



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

ALGORITMO DE PREDICCIÓN DE CRISIS RESPIRATORIA EN PACIENTES
PEDIÁTRICOS

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN GESTIÓN DE OPERACIONES

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

FABIÁN ANDRÉS GARCÍA TENORIO

PROFESOR GUÍA:
SEBASTIÁN ALEJANDRO RÍOS PÉREZ

MIEMBROS DE LA COMISIÓN:
DENIS SAURÉ VALENZUELA
LUCIANO VILLARROEL PARRA
FELIPE AGUILERA VALENZUELA

SANTIAGO DE CHILE

2018

RESUMEN

Las enfermedades respiratorias crónicas (ERC) afectan las vías respiratorias y otras estructuras del pulmón por más de tres meses, lo que en caso de infantes repercute tanto en su crecimiento como en su desarrollo normal. En Chile las enfermedades respiratorias son la primera causa de egreso hospitalario, siendo las ERC una de las razones más frecuentes. Diversas iniciativas se han implementado en apoyo a estas problemáticas, principalmente a través de Garantías Explícitas en Salud, pero éstas presentan un enfoque de atención médica necesaria y oportuna más que de prevención.

A partir de 2013 una iniciativa impulsada por el Ministerio de Salud busca reintegrar en sus domicilios a pacientes menores de 20 años que requieran apoyo respiratorio, lo que mejora su calidad de vida y la de sus familias. No obstante, el protocolo de detección de crisis y su consiguiente plan contingente recae en las familias y cuidadores. Por esto es fundamental generar estrategias que puedan prevenir y detectar de forma oportuna cualquier riesgo de salud una vez que llegan a sus casas.

Para esto, el objetivo general planteado es predecir el estado de riesgo de salud futuro del paciente en base a los signos vitales capturados, brindando el tiempo necesario para desencadenar el plan contingente asociado a la situación.

El resultado de este trabajo es un algoritmo capaz de conectarse a un sistema de captura de datos del paciente que logre detectar una situación adversa futura entregando este resultado en un tiempo mínimo de procesamiento. Considerando esto es fundamental contar con un repositorio de datos confiable, tanto para modelar el predictor como para su funcionamiento continuo.

Este algoritmo está dividido en dos etapas: elección de modelos óptimos y clasificación de riesgo futuro. La primera etapa busca escoger el conjunto de métodos de series de tiempo que sea capaz de predecir siguiente estado del paciente, y luego escoger el modelo de aprendizaje supervisado que clasifique de mejor forma el riesgo del paciente. La segunda etapa busca tomar los modelos ya seleccionados y utilizarlos para desencadenar la predicción. Con esto se logra que la respuesta del próximo estado de cada paciente se obtenga en segundos.

El resultado muestra que algunos signos vitales se pueden modelar de buena forma con la distancia de captura de los datos que se tiene (dos horas), como la temperatura, que para el paciente en que mejor se ajusta se puede modelar con un 1 % de error cuadrático medio. En contraste, la frecuencia cardiaca no es posible modelarla con la misma distancia entre datos, teniendo en el mejor caso un error cuadrático medio de 35, 90. Por otro lado, la clasificación de riesgo llega a resultados bastante buenos en términos de exactitud, cercano a 90 % en algunos casos, pero se tiene que para el mejor escenario el Valor-F alcanza sólo 67 %. Contrastando simultáneamente sensibilidad y especificidad se logra obtener modelos que están sobre 80 % en ambas métricas, o cercano a 90 % en una manteniéndose sobre 70 % en la otra.

A Karen. A Ximena.

Agradecimientos

Primero que todo, comenzar agradeciendo a mis padres. A Ximena, que me ha acompañado durante toda mi vida, desarrollando la persona que soy actualmente, y a Pedro, que me demostró que con convicción y trabajo uno puede alcanzar cualquier objetivo que se proponga. Ambos son modelos a seguir y los admiro mucho.

A mis hermanos, Bastián y Joaquín, quienes me hicieron crecer como persona. Gracias por escuchar mis reclamos, entender mis humores y siempre desafiarme a ser mejor. Espero que los caminos que están emprendiendo les permitan ser felices.

A Karen, mi compañera de vida, quien ha estado conmigo durante todo este proceso, siendo siempre mi pilar fundamental. Sin ti nada de esto sería posible.

A Graciela Donoso, quien es mi segunda madre, por ser un modelo de vida y por siempre saberlo todo. A Graciela y Sandra, por siempre estar ahí, por ser esa figura entre amigas y autoridades en quien uno puede confiar y que te ayudan a crecer. A Mauro, por ser siempre incondicional. Además gracias a todos ustedes por siempre comprender mis tiempos.

A la familia Villarroel Iasalvatore, quienes me acogieron en su familia, y ayudaron a crecer y formarme. Gracias especialmente a Manuel, quien generó un punto de inflexión en mi familia y gracias a él somos quienes somos actualmente. Gracias además por formar a la hermosa persona que tengo a mi lado.

Gracias a todos los amigos que me dejó esta etapa, los que espero me acompañen el resto de la vida: Alex, José, Diego, Javier, David, Felipe, Patrick, Jacqueline y Luis.

Agradecer especialmente al profesor Sebastián Ríos, por la confianza brindada y la guía que siempre me dio. Gracias por permitirme formar parte de un proyecto tan lindo y trabajar junto a un equipo tan completo, permitiéndome ganar una experiencia invaluable en mi formación profesional.

Tabla de Contenido

1. Introducción	1
1.1. Hipótesis	2
1.2. Objetivos	2
1.3. Resultados esperados	3
1.4. Metodología	3
1.5. Estructura de la tesis	6
2. Marco Teórico	7
2.1. Aplicaciones en salud	7
2.2. Problemas respiratorios en Chile	9
2.3. Aplicaciones en salud específicas al caso chileno	10
2.4. Algoritmo a desarrollar	11
2.4.1. Modelos de Series de tiempo	11
2.4.2. Algoritmos de Aprendizaje Supervisado	14
2.4.3. Métricas de Evaluación de Desempeño	17
3. Datos	20
3.1. Caracterización de los pacientes	21
3.2. Signos vitales	22
3.2.1. Temperatura	23
3.2.2. Frecuencia cardiaca	24
3.2.3. Frecuencia respiratoria	25
3.2.4. Saturación	26
3.3. Ventilación	27
4. Algoritmo de Predicción	28
4.1. Consideraciones y restricciones	29
4.2. Preparación de la data	30
4.2.1. Interpolación	30
4.2.2. Creación de variables	31
4.2.3. Variable objetivo	32
4.3. Etapa I: Elección de modelos	33
4.3.1. Elección de modelo de series de tiempo	33
4.3.2. Elección de modelo de aprendizaje supervisado	36
4.4. Etapa II: Clasificación de riesgo futuro	38
5. Resultados y Conclusión	40

5.1. Resultados	40
5.1.1. Resultados series de tiempo	41
5.1.1.1. Series de tiempo al modelar la temperatura	42
5.1.1.2. Series de tiempo al modelar saturación	43
5.1.1.3. Series de tiempo al modelar frecuencia cardiaca	44
5.1.1.4. Series de tiempo al modelar frecuencia respiratoria	44
5.1.2. Resultados aprendizaje supervisado	45
5.1.3. Evaluación global de resultados	47
5.2. Conclusiones y trabajo futuro	48
5.2.1. Recomendaciones y trabajo futuro	49

Bibliografía	51
---------------------	-----------

Índice de Tablas

3.1. Ejemplo datos utilizados. Fuente: Hospital Exequiel González Cortés	21
4.1. Construcción variable objetivo. Fuente: Elaboración propia	33
4.2. Salida 01 algoritmo ST, temperatura, extracto. Fuente: Algoritmo predicción	35
4.3. Salida 02 algoritmo ST, extracto. Fuente: Algoritmo predicción	36
4.4. Combinatoria variables, extracto. Fuente: Elaboración propia.	37
4.5. Rendimiento modelos iteración 04. Fuente: Algoritmo predicción.	38
5.1. Desempeño series de tiempo: ECM. Fuente: elaboración propia.	41
5.2. Desempeño series de tiempo: MMEC. Fuente: elaboración propia.	42
5.3. Series de tiempo en T: ECM. Fuente: elaboración propia.	42
5.4. Series de tiempo en T: MMEC. Fuente: elaboración propia.	43
5.5. Series de tiempo en Sat: ECM. Fuente: elaboración propia.	43
5.6. Series de tiempo en Sat: MMEC. Fuente: elaboración propia.	43
5.7. Series de tiempo en FC: ECM. Fuente: elaboración propia.	44
5.8. Series de tiempo en FC: MMEC. Fuente: elaboración propia.	44
5.9. Series de tiempo en FR: ECM. Fuente: elaboración propia.	45
5.10. Series de tiempo en FR: MMEC. Fuente: elaboración propia.	45
5.11. Aprendizaje supervisado por variables: exactitud. Fuente: elaboración propia.	46
5.12. Aprendizaje supervisado: rendimiento Valor-F. Fuente: elaboración propia. .	46
5.13. Clasificación modelos por métrica. Fuente: elaboración propia.	47
5.14. Exactitud de modelos por paciente. Fuente: elaboración propia.	48
5.15. Sensibilidad de modelos por paciente. Fuente: elaboración propia.	48

Índice de Ilustraciones

1.1.	Fases de CRISP-DM. Fuente: Shearer [44]	4
1.2.	Fases de CRISP-DM adaptado. Fuente: elaboración propia	5
2.1.	Diagrama árboles de decisión. Fuente: Wikimedia Commons [52]	15
2.2.	Diagrama máquina de soporte de vectores. Fuente: Wikimedia Commons [51]	17
3.1.	Edad de pacientes por género. Fuente: elaboración propia	21
3.2.	Edad de registros por género. Fuente: elaboración propia	22
3.3.	Signos vitales paciente id = 1. Fuente: elaboración propia	23
3.4.	Variación temperatura por género. Fuente: elaboración propia	24
3.5.	Variación frecuencia cardiaca por género. Fuente: elaboración propia	24
3.6.	Variación frecuencia cardiaca por edad. Fuente: elaboración propia	25
3.7.	Variación frecuencia respiratoria por género. Fuente: elaboración propia	25
3.8.	Variación frecuencia respiratoria por edad. Fuente: elaboración propia	26
3.9.	Variación saturación por género. Fuente: elaboración propia	26
4.1.	Descripción general algoritmo. Fuente: elaboración propia	29
4.2.	Elección serie de tiempo. Fuente: elaboración propia	34
4.3.	Elección modelo aprendizaje supervisado. Fuente: elaboración propia	38
4.4.	Proceso desencadenar predicción. Fuente: elaboración propia	39

Capítulo 1

Introducción

Las enfermedades respiratorias crónicas (ERC) afectan a las vías respiratorias y otras estructuras del pulmón por más de tres meses de duración, lo que afecta la oxigenación del cuerpo y, en caso de ser infantes, su crecimiento y desarrollo en los primeros años de vida [55, 6]. Algunas de las enfermedades más comunes son la Enfermedad Obstructiva Crónica (EPOC), enfermedades neuromusculares y el asma bronquial, sin embargo las consultas a servicios de urgencia son más bien por episodios agudos que complejizan el diagnóstico de base.

En Chile las enfermedades respiratorias son la primera causa de egreso hospitalario, siendo las ERC una de las razones más frecuentes después de las crisis agudas y otras enfermedades de vía aérea superior [12]. Así también, al hablar de atenciones de urgencia otorgadas por hospitales y Servicios de Atención Primaria de Urgencias (SAPU) las ERC lideran las estadísticas en los grupos etarios de 0 a 15 años de edad, por lo que es un problema de salud pública que el Estado ha intentado controlar [11].

Con este fin el Ministerio de Salud (MINSAL) ha generado propuestas a nivel de las Garantías Explícitas en Salud (GES) donde se ha integrado el EPOC, el asma bronquial moderada y severa y la displasia broncopulmonar del prematuro [45], sin embargo tienen como fin entregar la atención médica necesaria y oportuna, sin considerar aspectos de prevención y/o detección temprana.

A partir del año 2013 se integró a las estrategias del MINSAL el Programa Nacional de Asistencia Ventilatoria Invasiva (AVI) como No Invasiva (AVNI) que tienen como propósito reintegrar a su domicilio a pacientes menores de 20 años que requieren apoyo ventilatorio crónico como manejo de su patología de base, liberando camas hospitalarias y mejorando su calidad de vida relacionada a salud [48], ambos programas tienen coarticulación con la Atención Primaria de Salud (APS) que en conjunto con los profesionales, tanto del Hospital base como su Consultorio de referencia, generan planes de contingencia para crisis respiratorias. Para esto se le enseña a la familia los signos médicos de alarma y cómo activar el plan de urgencia [32, 33].

Sin embargo, se hace necesario generar estrategias que puedan prevenir y detectar de

manera más oportuna estas crisis, de manera que las acciones sean más eficaces.

Se propone en este estudio un modelo de predicción de crisis respiratorias en niños con ERC que estén asistidos en sus casas, basado en técnicas de aprendizaje supervisado y extrapolación temporal, con el fin de aportar a la toma de decisiones médicas y de futuras hospitalizaciones. Para esto se presenta la elaboración de un modelo de predicción a corto plazo de éstas crisis en pacientes pediátricos que presentan ERC, de esta forma tener una ventana de tiempo suficiente para activar el contingente médico necesario en éste tipo de urgencias.

1.1. Hipótesis

La hipótesis que desarrolla este estudio es dividida en dos aspectos, trabajados y desarrollados aisladamente en la solución propuesta.

Hipótesis 1: Existe una determinada forma de modelar las distintas señales biométricas del cuerpo humano, tales como temperatura, saturación de oxígeno, frecuencia cardiaca, frecuencia respiratoria, entre otras, para de esta forma prever cual será el estado siguiente de cada persona en base al conjunto de estados anteriores que el cuerpo de éste muestra. Esto se basa en que distintas señales del cuerpo humano siguen un ciclo circadiano, por lo que se podría modelar entendiendo que además de factores estacionales existe una notoria ciclicidad diaria [39, 8].

Hipótesis 2: Existe un determinado estado normal para pacientes pediátricos con ERC y un estado alterado para éstos, los que pueden ser claramente diferenciados en base a señales biométricas. Esto permitiría asignar “etiquetas” a cada estado de los pacientes y así desarrollar modelos de clasificación, que a posteriori servirían para definir cuál sería el estado en que se encontrará cada paciente a futuro.

1.2. Objetivos

A continuación se describe el objetivo general de este trabajo de tesis y sus respectivos objetivos específicos.

Objetivo general

Desarrollar un algoritmo que permita predecir el estado de riesgo de salud futuro que tendrán los pacientes pediátricos con ERC, brindando tiempo suficiente para realizar acciones preventivas y proactivas tanto por parte de los cuidadores como del equipo clínico asociado al paciente.

Objetivos específicos

1. Diseño, experimentación y evaluación de modelos de pronóstico en la modelación de signos vitales.
2. Diseño, experimentación y evaluación de modelos de predicción para ser utilizados en la modelación de riesgo.

1.3. Resultados esperados

1. Encontrar el modelo de serie de tiempo que extrapole de mejor forma cada signo vital de cada paciente, en una ventana de tiempo definida.
2. Encontrar el modelo de aprendizaje supervisado que discrimine de mejor forma el estado normal del estado de riesgo, para cada paciente.
3. Generar un algoritmo que sea capaz de utilizar los modelos de series de tiempo y de aprendizaje supervisado óptimos para cada paciente para predecir el estado futuro (normal o riesgo de salud) de éste.
4. Desarrollar el *framework* que permita agregar pacientes y modelos de manera simple a la investigación.

1.4. Metodología

La metodología utilizada en esta investigación es una variante de **CRoss Industry Standar Process for Data Mining** (CRISP-DM por sus siglas). Esta metodología organiza los proyectos de minería de datos en seis fases: entendimiento del negocio, entendimiento de la data, preparación de la data, modelamiento, evaluación e implementación [44].

La figura 1.1 muestra las distintas fases señaladas, detallando las iteraciones específicas que se propone realizar en la medida que cada fase es completada.

La fase inicial, **Entendimiento del Negocio**, tiene por objetivo determinar los objetivos a nivel de negocio, realizar un levantamiento general de la situación inicial en que se enmarcará el estudio (supuestos, requerimientos y restricciones), replantear los objetivos con la perspectiva de un estudio de minería de datos y generar un plan de trabajo.

Luego, la fase de **Entendimiento de la data** busca recopilar los datos a utilizar en el estudio, realizar análisis descriptivos de los mismos y verificar la calidad de éstos, previo a comenzar el trabajo directo con los datos. En el contexto de las iteraciones mostradas en la figura 1.1, desde esta fase se pueden replantear objetivos y restricciones definidos en la fase de **Entendimiento del Negocio** producto de las condiciones de los datos a trabajar, por lo que es necesario realizar una revisión de dicha fase.

Posteriormente, la fase de **Preparación de la data** busca seleccionar la sábana de datos con la que se trabajará, realizar las respectivas limpiezas a ésta, construir nuevas variables y realizar los cruces de datos necesarios con otras fuentes de información disponibles.

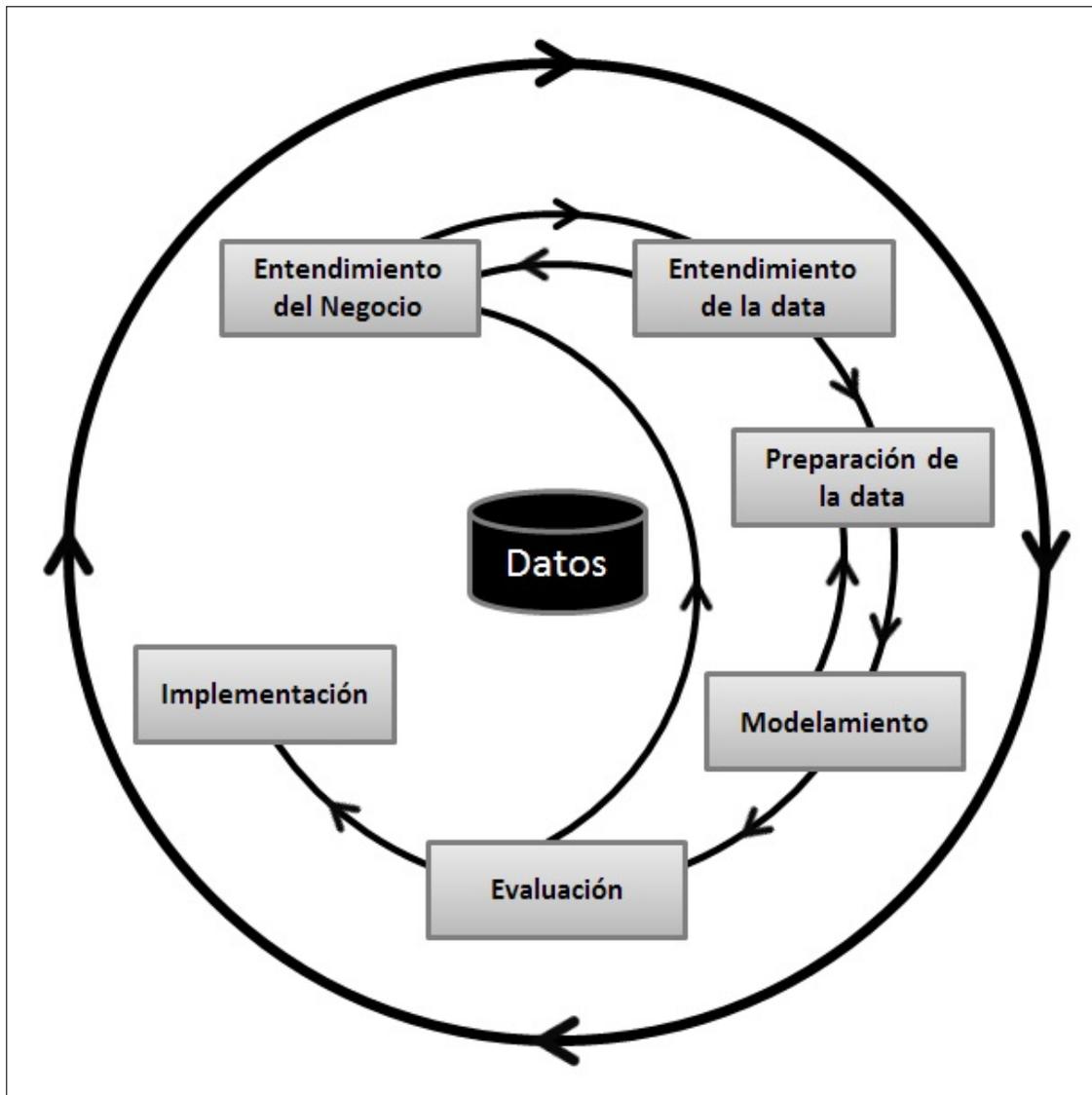


Figura 1.1: Fases de CRISP-DM. Fuente: Shearer [44]

La fase siguiente corresponde al **Modelamiento**, que es donde se evalúan los distintos modelos matemáticos que sean capaces de resolver el problema con las restricciones planteadas, se configuran parámetros globales y se desarrolla el modelo. Nuevamente en el contexto de las iteraciones planteado, en esta fase se debe reevaluar la fase de **Preparación de la data**, pues puede ser necesario la creación de nuevas variables o la parametrización con formato diferente de variables ya existentes.

