Facultad de Ciencias Médicas “Enrique Ortega”

Escuela de Medicina

APLICACIÓN DE APRENDIZAJE AUTOMÁTICO COMO PREDICTOR DE RESULTADO CLÍNICO EN PACIENTES CON SEPSIS

Trabajo de Investigación que se presenta como requisito para el título de:

Autor: Carlos Andres Vera

Tutor: Carlos Farhat

Co-tutor: Geovanny Alvarado

Samborondon, Septiembre de 2017

UEES
UNIVERSIDAD ESPIRITU SANTO
Postgrado

UNIVERSIDAD DE ESPECIALIDADES ESPIRITU SANTO

FACULTAD DE CIENCIAS MÉDICAS

ESCUELA DE MEDICINA

CERTIFICACION DEL TUTOR

En mi calidad de tutor(a) del trabajo de investigación de tesis para optar el título de Médico de la Facultad de Ciencias Médicas de la Universidad de Especialidades Espíritu Santo.

Certifico que: he dirigido el trabajo de titulación presentada por el señor **Carlos Andrés Vera Paz** con C.I. No. **0920016672**

Cuyo tema es **"Aplicación de Aprendizaje Automático como Predictor de Resultado Clínico en Pacientes con Sepsis".**

Revisado y corregido se aprobó en su totalidad, lo certifico:

.....

Dr. Carlos Farhat Zamora

TUTOR

Dedicatoria

A mi papá Carlos y mi mamá Lucía, con quienes estoy infinitamente endeudado y agradecido.

A mis hermanos José Antonio y Daniel y mi ahijado Jose Francisco, para que vean que todo es posible.

A Emilio, Erik y Federico quienes fueron mis eternos compañeros de estudio

A mi novia Yessie, quien hace que todo esto valga la pena.

Reconocimiento

Extiendo un reconocimiento a todos los docentes que han dejado una huella en mi formación. Especialmente a aquellos que son un modelo a seguir en el quehacer profesional: Dr. Carlos Orellana, Dr. Miguel Flor, Dr. Bolivar Zurita a quienes considero tanto maestros como amigos.

Agradezco especialmente a el Dr. Feliciano Ramos, quien me hizo sentir en casa en Zaragoza y me dejó ser parte de su práctica por tan solo un par de meses llenos de aprendizaje que no olvidare jamas.

Finalmente agradezco a mi amigo Geovanny Alvarado, quien fue la única persona que pudo entender lo que quería hacer con esta tesis, y quien supo escucharme cuando más lo necesitaba.

Indice General **6**

CAPÍTULO 1 **10**
Introducción 10
Antecedentes.- 11
Planteamiento del problema 15
Justificacion 16
Objetivo General 17
Hipotesis 17

CAPÍTULO 2 **18**
Fundamentación Teórica 18
Sepsis 18
SOFA 18
qSOFA 19
Big Data 20
La Historia Clinica Electronica. 22
MIMIC III 24
Herramientas Informáticas 25
PostgreSQL 25
R 26
Aprendizaje Automático 27
Avances clave en el AA 29
Tipos de Aprendizaje Automático 31
Aprendizaje Supervisado vs. No Supervisado 32
Aprendizaje Automático supervisado 33
Arboles de decision 34
Aprendizaje en conjunto 35
Bosques Aleatorios 35
Análisis de datos 36
Recolección y preprocesamiento 36
Discretización de las variables 40
Evaluacion 40
Validación cruzada (CV) k-Fold 41

CAPÍTULO 3 **42**

Metodologia 42
Operacionalización de las Variables 43
Estadística: 47
Resultados 51
Discusión 57
Conclusion 58

Bibliografia **60**

CAPÍTULO 1

Introducción

En la última década se ha registrado un rápido incremento de la capacidad de los sistemas informáticos en red y móviles para recoger y transportar grandes cantidades de datos, una fenómeno denominado "Big Data" (1). Los científicos e ingenieros que manejan estos datos han recurrido a menudo al Aprendizaje Automático(AA) para brindar soluciones al problema de la obtención de observaciones útiles y precisas, con el fin de generar predicciones y tomar decisiones a partir de estos enormes repositorios de datos(2) De hecho, el tamaño de los datos hace que sea esencial desarrollar procedimientos escalables que combinan datos computacionales y estadísticos con el fin de encontrar asociaciones que contribuyan a un modelo generalizable (3). El producto del AA es un modelo que puede ser utilizado y optimizado indefinidamente para una tarea.

La generación de modelos predictivos poseen mucho potencial para mejorar la atención de salud. La integración completa de los sistemas de AA y historia clínica universal crean un genoma digital de cada paciente. Esto implica un archivo que contiene la secuencia genética del paciente sumado todos los datos asociados a sus interacciones con el sistema de salud e información relevante de la analitica digital almacenado como datos clínicos estructurados(4). Esta información puede ser utilizada para distintas aplicaciones en varios ámbitos de la atención. por ejemplo automatizar la lista de espera de un servicio, con el fin de dar atención a las patologías más urgentes en un tiempo apropiado(5) o el uso de repositorios de imágenes para el entrenamiento de algoritmos de diagnóstico

automático(6). Esto es posible gracias a la unión de big data con las técnicas de AA como los árboles de decisión o modelos predictivos bayesianos en los cuales se entrenan a partir de un proceso iterativo para encontrar el modelo óptimo(7)

El presente trabajo busca aplicar técnicas de AA para generar un modelo predictivo de mortalidad en uci para pacientes con sepsis a partir de big data obtenida en la base de datos MIMIC-III(8). Para este fin se describirán varias técnicas utilizadas comúnmente en el análisis de datos tales como el uso de PSQL, la extracción de la base de datos MIMIC-III, la manipulación y limpieza de datos y su importancia, así como las bibliotecas de funciones en R y su papel en la generación de modelos predictivos. Finalmente se comparará el modelo generado mediante la métrica AUC con la herramienta predictiva utilizada en la actualidad (qSOFA)

Antecedentes.-

A Claude Shannon, un matemático de Bell Labs, se le atribuye haber sentado las bases de la digitalización en su artículo pionero de 1948, Una teoría matemática de la comunicación. En el cual define al bit como una unidad discreta de entropía, que representa una cantidad de información cuantificable y que sentó la base para la aplicación de la teoría de la información(9). Este concepto permitió por primera vez almacenar información de forma digital. Pronto el concepto del bit se aplicó en los sistemas de software donde se utilizaban como unidades programables en función de operadores lógicos permitiendo así generar operaciones complejas. Este desarrollo es la base de la computación moderna al permitir almacenar la información requerida dentro del ordenador. Previamente se tenía que configurar para cada función deseada a partir de circuitos electrónicos discretos(10).

Tan solo una década luego en 1959 Samuel, A.L. había programado un software que vencía a humanos en el juego de damas. El programa evaluaba distintos estados del juego y seleccionaba su movimiento acorde a la mejor opción. La repetición de juegos entrenaba a el sistema para identificar los estados favorables y tomar mejores decisiones. Ya entonces el autor consideraba las posibles aplicaciones de este tipo de algoritmos para el diagnóstico médico (11). En las décadas subsecuentes el campo del AA se divide en los sistemas de inteligencia artificial(AI) y el AA, el cual se ve restringido considerablemente por la falta de datos. Esto llevó a una pérdida de interés por las técnicas de AA mientras que el campo de la inteligencia artificial se concentraba en en el papel del conocimiento en inteligencia, independientemente de su origen(12).

En 1980 el campo del AA se reorganiza para buscar soluciones de optimización en respuesta a el advenimiento del internet. El producto de este nuevo enfoque es la creación de algoritmos que pueden interpretar grandes conjuntos de datos(12) Inicialmente utilizados como filtros de spam en e-mails la aplicación del reconocimiento de patrones el enfoque del AA se ha expandido paulatinamente junto con la magnitud de los datos computacionales disponibles y de el poder de procesamiento de las unidades de computación.

Este crecimiento ha sido posible gracias al aumento de la densidad de microprocesadores en las unidades de procesamiento. Los avances en el área de los microchips fueron predichos por la ley de moore que sostenía que la capacidad computacional total de la humanidad crece de forma exponencial, duplicando cada dos años (13). En 2014 la supercomputadora Tianhe-2 realizó 33.86 petaflops/s (33.86 cuatrillón de cálculos por segundo) (14)

Actualmente duplicar la capacidad computacional global implica un aumento monumental de los recursos disponibles. Mientras esta tendencia se mantenga los avances posibles gracias al poder computacional disponible serán cada vez más impactantes en nuestro dia a dia. Otro dato importante es que la gran mayoría del poder computacional se encuentra en dispositivos de consumidor(figura 1), por lo cual representan una red difusa de procesamiento a la que se puede acceder sistemáticamente

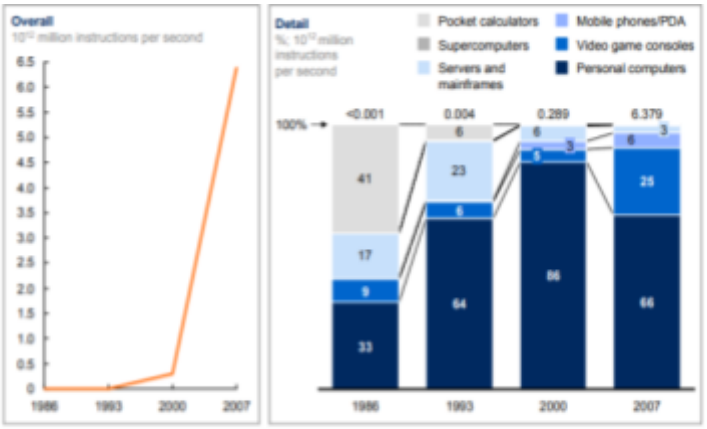


Figura Figura1: Capacidad computacional global media en 10^12 millones de instrucciones por segundo (Hilbert & Lopez, 2011)

En medicina se han realizado ya varios esfuerzos para aplicar el AA a problemas clínicos. Existen hoy sistemas que optimizan los niveles de glucosa de pacientes ambulatorios, mejoran la precisión de las interpretaciones automáticas de los electrocardiogramas y asisten en el diagnóstico imagenológico de nódulos pulmonares y cáncer de mama (15). Los modelos de decisión clínica para el

diagnóstico o manejo de varias patologías como la pancreatitis(16) o la depresión (17)han demostrado ser herramientas efectivas para reducir costos o mejorar desenlaces en estas entidades clínicas(18). Otros estudios combinan distintas modalidades de AA en un sistema aún más complejo que aplica diferentes algoritmos según la situación clínica del paciente en un flujo de manejo que sugiere el manejo óptimo para cada estado del paciente(19).

La miniaturización y masificación de los dispositivos electrónicos amplía aún más la capacidad de recolección de datos biométricos para el análisis automático(20). La monitorización constante de los signos vitales y otros datos biométricos aún por validar prometen jugar un papel cada vez más relevante en el manejo del paciente al formar parte de una constelación de datos que expresa un fenotipo. El uso de relojes inteligentes, medidores de actividad ligados al celular, dispositivos para medir la presión arterial y saturación de oxígeno conectados a la web así como el desarrollo de pruebas bioquímicas realizadas con dispositivos móviles(21) aumentaran en varios órdenes de magnitud la cantidad y la calidad de los datos recogidos de cada individuo lo que a su vez hace factible la aplicación extensa de algoritmos de AA para utilizar la información disponible en nuevas aplicaciones que mejoren la calidad de la atención de salud y reduzcan la morbimortalidad prevenible

Planteamiento del problema

Desde el 2016 el consenso de sepsis determinó que la sepsis es *la disfunción orgánica causada por una respuesta anómala del huésped a la infección que supone una amenaza a la supervivencia*. Se puede objetivizar la disfunción orgánica como un aumento de dos puntos o más en el score SOFA(Sequential[Sepsis-related] Organ Failure Assessment)(22,23). Desde el punto de vista clínico la sepsis es la reacción inflamatoria sistémica que produce un colapso circulatorio y la isquemia del sistema nervioso central, por lo cual debe ser manejado de forma inmediata para reducir la tasa de mortalidad asociada.

