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

ESTRATIFICACIÓN DE RIESGO DE INFECCIONES BACTERIANAS INVASORAS BASADO EN ALGORITMOS DE MACHINE LEARNING PARA PACIENTES PEDIÁTRICOS INGRESADOS POR NEUTROPENIA FEBRIL

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INDUSTRIAL

CARLOS ALBERTO VEGA HERNÁNDEZ

PROFESOR GUÍA:
ROCIO BELÉN RUIZ MORENO

MIEMBROS DE LA COMISIÓN:
EDGARDO JULIO SANTIBAÑEZ VIANI
FELIPE ANDRÉS VERA CID

SANTIAGO DE CHILE
2020

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL
POR: CARLOS ALBERTO VEGA HERNÁNDEZ
FECHA: 2020
PROF. GUÍA: ROCIO BELÉN RUIZ MORENO

ESTRATIFICACIÓN DE RIESGO DE INFECCIONES BACTERIANAS INVASORAS BASADO EN ALGORITMOS DE MACHINE LEARNING PARA PACIENTES PEDIÁTRICOS INGRESADOS POR NEUTROPENIA FEBRIL

La neutropenia febril es una condición que pueden sufrir las personas con cáncer debido a que su sistema inmune esta comprometido, tanto por el cáncer como por el tratamiento del mismo. Esta condición puede evolucionar hacia complicaciones como infecciones bacteriales e incluso la muerte. Para impedir lo anterior, los médicos han entregado un tratamiento preventivo, buscando evitar las posibles complicaciones derivadas de esta condición.

Los cuidados iniciales eran la entrega de antibióticos de amplio espectro con una hospitalización temprana que permitiera monitorear la evolución de los pacientes. Esta aproximación demostró ser efectiva al reducir las complicaciones pero poco eficiente, ya que entrega el mismo tratamiento agresivo a personas que no lo necesitan, disminuyendo la calidad de vida de estos, producto de las complicaciones derivadas del tratamiento, como las infecciones intrahospitalarias, interrupciones del tratamiento del cáncer y abuso de antibióticos.

Diversas investigaciones han buscado identificar los factores que separan a los pacientes de alto riesgo de complicaciones de los de bajo riesgo. Es por ello, que se han planteando modelos de clasificación, tanto en adultos como en niños. En particular, esta memoria se enfocara en la clasificación de niños, como la elaborada por Santolaya (2001). Esta clasificación se basa en reglas que podrían no ser las más óptimas para predecir el riesgo dado lo observado en los datos. Es por lo anterior, que se planteo la hipótesis de que el desarrollo de un modelo basado en machine learning pueda superar el rendimiento del modelo utilizado actualmente.

El modelo desarrollado tuvo un AUC de 0,834 en contraste con el 0,691 de la regla actual, siendo una diferencia que implica que más niños serán clasificados de mejor manera, tratándolos de la manera más óptimo a sus necesidades. El modelo desarrollado uso variables del modelo actual pero también incorporo variables de modelos encontrados en la literatura y algunas variables nuevas descubiertas en la base de datos. También se efectuó una evaluación económica de la implementación de este modelo y se estimó que el beneficio económico será de 230 millones de pesos para el sistema público de salud en los próximos 10 años y 1.014 días ahorrados para los pacientes en el mismo periodo. Por ultimo, se creo un prototipo funcional de clasificación, el cual ilustra el flujo de información para la clasificación del paciente, siendo un punto de partida para futuras plataformas.

Dentro de las conclusiones, se tiene que se verifico la hipótesis, es decir se puede crear un modelo basado en machine learning que tenga un mejor rendimiento del actual. También se confirmo que una solución de este tipo es efectiva en términos de costo-beneficio para el estado. No obstante, que esta solución sea exitosa no pasa exclusivamente por la construcción de la herramienta sino que de la implementación de esta, si esto no sucede, el modelo planteado en esta memoria no cumpliría su objetivo de mejorar la calidad de vida de los pacientes.

Dedico este trabajo a los menos privilegiados de nuestra sociedad, quienes necesitan que estemos ahí y que no hagamos la vista gorda al ver su lucha por temer una vida mejor.

Lo dedico también al Rafa Teo, mi sobrino, que es la motivación que me hace esforzarme el doble para que la sociedad en donde él viva sea mejor.

Agradecimientos

Agradezco a mis padres porque me enseñaron que la vida dedicada al otro es una vida que es bien usada. Les agradezco por la paciencia, los valores y el espacio que me entregaron para ser lo que soy y que me hace profundamente feliz, incluso con errores y falencias. Les agradezco por estar ahí y entregarme su incondicional amor.