En la fase de **Evaluación** se realiza una revisión del proceso general para determinar si el modelo propuesto cumple con las condiciones y objetivos definidos al inicio, así como también definir los siguientes pasos a desarrollar. De lo señalado en la figura 1.1 aquí la iteración se realiza con la primera fase, precisamente por el contraste de los resultados obtenidos con los esperados que se señaló anteriormente.

Finalmente, en la fase de **Implementación** se define un plan de puesta en marcha de la solución propuesta, con el respectivo plan de monitoreo y mantenimiento que este tipo de

soluciones requiere.

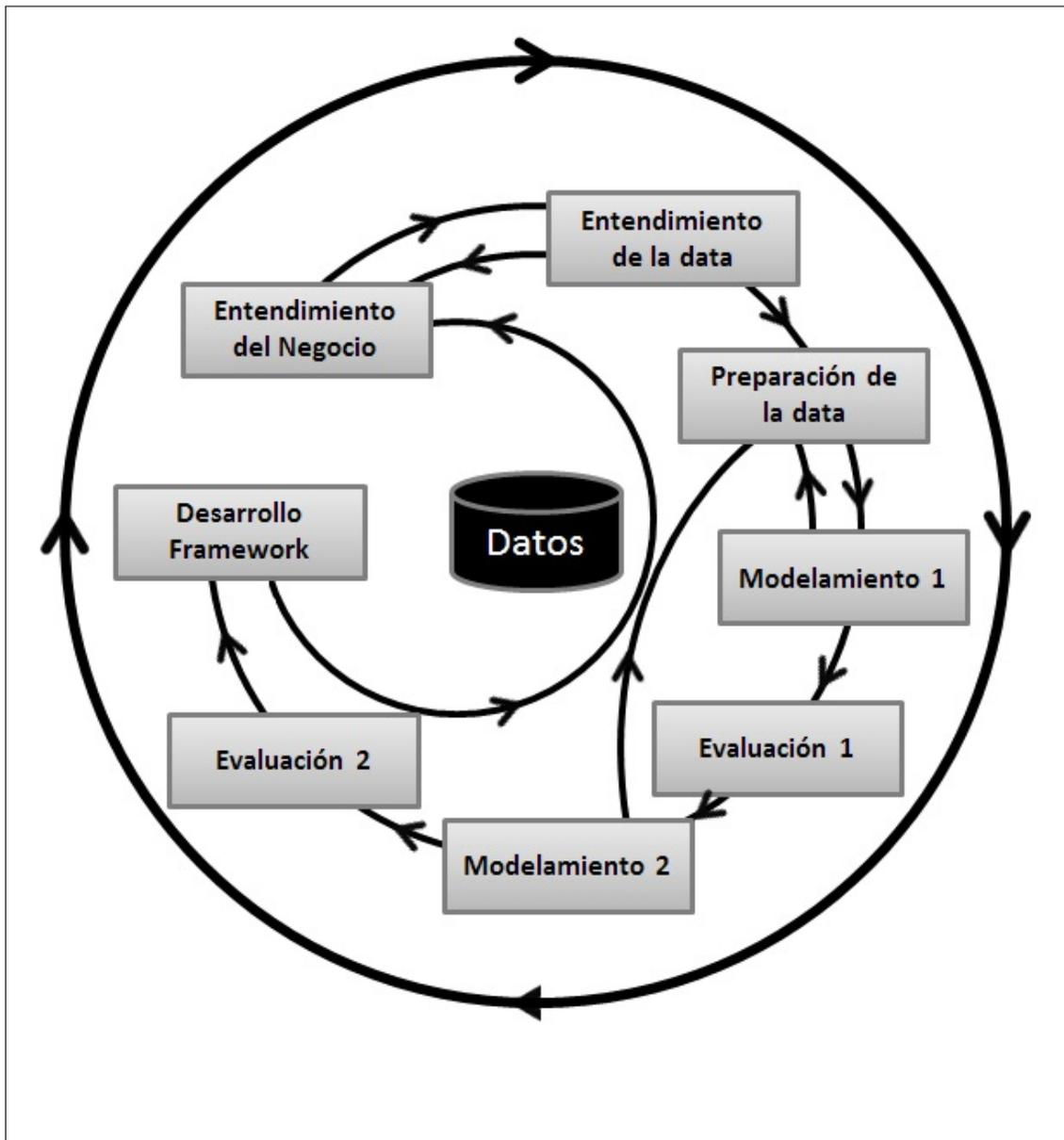


Figura 1.2: Fases de CRISP-DM adaptado. Fuente: elaboración propia

La figura 1.2 detalla la misma metodología pero adaptada a la problemática a abordar, considerando las iteraciones presentes en las distintas fases.

Esta variante se diferencia de lo propuesto por Shearer [44] principalmente en tres aspectos: división de las fases de **Modelamiento** y **Evaluación** en dos fases cada una, reemplazo de la fase de **Implementación** por la fase de **Desarrollo Framework** y modificación de la iteración final desde la fase **Evaluación** hacia la fase de **Entendimiento del Negocio** a la fase **Desarrollo Framework** hacia la misma fase inicial.

La primera variante, división de las fases de **Modelamiento** y **Evaluación**, responde a que el modelo a desarrollar es un modelo híbrido, donde primero se utilizará una solución

matemática que permita la extrapolación de una serie de tiempo y luego se utilizará una solución matemática que permita la clasificación del estado de un paciente. Es importante mencionar que desde el desarrollo de cada fase de **Modelamiento** se mantiene la iteración a la fase de **Preparación de la data**, pues se mantiene la lógica de ajuste de datos en función de mejorar el rendimiento de los modelos.

La segunda variante, el reemplazo de la fase de **Implementación** por la fase de **Desarrollo Framework**, responde al alcance en que se enmarca esta investigación. La implementación del algoritmo obtenido implica el desarrollo de tecnologías de captura y análisis de datos portable (que los pacientes puedan vestir), para así resolver la problemática expuesta. No obstante, dicho desarrollo escapa al alcance de este estudio.

Por otro lado, el desarrollo de un framework permite que en cuanto se cuente con las tecnologías necesarias para la implementación, ésta pueda realizarse sin tener que replicar esta investigación.

Finalmente la tercera variante, que considera los puntos anteriores, habla de la modificación de la iteración final, que responde a que el resultado plausible de esta investigación no se obtiene en la fase de **Evaluación**, que además está dividida, sino en el desarrollo de la solución tecnológica transversal que será utilizada para la implementación, es decir, el **Desarrollo Framework**, y por esto es importante que sea esta fase la que valide y contraste los resultados obtenidos con los esperados.

1.5. Estructura de la tesis

La estructura de este documento, que presenta los resultados y conclusiones del trabajo presentado anteriormente, se estructura como se detalla a continuación:

El Capítulo 2 detalla el marco teórico en que está contenido esta investigación, se muestra la bibliografía asociada a distintas soluciones propuestas a problemáticas similares a la planteada, tanto a nivel mundial como específicas para el caso chileno, además de estipular los sustentos matemáticos del trabajo posterior.

El Capítulo 3 realiza el análisis descriptivo de los datos, se señala las distintas variables disponibles para realizar este estudio y profundiza en cómo se comportan estas variables entre distintos grupos existentes en la muestra.

El Capítulo 4 especifica todo lo relacionado al desarrollo del algoritmo que da solución a la problemática planteada. Muestra especial énfasis en el funcionamiento del mismo y en los modelos utilizados para dicho propósito.

Finalmente, el Capítulo 5 muestra los principales resultados obtenidos de los experimentos de datos y da paso a la discusión que genera el estudio desarrollado.

Capítulo 2

Marco Teórico

A continuación se describen los aspectos relevantes del estado del arte en los que se enmarca este trabajo. El capítulo se divide en 4 partes, primero una descripción de distintas aplicaciones de los modelos señalados anteriormente en salud, después se detalla problemáticas de salud en Chile, específicamente las que atañen a este trabajo, luego se profundiza en los modelos señalados aplicados específicamente para el caso chileno, y, por último, se muestran los sustentos matemáticos que respaldaran el algoritmo a desarrollar.

2.1. Aplicaciones en salud

Técnicas de aprendizaje supervisado han sido utilizadas para diversas áreas de la medicina, ya sea para detectar enfermedades, anomalías, procesar datos, entre otros.

Los trabajos enfocados a la predicción de sucesos médicos han sido poco investigados por la complejidad de capturar datos para poblaciones específicas, especialmente cuando se trata de data sensible. Sin embargo, es un tema que se ha estado abriendo a la comunidad científica debido a la incorporación de tecnologías de información, por lo que actualmente es mayor la cantidad de artículos publicados de ésta índole [18].

Tal es el caso de Dai et al. [9] que generaron un algoritmo que considera la historia clínica de pacientes con enfermedades cardiacas, de tal manera que lograron predecir con 82% de exactitud si cada paciente será hospitalizado en el año siguiente, por lo que se alerta al equipo de salud y desencadena en acciones preventivas. Una ventaja de este tipo de modelos es que se puede monitorizar fácilmente a un gran número de pacientes, a diferencia de monitores humanos. Además, las tasas de precisión superan lo que es posible con las métricas de riesgo más empíricos. Respecto a la metodología de la investigación se compararon cinco algoritmos de inteligencia artificial, y se concluye que el que tuvo mayor precisión fue AdaBoost con árboles.

Wilke et al. [53] desarrolló algoritmos de minería de datos que, a través de data de genotipos, laboratorio clínico y su historia de medicación, generaron diagnósticos de Diabetes

Mellitus en personas mayores de 50 años. Se evidencia la utilización de antecedentes para detectar y clasificar adecuadamente a los pacientes, sin embargo también se utilizan modelos para predecir situaciones específicas.

Paxton et al. [36] utilizó el modelo SVM-Light con Kernel lineal para predecir el shock séptico, a través de data de la ficha médica de los pacientes que tenían antecedentes y signos de alerta. Además, con esta data se detectó el momento exacto en el que estaba en crisis, por lo que se podían generar planes de acción con los cuidadores y monitorear casos que pudiesen pasar desapercibidos. Este tipo de utilidad también fue utilizada por Himes et al. [19], quienes a través de redes bayesianas y data de registros médicos, identificaron el riesgo futuro de padecer EPOC entre pacientes con asma, quienes fueron observados durante 5 años.

En los últimos años los estudios referidos a usuarios pediátricos han ido en expansión, sin embargo aún son escasos, sobretodo por el acceso y manejo de la información por ser un grupo etario de mayor protección. Así se observa en el estudio de Sanders and Aronsky [43] quienes usaron redes bayesianas para identificar, a través de data recabada en el triage, a los pacientes pediátricos portadores de asma que necesitan cuidado especial. Se recabaron antecedentes de menores entre 2 a 18 años quienes estuvieron hospitalizados a lo menos 2 meses. Sin embargo sólo se utilizó ésta información con el fin de agrupar a los usuarios, sin la utilización de modelos de predicción.

Zhai et al. [57] desarrolla un modelo basado en regresiones logísticas que predice qué pacientes pediátricos necesitarán un traspaso a la unidad de cuidados intensivos durante su primer día internado, no obstante, el algoritmo resultante no logra señalar el momento en que ocurrirá dicho traspaso, sino sólo si ocurrirá o no.

Centrándose específicamente en las enfermedades respiratorias, Gorelick et al. [17] desarrolla un modelo basado en regresión logística y participación cursiva que busca definir que pacientes pediátricos que con asma aguda y que concurren a un servicio de urgencia necesitan hospitalización, a través de la utilización de la no readmisión de éstos en un plazo de 7 días como la variable de validación. Al igual que el ya señalado caso de Zhai et al. [57], no se aprecia la incorporación de una variable temporal en la readmisión utilizada para entender el episodio de riesgo específico de cada paciente pediátrico.

Radwan et al. [38] realiza un estudio en que aplica redes neuronales artificiales en la predicción de obtener una enfermedad respiratoria por parte de pacientes pediátricos mediante la incorporación de síntomas como también de signos vitales al modelo, sin embargo es necesario la utilización de demasiadas variables sin lograr dar un horizonte temporal de cuándo se tendrá la enfermedad respiratoria.

Por otro lado, la utilización de series de tiempo en el ámbito clínico, específicamente en las enfermedades respiratorias, ha respondido principalmente al estudio de la evolución de admisiones hospitalarias en el contexto general, más que en el ámbito específico de cada paciente.

Así es el caso de Braga et al. [3], quienes relacionan temporalmente los niveles diarios de PM_{10} , O_3 , SO_2 , CO y NO_2 con las admisiones hospitalarias respiratorias de niños. Weinberger et al. [50] utiliza un enfoque similar que contrasta la actividad del virus sincicial con la enfer-

medad neumocócica en infantes mediante el uso de series armónicas, siempre manteniéndose en aspectos macros.

Es importante considerar que si bien este tipo de soluciones está proliferando por el uso de nuevas tecnologías en áreas de la salud, las soluciones planteadas muchas veces se quedan en el espectro académico. Muro et al. [31] propone un cambio metodológico y de paradigmas en el área, donde los registros electrónicos de salud ya no sólo entreguen la información general del paciente y de su estado de salud a los tomadores de decisiones, sino también involucre este conjunto de modelos estudiados y validados para complementar el estado de cada persona.

2.2. Problemas respiratorios en Chile

Los nacimientos en Chile se han diversificado por su heterogeneidad en características, como también en el tipo de partos que se asisten. Según la Organización Mundial de la Salud (OMS) [54] el 7% de los nacimientos en Chile son de bebés prematuros, siendo el 1% partos prematuros extremos (menores de 1500 grs y/o de 32 semanas de edad gestacional), de los cuales un tercio presenta algún tipo de secuela.

En nuestro país las enfermedades respiratorias en niños se generan, en su mayoría, como consecuencia de anomalías congénitas secundarias a partos prematuros, siendo la más frecuente la Displasia Broncopulmonar (DBP). La DBP se entiende como una enfermedad pulmonar crónica caracterizada por la necesidad de oxígeno por más de 28 días, como consecuencia de un desarrollo inmaduro de los pulmones y alveolos expuestos a alguna noxa [22]. Se describe que cerca del 20% de los bebés prematuros extremos presentan DBP y alrededor de un 7% es dado de alta con necesidad de oxígeno domiciliario.

Si se observan las estadísticas de éste grupo estudio se evidencia que el 15% de los prematuros extremos fallecen en su primer año de vida, siendo la principal causa de muerte los problemas respiratorios agudos durante el periodo hibernal [34].

Es en ésta época además donde los niños con Enfermedades Pulmonares Crónicas, sean displásicos o no, son más vulnerables a tener enfermedades respiratorias agudas, de manera que aumenta en un 23% las consultas pediátricas, ya que son mucho más susceptibles a virus y patógenos en comparación a niños sanos, por lo que necesitan de forma más frecuente hospitalizarse y/o atenciones de urgencia, las cuales son atendidas por Atención Primaria de Salud [42].

En consecuencia según datos del Departamento de Estadísticas e Información de Salud (DEIS) del año 2011, las infecciones respiratorias son la primera causa de egresos hospitalarios de niños en el país con un 21,7% del total [56].

Así también la Organización Panamericana de la Salud y sus estudios en Chile indican que las neumonías en menores de 1 año son responsables del 90% del total de muertes por infecciones respiratorias [35].

El Ministerio de Salud de Chile ha generado políticas de estado para cubrir las necesidades

respiratorias de los niños en invierno, creando campañas y equipando salas de infecciones respiratorias agudas, ya que a pesar de localizar a los niños que tenían mayor riesgo respiratorio el 60% fallecía en su domicilio, a consecuencia de la incapacidad de reconocer los signos y alarmas de gravedad por parte de los cuidadores [29]

La hospitalización de usuarios pediátricos es una dificultad por su manejo y el alto costo de las familias, por lo que en la actualidad se ha avanzado en la hospitalización domiciliaria como una opción de menor costo, mayor calidad de vida para el usuario y que descongestiona a los hospitales permitiendo mayores hospitalizaciones con otras causas [24]. Sin embargo, con el fin de lograr una atención que realmente sea un beneficio para el usuario se hace necesario prevenir y predecir acontecimientos que pongan en peligro la vida del paciente.

2.3. Aplicaciones en salud específicas al caso chileno

En el caso chileno, la aplicación de este tipo de soluciones a problemas clínicos se ha visto retrasada por el poco desarrollo de disciplinas informáticas en la gestión médica, las que en caso de existir, responden principalmente a proyectos impulsados por instituciones individuales más que a una política de desarrollo nacional coordinada [7].

Además de lo anterior, la aplicación de este tipo de soluciones también se ve impactada por la data con que se trabaja, la cual tiende a ser muy desbalanceada considerando la población de control contra la de interés. Artetxe et al. [2] proponen un método que combina submuestreo aleatorio con aprendizaje supervisado, lo que mejora la adaptación de estos últimos a conjuntos de datos desbalanceados.

Así se observa en la investigación de Morales et al. [30] quien al estudiar 707 pacientes busca construir un indicador de riesgo de comportamiento suicida mediante la aplicación de árboles de decisión a un conjunto de instrumentos psiquiátricos (tests). La data asociada a la totalidad de los pacientes fue capturada de forma manual por los investigadores del estudio, lo que evidencia la restricción en los estudios de esta índole.

Por otro lado, Artetxe et al. [1] explora la readmisión de pacientes en el corto plazo (72 horas) en una muestra de 150.000 registros, capturados a lo largo de tres años en Santiago de Chile. Aplicando un algoritmo AHERF (*Anticipative Hybrid Extreme Rotation Forest*) alcanzan una sensibilidad cercana al 70% en la predicción de readmisiones para adultos y casos pediátricos. Profundizando en esta línea, Garmendia et al. [15] analizan las readmisiones cuya condición se ha agravado y requiere hospitalización del paciente, llegando a que utilizar redes neuronales artificiales en el alta de éstos es un buen método para reducir la readmisión, luego de contrastar entre 4 algoritmos mediante validación cruzada.

En el ámbito respiratorio, Erazo et al. [14] utilizó data capturada por polisomnogramas en pacientes menores a 15 años para prediagnosticar, mediante técnicas de minería de datos, apnea obstructiva del sueño. Ríos et al. [41] ahonda en este estudio obteniendo un algoritmo de detección automática de apnea obstructiva del sueño que, para la misma muestra anterior, lograba clasificar correctamente todos los episodios adversos de los individuos.

También en el ámbito respiratorio, Garmendia et al. [16] prueba que es posible predecir el triage que profesionales de la salud asignan a pacientes pediátricos con enfermedades respiratorias que asisten a servicios de urgencia en estado críticos, desarrollando además una aproximación a un mecanismo de supervisión automática para dichos pacientes en la unidad de cuidados intensivos.

Enfocándose en el aspecto preventivo, Echeverría et al. [13] propone una completa estructura ubicua que provee información en tiempo real para que el equipo clínico sea capaz de tomar decisiones informadas a distancia, lo que permite a pacientes crónicos y a sus familias mejorar mucho sus condiciones de vida.

2.4. Algoritmo a desarrollar

El algoritmo desarrollado en esta investigación consta de un modelo híbrido que combina dos modelos matemáticos: en primer lugar es necesario un modelo que permita extrapolar los signos vitales de cada paciente, para así generar la combinación de estados que define el estado futuro de la persona, y luego es necesario un modelo que sea capaz de tomar el conjunto de signos vitales generados para clasificar el estado de ese paciente en función de su probabilidad de riesgo futuro.

Para el desarrollo de todos los modelos descritos a continuación se utilizó la herramienta estadística R [37], empleándose librerías específicas en los casos en que se señale.

