta-analysis. *Journal of global health*, pages 1–11, 2021.

[18] Dhruv Patel, Vikram Kher, Bhushan Desai, Xiaomeng Lei, Steven Cen, Neha Nanda, Ali Gholamrezanezhad, Vinay Duddalwar, Bino Varghese, and Assad A Oberai. Machine learning based predictors for covid-19 disease severity. *Scientific Reports*, 11(1):1–7, 2021.

[19] Dimitris Bertsimas, Galit Lukin, Luca Mingardi, Omid Nohadani, Agni Orfanoudaki, Bartolomeo Stellato, Holly Wiberg, Sara Gonzalez-Garcia, Carlos Luis Parra-Calderon, Kenneth Robinson, et al. Covid-19 mortality risk assessment: An international multi-center study. *PloS one*, 15(12):e0243262, 2020.

[20] Li Yan, H Zhang, Jorge Goncalves, Yang Xiao, Maolin Wang, Yuqi Guo, Chuan Sun, Xiuchuan Tang, Liang Jin, Mingyang Zhang, et al. A machine learning-based model for survival prediction in patients with severe covid-19 infection. *Europe PMC*, 2020.

[21] Protocolo de seguimiento de casos y contactos covid 19 adulto e infantil. http://www.ancorauc.cl/2020/06/08/protocolo-de-seguimiento-de-casos-y-contactos-covid-19-adulto-e-infantil-2/.

[22] Machine learning, qué es y cómo funciona. https://www.bbva.com/es/machine-learning-que-es-y-como-funciona/. Accedido en Junio de 2021.

[23] Cómo desarrollar un modelo de machine learning desde cero. https://medium.com/datos-y-ciencia/machine-learning-c %C3 %B3mo-desarrollar-un-modelo-desde-cero-cc17654f0d48. Accedido en Junio de 2021.

[24] Validación de modelos predictivos. https://www.cienciadedatos.net/documentos/30$_{c}ross-validation_{o}neleaveout_{b}ootstrap. AccedidoenJuniode$2021.

# Apéndice A

# Algoritmos utilizados

En este apartado se explican en detalle los algoritmos de Machine learning que fueron utilizados para entrenar y validar los modelos desarrollados para este trabajo.

## A.1. Bagging

En el caso de algoritmos de clasificación, supongamos h(d,x) es un clasificador, por ejemplo un árbol, basado en el set de data D, produciendo una predicción dado el input x. Para hacer bagging sobre d, se generan sub-muestras mediante bootstraping de D (d1,...,dM), cada una de tamaño N, entonces

$$\hat{H}_{bag}(x) = VotoMayoritario\{h(d_m, x)\}_{m=1}^{M} \tag{A.1}$$

Este método puede reducir la varianza de predictores inestables como los árboles, pero se pierden las propiedades de interpretabilidad de estos.

| Algoritmo Bagging | |
|---|---|
| **Input** | M=Número de clasificadores débiles<br>$h = \{h_1...h_M\}$ Conjunto de clasificadores débiles.<br>$N$ Número de registros del set de datos utilizados para ajuste (entrenamiento)<br>$N'$ <N, Número de ejemplos generados por bootsrapping<br>d $\subset D$ Conjunto de entrenamiento<br>$g()$ Función de ensamble |
| **Output** | Modelo de ensamble $H_{bag}(X)$.<br>**para** $m = 1...M$<br>Generar conjunto $d_m$ seleccionando $N'$ muestras por reemplazo desde d<br>Entrenar clasificador $h_m$ sobre el subset $d_m$<br>Agregar $h_m$ al ensamble<br>**fin**<br>devolver $\hat{H}_{bag}(x) = g(h_1(x), ..., h_M(x))$ |

Figura A.1: Algoritmo Bagging

El algoritmo Bagging sigue los siguientes pasos:

1. Divide el set de entrenamiento en distintos subsets de datos, obteniendo como resultado diferentes muestras aleatorias con las siguientes características:
   - Muestra uniforme
   - Muestra con reemplazo (los individuos se pueden repetir en el mismo set de datos) o El tamaño de la muestra es igual al tamaño del set de entrenamiento, pero no contiene a todos los individuos ya que algunos se repiten
   - Si se usan muestras sin reemplazo, suele elegirse el 50 % de los datos como tamaño de muestra
2. Luego se crea un modelo predictivo con cada set, obteniendo modelos diferentes
3. Finalmente se construye o ensambla un único modelo predictivo, que es el promedio de todos los modelos

## A.2. Random Forest

El algoritmo Random Forest sigue los siguientes pasos:

1. Seleccionamos k features (columnas) de las m totales (siendo k menor a m) y creamos un árbol de decisión con esas k características.
2. Creamos n árboles variando siempre la cantidad de k features y también podríamos variar la cantidad de muestras que pasamos a esos árboles (esto es conocido como "bootstrap sample")
3. Guardamos el resultado de cada árbol obteniendo n salidas.
4. Se promedian los resultados de cada árbol para obtener la predicción final.