En cuanto al diagnóstico de sepsis se han desarrollado varios marcadores clínicos tales como la procalcitonina (24) que han aumentado la tasa de detección temprana de esta patología. Así mismo la evolución del consenso clínico en cuanto a la definición de la sepsis ha simplificado en gran medida el diagnóstico de esta entidad clínica. Sin embargo existe una alta variabilidad en el modo de manejo de la sepsis, lo cual produce resultados subóptimos y aumento de la utilización de recursos. Según Reade et. al “El manejo reportado de la sepsis temprana varía entre las especialidades y los países, y las respuestas no siguen las pautas del SSC(Surviving Sepsis Campaign). Las preocupaciones se relacionan con el conocimiento, las actitudes y los recursos”(25).

La carga de morbi mortalidad asociada con la sepsis es una de las más grandes del sistema de salud moderno. Se estima que al año se presentan 31.5 millones de casos de sepsis con 5.3 millones de muertes anuales.(26). En un estudio de las admisiones por sepsis entre 2010 y 2016 en varios centros de Estados Unidos(27) la estancia promedio fue 7.7 (SD: 10.2) días. El 38% de los pacientes

fue admitió a la UCI y de ellos, la estancia media en uci fue de 5.2 (SD: 6.6) días. Hubo una mortalidad del 11.1% en pacientes con ingreso por sepsis durante su encuentro índice. El costo promedio de la visita fue $ 18,023 (SD: $ 32,996), mediana $ 10,371. El 11.4% de estos pacientes fueron readmitidos dentro de los 30 días por cualquier causa con un promedio de estancia de 4 (SD: 6.3) días. El costo promedio de esas readmisiones fue $ 14,593 (SD: $ 22,212), mediana de $ 8,914 .

Justificacion

Este trabajo permitirá para generar una metodología para utilizar los datos recogidos diariamente por los sistemas de datos electrónicos como herramienta de pronóstico en patologías con alta carga de morbi mortalidad en nuestro sistema de salud. Este modelo podrá ser utilizado por todos los médicos de los centros que utilicen la metodología descrita ya que es posible adaptar herramientas web para interactuar con el modelo generado (28). . Es trascendente porque puede ser el punto de inicio para una red nacional informática que pueda utilizar los datos de todos los centros de atención para generar un modelo extremadamente robusto(29)

Los sistemas de AA permiten la homogeneización de la práctica clínica. Para este fin la minería de datos provee herramientas para unificar la información obtenida de varias fuentes como la sintomatología clínica y los estudios de laboratorio en un modelo computacional que puede ayudar a la toma de decisiones críticas en el cuidado de pacientes(30).Una vez generado un flujo de trabajo se pueden generar incontables modelos que se adapten a la información adquirida por el sistema hospitalario o que pueden ser centralizados para generar un consenso

nacional de la mejor práctica basándose en cientos de miles de instancias previas. La clave para generar un sistema óptimo radica en la recolección y el manejo de la base de datos y la selección del modelo estadístico apropiado para la función deseada (31).

Objetivo General

Desarrollar un modelo de predicción clínica a partir de datos obtenidos de pacientes con diagnóstico de Sepsis en la base de datos MIMIC-III

Objetivos Especificos

1. Describir técnicas de manipulación de bases de datos en SQL y análisis de datos en R
2. Generar un flujo de trabajo para la creación de modelos con distintos propósitos y/o datos
3. Comparar modelo generado con estandard de decisión clínica actual (qSOFA)

Hipotesis

El Aprendizaje Automático es superior que el qSOFA para predecir de el resultado clínico de la sepsis

CAPÍTULO 2

Fundamentación Teórica

Sepsis

Desde el 2016 el consenso de sepsis determinó que la sepsis es *la disfunción orgánica causada por una respuesta anómala del huésped a la infección que supone una amenaza a la supervivencia.* Se puede objetivizar la disfunción orgánica como un aumento de dos puntos o más en el score SOFA(*Sequential[Sepsis-related] Organ Failure Assessment) (23)*

La sepsis se presenta clínicamente como un continuo en el cual la severidad varía desde infección y bacteriemia hasta sepsis y shock séptico que puede producir disfunción multiorgánica y la muerte. Por esta razón se han generado y Criterios Sepsis-3 (23). enfatizaron el valor de un cambio de 2 o más puntos en el puntaje Sequential [Sepsis-related Failure Assessment (SOFA), introdujeron SOFA rápido (qSOFA) y eliminaron los criterios del síndrome de respuesta inflamatoria sistémica (SIRS) de la definición de sepsis.

SOFA

El score SOFA fue desarrollado en 1996 por (32). En el cual describen al fracaso multiorgánico como un evento secuencial cuando ocurre en el contexto de una respuesta inflamatoria sistémica. Este puntaje ha sido utilizado ampliamente para predecir el pronóstico de los pacientes con sepsis en varios ámbitos clínicos.Las variables reflejan la disfunción de distintos aparatos lo cual permite evaluar la gravedad del proceso inflamatorio (33).

qSOFA

El qSOFA es una revisión del score SOFA adaptado para ser aplicado con mayor facilidad en la práctica clínica. Este score presenta varias ventajas en comparación al score SOFA: Es simple (consta de tres elementos clínicos: hipotensión, taquipnea y conciencia alterada), se puede evaluar de manera fácil y repetida, se generó a través de un enfoque basado en datos y en un gran estudio retrospectivo fue más preciso que el puntaje SIRS para predecir la muerte y la transferencia a la UCI de pacientes con sospecha de sepsis fuera de la UCI(33,34).

El puntaje qSOFA es fácil de calcular ya que sólo tiene tres componentes, cada uno de los cuales son fácilmente identificables en la cabecera y se asignan un punto:

Velocidad respiratoria ≥ 22 / minuto

Alteración del estado mental

Presión arterial sistólica ≤ 100 mmHg

El recientemente introducido qSOFA proporciona una mejor discriminación que el SIRS para predecir la mortalidad y los días sin UCI(34). Sin embargo, Entre los adultos con sospecha de infección admitidos en una UCI, un aumento en la puntuación SOFA de 2 o más tuvo una mayor precisión pronóstica para la mortalidad hospitalaria que los criterios SIRS o la puntuación qSOFA.

Entre 184875 pacientes (edad media, 62,9 años [DE, 17,4), las mujeres, 82540 [44,6%], el diagnóstico más común de neumonía bacteriana, 32,634 [17,7%]), un total de 34578 pacientes (18,7%) murieron en el hospital, y 102976 pacientes

(55.7%) murieron o experimentaron una estadía en UCI de 3 días o más. La puntuación de SOFA aumentó en 2 o más puntos en 90.1%; El 86.7% manifestó 2 o más criterios SIRS, y el 54.4% tuvo una puntuación qSOFA de 2 o más puntos. SOFA demostró una discriminación significativamente mayor para la mortalidad hospitalaria (AUROC en bruto, 0,753 [IC 99%, 0,750-0,775]) que los criterios SIRS (AUROC bruto, 0,589 [IC 99%, 0,585-0,593]) o qSOFA (AUROC crudo, 0,607). [99% CI, 0,603-0,611]). Los hallazgos fueron consistentes para ambos resultados en múltiples análisis de sensibilidad(33).

Big Data

¿Qué es exactamente Big Data? Un informe entregado al Congreso de EE. UU. En agosto de 2012 define big data como "grandes volúmenes de datos de alta velocidad, complejos y variables que requieren técnicas y tecnologías avanzadas para permitir la captura, almacenamiento, distribución, administración y análisis de la información"(35).Se estima que el sistema de salud americano posee 150 exabytes de información y que este alcanzara los zettabytes (10^21 gigabytes) y no mucho después los yottabytes(10^24 gigabytes). (35).

Una de las características fundamentales del Big Data es el enorme volumen de datos representados por dimensionalidades heterogéneas y diversas.(7) Esto se debe a que los diferentes recopiladores de información usan sus propios esquemas para el registro de datos, y la naturaleza de las diferentes aplicaciones también da como resultado diversas representaciones de los datos.

Por ejemplo, cada ser humano en un mundo biomédico puede representarse mediante el uso de información demográfica simple, como sexo, edad, historial de enfermedad familiar, etc. Para el examen de rayos X y tomografía computarizada de cada individuo, las imágenes o los videos se utilizan para representa los

resultados porque proporcionan información visual para que los médicos lleven a cabo exámenes detallados. Para una prueba relacionada con el ADN o el genoma, las imágenes de expresión de microarrays y las secuencias se utilizan para representar la información del código genético porque esta es la forma en que nuestras técnicas actuales adquieren los datos. Bajo tales circunstancias, las características heterogéneas se refieren a los diferentes tipos de representaciones para los mismos individuos, y las diversas características se refieren a la variedad de características involucradas para representar cada observación individual.

La consecuencia de esta característica que en muchas ocasiones la diversidad o la magnitud de los datos supera la capacidad de los investigadores que generan la base de datos para explorarlos. Al generar nuevos modelos se pueden seleccionar las variables más relevantes luego de haber realizado una exploración general de las mismas (37).Por esta razón que muchos grupos de investigación han optado por otorgar libre acceso a los datos para que otros investigadores los utilizen en sus modelos (36). Este modelo de información “open source” permite generar modelos para todo tipo de aplicaciones y ha reinvigorado el campo del aprendizaje automático para resolver problemas cada vez más complejos

La realidad de los sistemas de salud que diferentes organizaciones (o profesionales de la salud) tienen sus propios esquemas para representar a cada paciente, la heterogeneidad de los datos y diversos problemas de dimensionalidad se convierten en grandes desafíos si intentamos habilitar la agregación de datos mediante la combinación de todas las fuentes. En la práctica la recolección de datos no puede aspirar a ser completa. La herramienta de recolección de datos ideal es flexible y permite parsear los datos de formas distintas dependiendo de la aplicación deseada (38) El objetivo de la recolección

de datos es capturar la mayor cantidad posible de información. Las características de las variables seleccionadas y el cohorte estudiado pueden ser modificadas en la etapa de preprocesamiento (39)Por esta razón nunca se pueden tener demasiados datos. Con una cantidad variada de fuentes de información se puede reutilizar un dataset para distintas aplicaciones utilizando combinaciones de variables y subgrupos de pacientes dependiendo de la patología estudiada o el segmento poblacional estudiado.

Este trabajo está motivado por un nuevo modelo de salud centrado en el paciente que crea un perfil de riesgo de enfermedad personalizado basándose en big data, así como un plan de manejo de enfermedad y de bienestar para un individuo.(40)Situamos nuestro trabajo en la observación de que las enfermedades no ocurren de forma aislada. Son el resultado de una interacción entre factores genéticos, moleculares, ambientales y de estilo de vida (41). A partir del análisis de datos poblacionales cada vez más complejos podremos combatir los problemas de salud pública a un nivel sistémico, modificando las variables que determinan su incidencia.

La Historia Clinica Electronica.

El término registro electrónico de salud (EHR) se refiere al conjunto completo de información que reside en forma electrónica y está relacionado con el estado de salud pasado, presente y futuro o la atención médica proporcionada a un sujeto de la atención.(42)El propósito principal de EHR es la documentación, recuperación, transmisión, enlace y procesamiento de información multimedia para usuarios legítimos para la entrega de conocimiento y soporte de decisiones que mejoran los servicios relacionados con la salud eficientes y seguros,

independientemente del modelo de atención médica aplicado(43).

Uno de los desafíos de trabajar con datos de EHR es la naturaleza heterogénea por la que se representa, con tipos de datos que incluyen: (1) cantidades numéricas como índice de masa corporal, (2) objetos de fecha y hora como fecha de nacimiento o de admisión, (3) valores categóricos como etnicidad o códigos de vocabularios controlados como ICD-10 (anteriormente ICD-9) diagnósticos o procedimientos CPT, y (4) texto libre de lenguaje natural como notas de progreso o resúmenes de alta. Además, estos tipos de datos se pueden ordenar cronológicamente para formar la base de (5) series de tiempo derivadas, como las señales de signos vitales perioperatorios o la historia del paciente multimodal(44).