2.4.1. Modelos de Series de tiempo

El primer modelo del algoritmo híbrido señalado anteriormente será resuelto con un modelo de extrapolación de serie de tiempo.

A continuación se resumirá los sustentos teóricos tras los distintos modelos contrastados y utilizados en este estudio.

Proceso Autoregresivo $AR(p)$

Un proceso autoregresivo $AR(p)$ es un modelo que describe cierta secuencia de datos variables en el tiempo, donde se define que cada valor obtenido en un determinado momento en el tiempo depende linealmente de sus propios valores anteriores.

Matemáticamente, el proceso autoregresivo se define como:

$$X_t = c + \sum_{i=1}^p \varphi_i X_{t-i} + \varepsilon_t \quad (2.1)$$

donde $\varphi_1, \dots, \varphi_p$ corresponde a los parámetros a calcular por el modelo, c corresponde a

un valor constante y ε_t corresponde a ruido blanco, es decir, un valor residual que no se correlaciona estadísticamente con sus valores anteriores [5].

Para el desarrollo de este modelo se utilizó la librería *tseries* [47].

Proceso de Medias Móviles $MA(q)$

Un proceso de medias móviles $MA(q)$ es un modelo que describe cierta secuencia de datos variables en el tiempo, donde se determina que cada valor depende de los valores residuales pasados de la serie.

Matemáticamente, el proceso de medias móviles se define como:

$$X_t = c + \varepsilon_t \sum_{i=1}^q \theta_i \varepsilon_{t-i} \quad (2.2)$$

donde $\theta_1, \dots, \theta_q$ corresponde a los parámetros a calcular por el modelo, c corresponde al valor medio de la serie y $\varepsilon_1, \dots, \varepsilon_t$ corresponde a ruido blanco o un valores residuales de la serie [5].

Para el desarrollo de este modelo se utilizó la librería *tseries* [47].

Proceso Autorregresivo de Medias Móviles $ARMA(p, q)$

Un proceso autorregresivo de medias móviles $ARMA(p, q)$ es un modelo que describe cierta secuencia de datos variables en el tiempo y divide el entendimiento de la serie de tiempo en dos partes: una autorregresiva $AR(p)$ y otra de medias móviles $MA(q)$.

La notación matemática de este tipo de procesos es:

$$X_t = \varepsilon_t + \sum_{i=1}^p \varphi_i X_{t-i} + \sum_{i=1}^q \theta_i \varepsilon_{t-i} \quad (2.3)$$

donde $\varphi_1, \dots, \varphi_p$ corresponde a los parámetros de la parte autorregresiva, $\theta_1, \dots, \theta_q$ a los parámetros de la parte de medias móviles y $\varepsilon_1, \dots, \varepsilon_t$ a los valores residuales de la serie [5].

Para el desarrollo de este modelo se utilizó la librería *tseries* [47].

Proceso Autorregresivo Integrado de Medias Móviles $ARIMA(p, d, q)$

Un proceso autorregresivo integrado de medias móviles $ARIMA(p, d, q)$ es una variante del proceso $ARMA(p, q)$ donde previamente se diferencia d veces la serie de tiempo original para eliminar la tendencia.

La notación matemática de este proceso es:

$$\Delta^d X_t = \sum_{i=1}^p \varphi_i \Delta^d X_{t-i} + \sum_{i=1}^q \theta_i \varepsilon_{t-i} \quad (2.4)$$

donde $\Delta^d X_t$ representa que sobre la serie original, X_t , se han aplicado d diferencias, $\varphi_1, \dots, \varphi_p$ corresponde a los parámetros de la parte autorregresiva y $\theta_1, \dots, \theta_q$ a los parámetros de la parte de medias móviles [26].

Para el desarrollo de este modelo se utilizó la librería *tseries* [47].

Modelo Armónico Estacional

Un modelo armónico estacional es un método que busca describir y modelar una serie de tiempo como un conjunto de funciones de seno y coseno, para así suavizar y profundizar la estacionalidad que la serie presenta.

Matemáticamente, el modelo armónico estacional se describe como:

$$X_t = m_t + \sum_{i=1}^{[s/2]} \left\{ s_i \sin\left(\frac{2\pi it}{s}\right) + c_i \cos\left(\frac{2\pi it}{s}\right) \right\} + \varepsilon_t \quad (2.5)$$

donde m_t es la tendencia que incluye un parámetro para el término constante, $s_1, \dots, s_{[s/2]}$ y $c_1, \dots, c_{[s/2]}$ son parámetros a calcular por el modelo, s es la estacionalidad de la serie de tiempo mientras $[s/2]$ explicita que sea la parte entera de la mitad de dicha estacionalidad, y ε_t corresponde al valor residual [26].

Método de Suavizamiento Exponencial

El método de suavizamiento exponencial primero supone que la serie de tiempo no posee una tendencia sistemática ni un efecto estacional, o en su defecto que éstos han sido removidos. Al considerar esto, se modela un punto extrapolado de la serie como la suma de la media no estacionara de la serie y un valor residual.

Matemáticamente lo anterior se describe como:

$$X_t = \mu_t + w_t \quad (2.6)$$

donde μ_t es la media no estacional, es decir, la media de los valores de la serie posterior a excluida la componente estacional en caso de existir, y w_t es una desviación aleatoria independiente con media 0 y desviación θ .

Para la estimación, a diferencia de los métodos anteriores, donde se trabajaba directamente con el modelo descrito, aquí se reestructura lo señalado por la ecuación 2.6, y por lo tanto queda la siguiente expresión:

$$a_t = X_{t-1} + \alpha (X_{t-1} - a_{t-1}) \quad (2.7)$$

donde a_t corresponde al valor medio a estimar, correspondiente a la extrapolación del período t , X_{t-1} corresponde al valor real obtenido en el período anterior y a_{t-1} corresponde a lo estimado en el período anterior, por lo que $(X_{t-1} - a_{t-1})$ corresponde al error de pronóstico del período anterior. Por otro lado, α corresponde al parámetro de suavizamiento, y se cumple que $0 \leq \alpha \leq 1$.

En la ecuación 2.7 se puede iterar sobre el valor estimado del período anterior, y se especifica $a_1 = X_1$ para evitar una secuencia infinita, por lo tanto se tiene:

$$a_t = X_{t-1} + \alpha (1 - \alpha) X_{t-2} + \alpha (1 - \alpha)^2 X_{t-3} + \dots \quad (2.8)$$

desde donde se desprende que a_t es una combinación lineal de los valores anteriores de la serie, dándole mayor peso a los valores recientes de la misma al considerar que $\alpha (1 - \alpha)^i$ se vuelve más pequeño en la medida que i crece [26].

Para el desarrollo de este modelo se utilizó la librería *forecast* [20].

Método de Holt-Winters

El método de Holt-Winters es una generalización del método de suavizamiento exponencial, pues no le exige a la serie de tiempo que no posea tendencia ni estacionalidad, sino más bien busca modelarlas.

Si se considera como referencia la ecuación 2.7, matemáticamente el modelo propuesto está expresado por:

$$\left. \begin{aligned} a_t &= \alpha (x_{t-1} - s_{(t-1)-p}) + (1 - \alpha) (a_{t-1} + b_{t-1}) \\ b_t &= \beta (a_t - a_{t-1}) + (1 - \beta) b_{t-1} \\ s_t &= \gamma (x_{t-1} - a_t) + (1 - \gamma) s_{t-p} \end{aligned} \right\} \quad (2.9)$$

donde a_t , b_t y s_t son los parámetros estimados para la media, tendencia y estacionalidad respectivamente. Además, α , β y γ son los parámetros de suavizamiento. Se considera que la serie cuenta con un período p [26].

2.4.2. Algoritmos de Aprendizaje Supervisado

El segundo modelo del algoritmo híbrido será resuelto con un algoritmo de aprendizaje supervisado, los que buscan poder predecir un patrón de comportamiento a partir de una base inicial de comportamientos o base de entrenamiento.

A continuación se resumirán los sustentos teóricos de los distintos modelos utilizados en este estudio.

Redes Neuronales Artificiales

Las redes neuronales artificiales (*Artificial Neural Network*, ANN) son un conjunto de algoritmos de aprendizaje supervisado en que se emula los sistemas nerviosos animales, interconectando neuronas para así generar una salida. Un conjunto de estas neuronas, o perceptrones como se les denomina dentro de las redes neuronales artificiales, son las que reciben las variables de entrada y, mediante combinaciones matemáticas que se realizan en otro conjunto de perceptrones, ubicado en una o varias capas ocultas, se entrega una o varias variables de salida que en definitiva genera la clasificación buscada [49].

Para el desarrollo de este modelo se utilizó la librería *nnnet* [49].

Árboles de Decisión

Los árboles de decisión (*Decision Trees*, DT) son un conjunto de algoritmos de aprendizaje supervisado que buscan predecir una variable objetivo en base a un conjunto de sentencias lógicas, a través del uso de condiciones sucesivas sobre las variables de entrada. Donde se determina cada condición se denomina nodo, mientras el camino que se sigue luego de definido como se resuelve el nodo se denomina ramificación. Donde se termina una ramificación se denomina nodo de término, y es en éstos donde se determina la probabilidad de ocurrencia de cada variable objetivo, lo que determina la condición buscada [4].

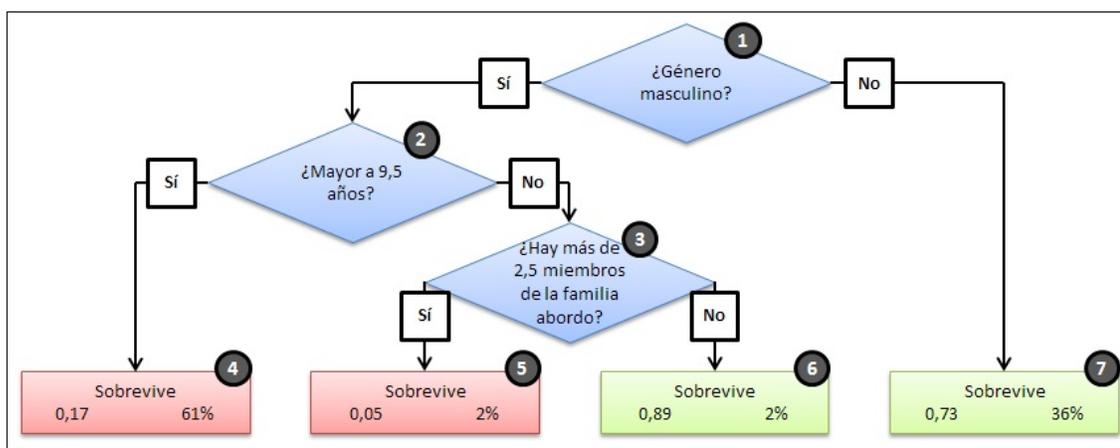


Figura 2.1: Diagrama árboles de decisión. Fuente: Wikimedia Commons [52]

La figura 2.1 representa gráficamente un árbol de decisión que ejemplifica la probabilidad de supervivencia de los pasajeros del Titanic. De esta forma, la primera sentencia lógica a evaluar (nodo 1) sería el género del individuo, dándose la situación que si era de género femenino tenía un 73 % de probabilidades de supervivencia, de acuerdo a lo señalado en el nodo 7 luego de haber seguido la ramificación "No" del nodo 1. El 36 % que se muestra en el nodo de término de esa ramificación hace alusión a la proporción de la muestra utilizada para entrenar el modelo. Al continuar con el ejemplo, se sigue por la otra ramificación desde

el nodo 1 se evidencia una segunda condición en el nodo 2, la que puede llevar a un nodo de término (nodo 4) o a una tercera condición (nodo 3), desde donde ambas ramificaciones terminan en nodos de término.

Para el desarrollo de este modelo se utilizó la librería *rpart* [46].

Regresiones Logísticas

La regresión logística es un tipo de regresión empleado para modelar la probabilidad de ocurrencia de un evento binario en función de diversos factores o variables independientes. La variaciones de esta regresión para modelar eventos donde la variable dependiente toma más de dos valores se denomina regresión logística multivariada, pero no será desarrollada en este estudio.

Matemáticamente, se busca un modelo que determine lo siguiente:

$$\mathbb{P}(Y = 1|X) = p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \quad (2.10)$$

donde Y es el evento a modelar, X es el vector de variables independientes o factores que buscan modelar el evento, β_0 es una constante a encontrar por el modelo que puede entenderse como el conjunto de factores no observables y β_1 representa un vector de peso de los distintos factores, también calculado por el modelo. La fórmula final de la ecuación 2.10 se denomina función **Logit** y busca evitar situaciones como $p(X) < 0$ o $p(X) > 1$, se considera que lo que se busca calcular es una probabilidad [21].

Máquina de Soporte de Vectores

Los algoritmos de máquinas de soporte de vectores (*Support Vector Machine*, SVM) buscan predecir una variable (o clasificarla) en base a la construcción de un hiperplano o un conjunto de hiperplanos en un espacio de muchas dimensiones, de tal forma que puedan ser utilizados como divisores de la muestra.

Para realizar lo anterior, y al tener en consideración las limitaciones actuales en términos computacionales para realizar curvas no lineales de separación o para conjuntos de datos que no pueden ser separados completamente, se trabaja con funciones Kernel, las que permiten proyectar los datos ingresados a un mayor espacio dimensional y simplificar la convergencia del algoritmo.

La figura 2.2 representa a la izquierda la entrada al algoritmo, mostrado como un espacio dimensional de dos dimensiones donde se busca clasificar entre dos grupos. Luego de aplicar una función Kernel $F = \{\phi(x) | x \in X\}$ se pasa a un espacio dimensional de tres dimensiones, donde ya podría clasificarse la variable objetivo en dos grupos mediante un hiperplano lineal.

Para el desarrollo de este modelo se utilizó la librería *e1071* [28].

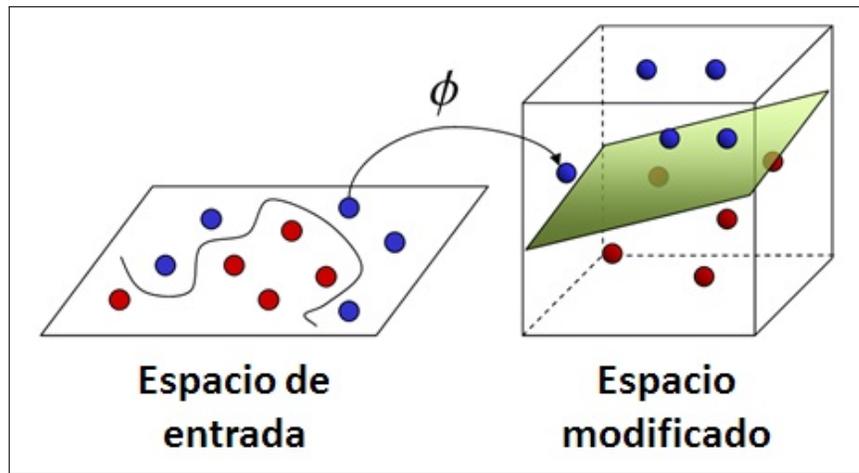


Figura 2.2: Diagrama máquina de soporte de vectores. Fuente: Wikimedia Commons [51]

2.4.3. Métricas de Evaluación de Desempeño

Los modelos desarrollados en cada sección del algoritmo híbrido serán evaluados mediante distintos indicadores, detallándose para cada caso a continuación.

Métricas para Modelos de Series de Tiempo

Para la evaluación de los modelos de series de tiempo se utilizarán dos métricas de medición: el error cuadrático medio y el menor máximo error cuadrático.

El **Error Cuadrático Medio**, ECM por sus siglas, busca entregar una medida comparativa de un estimador, y por lo tanto entrega el promedio de los del cuadrado de la diferencia entre el valor real y el valor estimado.

Matemáticamente lo anterior se define de la siguiente forma:

$$ECM = \frac{1}{n} \sum_{i=t+1}^{t+n} (\hat{X}_i - X_i)^2 \quad (2.11)$$

donde $\hat{X}_i, i \in \{t+1, \dots, t+n\}$ representa los valores estimados para los n períodos siguientes a t , $X_i, i \in \{t+1, \dots, t+n\}$ representa los valores observados para el mismo período, n corresponde a la cantidad de valores proyectados y t representa los valores no proyectados, es decir, los valores con los que se construye el estimador.

Entre los distintos modelos contrastados se busca el que tenga menor valor [23].

Por otro lado, el **Menor Máximo Error Cuadrático**, MMEC por sus siglas, es una métrica diseñada específicamente para este estudio, que busca definir al interior de un estimador cuál es su máximo error cuadrático, es decir, cuándo fue menos preciso y luego comparar este valor entre los distintos modelos contrastados, donde finalmente se elige al que obtenga el

menor valor entre éstos.

Matemáticamente se plantea como:

$$MMEC = \min_{\hat{X} \in M} \left\{ \max_{i \in \{1, \dots, n\}} \left(\hat{X}_{t+i} - X_{t+i} \right)^2 \right\} \quad (2.12)$$

donde \hat{X}_{t+i} y X_{t+i} representan los valores estimados y observados durante los períodos $\{t + 1, \dots, t + n\}$, y M representa el conjunto de modelos con que se estimará el vector X .

Esta métrica fue construida con la finalidad de conocer cuál modelo de serie de tiempo tuvo mejor comportamiento frente a cambios abruptos de la serie (situaciones anormales), para así entender cuál modelo se adapta mejor frente a posibles estados alterados.

Métricas para Modelos de Aprendizaje Supervisado

Para la evaluación de los modelos de aprendizaje supervisado se empleará 5 métricas comparativas: sensibilidad, especificidad, exactitud, precisión y valor F.

La sensibilidad (*sensitivity*) también denominada tasa de aciertos (*recall*), definida desde el punto de vista epidemiológico, se refiere a la probabilidad de clasificar correctamente a un individuo enfermo.

La especificidad (*specificity*) es la probabilidad de que un sujeto sano sea clasificado correctamente, es decir, con resultado negativo en la prueba.

La exactitud (*accuracy*) es la probabilidad de clasificar correctamente, independiente si se trate de individuos sanos o enfermos.

La precisión (*precision*) es la proporción de individuos enfermos que son correctamente clasificados.

Por último, el valor F (*F-score*) es una métrica que permite comparar que tan preciso y sensible es un modelo, en una única métrica.

Matemáticamente, las definiciones anteriores se traducen de la siguiente forma:

$$Sensibilidad = \frac{VP}{VP + FN} \quad (2.13)$$

$$Especificidad = \frac{VN}{VN + FP} \quad (2.14)$$

$$Exactitud = \frac{VP + VN}{VN + VN + FP + FN} \quad (2.15)$$

$$Presición = \frac{VP}{VP + FP} \quad (2.16)$$

$$ValorF = \frac{2 \times VP}{2 \times VP + FN + FP} \quad (2.17)$$

donde VP es verdadero positivo, es decir, clasificar como alguien con padecimiento cuando realmente está enfermo; VN es verdadero negativo, es decir, clasificar como alguien sano cuando efectivamente no tiene enfermedad; FP es falso positivo, que indica que fue clasificado como enfermo y estaba sano; y FN es falso negativo, que indica que el paciente fue clasificado como sano pero en realidad padecía la enfermedad [27].

Capítulo 3

Datos

Los datos utilizados en este trabajo corresponden a 77 pacientes pertenecientes a la Unidad de Pediatría y Unidad de Pacientes Críticos del Hospital Pediátrico Dr. Exequiel González Cortés, dependiente del Servicio de Salud Pública Metropolitano Sur de Santiago de Chile.

Como criterios de inclusión se considera tener diagnóstico de enfermedad respiratoria crónica, ser usuario pediátrico y haber necesitado monitorización y apoyo de oxígeno constante por parte del equipo médico durante los períodos de hospitalización observados. Se excluyen los usuarios con compromiso en la hemodinamia y/o que estuviera con alguna condición asociada que impide la estadía en su hogar (condición terminal, compromiso cerebral grave, trastorno de la deglución no resuelto, entre otros).

De estos pacientes se extrajo su historia médica anonimizada recopilada entre los años 2012 y 2015, y se seleccionó data general de su hospitalización:

- Caracterización del paciente: edad, sexo, diagnóstico y otras clasificaciones.
- Signos vitales: temperatura, frecuencia cardiaca y respiratoria, saturación de oxígeno, presión arterial y fracción de oxígeno inspirada (FiO₂).
- Ventilación: tipo y parámetros que se tengan, en caso de ser ventilación asistida mecánicamente.

A nivel de registros, se cuenta con un total de 13.814 observaciones, y se observa que los campos que varía entra cada paciente son el tipo de ventilación y sus signos vitales, además también se registra la edad del paciente medido en días para precisiones posteriores más específicas.