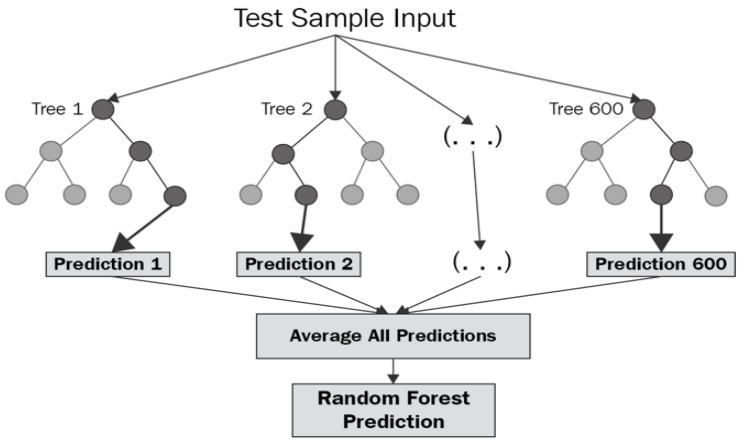



Figura A.2: Random Forest

## A.3. Boosting

El problema del aprendizaje predictivo puede ser caracterizado por un vector de variables explicativas o predictoras de un fenómeno x = {x1,...xm} y un vector de variable dependiente y. Dada una colección de M instancias {yi,xi} con i = 1..M observadas, el objetivo es usar esta data para lograr estimar una función de mapeo del vector x en y, de forma de utilizar esta función para predecir instancias donde solo valores de x son observados. Formalmente, se intenta estimar la función $\hat{f}(x) : x \rightarrow y$ que minimiza una función de pérdida L(y,f) sobre la distribución conjunta W de (y, x)

$$\hat{f}(x) = argmin_{f(x)} E_{y,x} L(y, f(x)) \tag{A.2}$$

| Algoritmo Boosting | |
|---|---|
| **Input** | M=Número de clasificadores débiles<br>$h = \{h_1 ... h_M\}$ Conjunto de clasificadores débiles.<br>$N$ Número de registros del set de datos utilizados para entrenamiento<br>d $\subset D$ Conjunto de entrenamiento<br>$g()$ Función de ensamble |
| **Output** | Modelo de ensamble $H_{boost}(X)$.<br>inicializar la distribución $W = \{w_1, ..., w_N\}$ de las muestras<br>**para** $m = 1...M$<br>Entrenar modelo $h_m$ sobre $d_m$ y su distribución $w_m$<br>Evaluar el error $\varepsilon_m \leftarrow P_{x \sim W}(\mathbb{I}(h_m(x) \neq y_i))$<br>Actualizar la distribución $W$ a partir del error $\varepsilon_m$.<br>**fin**<br>devolver $H_{boost}(x) = g(h_1(x), ..., h_M(x))$ |

Figura A.3: Algoritmo Boosting

## A.4. Extreme Gradient Boosting

Durante el entrenamiento, los parámetros de cada modelo débil son ajustados iterativamente tratando de encontrar el mínimo de una función objetivo, que puede ser la proporción de error en la clasificación, el área bajo la curva (AUC), la raíz del error cuadrático medio (RMSE) o alguna otra. Cada modelo es comparado con el anterior. Si un nuevo modelo tiene mejores resultados, entonces se toma este como base para realizar nuevas modificaciones. Si, por el contrario, tiene peores resultados, se regresa al mejor modelo anterior y se modifica ese de una manera diferente. Este proceso se repite hasta llegar a un punto en el que la diferencia entre modelos consecutivos es insignificante, lo cual nos indica que hemos encontrado el mejor modelo posible, o cuando se llega al número de iteraciones máximas definido por el usuario.

# Apéndice B

# Proceso seguimiento plataforma COVID19

En este apartado se encuentran imágenes de la plataforma COVID19 en funcionamiento, se agrega con el propósito de entender como es que se recolectan los datos utilizados para desarrollar los modelos. Además se agregan los criterios utilizados para definir el riesgo de los casos sospechosos y confirmados definidos por la Red de centros Ancora.

## B.1. Ingreso información del paciente

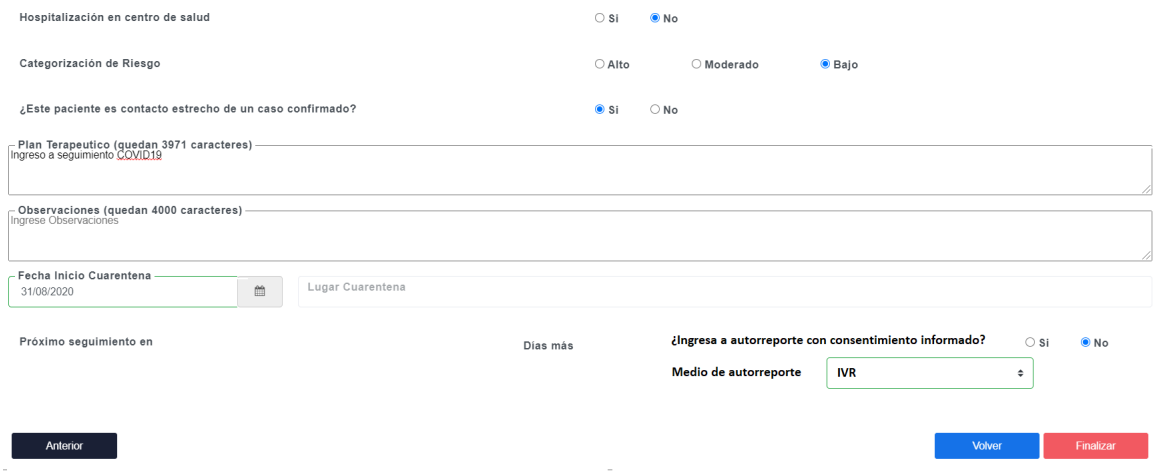



Figura B.1: Ingreso información pacientes plataforma COVID19

En esta imagen se puede observar el formulario que debe de rellenar la persona encargada de realizar el seguimiento. Esta tiene varios campos de opciones múltiples las cuales se usan posteriormente para construir los modelos y respuestas de campo libre para escribir los detalles del seguimiento.