Por este motivo es importante hacer énfasis en la importancia de la actualización y la exploración de nuevas metodologías de manejo de datos para aumentar la viabilidad de los sistemas de AA. La recolección longitudinal de datos permite extrapolar tendencias a nivel generacional.(5). Un paciente puede acumular cientos de entradas en la historia clínica a lo largo de la vida que dejan dejar una clara evidencia de las razones por las cuales se presentan distintas entidades clínicas(45). Si no se centralizan los datos a un identificador personal se vuelven mediciones inútiles ya que solo reflejan el estado de una variable en un punto en el tiempo. La agrupación de varios análisis en una historia universal permitirán extrapolar los datos biométricos encontrados para anticiparse a la aparición de patologías prevenibles y optimizar el manejo de aquellas enfermedades crónicas(29).

La historia clínica electrónica tiene que estar diseñada con el AA en mente. EL sistema tiene que ser intuitivo para los usuarios y almacenar la información en formatos fácilmente asequibles dentro de bases de datos a gran escala. Los datos

perdidos presentan un desafío adicional para el desarrollo de algoritmos predictivos, estos reducen la capacidad predictiva de un modelo y dificultan la generalización de las observaciones. Sin embargo con un volumen suficientemente grande de datos varios modelos obtienen un buen desempeño a pesar de las entradas faltantes (46) La historia clínica inteligente integra a los algoritmos de AA directamente en su diseño. Aquí las notas de evolución de los médicos pueden ser interpretadas por redes neurales que extraen conceptos que se integran en tiempo real en la base de datos del sistema. Esto sumado a la información biométrica almacenada continuamente y los resultados de laboratorio permite generar modelos en el punto de atención o demanda a nivel del médico individual, del servicio, el hospital o el sistema de salud. (47).

Una vez que se aplique una historia clínica universal electrónica y se centralizan los resultados en una base de datos desidentificada se podrá utilizar el potencial completo del AA(47). Hasta ese entonces se deben utilizar los datos disponibles en repositorios web y otras fuentes de datos como los buscadores de imágenes o los foros web que almacenan millones de datos creados por los usuarios(48). Esto no quiere decir que los datos obtenidos a partir de estas fuentes sean menos válidos, por el contrario son una herramienta fácilmente accesible para generar modelos para distintos tipos de aplicaciones de AA.

MIMIC III

MIMIC-III (Medical Information Mart for Intensive Care III) es una gran base de datos de libre disponibilidad que incluye datos relacionados con la salud relacionados con más de cuarenta mil pacientes que permanecieron en unidades de cuidados críticos del Beth Israel Deaconess Medical Center entre 2001 y

2012(8).

La base de datos incluye información como datos demográficos, mediciones de signos vitales en la cabecera (~ 1 punto de datos por hora), resultados de pruebas de laboratorio, procedimientos, medicamentos, notas del cuidador, informes de imágenes y mortalidad (tanto dentro como fuera del hospital).

MIMIC admite una amplia gama de estudios analíticos que abarcan la epidemiología, la mejora de las reglas de decisión clínica y el desarrollo de herramientas electrónicas. Es notable por tres factores:

1. Está disponible gratuitamente para investigadores de todo el mundo.
2. Abarca una población diversa y muy grande de pacientes de la UCI
3. Contiene datos de alta resolución temporal que incluyen resultados de laboratorio, documentación electrónica y tendencias y formas de onda del monitor de cabecera.

Herramientas Informáticas

PostgreSQL

PostgreSQL, a menudo simplemente Postgres, es un sistema de gestión de bases de datos relacionales de objetos (ORDBMS) con énfasis en la extensibilidad y el cumplimiento de estándares. Como servidor de base de datos, sus funciones principales son almacenar datos de forma segura y devolver esos datos en respuesta a las solicitudes de otras aplicaciones de software. Puede

manejar cargas de trabajo que van desde pequeñas aplicaciones de una sola máquina hasta grandes aplicaciones orientadas a Internet (o para el almacenamiento de datos) con muchos usuarios concurrente(49)(Es una plataforma ideal para el aprendizaje automático, ya que permite interrogar a una base de datos y obtener los resultados correspondientes a un identificador a través de distintas tablas individuales. Esto permite sintetizar los resultados obtenidos por varios observadores en una tabla de datos que puede ser exportada a cualquier sistema de análisis automático.

R

R es un lenguaje funcional para el cálculo estadístico y gráficos(50)(. Se puede ver como un dialecto del lenguaje S (desarrollado en AT & T) por el cual John Chambers recibió el premio del Software de la Asociación de Maquinaria Computacional (ACM) de 1998 que mencionaba que este lenguaje "alteró para siempre la forma en que las personas analizan, visualizan y manipulan datos" . R puede ser bastante útil simplemente utilizándolo de forma interactiva en su línea de comando. Aún así, los usos más avanzados del sistema llevarán al usuario a desarrollar sus propias funciones para sistematizar las tareas repetitivas, o incluso para agregar o cambiar algunas funcionalidades de los paquetes complementarios existentes, aprovechando el hecho de ser de código abierto.

R permite interactuar con los datos a través de sintaxis del lenguaje python. Esto permite crear código que aplica una serie de funciones matemáticas y pruebas estadísticas a los datos seleccionados para obtener los resultados esperados. El beneficio principal de este sistema es que la comunidad científica genera constantemente códigos dentro de bibliotecas que pueden ser adaptadas

fácilmente a los datos obtenidos. Mediante la implementación de los paquetes bnlearn(51), caret(52), ggplot2(53), pROC(54),entre otros, es posible generar modelos bajo distintas técnicas y compararlos en simultáneo para seleccionar al que presenta el mejor desempeño.

Aprendizaje Automático

El rendimiento y el análisis computacional de los algoritmos de aprendizaje automático es una rama de la estadística conocida como teoría de aprendizaje computacional.(55) El aprendizaje automático se enfoca en diseñar algoritmos que permitan que una computadora aprenda. El aprendizaje no implica necesariamente la conciencia, más bien es una cuestión de encontrar regularidades estadísticas u otros patrones en los datos. Por lo tanto, muchos algoritmos de aprendizaje automático apenas se parecerán a la forma en que el ser humano podría abordar una tarea de aprendizaje. Sin embargo, los algoritmos de aprendizaje pueden dar una idea de la dificultad relativa de aprendizaje en diferentes entornos(56).Los algoritmos de AA son capaces de identificar las asociaciones presentes en grandes conjuntos de datos y generar reglas mediante las cuales pueden predecir, por ejemplo, la clase de una observación nueva a partir de las variables conocidas y la experiencia adquirida durante el entrenamiento.

El término Aprendizaje Automático fue acuñado en 1959 por Arthur Samuel en un artículo en el cual describe un software que aprende a jugar damas a través de un árbol de decisión generado a partir de instancias de aprendizaje en las cuales el

algoritmo identifica los estados del juego que tienen una tendencia a la victoria(11). Desarrollado a partir del estudio del reconocimiento de patrones y la teoría de aprendizaje computacional en inteligencia artificial (9), el AA explora el estudio y la construcción de algoritmos que pueden aprender y hacer predicciones sobre datos; dichos algoritmos se superan al maximizar el desempeño de una variable determinada tomando decisiones basadas en datos, mediante la construcción de un modelo a partir de entradas de muestra (57).

El AA se emplea en una variedad de tareas informáticas en las que el diseño y la programación de algoritmos explícitos con un buen rendimiento es difícil o inviable; las aplicaciones de ejemplo incluyen el filtrado de correo electrónico, la detección de intrusos de red o información privilegiada maliciosa que trabaja en una violación de datos, reconocimiento óptico de caracteres (OCR), aprendizaje de rango y visión artificial (57).

Existen varias ventajas en el uso de AA Los parámetros de detección se pueden ajustar según las necesidades de la función deseada esto significa que los valores de sensibilidad y especificidad pueden ser optimizados utilizando el área bajo la curva (AUC) para captar la mayor cantidad posible de casos o par a minimizar los falsos positivos(58). Las variables pueden contener clases discretas o pueden ser objetos más complejos tales como documentos de texto , imágenes, secuencias de ADN o gráficos(59).El AA puede ser aplicado a problemas con características extremadamente distintas adaptándose para cumplir la función deseada: Las redes neurales pueden ser muy eficientes para procesar el significado del lenguaje e inferir el significado de un mensaje a partir de un sofisticado proceso de abstracción del significado(60), mientras que para la

discriminación de imágenes determina características que identifican a una clase de observaciones, permitiendo identificar a los miembros de la misma sin importar la posición o la iluminación de la imagen (61).

Avances clave en el AA

1. Nuevas formas de datos: Con los avances tecnológicos realizados en genética, imagen, monitorización de señales e identificación por radiofrecuencia (RFID) por nombrar algunos, la medicina está avanzando hacia la tutoría y el tratamiento personalizados. También se han generado una gama de formas no convencionales de datos. Por ejemplo el desarrollo de dispositivos como los relojes inteligentes permiten obtener mediciones de frecuencia cardiaca perpetuamente (15). Las futuras generaciones de accesorios digitales pueden incluir una gama más amplia de sensores Por lo tanto, existe la necesidad de que el diseño de sistemas de AA se adapte a la variedad de fuentes de información y una heterogeneidad en la calidad de los datos mediante tecnicas estadistica y selección de variables para conseguir resultados extrapolables a la práctica clínica

2. El alcance del análisis estadístico sobre los datos: El AA va más allá de la exploración estadística de los datos(19). EL objetivo final es crear una base de información sobre la cual evaluar nuevas instancias. La estadística no es el objetivo, sino el medio mediante el cual el algoritmo crea asociaciones y encuentra los patrones que le permiten discriminar entre distintas clases.(62) Las técnicas estadísticas clásicas cuantifican observaciones, el AA encuentra relaciones entre dichas observaciones para predecir una nueva instancia.

3. La escalabilidad de las técnicas: Con el crecimiento de los datos que alcanzan la tasa exponencial, hay una necesidad de algoritmos y técnicas que pueden explotar los datos generados y proporcionar predicciones que pueden escalar a los cambios en los datos, así como descubrir asociaciones ocultas y discriminar observaciones triviales, cosa que no puede realizarse manualmente una vez que los datos llegan a un volumen determinado (63).

El producto final de un proceso de aprendizaje automático es un modelo. Es importante aclarar lo que queremos decir con un modelo para definirlo apropiadamente como un objeto y no una abstracción. Esta palabra se usa obviamente en muchos contextos diferentes, pero en nuestro caso estamos hablando de alguna actividad científica basada en la observación de un fenómeno en forma de un conjunto de datos. Es interesante observar una definición de referencia de un modelo científico.

De acuerdo con Wikipedia (64) El modelado científico es una actividad científica, cuyo objetivo es hacer que una determinada parte o característica del mundo sea más fácil de comprender, definir, cuantificar, visualizar o simular al hacer referencia al conocimiento existente y usualmente comúnmente aceptado. Requiere seleccionar e identificar aspectos relevantes de una situación en el mundo real y luego usar diferentes tipos de modelos para diferentes objetivos, tales como modelos conceptuales para comprender mejor, modelos operacionales para operacionalizar, modelos matemáticos para cuantificar y modelos gráficos para visualizar el tema "

Tipos de Aprendizaje Automático

Los algoritmos de aprendizaje automático se organizan taxonómicamente, en función del resultado deseado del algoritmo. Los tipos de algoritmos comunes incluyen(55):

• Aprendizaje supervisado --- donde el algoritmo genera una función que asigna entradas a las salidas deseadas. Una formulación estándar de la tarea de aprendizaje supervisado es el problema de clasificación: se requiere que el alumno aprenda (para aproximar el comportamiento de) una función que mapea un vector en una de varias clases mirando varios ejemplos de entrada-salida de la función.