La tabla 3.1 posee un extracto, tanto a nivel de filas como de columnas, de la data que se tiene para cada paciente. Por supuesto no es necesario individualizar a los pacientes para el presente estudio, y es por ello que se trabaja con datos anonimizados, siendo el campo “id” el único identificador del paciente, que en el caso ejemplificado corresponde al paciente $id = 1$. El campo “Diag” hace alusión al diagnóstico que presenta el paciente. “Un” especifica la unidad en la que está internado, que puede ser la unidad de pacientes críticos, o UPC por sus siglas, o la unidad de pediatría. “Nac” consigna la fecha de nacimiento. Las

columnas “T”, “Sat”, “FC”, “FR” representa los signos vitales de temperatura, saturación, frecuencia cardiaca y respiratoria, respectivamente, registrados en la fecha y hora detallada en la columna “Medición”, mientras “Resp” detalla el tipo de apoyo respiratorio que tenía el paciente en ese momento.

Id	Diag	Un	Sexo	Nac	Medición	T	FC	FR	Sat	Resp
1	BNM	UPC	Masc	17/02/13	25/03/14 02:00	37.3	156	39	90	VMI
1	BNM	UPC	Masc	17/02/13	25/03/14 04:00	37	119	20	98	VMI
1	BNM	UPC	Masc	17/02/13	25/03/14 06:00	37.1	112	18	97	VMI
1	BNM	UPC	Masc	17/02/13	25/03/14 08:00	37.1	114	22	97	VMI
1	BNM	UPC	Masc	17/02/13	25/03/14 10:00	36.7	126	26	100	VMI
1	BNM	UPC	Masc	17/02/13	25/03/14 12:00	37.3	143	30	96	VMI

Tabla 3.1: Ejemplo datos utilizados. Fuente: Hospital Exequiel González Cortés

A continuación se realiza una descripción detallada de los datos que sustentan el estudio realizado. El capítulo se divide en 3 partes, con el fin de profundizar en los grupos de datos obtenidos durante la hospitalización de los pacientes, mencionados anteriormente.

3.1. Caracterización de los pacientes

De los 77 pacientes señalados, 46 son de género masculino, equivalente al 59,7% de los pacientes de la muestra, y 31 de género femenino, equivalente al 40,3% restante, con edades comprendidas entre los 0 y los 16 años.

La figura 3.1 muestra la distribución de las edades de los pacientes por género, donde se aprecia que para el género masculino se tiene una concentración mayor en los primeros tres años de edad, mientras que para el género femenino la concentración es más suave, manteniéndose una buena proporción de pacientes hasta los siete años.

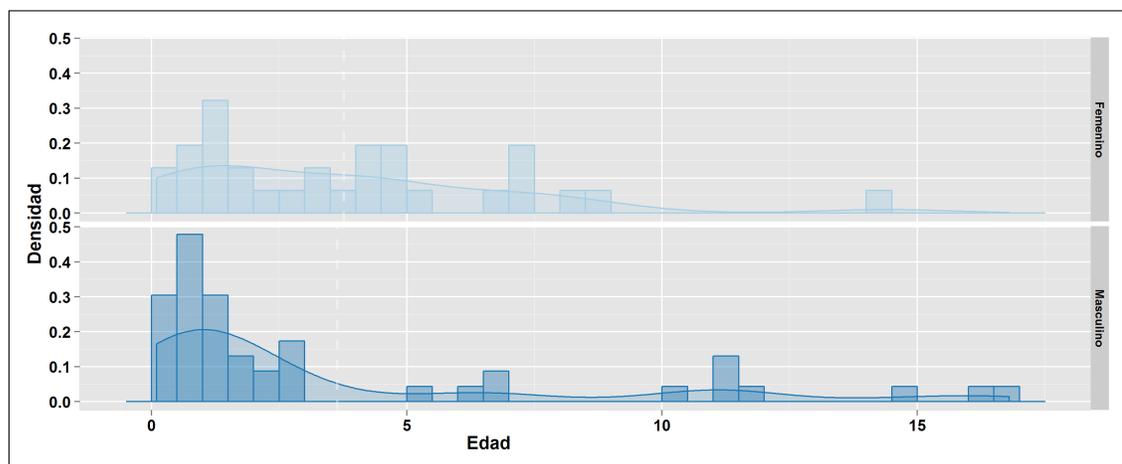


Figura 3.1: Edad de pacientes por género. Fuente: elaboración propia

Con el fin de complementar lo expuesto en la figura 3.1, y al considerar que cada paciente puede estar internado por distintos períodos de tiempo, la figura 3.2 muestra la distribución de género y edad por cada registro capturado. Se destaca que para el género masculino se acrecenta la concentración en los primeros tres años, mientras para el género femenino la mayor concentración de registros se tiene entre los tres y los cuatro años y medio, pero se mantiene una muestra importante en los primeros años.

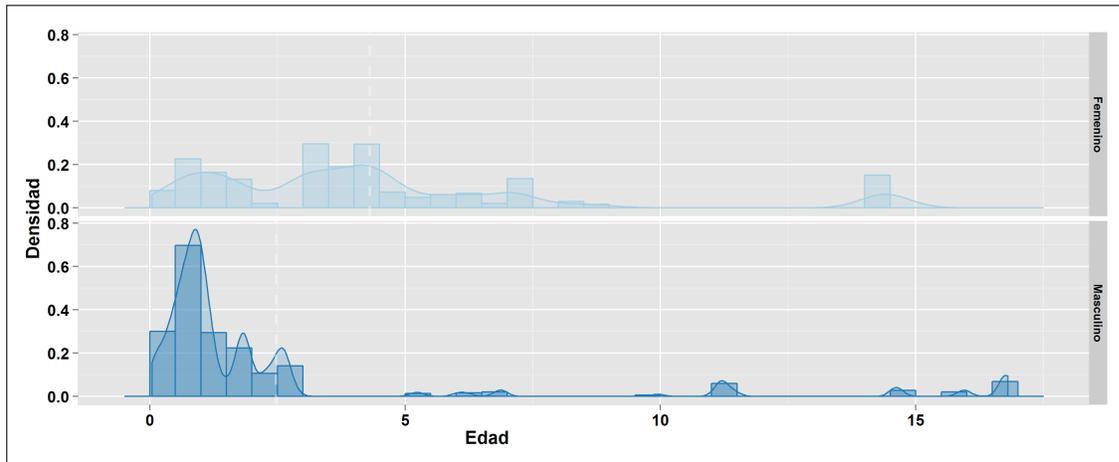


Figura 3.2: Edad de registros por género. Fuente: elaboración propia

Considerando la distribución de género a nivel de registro, la concentración en la muestra de registros masculinos se acrecenta respecto a lo ya expuesto a nivel de pacientes, pues de los 13.814 registros un 72,2% son asociados a pacientes género masculinos (9.973 registros), y sólo el 27,8% restante son asociados a pacientes de género femenino (3.841 registros).

Además, de éstos una gran proporción se concentra en los tres primeros años, por lo que es importante considerar dicha información previo a realizar hipótesis de la población a nivel general.

3.2. Signos vitales

Los signos vitales son registrados en fichas físicas por los profesionales de la salud a cargo de controlar la evolución de los pacientes, y posteriormente digitalizados específicamente para la ejecución de este estudio.

Éstos son registrados cada dos o cuatro horas, dependiendo de si el paciente se encuentra en la Unidad de Pacientes Críticos o en la Unidad de Pediatría, pues en la primera unidad dado las posibles mayores complicaciones es necesario mantener un control más constante, mientras en Pediatría se supervisa la evolución del paciente. Sin desmedro de lo anterior, existe ventanas de tiempo mayores sin captura de signos vitales debido a que el paciente se podía encontrar en procedimientos mayores, pero salvo estos casos específicos los signos vitales se capturan con la frecuencia detallada.

De cada paciente hospitalizado se obtiene, en promedio, 71 mediciones de sus signos vitales.

Se considera los plazos de captura de registros por unidad señalados, el período de tiempo de una hospitalización que contenga las 71 mediciones estaría entre 142 y 284 horas, equivalente a un lapso de entre 6 y 12 días.

La figura 3.3 ejemplifica algunos signos vitales, los tiempos en horas desde la hospitalización con que se capturan y la variabilidad que éstos muestran durante este lapso, específicamente para el paciente $id = 1$. En particular para este paciente se tiene un total de 40 registros, distribuidos en una estadía de 111 horas (4,65 días).

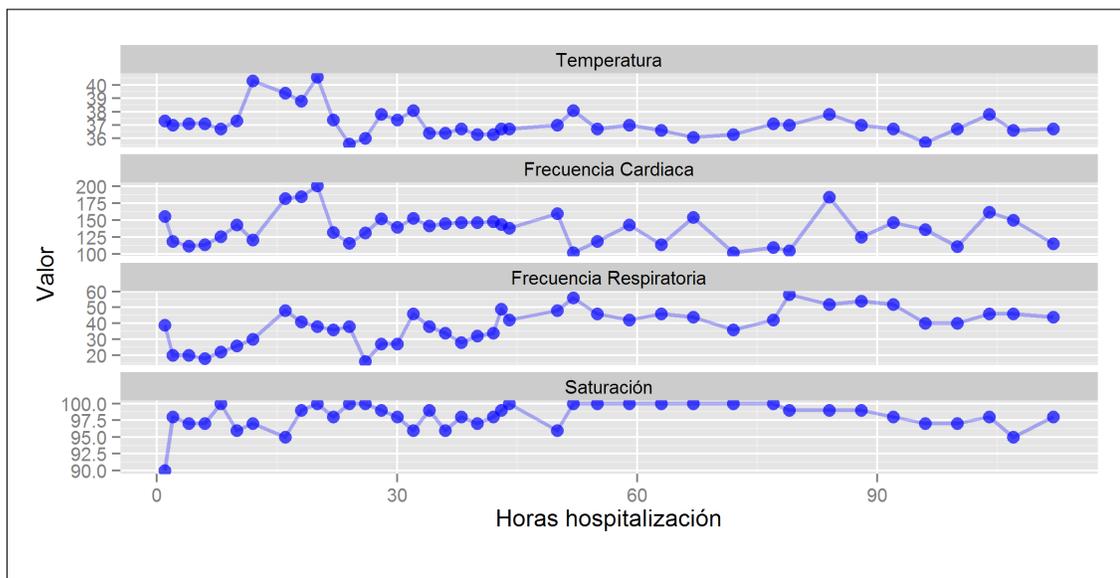


Figura 3.3: Signos vitales paciente $id = 1$. Fuente: elaboración propia

Al tener los datos señalados se busca profundizar en el comportamiento de los distintos signos presentes en la muestra. Para esto se consultó con los médicos del área pediátrica, levantando así el criterio experto asociado, con el fin de determinar tanto los indicadores que entregarían mayor información para la posterior predicción de riesgo como también cuales requerirían transformaciones para ajustarse a las distintas sub-poblaciones contenidas en el universo de estudio.

De esta forma se determina profundizar en la temperatura, frecuencia cardíaca, frecuencia respiratoria y saturación. Esta profundización se realizará según edad y género, pues son variables frente a las cuales los distintos signos vitales se comportan de forma distinta.

3.2.1. Temperatura

La figura 3.4 detalla el comportamiento de la temperatura por género de los pacientes. Si bien se aprecia una pequeña diferencia en el valor medio para los distintos géneros (T de los pacientes género masculino es en promedio $0,12^{\circ}\text{C}$ mayor que los pacientes femeninos), no se evidencia que este indicador varíe impulsado por el género del individuo.

En el caso de la edad, de acuerdo a lo señalado por el equipo clínico, la temperatura corporal no presenta variación entre distintas edades, al menos para los rangos que este

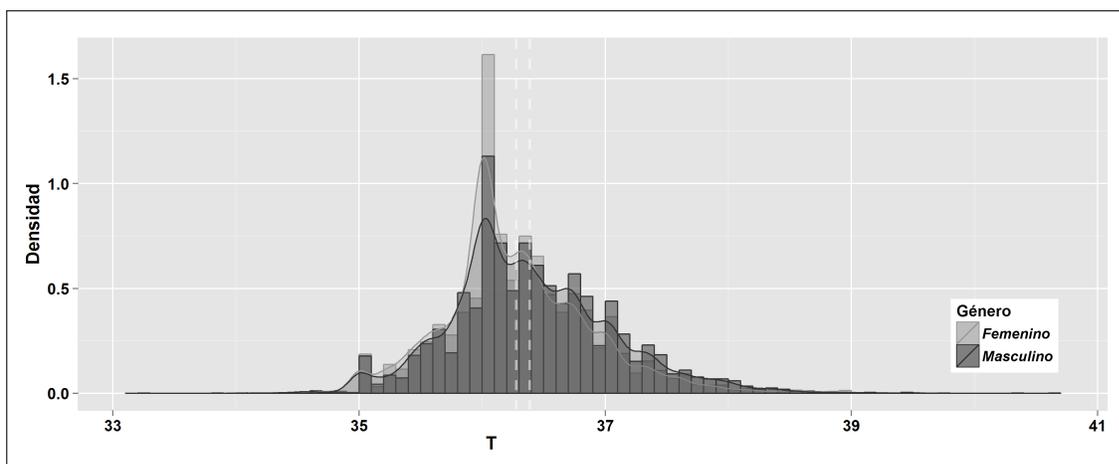


Figura 3.4: Variación temperatura por género. Fuente: elaboración propia

estudio abarca. Esto también es validado por Mancilla et al. [25], quienes profundizan en los distintos signos vitales para entender cómo se comportan éstos en los pacientes pediátricos, no destacándose el caso de la temperatura.

3.2.2. Frecuencia cardiaca

La figura 3.5 detalla cómo se comporta la frecuencia cardiaca por género. En este caso, a diferencia de la temperatura, se aprecia una pequeña diferencia entre las medias aritméticas siendo mayor el valor en el caso masculino. También se visualiza una pequeña diferencia en el comportamiento general de las subpoblaciones.

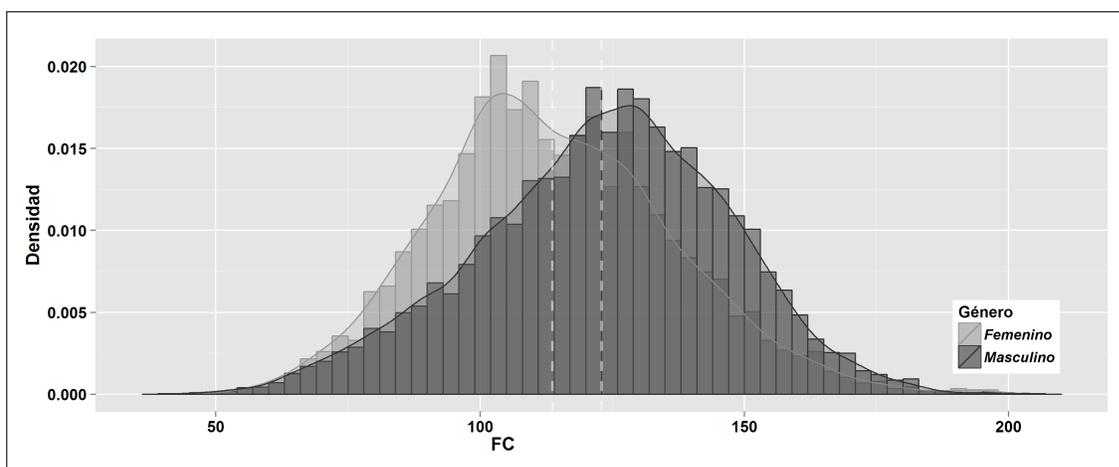


Figura 3.5: Variación frecuencia cardiaca por género. Fuente: elaboración propia

Distinto es el caso de la variación de la frecuencia cardiaca por edad, pues el equipo clínico señala que sí se evidencian diferencias claras en los valores normales de los pacientes de distintas edades. Al considerar esto, se utilizan los rangos de edad referenciales establecidos por Mancilla et al. [25] y se evalúa en la muestra a trabajar. La figura 3.6 establece que los niños mayores a diez años presentan una menor frecuencia cardiaca que los pacientes de

menor edad. Esta relación inversa (edad contra frecuencia cardiaca) se cumple en la mayoría de los casos, pero presenta una anomalía al contrastar a los pacientes entre cero y tres meses contra los de tres a seis meses, pues estos últimos muestran una frecuencia cardiaca mayor al grupo de menor edad (aquí la relación sería directa).

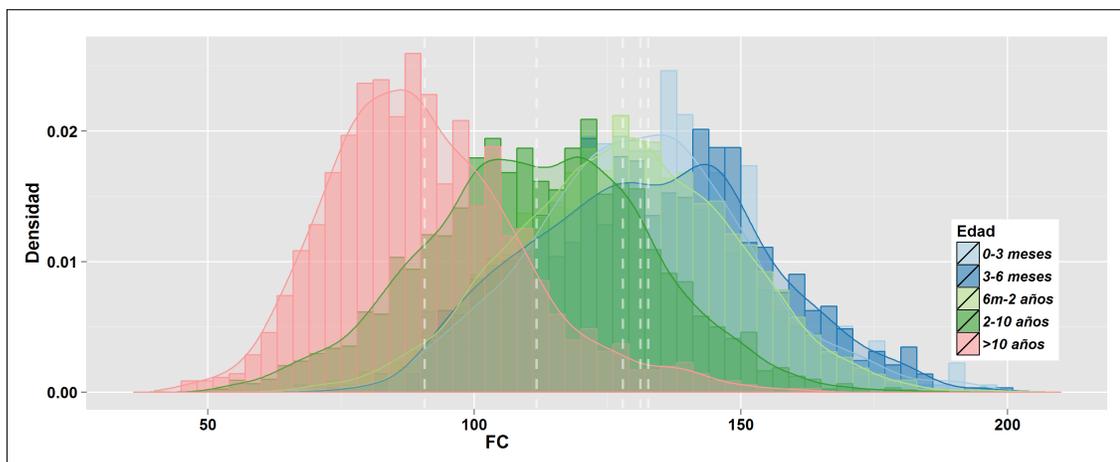


Figura 3.6: Variación frecuencia cardiaca por edad. Fuente: elaboración propia

3.2.3. Frecuencia respiratoria

La figura 3.7 detalla cómo se comporta la frecuencia respiratoria por género. Ésta, similar a lo ocurrido con la temperatura, no presenta diferencias al analizar cómo se comporta por el género del paciente, a pesar de existir una pequeña diferencia en la media aritmética entre género masculino y femenino.

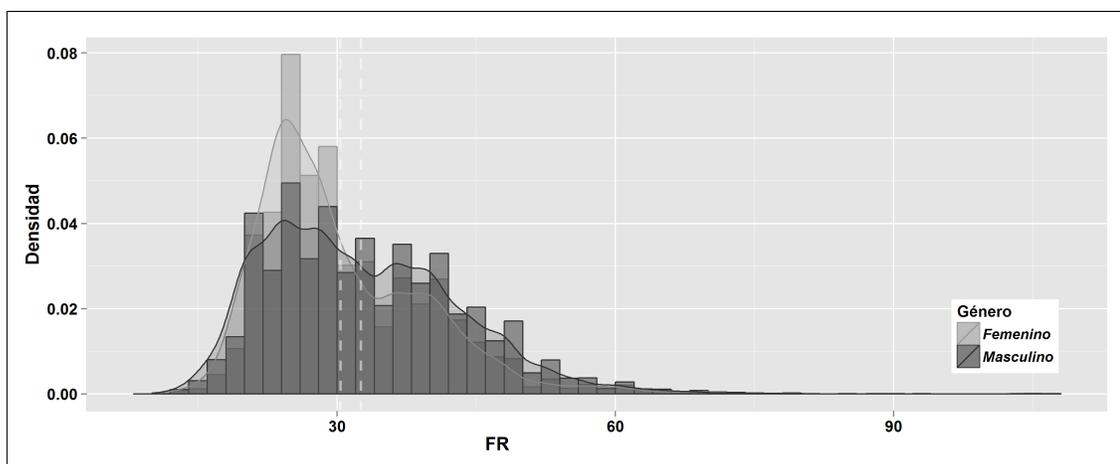


Figura 3.7: Variación frecuencia respiratoria por género. Fuente: elaboración propia

Al analizar la frecuencia respiratoria por edad, al igual que en el caso de la frecuencia cardiaca, el equipo clínico señala que sí se evidencia diferencias en los valores normales de los pacientes de distintas edades. En este caso los cinco rangos referenciales propuestos por Mancilla et al. [25] difieren respecto a los propuestos para la frecuencia cardiaca, por lo que

se evalúan estos nuevos rangos etarios en la muestra a trabajar. La figura 3.8 muestra que de los rangos propuestos, los primeros dos (cero a un año y de uno a dos años) no presentan una diferencia significativa en la muestra presente en este estudio, mientras que la diferencia para los tres rangos restantes, si bien existe, es inferior que lo evidenciado en el caso de la frecuencia cardiaca.