Los antecedentes clínicos se registran por cada ingreso de un nuevo paciente a la plataforma Covid. Como se puede ver en la figura siguiente, se marcan los antecedentes en un checkbox.


Figura B.2: Ingreso de antecedentes

Análogo a los antecedentes, los síntomas se registran en cada uno de los seguimientos, al pinchar la categoría "Otro", se abre un campo de texto libre para añadir por escrito los síntomas presentados que no aparecen en las opciones de la plataforma.

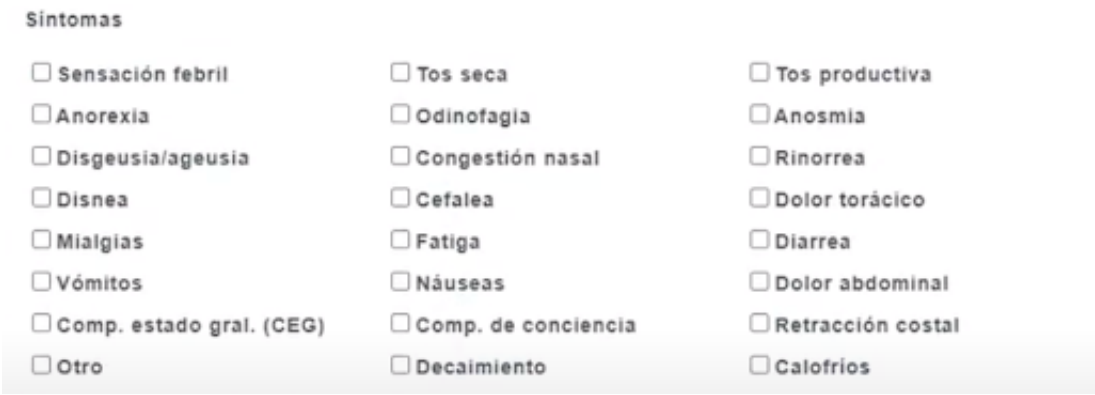



Figura B.3: Ingreso de síntomas

En el siguiente diagrama se observa el proceso del flujo de un paciente Covid desde que ingresa a la plataforma hasta que egresa de esta. En primera instancia, se ingresa al usuario al seguimiento en paralelo a su toma de examen, luego en el transcurso del seguimiento el usuario puede (o no) pasar por el hospital dependiendo de su gravedad y previa recomendación de un profesional (el cual debe ingresar los resultados de los exámenes a la plataforma Epivigilia, plataforma donde se concentra la información respecto al Covid a nivel Nacional), finalmente el paciente egresa del seguimiento, quedando registros de todos sus movimientos para un posterior análisis.