• Aprendizaje no supervisado --- que modela un conjunto de entradas: los ejemplos etiquetados no están disponibles.

• Aprendizaje semi-supervisado --- que combina ejemplos etiquetados y no etiquetados para generar una función o clasificador apropiado.

• Aprendizaje reforzado --- donde el algoritmo aprende una política de cómo actuar dada una observación del mundo. Cada acción tiene algún impacto en el medio ambiente, y el entorno proporciona retroalimentación que guía el algoritmo

de aprendizaje.

• Transducción --- similar al aprendizaje supervisado, pero no construye explícitamente una función: en cambio, trata de predecir nuevos productos basados en insumos de capacitación, productos de capacitación y nuevas entradas.

• Aprender a aprender --- donde el algoritmo aprende su propio sesgo inductivo basado en la experiencia previa.

Aprendizaje Supervisado vs. No Supervisado

Es importante entender la diferencia entre la agrupación (clasificación no supervisada) y el análisis discriminante (clasificación supervisada). En la clasificación supervisada, se nos proporciona una colección de patrones etiquetados (preclasificados); el problema es etiquetar un patrón recién encontrado, pero sin etiqueta(39). Normalmente, los patrones etiquetados (de entrenamiento) se utilizan para aprender las descripciones de las clases, que a su vez se utilizan para etiquetar un nuevo patrón. En el caso de la agrupación en clústeres, el problema es agrupar una determinada colección de patrones sin etiqueta en clusters significativos. En cierto sentido, las etiquetas también están asociadas a clústeres, pero estas etiquetas de categoría están basadas en datos; es decir, se obtienen únicamente a partir de los datos(65)

Aprendizaje Automático supervisado

EL AA supervisado es la técnica más común para entrenar redes neuronales y árboles de decisión. Ambos dependen en gran medida de la información dada por las clasificaciones predeterminadas. (66) En el caso de las redes neuronales, la clasificación se utiliza para determinar el error de la red y luego ajustar la red para minimizar, y en los árboles de decisión, las clasificaciones se utilizan para determinar qué atributos proporcionan la mayor cantidad de información que se puede usar para resolver el acertijo de clasificación(55)

Cada instancia en cualquier conjunto de datos utilizado por los algoritmos de AA se representa utilizando el mismo conjunto de características. Las características pueden ser continuas, categóricas o binarias. Si las instancias se dan con etiquetas conocidas (los resultados correctos correspondientes), entonces se denomina aprendizaje supervisado(39).

Los métodos de aprendizaje supervisados incluyen los clasificadores de E-mail, reconocedores de cara sobre imágenes y sistemas de diagnóstico para los pacientes(15). El objetivo de estos sistemas es generar una predicción en respuesta a una serie de variables basándose en el conocimiento adquirido a partir de observaciones de instancias previas. La generación de los algoritmos ocurre a partir de variables con clases conocidas (65)Los modelos son generados a partir de un proceso de selección y preprocesamiento de variables, en el cual se compartimentaliza los datos en subgrupos de entrenamiento y prueba dentro de los cuales se iterara numerosas veces para encontrar el modelo predictivo con el mejor desempeño para la función determinada (58)

Arboles de decision

El aprendizaje del árbol de decisión es un método comúnmente utilizado en la minería de datos. (67) El objetivo es crear un modelo que prediga el valor de una variable objetivo en función de varias variables de entrada.

El árbol de decisión consta de nodos que forman un árbol enraizado, lo que significa que es un árbol dirigido con un nodo llamado "raíz" que no tiene bordes entrantes. Todos los demás nodos tienen exactamente un borde entrante. Un nodo con bordes salientes se conoce como nodo "interno" o "de prueba". Todos los otros nodos se llaman "hojas" (también conocidos como nodos "terminales" o "de decisión").(68) En el árbol de decisiones, cada nodo interno divide el espacio de la instancia en dos o más subespacios de acuerdo con una cierta función discreta de los valores de los atributos de entrada.

En el caso más simple y más frecuente, cada prueba considera un único atributo, de modo que el espacio de la instancia se particione de acuerdo con el valor de los atributos. En el caso de los atributos numéricos, la condición se refiere a un rango. Cada hoja está asignada a una clase que representa el valor objetivo más apropiado. Alternativamente, la hoja puede contener un vector de probabilidad que indica la probabilidad de que el objetivo tenga un cierto valor característica de entrada. Los arcos que provienen de un nodo etiquetado con una característica de entrada están etiquetados con cada uno de los posibles valores de la característica de destino o salida o el arco conduce a un nodo de decisión subordinado en una característica de entrada diferente. Cada hoja del árbol está etiquetada con una clase o una distribución de probabilidad sobre las clases(68)

Aprendizaje en conjunto

El aprendizaje en conjunto se refiere a los métodos de AA que generan muchos clasificadores y agregan sus resultados.(69) Dos métodos bien conocidos son el refuerzo (véase, por ejemplo, (70) y el ensacado (71) de árboles de clasificación. En el refuerzo, los árboles sucesivos dan un peso adicional a los puntos correctamente pronosticados por predictores anteriores. Al final, se toma una votación ponderada para la predicción. En el ensacado, los árboles sucesivos no dependen de árboles anteriores: cada uno se construye de forma independiente utilizando una muestra de arranque del conjunto de datos. Al final, se toma una mayoría simple para la predicción.

Bosques Aleatorios

Los bosques aleatorios o los bosques de decisión aleatoria son un método de aprendizaje en conjunto para clasificación, regresión y otras tareas, que operan construyendo una multitud de árboles de decisión en el tiempo de entrenamiento y generando la clase que es el modo de las clases (clasificación) o predicción media (regresión) de los árboles individuales. El primer algoritmo para los bosques de decisión aleatoria fue creado por Tin Kam Ho (67)utilizando el método de subespacio aleatorio, (72) que, en la formulación de Ho, es una forma de implementar el enfoque de "discriminación estocástica" propuesto por Eugene Kleinberg. (73)

Breiman(74), propuso bosques aleatorios, que agregan una capa adicional de aleatoriedad al ensacado. Además de construir cada árbol usando una muestra

de arranque diferente de los datos, los bosques aleatorios cambian la forma en que se construyen los árboles de clasificación o regresión. En árboles estándar, cada nodo se divide utilizando la mejor división entre todas las variables. En un bosque al azar, cada nodo se divide utilizando el mejor entre un subconjunto de predictores elegidos al azar en ese nodo. Esta estrategia un tanto contraintuitiva resulta funcionar muy bien en comparación con muchos otros clasificadores, incluido el análisis discriminante, las máquinas de vectores de soporte y las redes neuronales, y es sólida contra el sobreajuste (74). Además, es muy fácil de usar en el sentido de que tiene solo dos parámetros (el número de variables en el subconjunto aleatorio en cada nodo y el número de árboles en el bosque), y generalmente no es muy sensible a sus valores. En R hay varias implementaciones de bosques aleatorios. El principal, basado en el código original de Leo Breiman, está disponible en el paquete randomForest (69).

Análisis de datos

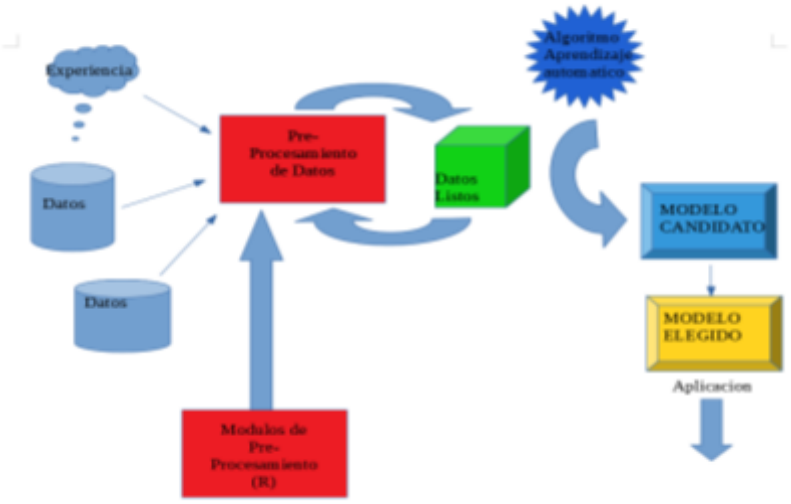
Recolección y preprocesamiento

El procedimiento de recolección de datos de pacientes se realiza a partir de la historia clínica electrónica que se genera al momento del nacimiento e incluye: a) el consentimiento del paciente, b) La creación de un identificador único encriptado para permitir el acceso a los investigadores sin comprometer la identidad del paciente (75) c) La captación de datos generados por las interacciones entre el paciente y el sistema de salud (76)

Una vez compilado en una base de datos inicia el proceso del descubrimiento de conocimiento. Este se realiza en varias etapas que son: Selección, Pre

procesamiento, transformación, minería de datos y evaluación/interpretación.(77) La selección de los datos inicia a partir de la generación de la hipótesis que el modelo buscará satisfacer. Los datos tienen que ser lo suficientemente extensos para enmarcar las asociaciones entre los factores, pero deben mantener así mismo la eficiencia computacional.

Figura 2: Esquema de flujo de informacion para analisis de aprendizaje automatico



Para comprender el proceso del AA es importante definir algunas propiedades de las tablas de datos estándar. Una tabla de datos (o conjunto de datos) es una estructura de datos bidimensional donde cada fila representa una entidad (por ejemplo, producto, persona, etc.) y las columnas representan las propiedades (nombre, edad, temperatura, etc.) que hemos medido en cada entidad. Los términos entidad y propiedad tienen muchos sinónimos. Las entidades se

denominan con frecuencia objetos, tuplas, registros, ejemplos o vectores de características, mientras que las propiedades a menudo se denominan características, atributos, variables, dimensiones o campos.

En términos de las filas de un conjunto de datos, podemos tener dos tipos principales de configuraciones: (i) cada fila es independiente de las demás; o (ii) hay alguna dependencia entre las filas.Ejemplos de posibles dependencias incluyen algún tipo de orden temporal (77)(por ejemplo, las filas representan las mediciones de un conjunto de variables en pasos de tiempo sucesivos), o algún orden espacial (por ejemplo, cada fila tiene una ubicación asociada y esto puede implicar alguna forma de relación de vecindad entre filas).

Con respecto a las columnas de un conjunto de datos podemos hablar sobre el tipo de valores de datos que almacenan. La distinción más frecuente es entre variables cuantitativas y categóricas. Una taxonomía más fina de los tipos de datos puede convertirlos en:

• intervalo: variables cuantitativas como fechas de ejemplo.

• ratio: variables cuantitativas como, por ejemplo, la altura de una persona o el precio de un producto.

• nominal: son variables categóricas cuyos valores son un tipo de etiquetas sin ningún orden entre ellas (por ejemplo, colores).

• ordinal: de nuevo variables categóricas, pero esta vez con algún orden implícito entre su conjunto finito de valores (por ejemplo, pequeño, mediano y grande).

A pesar de estas categorías, en la práctica las personas tienden a distinguir solo entre variables cuantitativas (también referidas como continuas) y categóricas

(también referidas como discretas). Las columnas de un conjunto de datos también pueden ser independientes entre sí o pueden estar correlacionadas de alguna manera, es decir, los valores de una variable pueden tener alguna forma de dependencia de los valores de otras variables. En R, las tablas de datos normalmente se almacenan en un marco de datos con columnas que almacenan variables cuantitativas como tipos de datos numéricos y variables categóricas como factores o cadenas de caracteres.

Wickham(78) presentó la noción de datos ordenados como un objetivo general que deberíamos seguir para facilitar nuestro análisis posterior en R. Las propiedades clave de los datos ordenados son que: (i) cada valor pertenece a una variable y a una observación; (ii) cada variable contiene todos los valores de una determinada propiedad medidos en todas las observaciones; y (iii) cada observación contiene todos los valores de las variables medidas para el caso respectivo. Estas propiedades conducen a un tipo de tabla de datos rectangular formada por filas (que representan las observaciones) y columnas (que representan las variables). A veces los datos no se proporcionan en un formato tan ordenado, que es requerido por la mayoría de las herramientas de modelado disponibles en R. El paquete tidyr (79) se puede usar para limpiar los datos y hacerlo más estándar.