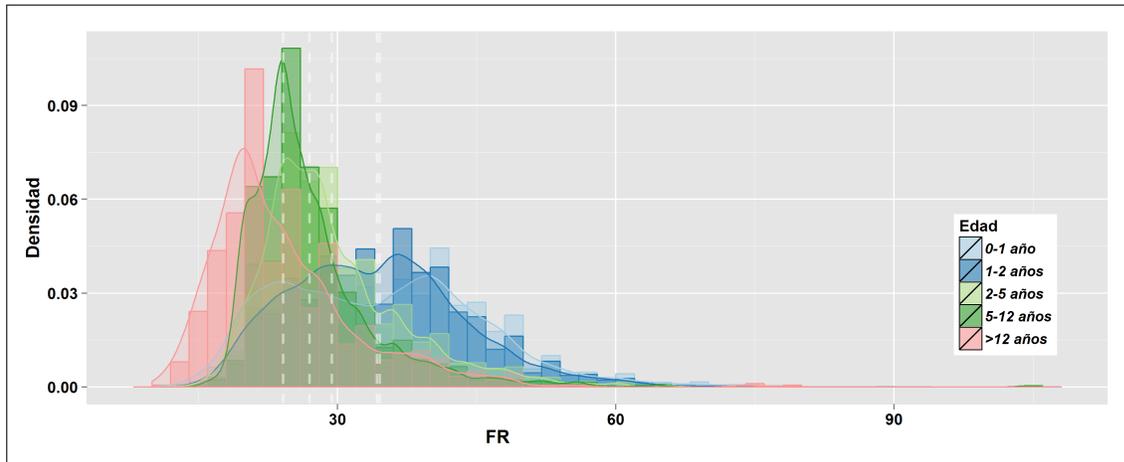


Figura 3.8: Variación frecuencia respiratoria por edad. Fuente: elaboración propia

3.2.4. Saturación

La figura 3.9 detalla el comportamiento de la saturación abierta por el género de los pacientes. La diferencia que se ve en el valor medio entre ambos géneros es prácticamente inexistente (0,4 de mayor saturación oxígeno para los pacientes de género masculino), no evidenciándose tampoco diferencia en las densidades del resto de los valores.

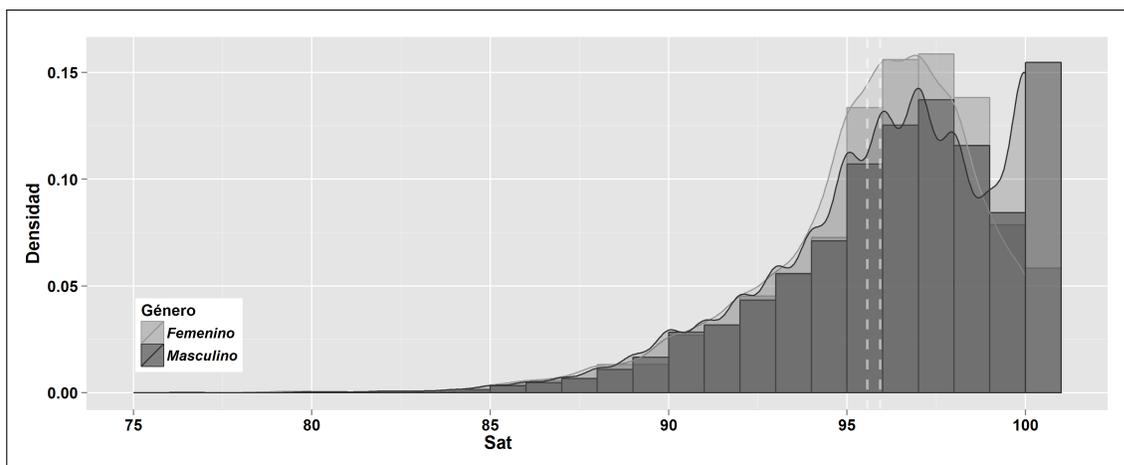


Figura 3.9: Variación saturación por género. Fuente: elaboración propia

En el caso de la edad, al igual que en el caso de la temperatura, no presenta variación de acuerdo a lo señalado por el equipo clínico, al menos para los rangos que este estudio abarca.

3.3. Ventilación

La variable de ventilación, o tipo de ventilación, es una variable consignada por los profesionales de la salud a cargo de los pacientes, tal como en el caso de la sección 3.2, en intervalos de tiempo dependiendo de la unidad en que se encuentre éste.

Al tener en consideración que la completitud de la información del estado del paciente se realiza en papel, es importante mencionar que esta variable no está estandarizada sino que contiene distintas nomenclaturas para referirse a un conjunto acotado de posibilidades.

Sin desmedro de lo anterior, y ayudados con el criterio experto de los equipos clínicos relacionados a la captura de estos datos, se identifican los siguientes estados distintos para esta variable, ordenados por gravedad creciente:

1. Respiración ambiental.
2. Naricera (0 a 3 litros).
3. Halo o Venturi 30 - 60.
4. Mascarilla con recirculación.
5. Naricera de alto flujo o NAF.
6. Ventilación mecánica no invasiva (VMNI).
7. Ventilación mecánica invasiva (VMI).

De esta forma es posible conocer el apoyo respiratorio que tenía cada paciente durante toda su estancia en el servicio de salud, y se considera también los cambios en dicho apoyo o en los parámetros asociados a éstos (en los casos que corresponda).

Capítulo 4

Algoritmo de Predicción

El algoritmo desarrollado es la aplicación automatizada de herramientas estadísticas y de inteligencia artificial, específicamente de modelos de series de tiempo y algoritmos de aprendizaje supervisado, para la predicción de crisis respiratoria en niños con ERC.

Éste busca conectarse con un dispositivo que vistan los niños que permita la captura instantánea de signos vitales y así alertar con un horizonte determinado de tiempo (entre media y dos horas) que el paciente podría presentar un cuadro adverso, lo que conectado a un sistema robusto en contacto con los padres o cuidadores y equipos clínicos permitirá tomar acciones preventivas en pro de reducir riesgos. Tanto el dispositivo de captura de signos vitales como el sistema mencionado, a los cuáles se conectará el algoritmo, están fuera del alcance de este estudio.

En virtud de entregar la respuesta en mejores tiempos el modelo híbrido (por tener componentes de series de tiempo y de aprendizaje supervisado) está dividido en dos etapas: elección de modelos y predicción de riesgo futuro. La primera etapa busca determinar el conjunto de modelos que mejor responde a cada paciente, mientras que la segunda etapa utiliza los modelos ya seleccionados y desarrollados (en términos de parámetros) y lo aplica a un conjunto más reducido de datos para así obtener respuestas instantáneas.

La figura 4.1 ilustra lo anterior. Se detalla además que está previsto que la primera etapa o de elección de modelos (parte inferior de la figura) se inicie de acuerdo a una programación determinada mediante requerimientos y restricciones del sistema, rendimiento del servidor, sobrecarga del mismo y otras variables asociadas al desempeño, mientras la segunda etapa o de predicción de riesgo futuro (parte superior de la figura) sea la respuesta a un requerimiento del sistema, el que además debe entregar en esta instancia la data necesaria para su procesamiento.

Si bien lo señalado anteriormente especifica el funcionamiento final en que se planea implementar el algoritmo, es importante mencionar que la data con que se desarrollará corresponde a lo expuesto en el capítulo 3, por lo que existe restricciones de desarrollo que quizás posteriormente sean resueltas de otra forma.

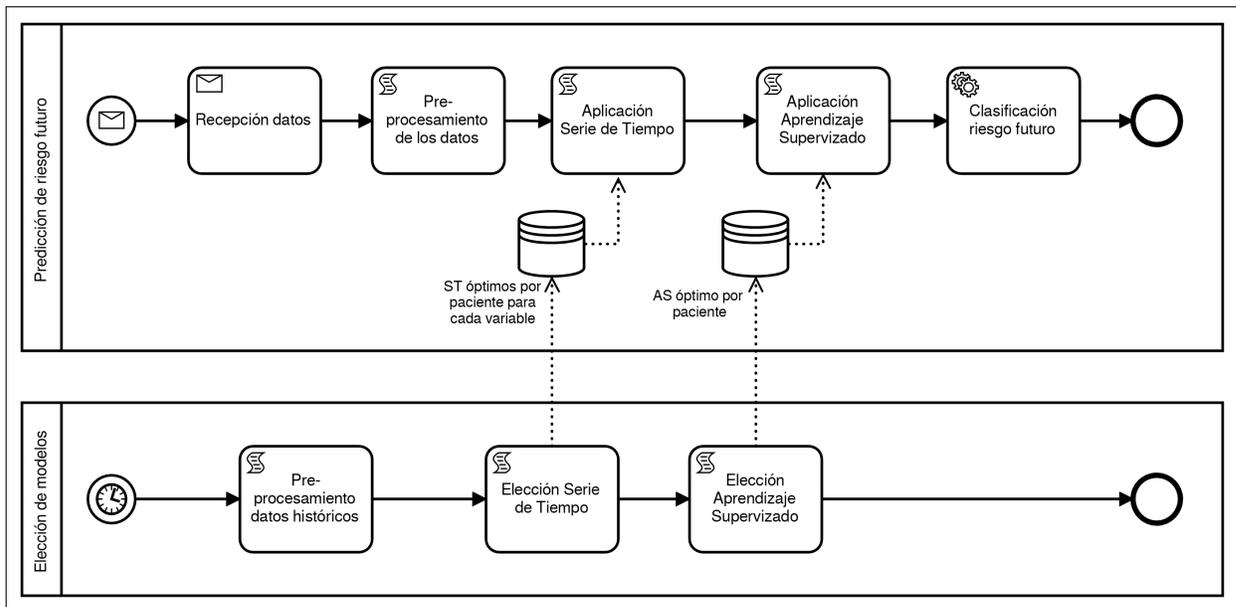


Figura 4.1: Descripción general algoritmo. Fuente: elaboración propia

A continuación se describe el trabajo realizado en el desarrollo del algoritmo, así como el funcionamiento del mismo. Este capítulo está dividido en cuatro secciones, primero las consideraciones y restricciones a tener en cuenta tanto en la concepción como en la elaboración del algoritmo; luego se profundiza sobre el preprocesamiento y transformaciones de los datos, es decir, toda la manipulación, adaptación y modificación que se debió realizar en virtud de tener mejores resultados con los modelos, finalmente, las dos secciones restantes hacen alusión al algoritmo mismo y como funcionan las etapas de elección de modelos y de predicción de riesgo futuro, ya explicitadas.

4.1. Consideraciones y restricciones

La lista que se detallada en esta sección engloba una serie de reglas que es necesario sopesar tanto en el desarrollo del algoritmo como en la futura implementación del mismo que, si bien no es parte de este estudio, es necesario considerar al buscar la utilización y aplicabilidad del mismo:

1. Los pacientes no pueden ser identificados individualmente, por lo que no se cuenta con información personal (nombre, rut, domicilio, etc.).
2. La ventana de predicción debe ser lo suficientemente amplia para desencadenar acciones preventivas (no debería ser menor a media hora).
3. El algoritmo debe ser capaz de continuar su funcionamiento frente a la pérdida de uno o más signos vitales, ya sea parcial o total, previendo la posibilidad de mal funcionamiento de un sensor o la desconexión del mismo.
4. El algoritmo debe ser capaz de restablecer su funcionamiento inmediatamente luego de desconexión de un paciente por un período acotado, en virtud de responder frente al

término de batería del dispositivo de captura o la desconexión total de éste por algún período aislado.

5. El algoritmo debe ser capaz de entregar una respuesta instantánea, no permitiéndose un tiempo de procesamiento mayor a algunos segundos.

Las secciones posteriores de este capítulo detallarán cómo es abordado cada uno de los puntos expuestos anteriormente, cuando corresponda.

4.2. Preparación de la data

Esta sección detalla todo el trabajo realizado con la data previo a la aplicación de los modelos, con el objetivo de aumentar la exactitud de éstos y reducir su tiempo de procesamiento.

4.2.1. Interpolación

En el contexto de las series de tiempo, si bien existen modelos que permiten la extrapolación de una serie de datos con intervalos irregulares, la generalidad se modela de forma regular en un conjunto de datos equidistantes entre ellos.

Al considerar que la captura de los datos de cada paciente no está espaciado de forma equidistante en el tiempo, pues por un lado los tiempos de captura en la unidad de pediatría y de pacientes críticos son distintos, y por otro existen ventanas sin información cuando a los pacientes se les realiza algún procedimiento más complejo, las condiciones ideales para modelar series de tiempo regulares no se cumplen, por lo que es necesario realizar una técnica matemática para regularizar artificialmente la distancia entre datos.

El algoritmo desarrollado resuelve lo anterior mediante la técnica de interpolación lineal, la que matemáticamente se define de la siguiente forma:

$$f(x | (x_1; x_2)) = f(x_1) + \frac{f(x_2) - f(x_1)}{x_2 - x_1} (x - x_1) \quad (4.1)$$

donde, para efectos de la data con que se trabajará, $f(x | (x_1; x_2))$ representa algún signo vital a estimar del período x , conociendo los valores próximos de dicho signo, $f(x_2)$ y $f(x_1)$, de los períodos x_1 y x_2 respectivamente.

Lo anterior además también resuelve la restricción número 4 de la sección 4.1.

4.2.2. Creación de variables

Según lo mostrado en el capítulo 3, específicamente en la sección 3.2, se aprecia que distintos signos vitales se comportan de forma distinta en distintas subpoblaciones de la muestra.

Para los modelos de aprendizaje supervisado es importante que estas diferencias sean capturadas por nuevas variables, las que en principio serían una combinación de variables categóricas, que definen las subpoblaciones, con variables numéricas, que definen los signos vitales.

De esta forma, de lo señalado en la figura 3.5, parte de la subsección 3.2.2, la frecuencia cardiaca presenta un comportamiento diferente entre el género masculino y femenino, por lo que se define la siguiente variable:

$$FC_{Genero} = f(\text{genero}) \times FC \quad \text{con } f(\text{genero}) = \begin{cases} \text{si } genero = Masculino & 1 \\ \text{si } genero = Femenino & 0 \end{cases} \quad (4.2)$$

donde FC_{Genero} representa la variable creada, que toma el valor de FC cuando se es de género masculino y 0 cuando se es de género femenino.

También en la subsección 3.2.2, la figura 3.6 muestra que la frecuencia cardiaca tiene un comportamiento distinto entre los distintos grupos de edades propuestos por Mancilla et al. [25], por lo que se define la siguiente variable:

$$FC_{Edad} = f(\text{edad}) \times FC \quad \text{con } f(\text{edad}) = \begin{cases} \text{si } edad \in [0\text{meses}, 3\text{meses}[& 1 \\ \text{si } edad \in [3\text{meses}, 6\text{meses}[& 2 \\ \text{si } edad \in [6\text{meses}, 2\text{años}[& 3 \\ \text{si } edad \in [2\text{años}, 10\text{años}[& 4 \\ \text{si } edad \in [10\text{años}, \infty[& 5 \end{cases} \quad (4.3)$$

donde FC_{Edad} representa la variable creada, FC es el valor de la frecuencia cardiaca para el paciente y $f(\text{edad})$ es el grupo etario en que quedaría clasificado. Es importante mencionar que sobre este grupo que dentro de \mathbb{R} es trabajado como variable tipo factor, lo que implica que no es una variable numérica de 5 valores sino que se trata como una variable categórica con 4 opciones posibles.

Finalmente, la figura 3.6 de la subsección 3.2.3 muestra que la frecuencia respiratoria también tiene un comportamiento distinto entre los grupos de edades propuestos por Mancilla et al. [25], por lo que se define la siguiente variable:

$$FR_{Edad} = f(\text{edad}) \times FR \quad \text{con } f(\text{genero}) = \begin{cases} \text{si edad} \in [0\text{año}, 1\text{año}[& 1 \\ \text{si edad} \in [1\text{año}, 2\text{años}[& 2 \\ \text{si edad} \in [2\text{años}, 5\text{años}[& 3 \\ \text{si edad} \in [5\text{años}, 12\text{años}[& 4 \\ \text{si edad} \in [12\text{años}, \infty[& 5 \end{cases} \quad (4.4)$$

donde FR_{Edad} representa la variable creada, FR es el valor de la frecuencia respiratoria y $f(\text{edad})$ representa el grupo etario. Nuevamente es importante mencionar que este último valor hace alusión a una variable de tipo factor.

4.2.3. Variable objetivo

La primera aproximación frente a la creación de la variable objetivo fue considerar sólo el apoyo respiratorio, capturado en la variable de ventilación explicitada en la sección 3.3, como el elemento para discernir la gravedad de un paciente. Los pacientes con bajo apoyo, es decir, con respiración ambiental, naricera o halo - venturi, son considerados en estado normal, mientras que el resto que utiliza mascarilla con recirculación, NAF, VMNI o VMI son los pacientes en estado alterado o de riesgo.

La gran inconsistencia de lo señalado es que no sopesa el trabajo del equipo clínico, el cual es estabilizar al paciente independiente del tipo de apoyo respiratorio con que esté, siendo éste sólo un elemento más en la labor de estabilización. De esta forma, es esperable que para cada tipo de apoyo respiratorio, independiente de su complejidad, se tengan estados normales y estados alterados del paciente.

Al tener esto en consideración, se construye la variable objetivo al definir como base el estado normal del paciente, a menos que exista un cambio en el apoyo respiratorio hacia uno de mayor asistencia.

Esto último se puede visualizar al tener en cuenta la disposición de la lista mostrada en la sección 3.3, la cual está ordenada según gravedad. De esta forma, cualquier cambio hacia un tipo de apoyo respiratorio que esté abajo en la lista implicaría que hubo un estado de riesgo de salud en el instante previo a dicho cambio, que fue lo que visualiza el equipo clínico. La única excepción a esto sería cuando el paciente tiene apoyo respiratorio de NAF, VMNI o VMI con un cambio en los parámetros de configuración (cantidad de oxígeno por ejemplo), donde se pueden alternar entre estados normales o alterados dependiendo de los parámetros mencionados sin que exista un cambio en el apoyo respiratorio.

La tabla 4.1 sintetiza la serie de reglas expuestas, al considerar el apoyo respiratorio que se tiene en un estado inicial (Estado 1) a un estado posterior (Estado 2). Se define 0 como estado normal y 1 como estado alterado o de riesgo. De esta forma y a modo de ejemplo, si en el Estado 1 se está con respiración ambiental cualquier cambio en el apoyo respiratorio para el Estado 2 implicará que el período anterior a dicho cambio es un estado alterado, pues indica que el equipo clínico visualizó algo en la condición del paciente que generaba

tener mayor resguardo. Por otro lado, cambios en el apoyo respiratorio a alguno de menor complejidad, como sería pasar de NAF en el Estado 1 a Naricera en el Estado 2, indica que el paciente mantuvo un estado de normalidad extendido.

		Estado 2						
		Amb.	Nar.	Halo	Masc.	NAF	VMNI	VMI
Estado 1	Ambiental	0	1	1	1	1	1	1
	Naricera	0	0	1	1	1	1	1
	Halo	0	0	0	1	1	1	1
	Mascarilla	0	0	0	0	1	1	1
	NAF	0	0	0	0	–	1	1
	VMNI	0	0	0	0	0	–	1
	VMI	0	0	0	0	0	0	–

Tabla 4.1: Construcción variable objetivo. Fuente: Elaboración propia

4.3. Etapa I: Elección de modelos

Esta primera etapa del algoritmo pretende realizar la elección del conjunto de modelos óptimos para cada paciente, para así reducir los tiempos de procesamiento en productivo, en función de poder cumplir con la restricción número 5 de la sección 4.1.

Los modelos seleccionados deben cumplir con criterios distintos dependiendo de si se trata de modelos de series de tiempo o de aprendizaje supervisado, dado por las métricas de medición detalladas en las subsecciones 2.4.3 y 2.4.3 respectivamente.

Nuevamente, esta etapa está dividida en dos fases, siguiendo la lógica híbrida del algoritmo, ya explicada.