Figura B.4: Diagrama proceso seguimiento usuario Covid

## B.2. Detalle de atención del paciente

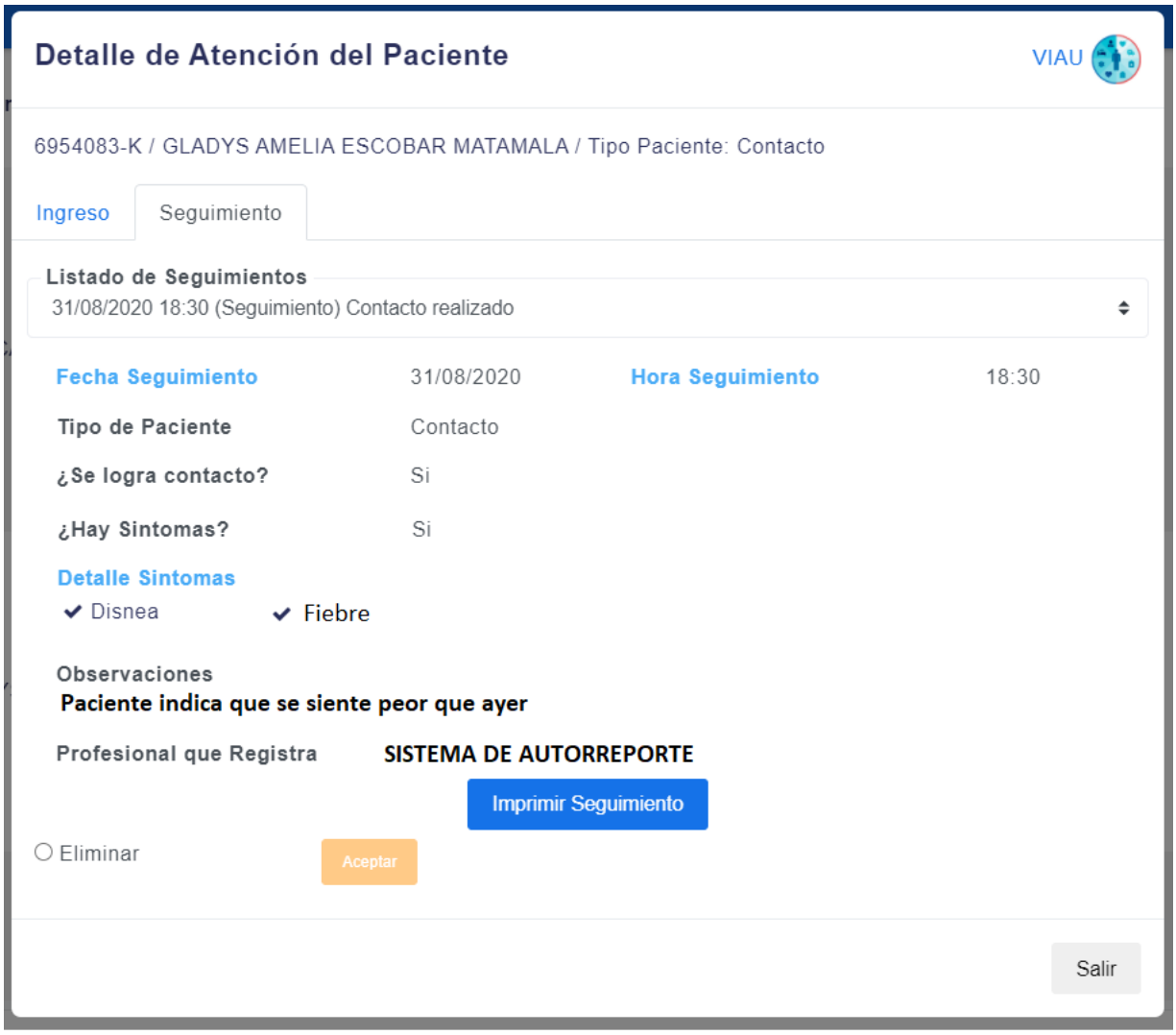



Figura B.5: Detalle de atención del paciente

En esta imagen se puede observar el detalle de atención de un seguimiento en especifico, el personal de salud puede usar esta sección como antesala al próximo seguimiento.

## B.3. Visualización pacientes en seguimiento

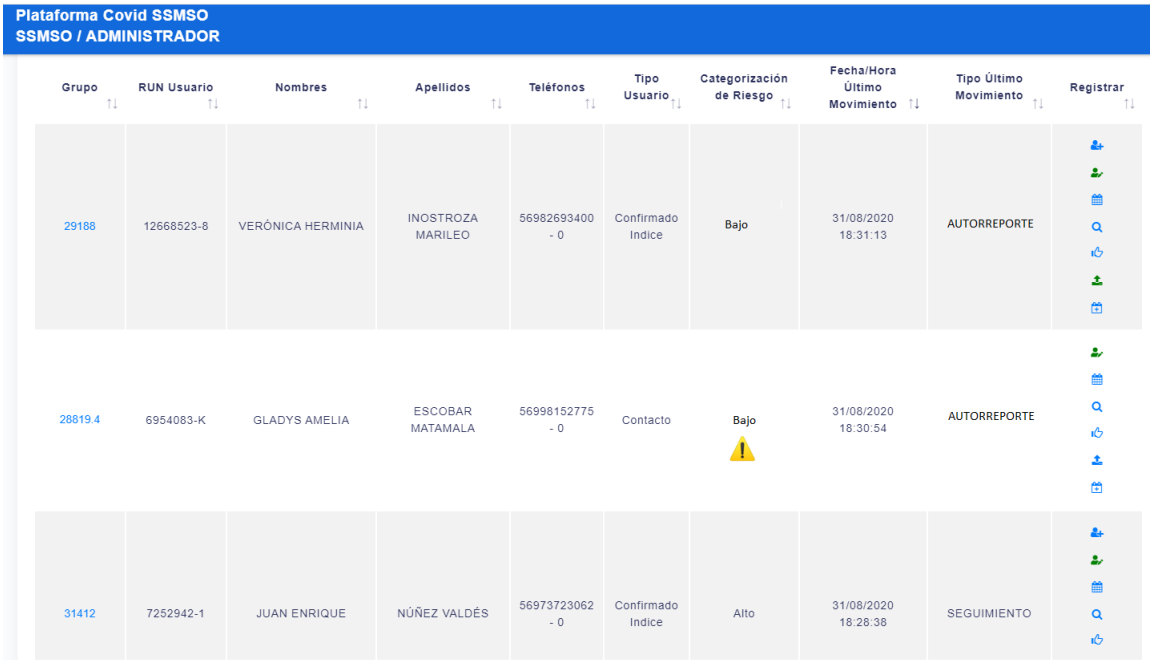



Figura B.6: Visualización pacientes en seguimiento

Aquí se observa el listado de pacientes que se encuentran en seguimiento activo, al pinchar en uno de los pacientes, se puede observar el detalle de los seguimientos de cada uno y ingresar un nuevo seguimiento.