A mi hermana Adriana, que me ha mostrado el valor de la resiliencia, que por muy mal estén las cosas, siempre se puede sonreír a quien te rodea. A mi familia paterna y materna, quienes siempre me han tenido fe y puesto más esperanzas de las que yo mismo me he puesto.

Le agradezco a Victoria Oporto por haber estado en casi todo mi recorrido universitario, por ser un constante apoyo y ayudarme a avanzar en ser una persona más plena, feliz, sana y fuerte. Quererte es una de las mejores cosas que me ha pasado en la vida.

Le agradezco a mis amigos del colegio: Arce, Diego, Víctor, Trejo, Sofia, Pepe y Marmota, que en ustedes encontré un lugar a donde pertenecer. A mis amigos de la U donde halle un lugar donde ser yo. En especial a Diego Gajardo, quien me aguanto mis lloriqueos, escucho mis ideas locas de un mejor país y que fue alguien que permanentemente me inspiro a ser mejor. Agradecer al Aichele y Alambrito por ser mis amigos, sacarme de mi zona de confort y enseñarme a programar.

Agradecerle al WIC, en especial a Rocío, quien confió en mi en un comienzo para realizar esta memoria ahí y apoyarme y ayudarme cuando era necesario. También al Pangui, al Felipe y la Fran quienes fueron los que me hicieron sentirme cómodo en un lugar desconocido, siempre ayudando, tirando la talla o simplemente con un saludo afectuoso al llegar.

También agradecerle al deporte, que sin él no termino la universidad, las horas dedicadas a correr, nadar, andar en bicicleta y jugar a la pelota fueron el escape perfecto para la tensión que esta facultad genera.

Agradecerle a Felipe Vildoso por los ramos que dicto y la oportunidad que me dio para ser ayudante de ellos. También agradecerle a Carlos Alvarado, tutor de práctica que me mostró el poder de la tecnología en un ambiente laboral. Ambas influencias crearon el camino laboral en el que estoy ahora y que me hace feliz.

Agradezco profundamente a la vida pues me ofreció la oportunidad de desarrollarme plenamente además de permitirme estar rodeado de la gente que hizo posible terminar esta memoria. Soy un afortunado.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

1.1. Antecedentes generales

Según datos de la Organización Mundial de la Salud (OMS) el cáncer se está convirtiendo en una de las patologías más mortíferas para la humanidad, solo siendo superada por las enfermedades cardiovasculares [7]. De acuerdo a la OMS, durante el año 2015 hubieron 8,8 millones de defunciones en el mundo por este causa, es decir, una de cada seis muertes en el mundo fueron consecuencia del cáncer.

Chile no escapa a esta tendencia. La prevalencia de cáncer en Chile es de 217 casos por 100 mil adultos y que según el Ministerio de Salud entre 1990 y 2015, el cáncer paso de provocar el 21 % del total de muertes a provocar el 26 % siendo en ese periodo la segunda causa de mortalidad en Chile, superando el 19 % promedio que hay en América del Sur durante el 2015 [8]. Es por lo anterior, que es urgente hacerse cargo de este incipiente problema de salud pública.

El tratamiento del cáncer usualmente implica la inmunosupresión del sistema inmune del paciente haciéndolo susceptible a infecciones de diferente tipo, aumentando la mortalidad, morbilidad y los costos asociados al tratamiento del cáncer. En particular, los pacientes con cáncer pueden desarrollar una condición llamada *neutropenia febril*, caracterizada por un bajo nivel de *neutrófilos* (células del sistema inmune) que es acompañada por un episodio de fiebre. La neutropenia febril hace vulnerable al paciente haciendo que sea más fácil que contraiga infecciones bacteriales, virales y fúngicas, enlenteciendo el tratamiento necesario para curar el cáncer [9].

Según [10, 11] entre el 10 % y 50 % de los pacientes con cáncer desarrollará neutropenia febril durante el tratamiento de su enfermedad, aumentando la cantidad y duración de las hospitalizaciones y retrasando los ciclos de quimioterapia. Incluso, el 10 % de los pacientes que presentan neutropenia febril al ingreso de urgencias mueren por complicaciones relacionadas a esta condición. Cabe destacar que la mortalidad aumenta entre un 23 % y 30 % en los casos de neutropenia febril con evidencia de bacterias en la sangre (*bacteriemia*), haciendo que la identificación de infecciones bacteriales sea vital para un buen tratamiento de esta condición.