4.3.1. Elección de modelo de series de tiempo

Esta primera fase de esta etapa busca modelar cada signo vital mediante cada modelo de serie de tiempo, para así encontrar cuales son los que mejores responden para cada signo vital para cada paciente y luego sólo aplicarlos.

Para esto, ya finalizada la normalización de la data de forma equidistante de acuerdo a lo expuesto en la subsección 4.2.1, se distribuye en conjuntos de datos que modelan 3 días (36 registros considerando la distribución de registros cada 2 horas), con lo que se busca extrapolar un dato con una ventada de dos horas, y así compararlo con el registro real. Se itera este proceso hasta haber probado con cada conjunto de datos que forme tres días consecutivos, lo que deja un registro extra para poder contrastar respecto al valor real.

Cada uno de estos conjuntos de datos es evaluado y se extrapola con los modelos detallados en la subsección 2.4.1, teniendo en consideración que para los modelos parametrizables se

toma cada uno de los siguientes valores y combinaciones que de éstos se desprendan:

- $AR(p)$: $p \in \{1, 2, 3, 4\}$.
- $MA(q)$: $q \in \{1, 2, 3, 4\}$.
- $ARMA(p, q)$: $p \in \{1, 2, 3, 4\}$ y $q \in \{1, 2, 3, 4\}$.
- $ARIMA(p, d, q)$: $p \in \{1, 2, 3, 4\}$, $d = 1$ y $q \in \{1, 2, 3, 4\}$.

De esta forma, al considerar todas las combinatorias de los modelos señalados en el listado anterior, cada signo vital de cada conjunto de datos se contrasta contra 43 modelos distintos de series de tiempo.

Se define que se realizará con tres días de historia del paciente pues de acuerdo a lo planteado por Richter et al. [40] diversos signos vitales siguen un ciclo circadiano, lo que determina una tendencia de comportamiento diario de éstos. Se ha estudiado que con al menos 3 períodos se puede obtener una aproximación del comportamiento estacional de una serie de tiempo.

Sin desmedro de lo anterior planteado, la cantidad de días utilizado para evaluar y posteriormente para extrapolar se puede cambiar en el archivo *Glob Var.R*, específicamente en la variable *VG.Dias*, ubicado en la carpeta raíz del algoritmo.

A modo de ejemplo, la figura 4.2 muestra el funcionamiento de esta fase del algoritmo, para el caso específico de la temperatura, para un paciente del cual se tiene 4 días de historia.

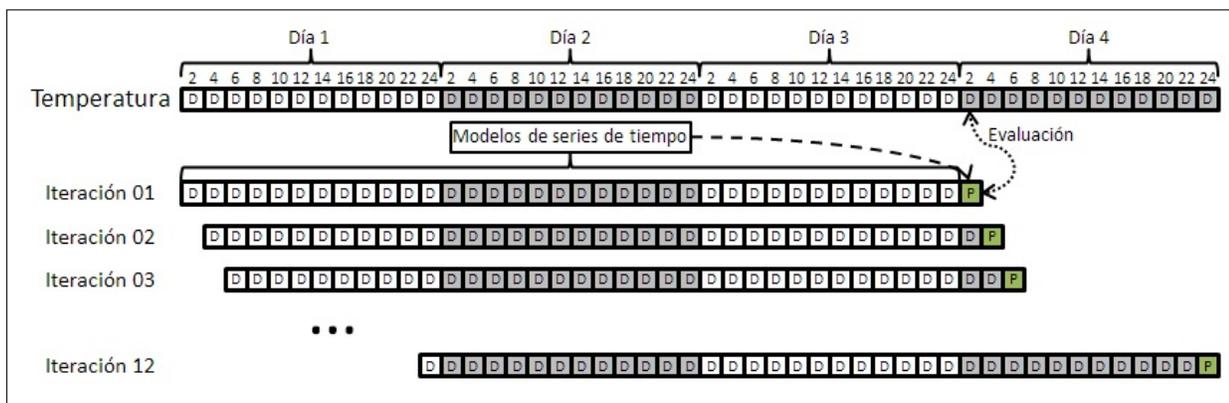


Figura 4.2: Elección serie de tiempo. Fuente: elaboración propia

En este caso, se tiene 48 registros de la temperatura, todos equidistantes y sin registros faltantes producto de la etapa previa de preparación de la data. La primera iteración del algoritmo considera los 36 registros iniciales, que constituyen los primeros tres días de datos. Éstos se modelan con los 43 métodos de serie de tiempo mencionados y se extrapola para obtener el valor proyectado, representado con un cuadro verde con la letra P en su interior. Dicho valor se contrasta con su valor real mediante el error cuadrático y se procede con la siguiente iteración, la cual contempla los valores que representen 3 días siguientes partiendo exactamente desde el segundo registro consignado para el paciente.

Una vez finalizada las 12 iteraciones, al considerar que la última contempla hasta un registro antes del registro final para así poder evaluar el valor extrapolado, se analiza cuales

son las series de tiempo que tuvieron un mejor desempeño tanto al evaluar con ECM como con MMEC, detallados en la subsección 2.4.3, y esas quedan consignadas como las dos series a utilizar para este paciente.

Si bien esta comparativa la realiza internamente el algoritmo, se deja dos archivos de respaldo que permiten revisar el detalle de la elección.

La tabla 4.2 es un extracto del primero de estos archivos mencionados, emulado para el caso ejemplificado en la figura 4.2. Ésta representa el detalle de lo realizado por el algoritmo, donde cada fila es el resultado de una iteración. Los campos mostrados comienzan con la caracterización de la iteración (“Id paciente”, “Fecha” y “T”), para luego tener conjuntos de tres campos que representan cada modelo contrastado. Específicamente la tabla ejemplifica con el método de alisamiento exponencial, donde se muestra primero el valor proyectado (“AlixExpo”), luego se consigna si la heurística generó algún error (“A.E. Error”) para terminar con el error cuadrático asociado a la proyección (“A.E. EC”).

Esta secuencia de tres campos se detalla para cada uno de los 43 modelos, por lo que en total se tiene 132 columnas (3 de identificación y 129 de resultados de los modelos).

Id Paciente	Fecha	T	AlisExpo	A.E. Error	A.E. EC
8	Día 4 02:00	37	37,799945	No error	0,63991199
8	Día 4 04:00	37,8	37,249945	No error	0,30256052
8	Día 4 06:00	36,7	36,1267448	No error	0,32862149
			⋮		...
8	Día 4 20:00	37	37,4753683	No error	0,22597499
8	Día 4 22:00	36,7	36,1267448	No error	0,32862149
8	Día 4 24:00	36,2	35,850316	No error	0,12227893

Tabla 4.2: Salida 01 algoritmo ST, temperatura, extracto. Fuente: Algoritmo predicción

El segundo archivo mencionado consolida lo realizado por las iteraciones, transversalmente a todos los signos vitales. La tabla 4.3 muestra un extracto de este archivo, donde se tiene en las filas los distintos modelos aplicados para cada signo vital (43 filas), y en las columnas el resultado consolidado de dicho modelo para cada signo, teniendo en cuenta que se consignan tres valores para evaluar los mejores modelos para cada caso: error cuadrático medio (**Signo ECM** en la tabla), máximo error cuadrático (**Signo MEC** en la tabla) y la cantidad de errores (**err.** en la tabla).

La elección de los modelos se realiza al evaluar el método que posea un menor error cuadrático medio y al buscar que se ajusten de mejor forma al signo vital. El que posea un menor máximo error cuadrático, lo que busca que se ajuste de mejor forma a los cambios específicos de algún signo vital, que sería lo esperable en escenarios adversos.

De esta forma a modo de ejemplo, si se considera sólo los signos vitales y métodos mostrados en la tabla 4.3, para el caso de la temperatura se escogería $AR(1)$ por el criterio de ECM y $ARIMA(4,1,4)$ por el criterio de MMEC. Para el caso de la saturación, se consideraría *Alisamiento Exponencial* tanto por el criterio de ECM como por el criterio de MMEC.

ST	T ECM	T MEC	err.		Sat ECM	Sat MEC	err.
AlisExpo	0,28259902	1,18195478	0		0,81182843	2,94115493	0
HoltWinters	0,84078295	4,13410668	0		1,58960572	8,03897648	0
MovArmon	3,30524505	13,3811948	1		5,49101939	14,9289417	1
AR(1)	0,28041589	1,36791092	0	...	1,16259798	5,63861709	0
AR(2)	0,28215455	1,30772016	0		1,07377035	4,55637882	0
AR(3)	0,2988947	1,30806452	0		1,05743588	4,01721181	0
	⋮					⋮	
ARIMA(4,1,4)	0,37328514	1,13708438	8		0,96227944	4,4127333	2

Tabla 4.3: Salida 02 algoritmo ST, extracto. Fuente: Algoritmo predicción

Ambas elecciones se consignan en una base de datos, que registra el id del paciente, la fecha en que se evaluó la elección de los mejores signos vitales para el paciente y los dos modelos de series de tiempo para cada signo vital.

Este proceso se repite para cada signo vital de este paciente, y luego para cada paciente presente en el estudio.

4.3.2. Elección de modelo de aprendizaje supervisado

Esta fase de esta etapa busca utilizar la combinación de variables disponibles y con éstas determinar que modelo es el que mejor clasifica si el paciente se encuentra en un estado normal o alterado.

En esta fase también es importante ser capaz de cumplir la restricción número 3 de la sección 4.1, sobre todo el apartado que hace alusión a una pérdida total de uno o más signos vitales.

La primera aproximación para resolver este apartado es considerar todas las variables independientes presentes, es decir, los signos vitales registrados por el equipo clínico, el género y edad como caracterización del paciente, y las variables creadas en la subsección 4.2.2, y con éstas predicen la variable dependiente, es decir, el estado del paciente.

No obstante lo anterior no permite abordar la pérdida de signos vitales de la restricción ya señalada, pues el objetivo de dividir el algoritmo en dos etapas es reducir los tiempos de procesamiento en función de la restricción número 5 de la misma lista. Por tanto de esta fase se busca obtener el modelo de aprendizaje supervisado que mejor se comporta para cada paciente y dejarlo construido con sus parámetros establecidos para luego sólo aplicarlo, sin embargo esto no permitiría luego cambiar el conjunto de variables independientes que se le cargaría al modelo (al considerar la pérdida de un signo vital).

Es por esto que se busca resolver ambas problemáticas al evaluar el modelo que tiene mejor comportamiento para cada combinatoria de variables posibles. Es decir, se busca el modelo que mejor capacidad predictiva tenga cuando se cuente sólo con un signo vital (sea cual sea éste), cuando se cuente con dos signos (cualquier combinación de éstos), con tres o

con los cuatro establecidos como base del estudio.

El algoritmo resuelve lo anterior asignándole un determinado número primo a cada signo vital (2 a temperatura, 3 a saturación, 5 a frecuencia cardiaca y 7 a frecuencia respiratoria), y luego evalúa y selecciona un modelo de aprendizaje supervisado para cada producto de la combinación de estos números. La tabla 4.4 muestra algunas de las 15 combinaciones posibles considerando los cuatro signos vitales que pudiesen estar ausentes en el análisis, donde también se muestra las variables que estarían incluidas.

Comb.	Edad	Gén.	T	Sat	FC	FC x Ed.	FR	FR x Ed.	FR x Gén.
2	x	x	x						
3	x	x		x					
5	x	x			x	x			
6	x	x	x	x					
7	x	x					x	x	x
10	x	x	x		x	x			
30	x	x	x	x	x	x			
35	x	x			x	x	x	x	x
210	x	x	x	x	x	x	x	x	x

Tabla 4.4: Combinatoria variables, extracto. Fuente: Elaboración propia.

De esta forma el algoritmo itera 15 veces en la búsqueda de cada modelo que discrimine de mejor forma el estado normal del alterado para cada conjunto de variables de cada paciente.

Cada iteración tiene dos partes: validación y construcción del modelo.

La validación se encarga de contrastar cada modelo contra el conjunto de datos correspondiente (de acuerdo a la combinación de variables) mediante la técnica de validación cruzada, donde se considera el 80 % de los datos para entrenar y 20 % de los datos para evaluar. Este circuito se realiza 5 veces hasta que todos los datos hayan sido usados como base de validación. La figura 4.3 representa este procedimiento, específicamente la cuarta iteración del modelo, es decir, donde están presente las variables de temperatura y saturación, donde se muestra la primera y última validación cruzada realizada.

En cada validación se contrasta cada estado predicho con su equivalente valor real, y con esto se obtiene la cantidad de casos verdaderos positivos (*VP*), verdaderos negativos (*VN*), falsos positivos (*FP*) y falsos negativos (*FN*). Al término de la quinta validación, ya teniendo todos los valores predichos contrastados, se procede a evaluar cada modelo utilizado con las métricas detalladas en la subsección 2.4.3, para de esta forma tener el rendimiento general de cada modelo para el conjunto de variables definidos por la iteración.

Si bien tanto la validación como la elección del modelo con mejor rendimiento el algoritmo lo realiza automáticamente, en la validación de cada iteración se registra el rendimiento de los modelos contrastados. La tabla 4.5 muestra el resultado de la iteración 04, también detallada en la figura señalada anteriormente para un paciente determinado ($id = 13$), y se muestra así

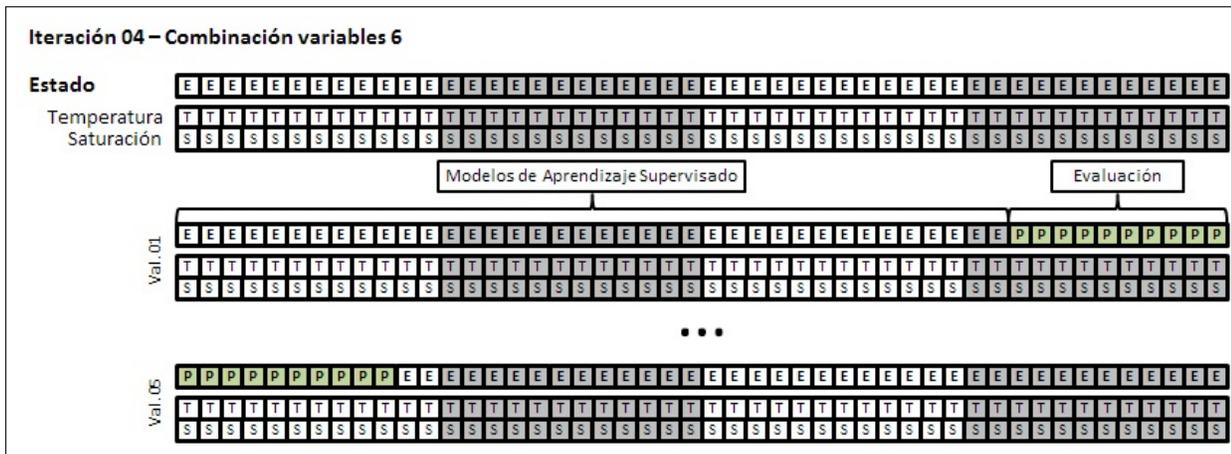


Figura 4.3: Elección modelo aprendizaje supervisado. Fuente: elaboración propia

que en este caso el modelo con mejor comportamiento sería el árbol de decisión, al considerar como métrica discriminante la exactitud.

Modelo	Comb.	Exactitud	Precisión	Sensib.	Valor F	Especif.
Logit	6	0.8412371	0.2	0.2926829	0.2376237	0.8918918
DT	6	0.90515463	0.3913043	0.2195121	0.28125	0.9684684
ANN	6	0.8154639		0		1
SVM	6	0.8412371	0.125	0.1463414	0.1348314	0.9054054

Tabla 4.5: Rendimiento modelos iteración 04. Fuente: Algoritmo predicción.

La construcción del modelo toma todo lo calculado en la parte de validación y considera alguna métrica (parametrizable pero fijada por defecto en la exactitud) para seleccionar el modelo a construir por cada iteración para cada paciente, es decir, construye 15 modelos por paciente.

Para la construcción de éstos se combina los modelos generados en cada validación de la etapa anterior y con esto se genera un modelo final, lo que busca reducir el sobre-ajuste que podría darse si se utilizara toda la data para la construcción de éstos.

Los modelos quedan almacenados en una carpeta distinta para cada paciente en formato *.rda, que permite a R cargarlo y utilizarlo sin procesamiento adicional.

4.4. Etapa II: Clasificación de riesgo futuro

Esta segunda etapa busca utilizar la información y modelos generados en la *Etapa I* y así entregar una respuesta del estado del paciente con un tiempo mínimo de procesamiento.

Para lo anterior es necesario contar con tres elementos: tres días de datos del paciente, información sobre los métodos de serie de tiempo con mejores desempeños para cada signo

vital y los modelos de aprendizaje supervisado construidos para cualquier combinación de variables de entrada.

La figura 4.4 detalla el proceso de clasificación de riesgo futuro usando el modelo y notación de procesos de negocio, o BPMN por sus siglas en inglés (*Business Process Model and Notation*) [10], donde se destaca que esta parte del algoritmo se desencadenaría con la llegada de la información señalada antes (el primero de los elementos necesarios), y rescata desde las bases internas los métodos de serie de tiempo y carga los modelos de aprendizaje supervisado a utilizar (los dos elementos necesarios restantes).

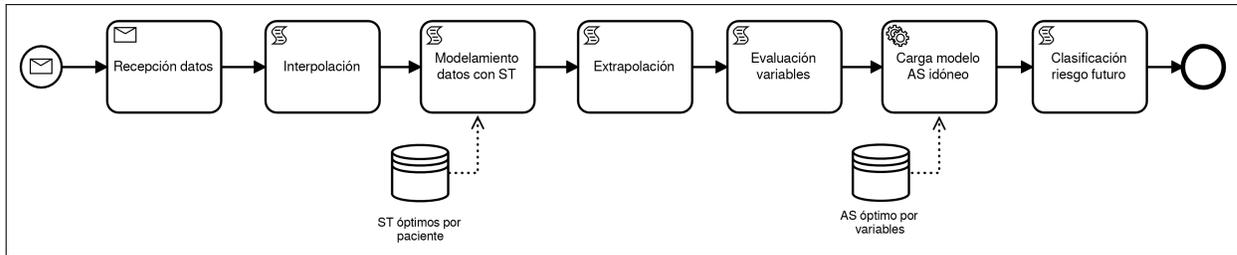


Figura 4.4: Proceso desencadenar predicción. Fuente: elaboración propia

1. **Recepción de los datos:** básicamente es la captura e incorporación de la data referente a los últimos tres días del paciente por parte del algoritmo.
2. **Interpolación:** como su nombre lo indica, aquí se interpola para homologar los datos ingresados cuando existan ventanas sin datos. Además se realizan otros preprocesamientos básicos asociados.
3. **Modelamiento con ST:** se captura los dos modelos con mejor rendimiento por signo vital para el paciente en ejecución y con estos se modela los signos recibidos.
4. **Extrapolación:** se proyecta cada signo vital un período posterior al último registro ingresado, con ambos modelos de serie de tiempo. Se obtiene dos conjuntos de signos vitales proyectados.
5. **Evaluación de variables:** se define qué variables se tiene disponibles (que hayan llegado en la recepción y que hayan sido proyectadas sin errores), y en caso de ser necesario, se crean las variables adicionales detalladas en la subsección 4.2.2.
6. **Carga modelo AS idóneo:** de acuerdo a la combinatoria de variables disponible, se carga el modelo correspondiente.
7. **Clasificación riesgo futuro:** se considera el conjunto de signos vitales proyectados y con éstos se calcula si esta combinación corresponde a un estado normal o alterado.

Capítulo 5

Resultados y Conclusión

A continuación se detallan los principales resultados obtenidos de la aplicación del algoritmo desarrollado con la data base con que se trabajó en este estudio. También se plantean las conclusiones en torno al mismo y se definen los aspectos que se visualizan como fundamentales a considerar de cara al trabajo futuro.

5.1. Resultados

Si bien el resultado final que se busca contrastar en este estudio es la capacidad predictiva de un estado de riesgo futuro por parte del algoritmo, al tratarse de un modelo híbrido y donde a su vez la etapa relacionada con las series de tiempo tiene múltiples modelos en paralelo (dos por cada signo vital), el generar una evaluación global puede ocultar errores de desempeño individual entre éstos que se compensen y que en definitiva generen una correcta predicción, sin necesariamente estar frente a un conjunto de modelos con buena capacidad predictiva de forma individual.