# B.4. Informe de seguimientos

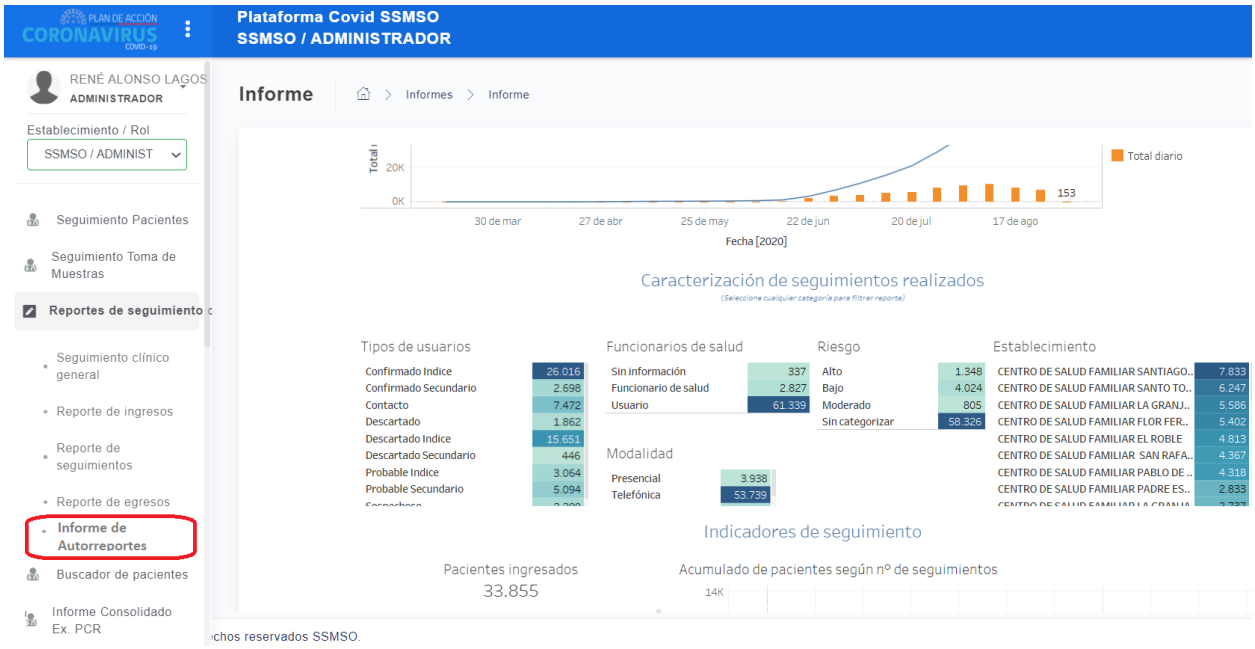



Figura B.7: Informe de seguimientos

Finalmente, observamos el Dashboard en donde el equipo del SSMSO puede analizar la situación actual en cuanto al seguimiento domiciliario de pacientes COVID de manera general. Aquí se puede conocer el numero de usuarios en seguimiento por tipo de ingreso, por establecimiento y por la clasificación de riesgo entregada por el personal médico.