El pre procesamiento consiste en aplicar las nociones de datos ordenados y los objetivos de la investigación para determinar qué datos deben ser rescatados para generar el modelo. Las características seleccionadas alterarán tanto el tiempo y la complejidad del análisis como el desempeño del modelo para cumplir su función y ser aplicable en situaciones reales(80,81).

Discretización de las variables

Otra técnica de preprocesamiento de datos utilizada con frecuencia es la discretización de variables numéricas. al transformarlas en factores con intervalos significativos. Esto puede ser motivado por los objetivos del análisis o incluso por la reducción de la complejidad computacional de algunas herramientas de modelado que tienen este factor que depende del número de valores diferentes de las variables. Cualquiera que sea la motivación, esto implica cambiar una variable numérica en una variable ordinal, aunque con frecuencia termina siendo tratada como una variable nominal porque algunas herramientas no distinguen entre estos dos tipos diferentes de variables.

La mayor parte del tiempo la discretización se lleva a cabo en función de algunos conocimientos de dominio que proporcionan los intervalos en el rango de la variable original que tienen sentido para los usuarios finales. Para el propósito de este estudio se discretizan las variables de laboratorio y signos vitales de los pacientes según los puntos de corte establecidos en el score SOFA y qSOFA que descritos anteriormente utilizando código disponible en el repositorio de github del proyecto MIMIC-III. Este código está diseñado para extraer los valores de laboratorio y datos clínicos de las evoluciones del paciente para calcular los valores de SOFA y qSOFA para luego compilarlos en un dataframe en R .

Evaluacion

Uno de los problemas clave para un proyecto exitoso de minería de datos es poder evaluar correctamente el rendimiento de los modelos propuestos(77). Las estimaciones confiables del rendimiento son de suma importancia cuando se trata de decidir implementar sus modelos en producción(81). Si no se proporcionan

estas estimaciones confiables, es posible que se obtengan resultados decepcionantes que pueden comprometer seriamente la aplicación del aprendizaje automático en el sistema de salud.

Validación cruzada (CV) k-Fold

La validación cruzada (CV) k-Fold es uno de los métodos más comunes para evaluar el desempeño predictivo de un modelo. Consiste en repetir k veces un ciclo de tren / prueba, pero donde el conjunto de prueba se elige cuidadosamente en lugar de seleccionar al azar como en k repeticiones de submuestreo aleatorio. Comenzamos por organizar aleatoriamente los datos de entrenamiento para evitar cualquier efecto de ordenamiento. Luego dividimos el conjunto de datos en k particiones de igual tamaño. Estos serán los k conjuntos de prueba . Para cada uno de los conjuntos de prueba, el conjunto de entrenamiento respectivo estará formado por las restantes k - 1 particiones . La estimación de validación cruzada k-fold será el promedio de los k puntajes individuales obtenidos en cada partición de prueba. k-fold CV es con frecuencia el procedimiento seleccionado para estimar el rendimiento de un modelo. Es la recomendación para conjuntos de datos medianos (pocos cientos a pocos miles de casos). Algunas veces se repite el proceso varias veces y hacemos un promedio de estas repeticiones para aumentar la confiabilidad de las estimaciones.(82,83).

CAPÍTULO 3

Metodologia

Estudio retrospectivo, longitudinal

Universo: Pacientes incluidos en base de datos MIMIC-III

Muestra: Todos los pacientes del universo que cumplan los criterios de inclusión y exclusión.

Criterios de inclusión:

- Pacientes con código de Sepsis severa(ICD9 995.92) o Sepsis(ICD9 995.91) de la base de datos MIMICIII

Criterios de exclusión:

- Pacientes con información incompleta o variables no disponibles

Método de recolección de datos: Para acceder a la base de datos mimic el autor completó un curso web sobre la importancia de la privacidad y el manejo seguro de datos tras lo cual se generaron las claves de acceso. La base de datos fue descargada en formato .csv y consta de 27 tablas de excel que contienen diferentes datos con identificadores individuales randomizados para proteger la identidad de los pacientes. La base de datos está diseñada para ser ensamblada en el lenguaje PostgresSQL

Operacionalización de las Variables

Variable	Definición	Dimensión	Indicador	Nivel de medición	Instrumentos de medición de datos	Estadística
Temperatura Corporal	Medida de la capacidad del cuerpo para la termorregulación	Medida de la capacidad del cuerpo para la termorregulación en pacientes MIMIC-III	Hipotermia (<36.5 grados centígrados) Normotermia (36.5-37.5) Hipertermia (> 37.5)	Ordinal	Historia Clinica	Frecuencia, Porcentaje
Conteo Leucocitario	Número de glóbulos	Número de glóbulos	Leucocitos >12,000/	Ordinal	Historia Clinica	Frecuencia, Porcentaje

	blancos en una muestra de sangre	blancos en una muestra de sangre de pacientes MIMIC-III	mm3(Según criterios de sirs) Si o No			
Frecuencia Respiratoria	Número de respiraciones realizadas en un minuto	Número de respiraciones por minuto de pacientes MIMIC-III	FR > 20 Si o No	Ordinal	Historia Clinica	Frecuencia, Porcentaje
Presion Arterial Media	Promedio de presión arterial durante un ciclo cardiaco	Promedio de presión arterial durante un ciclo cardíaco en pacientes MIMIC-III	PAM < 70 Si o No	Ordinal	Historia Clinica	Frecuencia Porcentaje
Nivel de Bilirrubina	Prueba utilizada para evaluar la	Prueba utilizada para evaluar la	Alto(>= 3.5mg/dL) o muy alto (>5	Ordinal	Historia clinica	Frecuencia, Porcentaje

	función hepática	función hepática en pacientes MIMIC-III	mg/ dL)			
Conteo Plaquetario	Prueba utilizada para evaluar la hemostasia y la función de la médula ósea	Prueba utilizada para evaluar la hemostasia y la función de la médula ósea en pacientes MIMIC-III	<150 bajo <100 muy bajo <50 depleción	Ordinal	Historia Clinica	Frecuencia, Porcentaje
Nivel de Conciencia	Prueba que evalúa el estado neurocognitivo de un individuo según criterio qSOFA	Prueba que evalúa el estado neurocognitivo en un paciente MIMIC-III según criterio qSOFA	Normal, alterado (<13 GCS)	Ordinal	Historia Clinica	Frecuencia, Porcentaje

Creatinina	Prueba que evalua niveles de creatinina serica en un individuo según criterio SOFA	Prueba que evalúa niveles de creatinina sérica en un paciente MIMIC-II según criterio SOFA	Moderada 1.2-2 mg/dl Alta 2.3-5 mg/dl Muy alta >5 mg/dl	Ordinal	Historia Clinica	Frecuencia, Porcentaje
Edad	Tiempo transcurrido desde el nacimiento de un individuo	Tiempo transcurrido desde el nacimiento de un individuo	(18-30)(30-50)(50-70)(70-90	Ordinal	Historia Clinica	Frecuencia, Porcentaje
Mortalidad	Cese de homeostasis y actividad nerviosa central	Cese de homeostasis y actividad nerviosa central en pacientes MIMIC-III	Si o No	Ordinal	Historia Clinica	Frecuencia, Porcentaje

Ocurrencia de Sepsis según código CIE 9	Respuesta inflamatoria que amenaza la vida del paciente en respuesta a una infección	Respuesta inflamatoria que amenaza la vida en respuesta a una infección en pacientes MIMIC-III	995.92 o 995.91	nominal	Historia Clinica	Frecuencia, Porcentaje

Estadística:

Las variables utilizadas para construir el modelo son las descritas en la operacionalización de las variables. Los datos fueron extraídos de la base de datos MIMIC-III utilizando los repositorios encontrados en la página de GitHub asociada a los investigadores participantes. Los algoritmos disponibles extraen los datos necesarios para calcular las variables de SOFA y qSOFA automáticamente a partir de un cohorte seleccionado en base al código CIE 9 lo que incluye 4182 pacientes.

Para ser evaluadas las variables de laboratorio fueron agrupadas según valores de corte establecidos en el score SOFA. los niveles clínicos en la escala q sofa fueron codificados como factores con niveles numéricos según los puntos de corte. los datos clínicos necesarios para calcular fueron extraídos en SQL

utilizando el script “vitalsfirstday” que se encuentra en el anexo 1. Las observaciones utilizadas corresponden únicamente al primer día de internación.

El siguiente paso una vez completada la limpieza de los datos consiste en explorar la distribución de las variables de estudio. Al mismo tiempo se procede a dividir el cohorte en un grupo de entrenamiento, que representa al 75% de los casos, elegidos de manera aleatoria y un conjunto de prueba que consiste en el resto de las observaciones. Se generaron dos grupos de modelos en relación a la mortalidad en cuidados intensivos e intrahospitalaria

Una de las características de este cohorte es que la mortalidad está representada de forma inbalanceada. Esto implica que hay más pacientes pertenecientes en una de las clases . El problema con esta situación radica en que los algoritmos de clasificación pueden seleccionar la variable mayoritaria indistintamente de las variables disponibles, ya que es más probable obtener un buen resultado si todos las predicciones pertenecen al grupo de mayoría. Para corregir esto se aplicó un algoritmo de downsampling que seleccionó un subconjunto en el cual la prevalencia de pacientes que sobreviven y fallecen es 1:1.

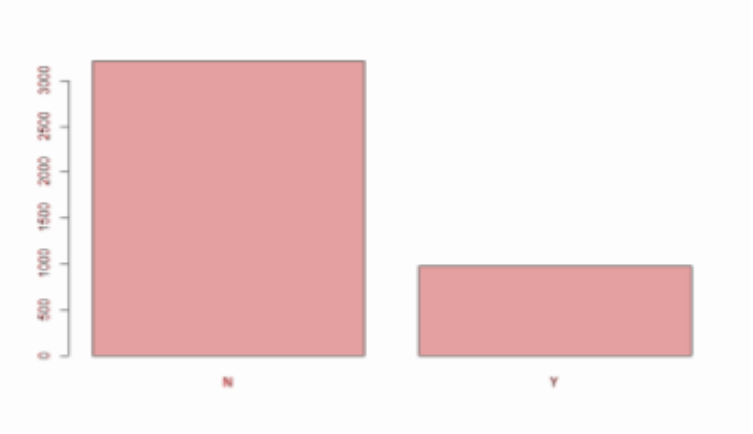


Grafico 1: Mortalidad
En UCI

Se construyó un modelo lineal generalizado que estudia cada variable seleccionada en relación a la mortalidad. En la tabla 1 se puede observar el impacto de cada una de las variables dentro del modelo. Luego de preparar nuestros subgrupos se procede a construir 4 modelos de clasificación binaria basándose en las variables antes descritas y con el objetivo de clasificar a los pacientes según la variable "Class" con niveles 0 o 1 que corresponde a la mortalidad en UCI de ese paciente.

Todos los modelos fueron generados con el paquete ("Caret") en R. El primer modelo generado (modelo A) es un Bayes tonto, quien optimizar su puntaje al seleccionar en todas las instancias a la clase mayoritaria.

El modelo B fue generado a partir de una máquina de soporte de vectores. La selección de este modelo se realizó a partir de la comparación de una serie de algoritmos generados con las mismas constantes de entrenamiento para comparar su desempeño. En la gráfica 2 podemos observar el rango de desempeño de distintos algoritmos.

El modelo C fue generado a partir de un bosque aleatorio optimizado por un proceso de retroalimentación. Este proceso consiste en una validación k-fold en el cual se optimizan las nuevas instancias en base a los resultados obtenidos anteriormente el subconjunto de los datos utilizados es obtenido mediante la función downSample para conseguir una prevalencia de mortalidad de 0.5. El punto de corte del modelo es determinado automáticamente a partir de La distancia del modelo perfecto. Esto se determina optimizando la sensibilidad y la especificidad en función a la curva roc. Por lo cual se puede elegir un punto de corte que maximiza el desempeño del modelo.