De esta forma, y para evitar situaciones en las que, por ejemplo, una incorrecta extrapolación del siguiente estado del paciente (dado por el conjunto de signos vitales extrapolados) sumado a un modelo de clasificación de riesgo con bajo rendimiento den resultados acertados, la evaluación de los modelos se realizará de forma independiente.

Es importante señalar que dado las restricciones de construcción del modelo, específicamente la que hace alusión a que se necesitan tres días para predecir el siguiente estado y por ende un tiempo mayor a este para determinar los modelos con mejor rendimiento por signo vital, además de la complejidad para construir la variable objetivo, que necesita que exista al menos un cambio del apoyo respiratorio del paciente hacia uno de mayor gravedad durante su estadía en el centro de asistencia médica, no todos los pacientes detallados en el capítulo 3 son contrastados en estos resultados, sino que la lista se reduce a 51, quienes cumplieran los requisitos anteriores.

5.1.1. Resultados series de tiempo

Como se señala en la subsección 4.3.1, de cada ejecución del algoritmo contra la data del paciente se obtienen dos series de tiempo por signo vital: la que tuvo mejor desempeño evaluando el ECM y la que presentó el MMEC.

La tabla 5.1 detalla la cantidad de veces que cada método de serie de tiempo fue seleccionado por cada signo vital para algún paciente, de acuerdo al criterio de ECM.

Se observa que de manera transversal entre los signos vitales el método que tuvo mejor rendimiento fue alisamiento exponencial, siendo el método con mejor ECM en el 76 % de los pacientes para la temperatura, 73 % para la saturación de oxígeno, 57 % para la frecuencia cardiaca y 78 % para la frecuencia respiratoria; seguido por el proceso autorregresivo de medias móviles, con parámetros $p = 1$ y $q = 1$, cuyo mejor desempeño fue al modelar la frecuencia cardiaca, donde para un 18 % de los pacientes presentaba el menor ECM.

Doce son los casos no mostrados en la tabla, aglutinados en otros, los que tuvieron el mejor desempeño sólo para un signo vital de un paciente.

	T	Sat	FC	FR
Alisamiento exponencial	39	37	29	40
ARMA(1,1)	2	4	9	3
MA(2)	1	3	1	
ARMA(2,1)	2		2	
AR(1)		2	2	
AR(3)	3			
MA(1)	2			1
ARMA(1,2)		2		1
ARMA(3,1)		1	2	
ARMA(3,2)	1	1		
ARIMA(1,1,2)			1	1
Otros	1	1	5	5

Tabla 5.1: Desempeño series de tiempo: ECM. Fuente: elaboración propia.

Por otro lado, la tabla 5.2 detalla la cantidad de veces que cada serie de tiempo fue seleccionado para algún signo vital de los pacientes, pero considerando el criterio MMEC.

Al igual que en el caso anterior el método de alisamiento exponencial fue el que tuvo mejor rendimiento transversalmente entre los signos vitales. Al modelar la temperatura fue el con mejor rendimiento en el 29 % de los pacientes, en el 31 % de los pacientes para la saturación, mientras para frecuencia cardiaca y respiratoria en un 27 % y 29 % respectivamente.

El método que lo sigue bajo la misma métrica es el proceso de medias móviles, con parámetros $q = 1$ y $q = 2$. Dichos métodos fueron capaces de modelar correctamente el 6 % de los signos vitales de todos los pacientes, destacándose $MA(q=1)$ modelando la frecuencia cardiaca, donde para el 12 % de los pacientes es el modelo con mejores resultados específicamente en ese signo vital.

En este caso los métodos que modelan de mejor forma los signos vitales están más disgregados que en el caso anterior, quedando fuera de la tabla 28 métodos que destacaban en 3 o menos signos vitales de pacientes distintos.

	T	Sat	FC	FR
Alisamiento exponencial	15	16	14	15
MA(1)	3	6	3	2
MA(2)	3	3	4	3
ARMA(1,1)	1	2	3	3
AR(1)		4	4	1
ARMA(3,3)	3	1	2	
ARMA(1,3)	1	1	3	1
ARMA(1,2)	2	1		2
ARIMA(3,1,4)			3	2
AR(4)	2		2	
AR(2)	1		2	1
Holt Winters	1		1	2
ARMA(3,2)	1	2	1	
ARIMA(1,1,2)		1	2	1
Otros	18	14	7	18

Tabla 5.2: Desempeño series de tiempo: MMEC. Fuente: elaboración propia.

A continuación se presenta el detalle de los resultados abierto por signo vital modelado.

5.1.1.1. Series de tiempo al modelar la temperatura

La tabla 5.3 muestra los 5 mejores desempeños de series de tiempo al modelar la temperatura al utilizar la métrica de ECM.

Si se considera el mejor caso mostrado, el del paciente $id = 27$, se aprecia que tiene un error cuadrático de 0,011, por lo que al evaluar el error promedio se observa que la capacidad de extrapolación de esa serie en ese signo vital para ese paciente tiene un error promedio de $0,1048^{\circ}C$.

Serie de tiempo	Paciente	ECM	Errores
Alisamiento exponencial	27	0,0110	0
Alisamiento exponencial	31	0,0393	0
MA(1)	48	0,0441	0
Alisamiento exponencial	43	0,0452	0
Alisamiento exponencial	36	0,0467	0

Tabla 5.3: Series de tiempo en T: ECM. Fuente: elaboración propia.

La tabla 5.4 muestra los 5 mejores desempeños de serie de tiempo al modelar la temperatura utilizando la métrica MMEC.

Nuevamente al tomar el mejor caso expuesto, para el paciente $id = 14$ se tendría un error máximo en alguna extrapolación de $0,4667^{\circ}C$, es decir, es esperable que cualquier extrapolación esté bajo ese error.

Serie de tiempo	Paciente	MMEC	Errores
Alisamiento exponencial	14	0,2178	0
ARMA(4,4)	36	0,2304	11
ARMA(3,4)	48	0,2369	2
ARIMA(3,1,1)	43	0,3716	0
MA(1)	41	0,3724	0

Tabla 5.4: Series de tiempo en T: MMEC. Fuente: elaboración propia.

De lo expuesto se aprecia que la temperatura puede ser modelada correctamente con los métodos aplicados, teniendo un espaciado de dos horas.

5.1.1.2. Series de tiempo al modelar saturación

La tabla 5.5 muestra los 5 mejores desempeños de series de tiempo al modelar la saturación, al contrastar con la métrica de ECM.

Del caso del paciente $id = 27$ se tiene un ECM de 0,2178, por lo que el error promedio está en torno a $0,44pp$ de saturación de oxígeno.

Serie de tiempo	Paciente	ECM	Errores
Alisamiento exponencial	27	0,1970	0
AR(1)	48	0,3703	0
ARMA(1,1)	30	0,4653	0
Alisamiento exponencial	49	0,4824	0
Alisamiento exponencial	25	0,6426	0

Tabla 5.5: Series de tiempo en Sat: ECM. Fuente: elaboración propia.

La tabla 5.6 expone los 5 mejores desempeños de series de tiempo al modelar saturación cuando se emplea la métrica MMEC.

Serie de tiempo	Paciente	MMEC	Errores
Alisamiento exponencial	48	1,2981	0
ARMA(4,4)	1	2,9412	0
ARMA(3,4)	30	3,4144	0
ARIMA(3,1,1)	49	4,0000	0
MA(1)	70	7,7339	0

Tabla 5.6: Series de tiempo en Sat: MMEC. Fuente: elaboración propia.

Si en este caso se considera el quinto mejor desempeño, tenemos que ya para el paciente $id = 70$ se tiene un error cuadrático de 7,7339, lo que significa que el máximo error en este caso presentó una diferencia de $2,78pp$ de saturación de oxígeno.

De lo mostrado, si bien se tiene series que se ajustan de muy buena forma a la saturación de los pacientes a nivel general, no son capaces de responder con el mismo buen desempeño frente a cambios más abruptos (capturado por la métrica de MMEC), por lo que se recomendaría reevaluar el desempeño con un período menor a las dos horas.

5.1.1.3. Series de tiempo al modelar frecuencia cardiaca

La tabla 5.7 muestra los 5 mejores desempeños de series de tiempo al modelar la frecuencia cardiaca, al contrastar con la métrica de ECM.

El paciente $id = 27$, para el cual el modelado de la frecuencia cardiaca se ajusta de mejor forma, tiene un ECM de 35,9022, por lo que el error promedio de la serie al extrapolar la frecuencia cardiaca está en torno a los 5,99 latidos por minutos.

Serie de tiempo	Paciente	ECM	Errores
Alisamiento exponencial	27	35,9022	0
MA(2)	33	41,2811	0
ARMA(1,1)	4	44,8249	1
Alisamiento exponencial	30	48,5372	0
Alisamiento exponencial	15	49,3166	0

Tabla 5.7: Series de tiempo en FC: ECM. Fuente: elaboración propia.

La tabla 5.8 expone los 5 mejores desempeños de series de tiempo al modelar frecuencia cardiaca cuando se emplea la métrica MMEC.

El paciente $id = 48$ presenta un MMEC de 161,6910, lo que significa que el máximo error en este caso presentó una diferencia de 12,71 latidos por minutos.

Serie de tiempo	Paciente	MMEC	Errores
ARMA(4,2)	48	161,6910	2
ARMA(3,1)	49	272,4068	1
ARMA(1,1)	9	404,1390	0
MA(2)	33	451,1210	0
ARMA(1,3)	45	475,3795	0

Tabla 5.8: Series de tiempo en FC: MMEC. Fuente: elaboración propia.

De lo expuesto se aprecia que la frecuencia cardiaca, con datos registrados con dos horas de diferencia, no es posible modelarla de forma aceptable con alguno de los 43 modelos de series de tiempo expuestos.

5.1.1.4. Series de tiempo al modelar frecuencia respiratoria

La tabla 5.9 muestra los 5 mejores desempeños de series de tiempo al modelar la frecuencia respiratoria, al contrastar con la métrica de ECM.

Si se revisa el detalle del paciente $id = 43$, quinto en la tabla, se tiene un ECM de 2,2556, por lo que el error promedio de la serie al extrapolar la frecuencia respiratoria está en torno a los 1,5 respiraciones por minutos.

Serie de tiempo	Paciente	ECM	Errores
Alisamiento exponencial	49	1,5595	0
Alisamiento exponencial	48	1,7608	0
Alisamiento exponencial	36	2,1654	0
ARMA(1,1)	9	2,2499	2
Alisamiento exponencial	43	2,2556	0

Tabla 5.9: Series de tiempo en FR: ECM. Fuente: elaboración propia.

La tabla 5.10 expone los 5 mejores desempeños de series de tiempo al modelar frecuencia respiratoria cuando se emplea la métrica MMEC.

El paciente $id = 45$, quinto en la tabla, presenta un MMEC de 30,3119, lo que significa que el máximo error en este caso presentó una diferencia de 5,5 respiraciones por minuto.

Serie de tiempo	Paciente	MMEC	Errores
Alisamiento exponencial	48	9,0011	0
Alisamiento exponencial	36	16,0032	0
ARIMA(3,1,4)	9	17,5500	26
MA(2)	51	20,2812	0
ARMA(1,3)	43	20,3119	3

Tabla 5.10: Series de tiempo en FR: MMEC. Fuente: elaboración propia.

Al igual que lo concluido para la saturación, en este caso existen series que se ajustan de buena forma al modelado de la frecuencia respiratoria, pero no responden de buena forma frente a los cambios más grandes del signo vital.

5.1.2. Resultados aprendizaje supervisado

Como se señala en la subsección 4.3.2, específicamente lo mostrado en la tabla 4.4, para cada paciente debe ser seleccionado un total de 15 modelos, uno por cada combinación de variables seleccionadas.

La tabla 5.11 detalla la cantidad de veces que cada modelo de aprendizaje supervisado es seleccionado como el modelo con mejor ajuste para cada cantidad de variables disponibles, medido basándose en la métrica de exactitud.

Si se considera los 51 pacientes y los 15 modelos óptimos a elegir por paciente, se tiene un total de 765 modelos distintos elegidos (y construidos) por el algoritmo. De éstos, el 55% corresponde a SVM, el 28% corresponde a Logit, el 14% a DT y el 3% restante a ANN. Además se observa que en los primeros conjuntos de variables la tendencia hacia SVM no es tan clara, pero posterior al conjunto de variables '10' (que incluye **T** y **FC**) ya es este el

modelo predominante entre los pacientes. Sin desmedro de lo anterior, es importante notar que la exactitud media de este modelo es bastante estable y relativamente menor que el resto de los modelos, alcanzando una exactitud del 67%, mientras que para Logit, que es el siguiente en la cantidad de pacientes para los que es el modelo óptimo, la exactitud media es de 87%.

Conjunto de Variables	ANN		DT		Logit		SVM	
	Cant.	Exact. media	Cant.	Exact. media	Cant.	Exact. media	Cant.	Exact. media
2	4	72 %	15	78 %	21	89 %	11	67 %
3	2	82 %	8	77 %	24	87 %	17	69 %
5	1	73 %	9	73 %	19	88 %	22	65 %
6	2	85 %	12	74 %	21	87 %	16	73 %
7	3	74 %	6	86 %	17	84 %	25	68 %
10	1	73 %	9	72 %	14	88 %	27	68 %
14	4	82 %	6	76 %	12	88 %	29	64 %
15	1	73 %	8	71 %	15	89 %	27	68 %
21			6	79 %	15	84 %	30	70 %
30	2	84 %	10	69 %	11	88 %	28	72 %
35			4	84 %	12	84 %	35	66 %
42	1	76 %	4	72 %	14	87 %	32	66 %
70	1	97 %	4	76 %	6	84 %	40	66 %
105			4	73 %	8	88 %	39	67 %
210			2	79 %	9	88 %	40	66 %
Total	22	79 %	107	75 %	218	87 %	418	67 %

Tabla 5.11: Aprendizaje supervisado por variables: exactitud. Fuente: elaboración propia.

De la ecuación 2.17 se desprende que el Valor-F es un indicador que matemáticamente combina la Precisión y la Sensibilidad. La tabla 5.12 expone los 5 mejores modelos, de los seleccionados en base a la exactitud, ordenados por este indicador. Es importante considerar que pueden existir modelos con un Valor-F superior no considerados, pero dado que el algoritmo está parametrizado para realizar la elección en base a la Exactitud y no a la Sensibilidad (por ejemplo), éstos pudieron quedar fuera del listado final.

Pac.	Var. disp.	Modelo	Var. util.	Exact.	Precis.	Valor-F	Sensib.	Especif.
49	15	Logit	15	98 %	100 %	0,67	50 %	100 %
49	105	Logit	15	98 %	100 %	0,67	50 %	100 %
43	15	Logit	15	93 %	50 %	0,60	75 %	95 %
43	30	Logit	15	93 %	50 %	0,60	75 %	95 %
70	2	DT	2	78 %	50 %	0,58	69 %	81 %

Tabla 5.12: Aprendizaje supervisado: rendimiento Valor-F. Fuente: elaboración propia.

La tabla además posee el contraste entre variables disponibles, 'Var. disp.', y variables utilizadas, 'Var. util.'. Las variables disponibles define en qué iteración estaba el algoritmo para definir riesgo, y una vez contrastado el mejor modelo con ese conjunto de variables, el

algoritmo revisa entre los modelos anteriores de conjuntos de variables que sean múltiplos del valor que define la combinación de variables actual, y en caso de existir un modelo con mejor desempeño se deja este (con menos variables).

A modo de ejemplo, en la segunda fila se expone que el paciente $id = 49$, en la iteración definida por el conjunto de variables '105', es decir, las variables Sat, FC y FR (pues $3 \times 5 \times 7 = 105$), no logró mejorar el rendimiento del modelo obtenido sólo con las variables Sat y FC ($3 \times 5 = 15$), por ende se deja el mismo modelo como el óptimo de esta iteración, el que utiliza sólo el conjunto de variables '15'.

5.1.3. Evaluación global de resultados

Si bien se especificó al comienzo de esta sección que el generar un criterio de evaluación global de resultados podía restringir la capacidad predictiva de los modelos de forma aislada y/o esconder resultados deficientes de los modelos individuales, es importante generar una métrica comparativa para vislumbrar si el algoritmo híbrido planteado logra cumplir con los objetivos planteados.

Para esto se plantea una clasificación preliminar de los modelos dependiendo de la cantidad de casos clasificados correctamente para cada paciente, definido por alguna métrica definida. La tabla 5.13 detalla la nomenclatura expuesta para lo anterior, donde si por ejemplo se considerara como métrica la exactitud, se definiría que en los casos en que el modelo final sea capaz de predecir correctamente al menos el 80 % de los casos el conjunto óptimo de modelos sería considerado muy bueno, entre 70 % y 80 % el conjunto de modelos se consideraría bueno, y así hasta los conjuntos de modelos bajo 60 % que serían considerados malos.

Clasificación	Promedio métrica
Muy bueno	$80 \% \leq x$
Bueno	$70 \% \leq x < 80 \%$
Regular	$60 \% \leq x < 70 \%$
Malo	$x < 60 \%$

Tabla 5.13: Clasificación modelos por métrica. Fuente: elaboración propia.

De esta forma, se define el 80 % de la data capturada inicialmente por paciente como entrenamiento, con la cuál se construyen los modelos y los parámetros de éstos, y el 20 % restante como data de validación. Esto para aproximarse a la realidad de la ejecución, donde se tendrá data inicial de cada paciente y se buscará predecir la data incorporada recientemente.

La tabla 5.14 muestra lo anterior para la métrica de exactitud. En este caso, el algoritmo es capaz de proveer un modelo *Muy Bueno* en el 63 % de los casos, mientras no se entregan modelos que sean *Malos*. Por otro lado, sólo un 21 % de ellos *Regulares*.

Adicional a lo anterior, la tabla 5.15 muestra el mismo escenario para la métrica de sensibilidad. En este caso, el algoritmo es capaz de proveer un modelo *Muy Bueno* en el 50 % de los casos, mientras modelos *Malos* son entregados en un 29 % de los casos. Los modelos *Buenos* y *Regulares* son entregados para un 21 % de los pacientes a quienes se les aplicó.

Clasificación	Proporción de pacientes
Muy bueno	63 %
Bueno	16 %
Regular	21 %
Malo	0 %

Tabla 5.14: Exactitud de modelos por paciente. Fuente: elaboración propia.

Clasificación	Proporción de pacientes
Muy bueno	50 %
Bueno	8 %
Regular	13 %
Malo	29 %

Tabla 5.15: Sensibilidad de modelos por paciente. Fuente: elaboración propia.

5.2. Conclusiones y trabajo futuro

Al trabajar en soluciones de aprendizaje supervisado, la confiabilidad en los datos de entrenamiento es fundamental. Proyectos y estudios que involucren este tipo de soluciones vinculados al área de la salud han crecido en los últimos años por la implementación de fichas clínicas electrónicas y, en casos más avanzados, la integración directa de los sistemas de medición de signos vitales con repositorios de datos.

Por otro lado, al trabajar con herramientas que modelan secuencias temporales de datos existen factores como la distancia entre las mediciones, elementos comparativos y otros elementos externos que son fundamentales para un correcto modelado y posterior análisis.

El estudio desarrollado tiene como base datos capturados de forma manual, no equidistantes entre sí y con lagunas de información debido a protocolos internos del centro asistencial desde donde se capturó, dependiendo de fluctuaciones en el estado de salud del paciente.

De esta forma, y a modo de ejemplo, la variable objetivo, elemento fundamental en este tipo de estudios, se genera entendiendo cuándo el paciente presenta un cambio hacia un estado de mayor gravedad, definido por un aumento en el apoyo respiratorio, y se consigna el conjunto de signos vitales anterior a dicho cambio para determinar que ese estado presentaba alteración. No obstante, esto no considera que el problema el paciente lo pudo sufrir en un intervalo intermedio entre dos mediciones, y por ende los signos asociados a ambas mediciones corresponden a estados normales (el inicial por estabilidad previa y el posterior por las acciones clínicas aplicadas al paciente). Tampoco es considerado el uso de fármacos durante el período de hospitalización, que claramente afecta uno o más signos vitales.

Si bien lo anterior acusa y asume errores en la medición misma, el aporte y diferenciación de este estudio es la estructura lógica del algoritmo, que busca proyectar un siguiente estado del paciente con herramientas estadísticas para luego clasificar su riesgo con herramientas de aprendizaje supervisado, facilitando la inclusión de más modelos en ambas líneas investigativas y definiendo de forma paramétrica la distancia entre los datos capturados.