## B.5. Criterios de clasificación de riesgo

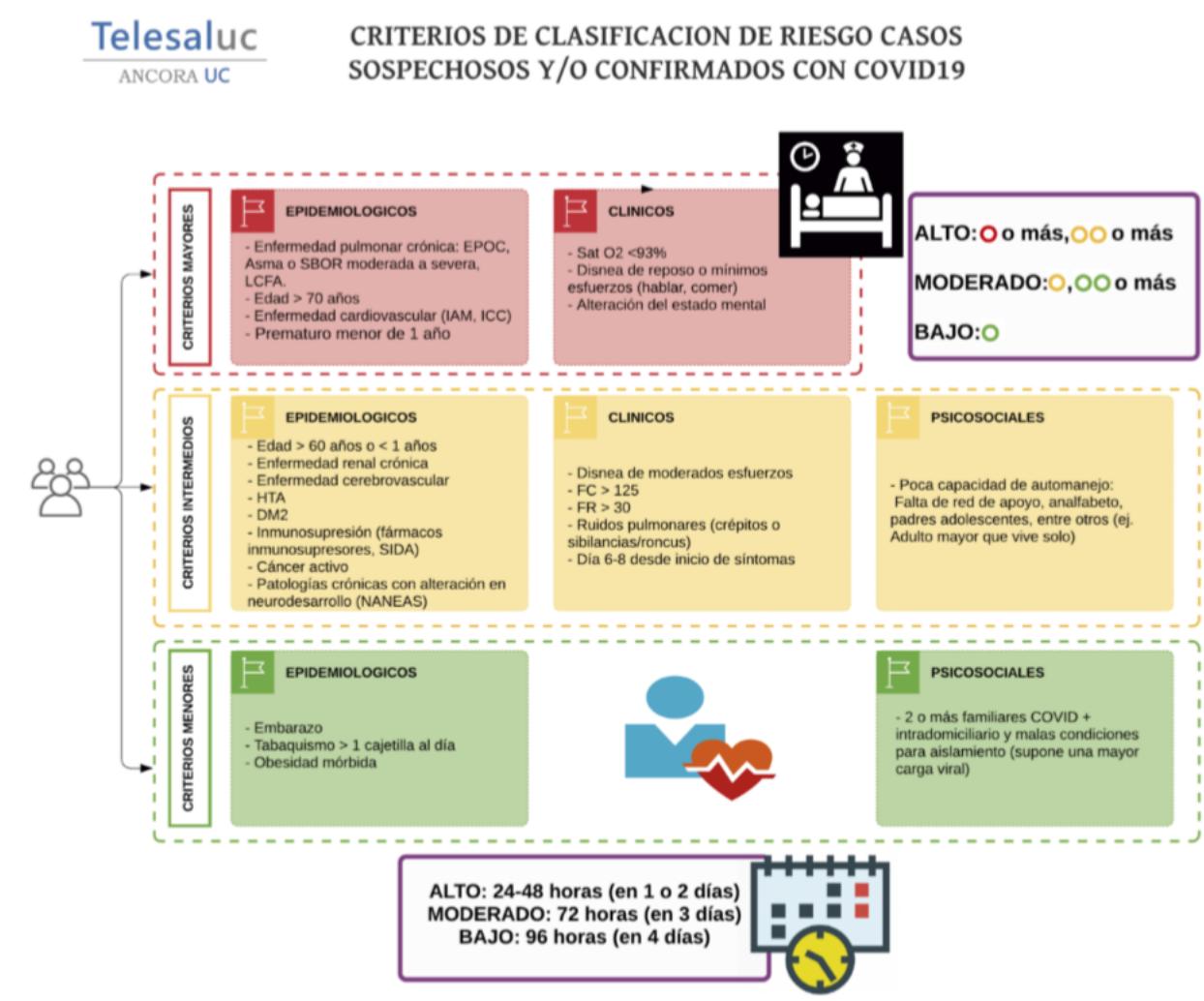



Figura B.8: Criterios de clasificación riesgo Ancora UC

# Apéndice C

## C.1. Glosario

| | |
|---|---|
| AUC | Area under the curve |
| ANID | Asociación Nacional de Investigación y Desarrollo |
| CART | Classification and regression trees |
| DEIS | Departamento de estadísticas e información de salud |
| FP | Falso positivo |
| FN | Falso negativo |
| IVR | Interactive Voice Response |
| MIO | Mixed Interger Operations |
| SDRA | Sindrome de diestrés respiratorio agudo |
| SIDRA | Sistema de información de redes asistenciales |
| SMOTE | Synthetic minority oversampling technique |
| SSMSO | Servicio de salud metropolitano sur oriente |
| UCI | Unidad de cuidados intensivos |
| UTI | Unidad de tratamientos intensivos |
| UGC | Unidad de gestión de camas |
| USD | Unidad de salud digital |
| VN | Verdaderos negativos |
| VP | Verdaderos positivos |
| XGBoost | Extreme gradient boosting |

## C.2. Definiciones síntomas

Con el objetivo de que el lector pueda entender de manera coloquial los síntomas que se registran en el seguimiento, se confecciona un glosario a partir de las definiciones entregadas por un médico general de el Hospital Metropolitano de Santiago.

- Anorexia: Persona sin ganas de comer
- Anosmia: Falta de sensación del olfato
- Calofríos: Sensación de calor y frió
- Cefalea: Cualquier tipo de dolor de cabeza
- Cianosis: Coloración azul de la piel, puede ser distal o perioral (piel y mucosa), relación con baja cantidad de oxigeno en la sangre, síntoma grave
- Compromiso de conciencia: Pueden ser cuantitativos (que tan despierta esta la persona) o cualitativos (Si la persona esta orientada en tiempo y espacio)
- Compromiso estado general. (CEG): 3 síntomas que van en conjunto astenia (falta de fuerza o animo para hacer cosas), adinamia y anorexia
- Congestión nasal: Tener mucosa, en los niños puede ser mas grave
- Decaimiento: Ligado al CEG (prácticamente lo mismo), astenia y adinamia, no se considera síntoma grave por si solo
- Diarrea: Puede ser clasificado en varios tipos según si tiene elementos patológicos (sangre, pus), aumento de la frecuencia de las disposiciones mas acuosas o blandas. Si es muy seguida mas una deshidratación puede ser considerado un síntoma grave. Por si sola lo considera un síntoma leve
- Disgeusia/ageusia: Perdida del gusto o sentir gustos de forma extraña, no se considera síntoma grave
- Disnea: Síntoma importante, requerimientos de oxigeno, sensación de ahogo. Existen distintos grados de disnea (gente con disnea realizando deporte, gente con disnea al momento de levantarse, gente con disnea por vestirse)
- Dolor abdominal: Puede tener características de no gravedad, pero hay otros dolores que pueden ser muy graves (depende del tipo del dolor). Retorcijones se considera poco grave, pero gente con dolor abdominal continuo no tipo retorcijón, en lo general puede ser considerado mas grave.
- Dolor torácico: Siempre se categoriza en la urgencia como síntoma de gravedad, urge hacerle exámenes, imágenes, exámenes físicos. Gran mayoría no tienen nada, pero abajo del dolor torácico puede aparecer patología mas importante como infarto, neumotórax, etc.
- Fatiga: Parecido a decaimiento y CEG, sensación de decaimiento, de no tener ganas de nada, no es un síntoma que se considere grave

- Mialgias: Dolor de musculo, puede ser por deporte, en infecciones virales, la influenza da mucho dolor muscular y arterial (al igual que el covid). Orienta a que una persona pueda tener una enfermedad viral
- Náuseas: Ganas de vomitar sin vomitar
- Odinofagia: Dolor de garganta
- Postración: Ocurre en las personas de mayor edad en general. Incapacidad de levantarse y no moverse por sentirse enfermo, si es muy prolongada puede ser grave
- Retracción costal: Asociado a la disnea, de por si sola no es un síntoma, sino que un signo. Se asocia con disneas de mayor gravedad
- Rinorrea: Muy ligada a la Congestión nasal, cuando los mocos son muy líquidos
- Sensación febril: Cuando el paciente declara que se siente afiebrado sin tomarse la temperatura, fiebre es cuando se toma la temperatura y la temperatura esta sobre $38^{\text{o}}$
- Taquipnea: Relación con la disnea, frecuencia respiratoria alta, persona normal tiene una frecuencia respiratoria de 10 a 20 respiraciones por minuto, una persona con taquipnea tiene sobre 20, grave sobre 30-35.
- Tos productiva: Tos que tiene secreciones
- Tos seca: Tos sin secreciones en el pecho
- Vómitos: Vómitos con o sin nauseas