A partir de los modelos generados se realizaron predicciones sobre la mortalidad de los pacientes del grupo de prueba con lo cual se configuró una tabla de contingencia 2x2 para evaluar el desempeño de cada uno de ellos. El grupo de prueba que consistió en 1048 pacientes separados en la etapa de preprocesamiento de datos. Se obtuvieron curvas ROC y se graficaron utilizando gráficos agregados para visualizar la variabilidad de los modelos.

Resultados

El análisis exploratorio de los datos mediante regresión lineal generalizada encontró que presentar un nivel adecuado de plaquetas y ser catalogado como sepsis severa resultaron ser los dos factores más importantes para determinar la mortalidad (OR 0.34 y 3.88 respectivamente). Algunas otras variables también demostraron ser predictores estadísticamente significativos de mortalidad. Entre ellas los niveles de bilirrubina mayores a 5 se asociaron con un OR de 2.47 mientras que la alteración del estado de conciencia obtuvo un OR de 1.25. El resto de las variables se resumen en el gráfico 2

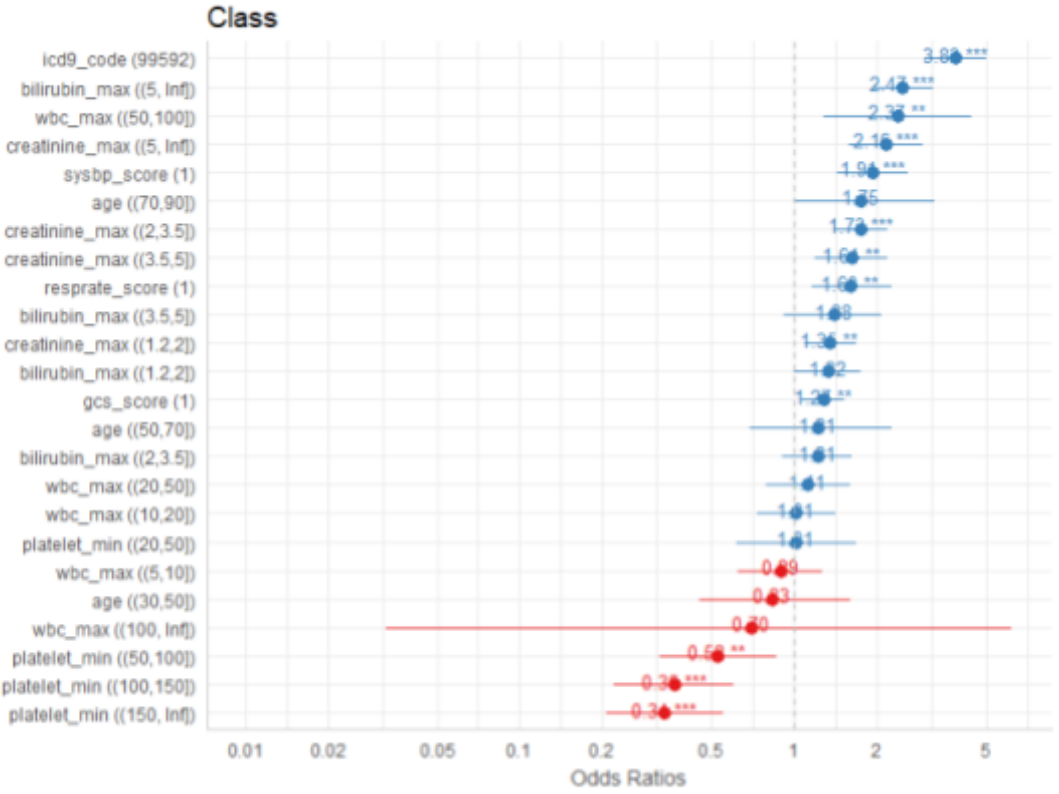
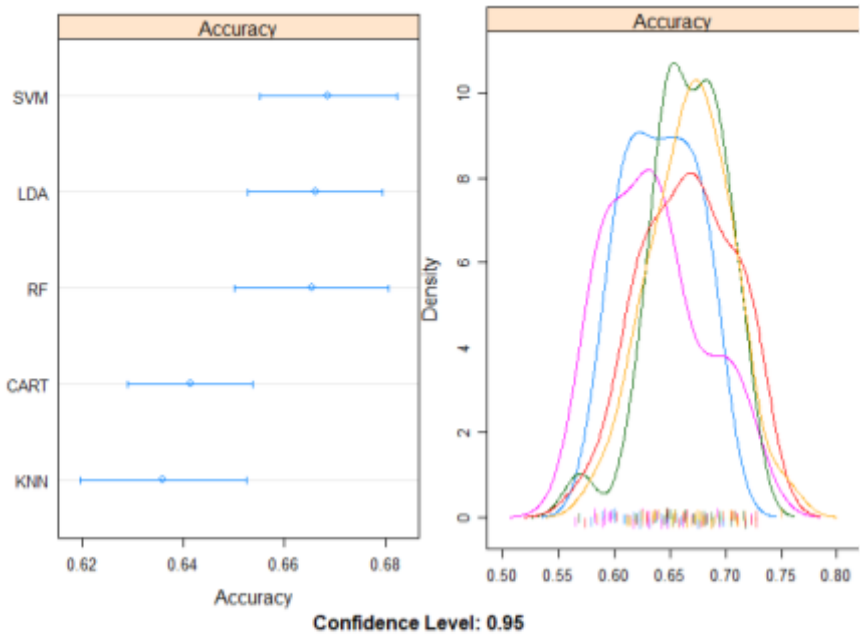


Grafico 2 Forest Plot OR por variables

La generación en paralelo de 5 modelos de AA construyó una máquina de soporte de vectores con una precisión de 0.75 en la predicción de mortalidad como el modelo óptimo en este cohorte. Entre los algoritmos generados destaca el bosque aleatorio quien obtuvo una presión máxima de 0.72 En la gráfica 3 podemos observar la distribución de precisión de los algoritmos en cada ronda de entrenamiento, El promedio de estas instancias es resumido en el modelo final. Los resultados del entrenamiento de los algoritmos se visualiza en la parte inferior del gráfico 3 representan los rangos de precisión obtenidos por cada modelo



Confidence Level: 0.95

Accuracy

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
CART	0.5804598	0.6141618	0.6408046	0.6412850	0.6704953	0.6994220	0
LDA	0.5689655	0.6473988	0.6647399	0.6661229	0.6887582	0.7183908	3
SVM	0.5862069	0.6445087	0.6724694	0.6686168	0.6913091	0.7514451	0
KNN	0.5664740	0.5959571	0.6311209	0.6359122	0.6594493	0.7241379	0
RF	0.5747126	0.6315029	0.6657033	0.6653579	0.7011494	0.7283237	0

Grafico 3 Resultados de generacion en paralelo de modelos

La Tabla 1 resume los desempeños de predicción de mortalidad para cada resumen de par estadístico utilizando los clasificadores RF, SVM y NB. La mejor predicción de mortalidad (AUC = 0.703 SN = 0.8754, SP = 0.3416) entre el conjunto de datos de pacientes con sepsis se proporcionó utilizando RF con las diez variables descritas en la operacionalización de variables resumidas según los puntos de corte establecidos.

Bosque Aleatorio	Máquina de Soporte de Vectores	Clasificador Bayesiano
Reference Prediction N Y N 611 87 Y 451 234	Reference Prediction N Y N 593 71 Y 469 250	Reference Prediction N Y N 787 156 Y 275 165
Sensitivity : 0.7290 Specificity : 0.5753 Pos Pred Value : 0.3416 Neg Pred Value : 0.8754	Sensitivity : 0.7788 Specificity : 0.5584 Pos Pred Value : 0.3477 Neg Pred Value : 0.8931	Sensitivity : 0.5140 Specificity : 0.7411 Pos Pred Value : 0.3750 Neg Pred Value : 0.8346

Tabla 1

Una vez que se completaron los modelos se hizo un análisis mediante una curva ROC para comparar su capacidad predictiva. Entre ellos se seleccionó al modelo de bosque aleatorio utilizando un valor de corte de 0.59 que obtiene el mejor resultado para este modelo calculando la distancia del modelo ideal como se observa en la gráfica 3. En este punto la sensibilidad es de 0.52 y la especificidad de 0.79 con AUROC de 0.703.

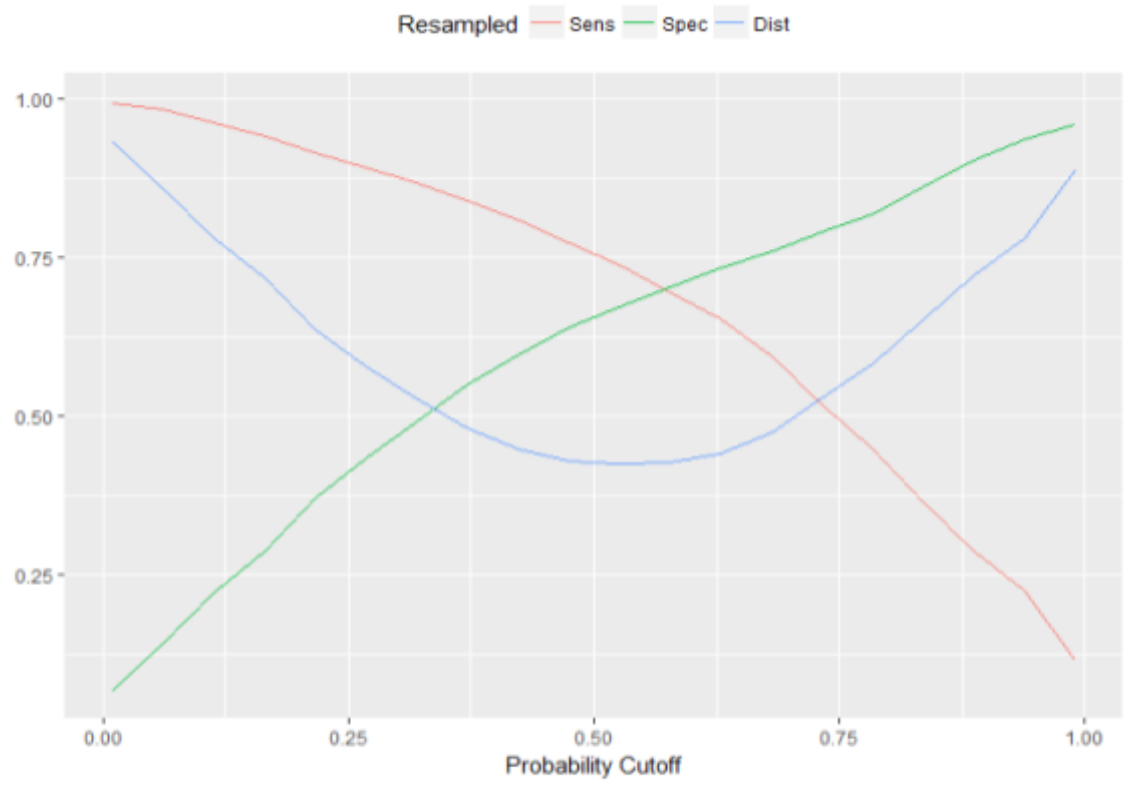
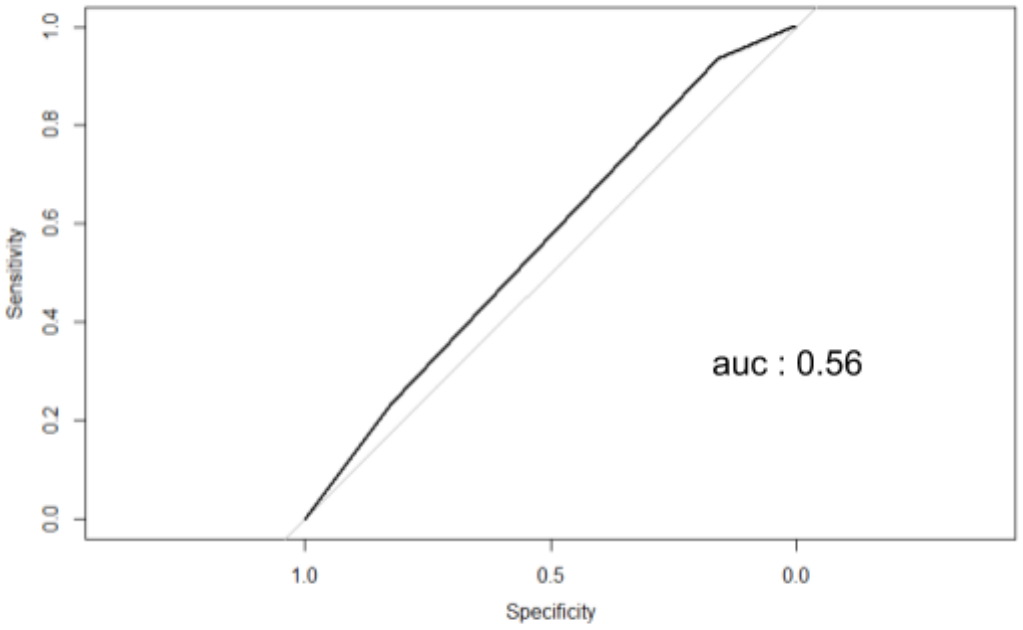