Es así como se esperaría que esta solución conectado a un sistema de captura oportuna de datos y estados del paciente presente mejores resultados que los expuestos, y así en trabajos futuros lograr mayor especificidad en la predicción de riesgos.

Al considerar el resultado final de este trabajo, si bien el estado de riesgo se mide con una sensibilidad bastante baja, el algoritmo está diseñado para mejorar en función de datos capturados de forma menos espaciada, probablemente llegando a mediciones cada 30 o 15 minutos y definiendo mejor el estado de riesgo mismo de cada paciente. Por lo que el objetivo general se cumple, siempre y cuando el modelo diseñado cumpla con la cantidad y tiempo suficiente de datos para analizar.

5.2.1. Recomendaciones y trabajo futuro

Respecto a la captura de datos, sería fundamental lograr probar los modelos estadísticos con una distancia menor a dos horas, especialmente para el caso de la frecuencia cardiaca, pues los resultados muestran que no es posible modelarla de forma correcta. Este factor influye en que el modelo de predicción cumpla de modo menos específico el objetivo de pronóstico de una situación de salud actual.

En la misma línea de datos, actualmente se modela sólo el ciclo circadiano diario como estacionalidad, no obstante, pueden existir comportamientos semanales o mensuales que de momento no es posible capturar. Esto además de influir en la proyección del siguiente estado corporal, también influiría en la clasificación de riesgo, pues claramente la temperatura corporal normal en verano es distinta a la de invierno.

Por último en lo relacionado a los datos, actualmente todos los experimentos se realizan con data capturada en un ambiente controlado mientras los pacientes pediátricos están hospitalizados, no obstante, es importante entender cómo se comportan los signos vitales normales en un ambiente de desarrollo normal, donde los niños juegan y corren.

Esto último se hace fundamental, ya que la calidad de vida es un punto a considerar en el crecimiento de una persona, de modo que el desarrollo de un algoritmo de predicción con el objetivo de brindar tiempo suficiente para reaccionar a una eventual crisis, debería considerar el espacio natural y el modelo aprender de estas conductas.

Por lo mismo respecto al funcionamiento del algoritmo, es importante desarrollar una métrica que sea capaz de unificar ambas etapas para poder realizar comparativas futuras con otros algoritmos que utilicen un enfoque híbrido, similar a lo expuesto.

Referente también a la validación, pero específicamente al funcionamiento actual, no está previsto que la cantidad de errores en la heurística de extrapolación de una serie de tiempo se considere en la elección de la misma, resolviendo posibles errores con cambios en la cantidad de variables disponibles de cara a la etapa II del mismo.

Finalmente, y respecto a un funcionamiento general del sistema conectado con la plataforma que se tiene prevista, actualmente un nuevo paciente o debe pasar por un período de

calibración o se debe configurar el paciente más parecido que esté en el estudio y se define como base de entrenamiento, pero no existe un modelo capaz de generar una base robusta para la incorporación de estos nuevos pacientes, especialmente en lo relacionado a las situaciones de riesgo.

Esto hace reflexionar la necesidad de obtener datos de manera más rápida, objetiva y constante posible, de modo que las investigaciones y tecnologías futuras sean utilizadas por los pacientes y sus familias como un beneficio que aporte al desarrollo integral del niño y se pueda extrapolar a su vida fuera del recinto hospitalario.

Bibliografía

- [1] Arkaitz Artetxe, Borja Ayerdi, Manuel Graña, and Sebastian Rios. Using anticipative hybrid extreme rotation forest to predict emergency service readmission risk. *Journal of Computational Science*, 20:154 – 161, 2017. ISSN 1877-7503. doi: <https://doi.org/10.1016/j.jocs.2016.12.008>. URL <http://www.sciencedirect.com/science/article/pii/S1877750316305130>.
- [2] Arkaitz Artetxe, Manuel Graña, Andoni Beristain, and Sebastián Ríos. Balanced training of a hybrid ensemble method for imbalanced datasets: a case of emergency department readmission prediction. *Neural Computing and Applications*, Oct 2017. ISSN 1433-3058. doi: [10.1007/s00521-017-3242-y](https://doi.org/10.1007/s00521-017-3242-y). URL <https://doi.org/10.1007/s00521-017-3242-y>.
- [3] Alfésio L.F. Braga, Gleice M.S. Conceição, Luiz A.A. Pereira, Humberto S. Kishi, Júlio C.R. Pereira, Maria F. Andrade, Fábio L.T. Gonçalves, Paulo H.N. Saldiva, and Maria R.D.O. Latorre. Air pollution and pediatric respiratory hospital admissions in são paulo, brazil. *Journal of Environmental Medicine*, 1(2), 1999. doi: [10.1002/\(SICI\)1099-1301\(199904/06\)1:2<95::AID-JEM16>3.0.CO;2-S](https://doi.org/10.1002/(SICI)1099-1301(199904/06)1:2<95::AID-JEM16>3.0.CO;2-S).
- [4] Max Bramer. *Principles of Data Mining*. Springer London, 2013. doi: [10.1007/978-1-4471-4884-5](https://doi.org/10.1007/978-1-4471-4884-5). URL <https://doi.org/10.1007/978-1-4471-4884-5>.
- [5] Peter Brockwell and Richard Davis. *Introduction to Time Series and Forecasting*. Springer, 2002. ISBN 0387953515. URL <https://www.amazon.com/Introduction-Time-Forecasting-INTRODUCTION-FORECASTING/dp/B00DIL03UK?SubscriptionId=0JYN1NVW651KCA56C102&tag=techkie-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=B00DIL03UK>.
- [6] Patricia Cabezas, Carlos Toro, and María L. Boza. Enfermedades respiratorias crónicas y recurrentes de la infancia y la niñez en un hospital general. *Revista chilena de pediatría*, 68(4), aug 1997. doi: [10.4067/s0370-41061997000400005](https://doi.org/10.4067/s0370-41061997000400005). URL <https://doi.org/10.4067/s0370-41061997000400005>.
- [7] Daniel Capurro, Mauricio Soto, Macarena Vivent, Marcelo Lopetegui, and Jorge R Herskovic. Informática biomédica. *Revista médica de Chile*, 139(12):1611–1616, dec 2011. doi: [10.4067/s0034-98872011001200013](https://doi.org/10.4067/s0034-98872011001200013). URL <https://doi.org/10.4067/s0034-98872011001200013>.

- [8] C. A. Czeisler. Stability, precision, and near-24-hour period of the human circadian pacemaker. *Science*, 284(5423):2177–2181, jun 1999. doi: 10.1126/science.284.5423.2177. URL <https://doi.org/10.1126/science.284.5423.2177>.
- [9] Wuyang Dai, Theodora S. Brisimi, William G. Adams, Theofanie Mela, Venkatesh Saligrama, and Ioannis Ch. Paschalidis. Prediction of hospitalization due to heart diseases by supervised learning methods. *International Journal of Medical Informatics*, Mar 2015. ISSN 1386-5056. doi: 10.1016/j.ijmedinf.2014.10.002. URL <http://dx.doi.org/10.1016/j.ijmedinf.2014.10.002>.
- [10] Tom Debevoise and Rick Geneva. *The Microguide to Process Modeling in BPMN*. BookSurge Publishing, 2008. ISBN 1419693107. URL <https://www.amazon.com/Microguide-Process-Modeling-BPMN/dp/1419693107?SubscriptionId=0JYN1NVW651KCA56C102&tag=techkie-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=1419693107>.
- [11] Ministerio de Salud Chile DEIS, Departamento de Estadísticas e Información de Salud. Atenciones de urgencia, 2011. URL http://deis.minsal.cl/AtencionesUrgencia/Total_Atenciones_urgencia_2011.htm.
- [12] Ministerio de Salud Chile DEIS, Departamento de Estadísticas e Información de Salud. Egresos hospitalarios 2012-2014, 2014. URL http://intradeis.minsal.cl/egresoshospitalarios/menu_publica_nueva/menu_publica_nueva.htm.
- [13] Matías Echeverría, Angel Jimenez-Molina, and Sebastián A. Ríos. A semantic framework for continuous u-health services provisioning. *Procedia Computer Science*, 60: 603–612, 2015. doi: 10.1016/j.procs.2015.08.187. URL <https://doi.org/10.1016/j.procs.2015.08.187>.
- [14] Lili Erazo, Sebastián A. Ríos, and . A benchmark on automatic obstructive sleep apnea screening algorithms in children. *Procedia Computer Science*, 35:739–746, 2014. doi: 10.1016/j.procs.2014.08.156. URL <https://doi.org/10.1016/j.procs.2014.08.156>.
- [15] Asier Garmendia, Manuel Graña, Jose Manuel Lopez-Guede, and Sebastian Rios. Predicting patient hospitalization after emergency readmission. *Cybernetics and Systems*, 48(3):182–192, 2017. doi: 10.1080/01969722.2016.1276772. URL <https://doi.org/10.1080/01969722.2016.1276772>.
- [16] Asier Garmendia, Sebastian A. Rios, Jose M Lopez-Guede, and Manuel Graña. Triage prediction in pediatric patients with respiratory problems. *Neurocomputing*, 2017. ISSN 0925-2312. doi: <https://doi.org/10.1016/j.neucom.2017.01.122>. URL <http://www.sciencedirect.com/science/article/pii/S0925231217315230>.
- [17] Marc Gorelick, Philip V. Scribano, Martha W. Stevens, Theresa Schultz, and Justine Shults. Predicting need for hospitalization in acute pediatric asthma. *Pediatric Emergency Care*, 24(11):735–744, Nov 2008. doi: 10.1097/pec.0b013e31818c268f. URL <https://doi.org/10.1097/pec.0b013e31818c268f>.
- [18] Manuel Graña, Darya Chyzhyk, Carlos Toro, and Sebastian Rios. Innovations in

- healthcare and medicine editorial. *Computers in Biology and Medicine*, 72:226 – 228, 2016. ISSN 0010-4825. doi: <https://doi.org/10.1016/j.compbimed.2016.03.003>. URL <http://www.sciencedirect.com/science/article/pii/S0010482516300579>.
- [19] B. E. Himes, Y. Dai, I. S. Kohane, S. T. Weiss, and M. F. Ramoni. Prediction of chronic obstructive pulmonary disease (copd) in asthma patients using electronic medical records. *Journal of the American Medical Informatics Association*, May 2009. ISSN 1527-974X. doi: 10.1197/jamia.m2846. URL <http://dx.doi.org/10.1197/jamia.M2846>.
- [20] Rob Hyndman, George Athanasopoulos, Slava Razbash, Drew Schmidt, Zhenyu Zhou, Yousaf Khan, and Christoph Bergmeir. *forecast: Forecasting functions for time series and linear models*, 2013. URL <http://CRAN.R-project.org/package=forecast>. R package version 4.8.
- [21] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. Springer New York, 2013. doi: 10.1007/978-1-4614-7138-7. URL <https://doi.org/10.1007/978-1-4614-7138-7>.
- [22] Alan Jobe and Eduardo Bancalari. Bronchopulmonary dysplasia. *American Journal of Respiratory and Critical Care Medicine*, 163(7):1723–1729, jun 2001. doi: 10.1164/ajrccm.163.7.2011060. URL <https://doi.org/10.1164/ajrccm.163.7.2011060>.
- [23] Erich L. Lehmann and George Casella. *Theory of Point Estimation (Springer Texts in Statistics)*. Springer, 2003. ISBN 0387985026. URL <https://www.amazon.com/Theory-Point-Estimation-Springer-Statistics/dp/0387985026?SubscriptionId=0JYN1NVW651KCA56C102&tag=techkie-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=0387985026>.
- [24] Constanza Lizama and Silvana Santander. Programa de hospitalización domiciliaria, apr 2010. URL <http://ciperchile.cl/pdfs/2014/01/sanjose/programaHosp.pdf>.
- [25] Enrique Paris Mancilla, Ignacio Sánchez, Daniel Beltramino, and Alfonso Copto García. *Meneghello. Pediatría. 2 Vol (Spanish Edition)*. Zagier & Urruty Pubns, 2015. ISBN 9500618818. URL <https://www.amazon.com/Meneghello-Pediatr%C3%ADa-2-Vol-Spanish/dp/9500618818?SubscriptionId=0JYN1NVW651KCA56C102&tag=techkie-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=9500618818>.
- [26] Andrew V. Metcalfe and Paul S.P. Cowpertwait. *Introductory Time Series with R*. Springer New York, 2009. doi: 10.1007/978-0-387-88698-5. URL <https://doi.org/10.1007/978-0-387-88698-5>.
- [27] Charles Metz. Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8(4): 283–298, Oct 1978.
- [28] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2014. URL <http://CRAN.R-project.org/package=e1071>. R package version 1.6-3.

- [29] Ministerio de Salud de Chile. Campaña de invierno 2017, may 2017. URL <http://web.minsal.cl/campana-de-invierno-2017/>.
- [30] Susana Morales, Jorge Barros, Orietta Echávarri, Fabián García, Alex Osses, Claudia Moya, María Paz Maino, Ronit Fischman, Catalina Núñez, Tita Szmulewicz, and Alemka Tomicic. Acute mental discomfort associated with suicide behavior in a clinical sample of patients with affective disorders: Ascertaining critical variables using artificial intelligence tools. *Frontiers in Psychiatry*, 8, feb 2017. doi: 10.3389/fpsy.2017.00007. URL <https://doi.org/10.3389/fpsy.2017.00007>.
- [31] Naiara Muro, Eider Sanchez, Carlos Toro, Eduardo Carrasco, Sebastián A. Ríos, Frank Guijarro, and Manuel Graña. Experience-based electronic health records. *Cybernetics and Systems*, 47(1-2):126–139, 2016. doi: 10.1080/01969722.2016.1128774. URL <https://www.tandfonline.com/doi/abs/10.1080/01969722.2016.1128774>.
- [32] Rebeca Paiva and Sandra Navarro. *Protocolo de Ventilación Mecánica Invasiva en APS*. Subsecretaría de Redes Asistenciales - Ministerio de Salud Chile, 2013.
- [33] Rebeca Paiva and Sandra Navarro. *Programa de Asistencia Ventilatoria No Invasiva (AVNI) en APS - Normativa Técnica*. Subsecretaría de Redes Asistenciales - Ministerio de Salud Chile, 2013.
- [34] Maria Angélica Palomino, Mónica Morgues, and Fernando Martínez. Management of infants with chronic lung disease of prematurity in chile. *Early Human Development*, 81(2):143–149, feb 2005. doi: 10.1016/j.earlhumdev.2004.12.003. URL <https://doi.org/10.1016/j.earlhumdev.2004.12.003>.
- [35] Pan American Health Organization. *Health in the Americas*. Pan American Health Organization, 2 edition, 2007. ISBN 978-92-75-11622-9. URL http://www.paho.org/hq/dmdocuments/2010/Health_in_the_Americas_Vol_1_Regional_2007.pdf.
- [36] Chris Paxton, Alexandru Niculescu-Mizil, and Suchi Saria. Developing predictive models using electronic medical records: challenges and pitfalls. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, 2013.
- [37] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL <http://www.R-project.org/>. ISBN 3-900051-07-0.
- [38] Ahmed A. Radwan, Hazem M. El-Bakry, and Hager M. El Hadad. A new expert system for pediatric respiratory diseases by using neural networks. In *Proceedings of the 2Nd International Conference on Applied Informatics and Computing Theory*, pages 296–306. World Scientific and Engineering Academy and Society (WSEAS), 2011. ISBN 978-1-61804-034-3. URL <http://dl.acm.org/citation.cfm?id=2047895.2047949>.
- [39] Roberto Refinetti and Michael Menaker. The circadian rhythm of body temperature. *Physiology & Behavior*, 51(3):613–637, mar 1992. doi: 10.1016/0031-9384(92)90188-8. URL [https://doi.org/10.1016/0031-9384\(92\)90188-8](https://doi.org/10.1016/0031-9384(92)90188-8).

- [40] H. G. Richter, C. Torres-Farfan, P. P. Rojas-Garcia, C. Campino, F. Torrealba, and M. Seron-Ferre. The circadian timing system: making sense of day/night gene expression. *Biol. Res.*, 37(1):11–28, 2004.
- [41] Sebastián A. Ríos, Lili Erazo, and . An automatic apnea screening algorithm for children. *Expert Systems with Applications*, 48:42 – 54, 2016. ISSN 0957-4174. doi: <https://doi.org/10.1016/j.eswa.2015.11.013>. URL <http://www.sciencedirect.com/science/article/pii/S0957417415007794>.
- [42] Fernando Saldías. Recomendaciones para el diagnóstico, manejo y prevención de la influenza en Chile: Período 2013. *Revista chilena de enfermedades respiratorias*, 29(1): 31–38, mar 2013. doi: 10.4067/s0717-73482013000100005. URL <https://doi.org/10.4067/s0717-73482013000100005>.
- [43] David L. Sanders and Dominik Aronsky. Detecting asthma exacerbations in a pediatric emergency department using a bayesian network. *AMIA Annu Symp Proc*, 2006. URL <http://www.ncbi.nlm.nih.gov/pubmed/17238428>.
- [44] C. Shearer. The crisp-dm model: The new blueprint for data mining. *Journal of Data Warehousing*, 5(4):13–22, fall 2000. URL <https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf>.
- [45] Chile Superintendencia de Salud. Garantías explícitas en salud (auge o ges), 2017. URL <http://www.supersalud.gob.cl/difusion/572/w3-propertyname-501.html>.
- [46] Terry M Therneau and Beth Atkinson. R port by Brian Ripley. Note that maintainers are not available to give advice on using a package they did not author. *rpart: Recursive Partitioning*, 2012. URL <http://CRAN.R-project.org/package=rpart>. R package version 3.1-53.
- [47] Adrian Trapletti and Kurt Hornik. *tseries: Time Series Analysis and Computational Finance*, 2013. URL <http://CRAN.R-project.org/package=tseries>. R package version 0.10-32.
- [48] Irma Vargas, Rebeca Paiva, Raúl Valenzuela, and Paola Pontoni. *Manual Operativo Programas de Salud Respiratoria*. División de Atención Primaria - Unidad de Salud Respiratoria, 2015.
- [49] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL <http://www.stats.ox.ac.uk/pub/MASS4>. ISBN 0-387-95457-0.
- [50] Daniel M. Weinberger, Keith P. Klugman, Claudia A. Steiner, Lone Simonsen, and Cécile Viboud. Association between respiratory syncytial virus activity and pneumococcal disease in infants: A time series analysis of US hospitalization data. *PLoS Medicine*, 12(1), jan 2015. doi: 10.1371/journal.pmed.1001776.
- [51] Wikimedia Commons. Kernel yontemi ile veriyi daha fazla dimensiyonlu uzaya tasima

- islemi, 2016. URL https://upload.wikimedia.org/wikipedia/commons/d/d8/Kernel_yontemi_ile_veriyi_daha_fazla_dimensiyonlu_uzaya_tasima_islemi.png.
- [52] Wikimedia Commons. Titanic survival decison tree, 2017. URL https://upload.wikimedia.org/wikipedia/commons/6/6b/Titanic_Survival_Decison_Tree_SVG.png.
- [53] R. A. Wilke, R. L. Berg, P. Peissig, T. Kitchner, B. Sijercic, C. A. McCarty, and D. J. McCarty. Use of an electronic medical record for the identification of research subjects with diabetes mellitus. *Clinical Medicine & Research*, Mar 2007. ISSN 1554-6179. doi: 10.3121/cmr.2007.726. URL <http://dx.doi.org/10.3121/cmr.2007.726>.
- [54] World Health Organization. Nacidos demasiado pronto: Informe de acción global sobre nacimientos prematuros, 2012. URL http://www.who.int/pmnch/media/news/2012/borntoosoon_execsum_es.pdf.
- [55] World Health Organization. About chronic respiratory diseases, apr 2017. URL http://www.who.int/respiratory/about_topic/en/.
- [56] Alexis Yañez, Lorena Tapia, Dona Benadof, and María Angélica Palomino. Infecciones respiratorias agudas bajas: descripción de los egresos del hospital roberto del río en el año 2016. *Revista de Pediatría Electrónica*, apr 2017. ISSN 0718-0918.
- [57] Haijun Zhai, Patrick Brady, Qi Li, Todd Lingren, Yizhao Ni, Derek S. Wheeler, and Imre Solti. Developing and evaluating a machine learning based algorithm to predict the need of pediatric intensive care unit transfer for newly hospitalized children. *Resuscitation*, Aug 2014. doi: 10.1016/j.resuscitation.2014.04.009. URL <https://doi.org/10.1016/j.resuscitation.2014.04.009>.