## C.3. Matriz de correlación

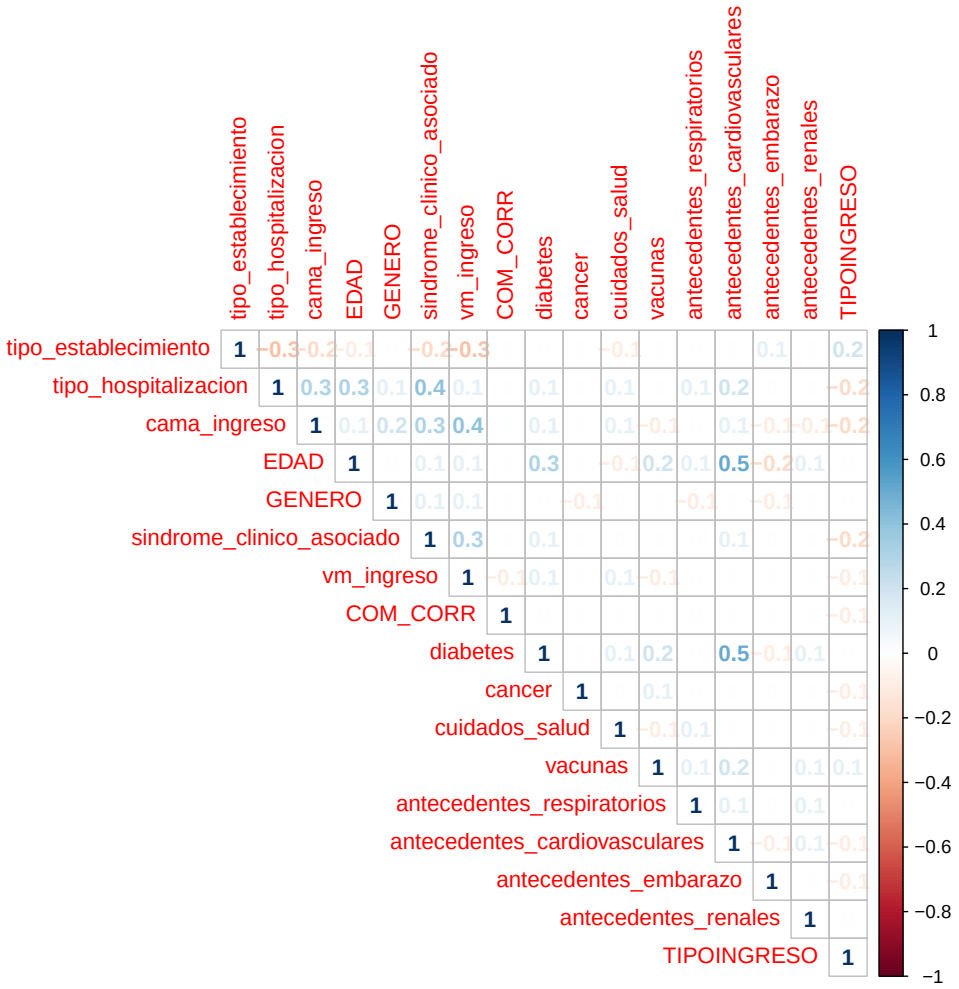



Figura C.1: Matriz de correlación variables modelo hospital

## C.4. Aplicación web

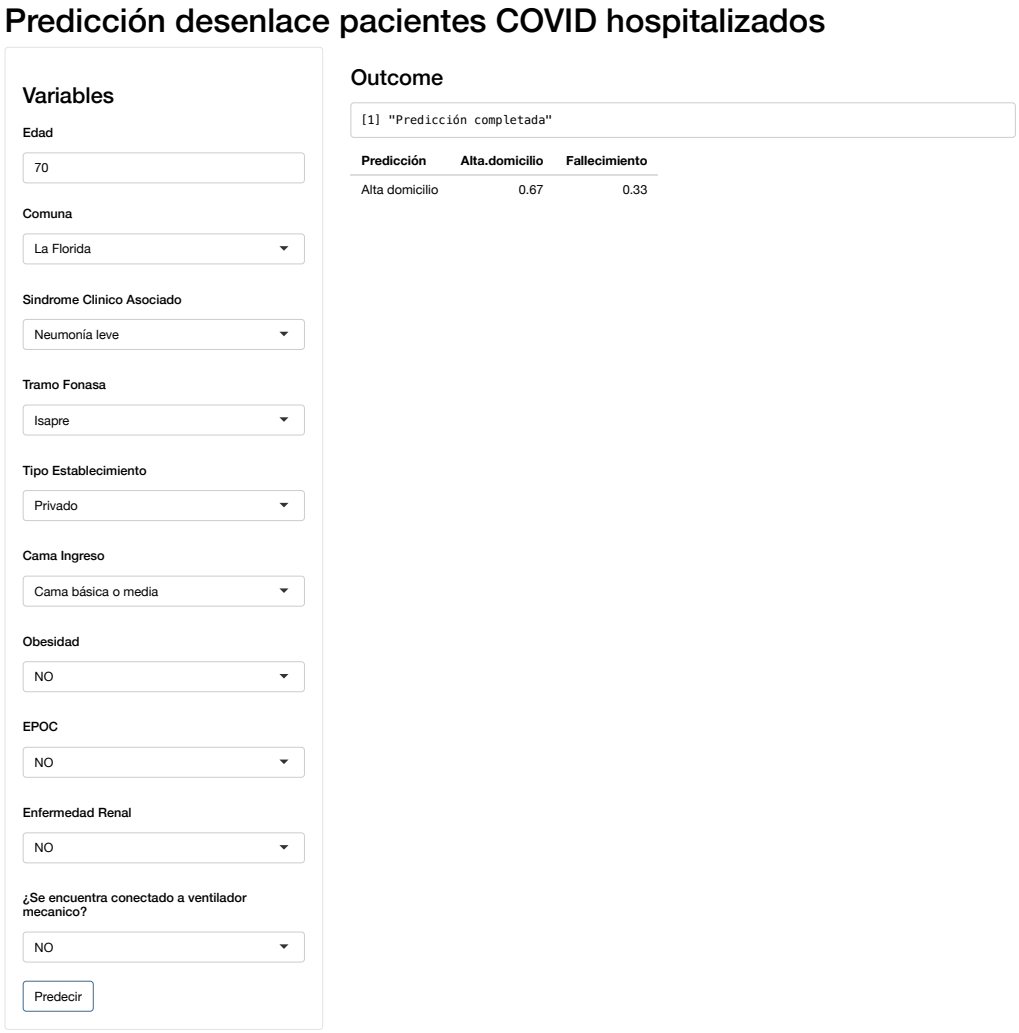



Figura C.2: Aplicación web

# Apéndice D

# Análisis pacientes vacunados

- Se han inoculado a 62.050 personas, de las cuales 25.754 tienen una dosis y 36.296 dos dosis, además 55,694 se han vacunado con la vacuna de Sinovac y 6357 con la vacuna de Pfizer
- Para realizar este análisis, seleccionamos a las personas que ingresaron al seguimiento en una fecha posterior a cuando se vacunaron por ultima vez (6,561), además nos quedamos con las personas Confirmadas y probables, dándonos una submuestra de 1,631 pacientes

Algunas estadísticas de la muestra al ingreso al seguimiento:

| **Numero de vacunas** | **Pacientes** | **Edad promedio** | **Genero F** | **Hospitalizado** |
|---|---|---|---|---|
| **1 dosis** | 1,038 | 53 | 60 % | 1,3 % (13) |
| **2 dosis** | 593 | 60 | 60 % | 1,3 % (8) |

Tabla D.1: Pacientes vacunados por numero de dosis

Ahora nos interesa conocer como le fue a este grupo de personas a lo largo del seguimiento, para eso separamos el análisis en: personas en seguimiento, personas que ya egresaron y las personas que recién ingresaron.

**Pacientes egresados (851)**

| Vacunas | Personas | Causal egreso | Hospitalizado | Fallecimiento |
|---|---|---|---|---|
| **1 dosis** | 526 | 81 % (422), alta<br>17 % (91), traslado<br>2 % (13), fallecen | 10 % (54) | 2 % (13) |