Grafico 3: Rango de Sens, Esp y distancia de modelo perfecto segun punto de corte para modelo RF

Al finalizar se aplicaron los modelos confeccionados a el subgrupo de pacientes de prueba y se obtuvieron los siguientes resultados: Los modelos de mortalidad hospitalaria tuvieron un mejor desempeño, obteniendo medidas de AUC entre 7.22 para el modelo SVM y 0.69 para el modelo lineal. como se disponen en el gráfico 4. Para la predicción de mortalidad hospitalaria el modelo de árbol de decisión obtuvo una sensibilidad de 0.751 y especificidad de 0.555.

Se realizó un análisis ROC utilizando score QSOFA obteniéndose área bajo la curva de 0.56 cómo se grafica a continuación



Curva ROC de score QSOFA

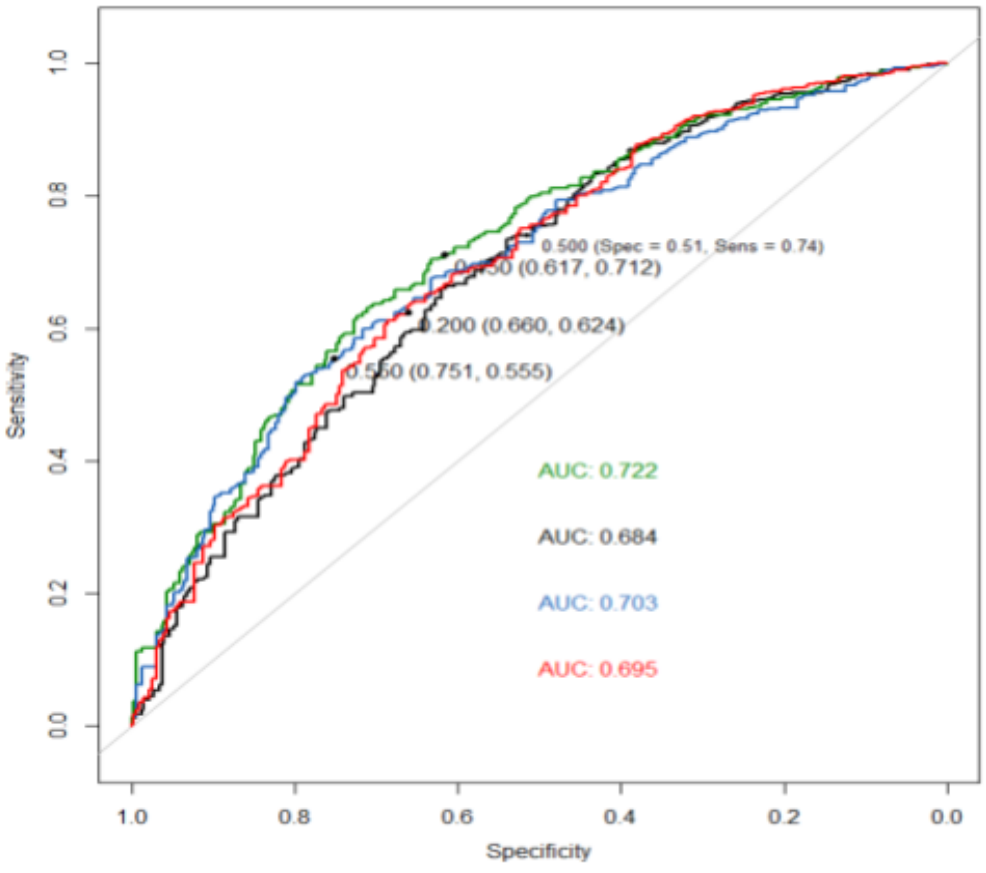


Grafico 4: Analisis de Curva ROC de modelos de estudio

Discusión

El presente trabajo demuestra la capacidad de los algoritmos de AA para predecir el resultado de procesos complejos como la mortalidad por sepsis. El modelo construido para esta tesis con AUROC de 0.703 es superior que el score qSOFA AUC 0.56 obtenido en este estudio y de los puntajes reportados en la literatura para grupos de pacientes similares(33).

Este estudio demuestra la viabilidad de utilizar la clasificación de RF con la selección de características para predecir el riesgo de mortalidad para pacientes con sepsis. Con las diez características seleccionadas precisión y capacidad de discernimiento logradas fueron de 0.703 AUC. Este resultado es comparable al de Ribas(84) En su estudio, se empleó la regresión logística con análisis factorial para lograr una precisión de 0,78 y la discriminación de 0,75 AUC con 156 pacientes con UCI con sepsis grave. La diferencia clave entre nuestro trabajo y el de Ribas es la naturaleza de las observaciones. Las variables utilizadas en este estudio incluyen laboratorios y variables clínicas obtenidas durante el ingreso del paciente. Ribas realizó su clasificación final con muchas más variables clínicas, incluyendo el número de órganos disfuncionales, ventilación mecánica, puntaje APACHE II y paquetes de resucitación. Estas variables son todas derivadas de los sistemas agregados de puntuación utilizados en los entornos de cuidados intensivos del hospital y, en general, no están disponibles en entornos que no son de la UCI.

Como la mayoría de los casos de sepsis se identifican inicialmente en entornos que no pertenecen a la UCI (departamento de emergencia o sala de hospital), y la estratificación de riesgo rápida es esencial para el tratamiento, nuestro enfoque

tiene una aplicabilidad potencial mucho más amplia. Dado que las puntuaciones agregada no se utilizó como datos de entrada en este estudio, nuestros datos de entrada pueden ser más específicos para un paciente individual.

Conclusion

Los sistemas clínicos se abruman con pruebas innecesarias y los pacientes pueden estar expuestos a intervenciones de alto riesgo que pueden no beneficiarlos. Un sistema dinámico de apoyo a la decisión que predice con precisión el riesgo de mortalidad podría enfocar recursos limitados, evitar la "fatiga de alerta" que puede acompañar a los sistemas de apoyo a la decisión, y disminuir las complicaciones de procedimiento para los pacientes. El éxito de nuestra clasificación de mortalidad podría extenderse al incluir otras variables o patologías para optimizar los patrones de respuesta

Otras ventajas de este modelo incluyen la adaptabilidad a nuevas variables y requerimientos. El proceso aplicado puede ser reproducido y modificado para otras patologías y poblaciones con bastante facilidad.

El enfoque de este estudio no radica en encontrar el modelo perfecto para predecir la mortalidad por sepsis, sino más bien busca demostrar el potencial del aprendizaje automático como herramienta para múltiples problemas del sistema de salud tanto a nivel del proveedor como a nivel institucional y sistemático permitiendo utilizar los datos recogidos a nivel regional para informar toma de decisiones y permitir una respuesta homogénea en los distintos puntos del sistema de salud.

La clave está en optimizar las técnicas de obtención y procesamiento de datos.

Una vez que se ponga en marcha un sistema informático moderno se pueden adaptar los datos obtenidos para cualquier problema de alta complejidad. Los modelos pueden servir como herramientas que materializan la información disponible en un predictor tangible en el que se puede apoyar para tomar una decisión clínica.

El futuro de estos sistemas informáticos es extremadamente prometedor. No es difícil visualizar sistemas a nivel hospitalario que sean manejados en conjunto mediante los equipos de informática y los médicos de la institución, donde la toma de decisiones del día a día alimenta a un sistema informático que encuentra patrones escondidos en los datos y los pone en servicio de la práctica clínica. El desafío radica en renovar los sistemas informáticos médicos con el objetivo de facilitar el flujo de datos para realizar estos modelos diariamente con el objetivo de optimizar el uso de recursos y los resultados clínicos.’

Bibliografia

1. Lohr S. The Origins of “Big Data”: An Etymological Detective Story [Internet]. Bits Blog. 2013 [cited 2018 Mar 11]. Available from: //bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/

2. Sarikaya R, Hinton GE, Deoras A. Application of Deep Belief Networks for Natural Language Understanding. IEEEACM Trans Audio Speech Lang Process. 2014 Apr;22(4):778–84.

3. Cox MT, Ram A. Introspective multistrategy learning: On the construction of learning strategies. Artif Intell. 1999 Aug;112(1–2):1–55.

4. Beaulieu-Jones B. Machine Learning for Structured Clinical Data. In: Holmes DE, Jain LC, editors. Advances in Biomedical Informatics [Internet]. Cham: Springer International Publishing; 2018 [cited 2018 Mar 11]. p. 35–51. Available from: http://link.springer.com/10.1007/978-3-319-67513-8_3

5. Gunter TD, Terry NP. The Emergence of National Electronic Health Record Architectures in the United States and Australia: Models, Costs, and Questions. J Med Internet Res. 2005 Mar 14;7(1):e3.

6. Giger ML. Machine Learning in Medical Imaging. J Am Coll Radiol. 2018 Mar;15(3):512–20.

7. Wu X, Zhu X, Wu G-Q, Ding W. Data Mining with Big Data. :26.

8. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016 May 24;3:160035.

9. Shannon CE. A Mathematical Theory of Communication. Bell Syst Tech J. 1948 Oct;27(4):623–56.

10. Before the Computer [Internet]. Princeton University Press. [cited 2018 Mar 11].

Available from: https://press.princeton.edu/titles/5134.html

11. Samuel AL. Some studies in machine learning using the game of checkers. IBM J Res Dev. 2000 Jan;44(1.2):206–26.

12. Langley P. The changing science of machine learning. Mach Learn. 2011 Mar;82(3):275–9.

13. Moore GE. Cramming More Components Onto Integrated Circuits. Proc IEEE. 1998 Jan;86(1):82–5.

14. November 2014 | TOP500 Supercomputer Sites [Internet]. [cited 2018 Mar 11]. Available from: https://www.top500.org/lists/2014/11/

15. Dua S, Acharya UR, Dua P, editors. Machine Learning in Healthcare Informatics [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2014 [cited 2018 Mar 11]. (Intelligent Systems Reference Library; vol. 56). Available from: http://link.springer.com/10.1007/978-3-642-40017-9

16. van den Heever M, Mittal A, Haydock M, Windsor J. The use of intelligent database systems in acute pancreatitis – A systematic review. Pancreatology. 2014 Jan;14(1):9–16.

17. Kessler RC, van Loo HM, Wardenaar KJ, Bossarte RM, Brenner LA, Cai T, et al. Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. Mol Psychiatry. 2016 Oct;21(10):1366–71.

18. Mani S, Ozdas A, Aliferis C, Varol HA, Chen Q, Carnevale R, et al. Medical decision support using machine learning for early detection of late-onset neonatal sepsis. J Am Med Inform Assoc. 2014 Mar;21(2):326–36.