| **2 dosis** | 252 | 83 % (210), alta<br>16 % (40), traslado<br>1 % (2), fallecen | 8 % (20) | 1 % (2) |
| **2 dosis + 14 días** | 73 | 57 % (42), alta<br>43 % (31), traslado<br>0 % (0), fallecen | 4 % (3) | 0 % (0) |

Tabla D.2: Vacunados egresados

**Pacientes en seguimiento (752)**

- Cat1= Categoría de mayor riesgo
- S/I = Sin información

| **Vacunas** | **Personas** | **Presenta síntomas** | **Hospitalizado** | **Ultima categorización** |
|---|---|---|---|---|
| **1 dosis** | 494 | 63 % (310) | 1 % (6) | 23 % (112), Cat1<br>18 % (91), Cat2<br>49 % (243), Cat3<br>10 %, (48), S/I |
| **2 dosis** | 118 | 49 % (58) | 3 % (3) | 18 % (21), Cat1<br>14 % (16), Cat2<br>53 %, (62), Cat3<br>16 % (19), S/I |
| **2 dosis + 14 días** | 140 | 54 % (75) | 0 % (0) | 23 % (32), Cat1<br>22 % (31), Cat2<br>36 % (50), Cat3<br>19 % (27), S/I |

Tabla D.3: Vacunados en seguimiento

**Pacientes al ingreso (28)**

| Vacunas | Personas | Hospitalizado | Tipo ingreso |
|---|---|---|---|
| **1 dosis** | 18 | 1 | 9, Confirmado<br>9, Probable |
| **2 dosis** | 7 | 1 | 5, Confirmado<br>2, Probable |
| **2 dosis + 14 días** | 3 | 0 | 2, Confirmado<br>1, Probable |

Tabla D.4: Vacunados al ingreso

**Observaciones**

- Se puede notar como a medida que aumenta el numero de vacunas y se cumple el tiempo estimado para obtener la humanidad el como disminuye el porcentaje de hospitalizados y de fallecidos en cada uno de los grupos