19. Tsoukalas A, Albertson T, Tagkopoulos I. From Data to Optimal Decision Making: A Data-Driven, Probabilistic Machine Learning Approach to Decision Support for

Patients With Sepsis. JMIR Med Inform. 2015 Feb 24;3(1):e11.

20. Zheng Y-L, Ding X-R, Poon CCY, Lo BPL, Zhang H, Zhou X-L, et al. Unobtrusive Sensing and Wearable Devices for Health Informatics. IEEE Trans Biomed Eng. 2014 May;61(5):1538–54.

21. Zangheri M, Cevenini L, Anfossi L, Baggiani C, Simoni P, Di Nardo F, et al. A simple and compact smartphone accessory for quantitative chemiluminescence-based lateral flow immunoassay for salivary cortisol detection. Biosens Bioelectron. 2015 Feb;64:63–8.

22. Rhee C, Dantes R, Epstein L, Murphy DJ, Seymour CW, Iwashyna TJ, et al. Incidence and Trends of Sepsis in US Hospitals Using Clinical vs Claims Data, 2009-2014. JAMA. 2017 Oct 3;318(13):1241.

23. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016 Feb 23;315(8):801.

24. Jin M, Khan AI. Procalcitonin: Uses in the Clinical Laboratory for the Diagnosis of Sepsis. Lab Med. 2010 Mar;41(3):173–7.

25. Reade MC, Huang DT, Bell D, Coats TJ, Cross AM, Moran JL, et al. Variability in management of early severe sepsis. Emerg Med J. 2010 Feb 1;27(2):110–5.

26. Fleischmann C, Scherag A, Adhikari NKJ, Hartog CS, Tsaganos T, Schlattmann P, et al. Assessment of Global Incidence and Mortality of Hospital-treated Sepsis. Current Estimates and Limitations. Am J Respir Crit Care Med. 2016 Feb;193(3):259–72.

27. Paoli C, Reynolds M, Sinha M, Gitlin M. 1486: CURRENT BURDEN, OUTCOMES, AND COSTS OF SEPSIS MANAGEMENT IN UNITED STATES HOSPITALS. Crit Care Med. 2018 Jan;46:727.

28. Rampasek L, Goldenberg A. TensorFlow: Biology’s Gateway to Deep Learning? Cell Syst. 2016 Jan;2(1):12–4.

29. Obermeyer Z, Emanuel EJ. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. N Engl J Med. 2016 Sep 29;375(13):1216–9.

30. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J. 2015;13:8–17.

31. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: A System for Large-Scale Machine Learning. In: OSDI. 2016. p. 265–283.

32. Vincent J-L, Moreno R, Takala J, Willatts S, Mendon&#xfffd;a AD, Bruining H, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. Intensive Care Med. 1996 Jul 1;22(7):707–10.

33. Raith EP, Udy AA, Bailey M, McGloughlin S, MacIsaac C, Bellomo R, et al. Prognostic Accuracy of the SOFA Score, SIRS Criteria, and qSOFA Score for In-Hospital Mortality Among Adults With Suspected Infection Admitted to the Intensive Care Unit. JAMA. 2017 Jan 17;317(3):290–300.

34. Finkelsztein EJ, Jones DS, Ma KC, Pabón MA, Delgado T, Nakahira K, et al. Comparison of qSOFA and SIRS for predicting adverse outcomes of patients with suspicion of sepsis outside the intensive care unit. Crit Care. 2017 Mar 26;21:73.

35. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inf Sci Syst [Internet]. 2014 Dec [cited 2018 Mar 11];2(1). Available from: http://link.springer.com/10.1186/2047-2501-2-3

36. A Complete Tutorial which teaches Data Exploration in detail [Internet]. Analytics Vidhya. 2016 [cited 2018 Mar 11]. Available from: https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/

37. Poldrack RA, Gorgolewski KJ. Making big data open: data sharing in neuroimaging. Nat Neurosci. 2014 Nov;17(11):1510–7.

38. Gavves E, Mensink T, Tommasi T, Snoek CGM, Tuytelaars T. Active Transfer Learning with Zero-Shot Priors: Reusing Past Datasets for Future Tasks. In IEEE; 2015 [cited 2018 Mar 11]. p. 2731–9. Available from: http://ieeexplore.ieee.org/document/7410670/

39. Kotsiantis SB. Supervised Machine Learning: A Review of Classification Techniques. 2007;20.

40. Davis DA, Chawla NV, Blumm N, Christakis N, Barabasi A-L. Predicting individual disease risk based on medical history. In ACM Press; 2008 [cited 2018 Mar 11]. p. 769. Available from: http://portal.acm.org/citation.cfm?doid=1458082.1458185

41. Schrider DR, Kern AD. Supervised Machine Learning for Population Genetics: A New Paradigm. Trends Genet [Internet]. 2018 Jan [cited 2018 Mar 11]; Available from: http://linkinghub.elsevier.com/retrieve/pii/S0168952517302251

42. Katehakis DG, Tsiknakis M. Electronic Health Record. In: Wiley Encyclopedia of Biomedical Engineering [Internet]. John Wiley & Sons, Inc.; 2006 [cited 2018 Mar 11]. Available from: http://onlinelibrary.wiley.com/doi/10.1002/9780471740360.ebs1440/abstract

43. Charles A. S. Developing Universal Electronic Medical Records. Gastroenterol Hepatol. 2008 Mar;4(3):193–5.

44. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE J Biomed Health Inform. 2017;PP(99):1–1.

45. Hood L, Balling R, Auffray C. Revolutionizing medicine in the 21st century through systems approaches. Biotechnol J. 2012 Aug;7(8):992–1001.

46. Jerez JM, Molina I, García-Laencina PJ, Alba E, Ribelles N, Martín M, et al. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif Intell Med. 2010 Oct;50(2):105–15.

47. Arel I, Rose DC, Karnowski TP. Deep Machine Learning - A New Frontier in Artificial Intelligence Research [Research Frontier]. IEEE Comput Intell Mag. 2010 Nov;5(4):13–8.

48. Okuma K, Taleghani A, Freitas N, Little JJ, Lowe D. Computer Vision - ECCV 2004. 2004. 28 p.

49. Dey D, Sarkar S. PSQL: A query language for probabilistic relational data. Data Knowl Eng. 1998 Oct;28(1):107–20.

50. Ihaka R, Gentleman R. R: A Language for Data Analysis and Graphics. J Comput Graph Stat. 1996 Sep 1;5(3):299–314.

51. Scutari M. Learning Bayesian Networks with the bnlearn R Package. ArXiv09083817 Stat [Internet]. 2009 Aug 26 [cited 2018 Mar 13]; Available from: http://arxiv.org/abs/0908.3817

52. Kuhn M. Building Predictive Models in R Using the caret Package. J Stat Softw [Internet]. 2008 [cited 2018 Mar 13];028(i05). Available from: https://ideas.repec.org/a/jss/jstsof/v028i05.html

53. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer; 2016. 266 p.

54. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011 Mar 17;12:77.

55. Oladipupo T. Types of Machine Learning Algorithms. In: Zhang Y, editor. New Advances in Machine Learning [Internet]. InTech; 2010 [cited 2018 Mar 12]. Available from:

http://www.intechopen.com/books/new-advances-in-machine-learning/types-of-machine-learning-algorithms

56. Koza JR. Genetic Programming: Proceedings of the Third Annual Conference, July 22-25, 1998, University of Wisconsin, Madison. M. Kaufmann Publishers; 1998. 924 p.

57. Mohri M, Rostamizadeh A, Talwalkar A. Foundations of machine learning. Cambridge, MA: MIT Press; 2012. 414 p. (Adaptive computation and machine learning series).

58. Connors C, Vatsavai RR. Semi-supervised deep generative models for change detection in very high resolution imagery. In: 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). 2017. p. 1063–6.

59. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015 Jun;16(6):321–32.

60. Lee H-Y, Tseng B-H, Wen T-H, Tsao Y, Lee H-Y, Tseng B-H, et al. Personalizing Recurrent-Neural-Network-Based Language Model by Social Network. IEEEACM Trans Audio Speech Lang Proc. 2017 Mar;25(3):519–530.

61. Real E, Aggarwal A, Huang Y, Le QV. Regularized Evolution for Image Classifier Architecture Search. ArXiv180201548 Cs [Internet]. 2018 Feb 5 [cited 2018 Mar 11]; Available from: http://arxiv.org/abs/1802.01548

62. Bzdok D, Krzywinski M, Altman N. Machine learning: A primer. Nat Methods [Internet]. 2017 Nov [cited 2018 Mar 11]; Available from: https://hal.archives-ouvertes.fr/hal-01598285

63. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010 Oct;467(7319):1061–73.

64. Scientific modelling. In: Wikipedia [Internet]. 2018 [cited 2018 Mar 13]. Available

from:
https://en.wikipedia.org/w/index.php?title=Scientific_modelling&oldid=819852590

65. Jain AK, Murty MN, Flynn PJ. Data Clustering: A Review. ACM Comput Surv. 1999 Sep;31(3):264–323.

66. Fritzke B. Growing cell structures—A self-organizing network for unsupervised and supervised learning. Neural Netw. 1994 Jan;7(9):1441–60.

67. Ho T. Random decision forests. In: Document Analysis and Recognition, International Conference on. 1995. p. 278–82 vol.1.

68. Rokach L, Maimon O. Data Mining with Decision Trees: Theory and Applications [Internet]. 2nd ed. WORLD SCIENTIFIC; 2014 [cited 2018 Mar 12]. (Series in Machine Perception and Artificial Intelligence; vol. 81). Available from: http://www.worldscientific.com/worldscibooks/10.1142/9097

69. Liaw A, Wiener M. Classification and Regression by RandomForest. Forest. 2001 Nov 30;23.

70. Schapire RE, Freund Y, Bartlett P, Lee WS. Boosting the margin: a new explanation for the effectiveness of voting methods. Ann Stat. 1998 Oct;26(5):1651–86.

71. Breiman L. Bagging predictors. Mach Learn. 1996 Aug 1;24(2):123–40.

72. Ho TK. The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell. 1998 Aug;20(8):832–44.

73. KLEINBERG EM. STOCHASTIC DISCRIMINATION. :33.

74. Breiman L. Random Forests. Mach Learn. 2001 Oct 1;45(1):5–32.

75. Mamoshina P, Ojomoko L, Yanovich Y, Ostrovski A, Botezatu A, Prikhodko P, et al. Converging blockchain and next-generation artificial intelligence technologies to

decentralize and accelerate biomedical research and healthcare. Oncotarget. 2017 Nov 9;9(5):5665–90.

76. Ohno-Machado L, Silveira PSP, Vinterbo S. Protecting patient privacy by quantifiable control of disclosures in disseminated databases. Int J Med Inf. 2004 Aug;73(7–8):599–606.

77. Torgo L. Data Mining with R: Learning with Case Studies [Internet]. Chapman and Hall/CRC; 2010 [cited 2018 Mar 13]. (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series; vol. 20105341). Available from: https://www.taylorfrancis.com/books/9781439885994

78. Hadley Wickham. Tidy Data. J Stat Softw. 2014 Jan 1;59(1):1–23.

79. Boehmke B. Data processing with dplyr & tidyr. Retrieved RPub Com. 2014;

80. Famili A, Shen W-M, Weber R, Simoudis E. Data Preprocessing and Intelligent Data Analysis. Intell Data Anal. 1997 Jan 1;1(1):3–23.

81. Kotsiantis SB, Kanellopoulos D, Pintelas PE. Data Preprocessing for Supervised Leaning. 2006;1(1):7.

82. Dietterich TG. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Comput. 1998 Oct 1;10(7):1895–923.

83. Nadeau C, Bengio Y. Inference for the Generalization Error. :49.

84. Ribas VJ, Vellido A, Ruiz-Rodríguez JC, Rello J. Severe sepsis mortality prediction with logistic regression over latent factors. Expert Syst Appl. 2012 Feb;39(2):1937–43.