Dado que la neutropenia febril es una emergencia clínica para los pacientes con cáncer, la solución inicial para tratarla fue un aproximamiento agresivo con antibióticos de amplio espectro con el objetivo de evitar las complicaciones derivadas de una infección bacteriana severa [12], que, como se ha dicho aumenta la probabilidad de muerte del paciente. Dicho enfoque ha resultado no ser suficiente para evitar las complicaciones e incluso, se evidencio que podría ser innecesario para pacientes con bajo nivel de riesgo de complicaciones [13].

Esto demostró ser innecesario porque hay casos en los cuales el paciente tienen una salud estable y aún así se prosigue con el tratamiento intra-hospitalario estándar, lo cual podría generar alguna de las siguientes complicaciones colaterales de la hospitalización [14]:

1. **Toxicidad por antibióticos.** El abuso de antibióticos podría hacer que las cantidades sugeridas para tratar la condición sean superiores a los que el organismo puede manejar, haciendo que el cuerpo las rechace [15].
2. **Resistencia antimicrobial** Suministrar antibióticos de manera indiscriminada podría hacer que las bacterias generen resistencia a estos, haciendo que el tratamiento antimicrobiano sea inefectivo y se prolongue en el tiempo [16].
3. **Infecciones intrahospitalarias (nosocomiales).** Una complicación del internamiento hospitalario es que el paciente se expone a las bacterias propias de un hospital, poniendo en riesgo al paciente al exponerlo a una infección que no tenía cuando ingreso al hospital [17].
4. **Infecciones fúngicas.** Los tratamientos intra-hospitalarios al utilizar de manera intensiva antibióticos, eliminan las bacterias que compiten con los hongos por el "alimento" disponible en el cuerpo humano, haciéndolo vulnerable a la aparición de infecciones fúngicas [18].
5. **Impacto psicológico.** Otra complicación, usualmente sub-estimada, es el impacto psicológico en los pacientes hospitalizados, que pueden desarrollar estados emocionales de tristeza, ansiedad y desencadenar efectos negativos como la baja expresión de sus emociones [19].

Es por lo anterior que numerosas investigaciones han buscado complementar el tratamiento antimicrobiano, creando una clasificación de riesgo de complicaciones en adultos con el fin de separar entre los pacientes sin riesgo de complicaciones de aquellos con riesgo evidente de complicaciones. Talcott fue uno de los primeros en desarrollar tales reglas, le siguieron los puntajes de riesgo MASCC y CISNE que complemento el puntaje MASCC, al enfocarse en pacientes con una aparente estabilidad en su condición médica.

La investigación de la estratificación de riesgo de complicaciones de adultos ha sido prolífica pero los resultados encontrados no pueden ser extrapolados a los niños dado que, la biología tanto del niño como de los cánceres que afectan a los niños son diferentes a la de los adultos. Entendiendo esto, médicos y investigadores del mundo se han dedicado a crear puntajes asociados a las características de los niños y de sus canceres.

El cáncer infantil en Chile es una enfermedad de baja frecuencia pero de alta carga emocional. Se estima que la incidencia del cáncer infantil es entre 113 a 136 casos por cada 1.000.000 de niños menores de 15 años, es decir se esperan entre 500 a 600 casos al año. De esto casos se estima que el 80 % será atendido en el sistema de salud público a través del

Programa Nacional Infantil de Drogas Antineoplásicas[1] (PINDA) [20]. Los tipos de cáncer que más afectan a los niños son las *leucemias, tumores del sistema nervioso central y los linfomas y tumores reticuloendoteliales*

Además, se espera que un niño en tratamiento con quimioterapia sea ingresado por neutropenia febril entre 8 a 10 veces en el transcurso del tratamiento. Asimismo, la sobre vida al 5to año de diagnosticado el cáncer es mayor al 78 %, es por esto que es relevante entregar un tratamiento racional a las pacientes infantiles evitando hospitalizaciones innecesarias, que pudieran afectar desarrollo psico-emocional normal del niño, entendiendo que lo normal en un niño con cáncer no es lo mismo que para uno sin cáncer.

Al igual que en el caso de los adultos, la investigación en la estratificación de riesgo en niños ha sido vasta pero dada las características particulares de cada uno de los estudios, es difícil extrapolar los resultados. Es por lo anterior que un grupo de investigadores chilenos ya ha realizado estudios para identificar los factores de riesgo en los casos chilenos y elaborar sus propias reglas y aplicarlas en la realidad nacional. Es así como hoy en día para definir el riesgo de un paciente pediátrico, se utiliza factores de riesgo calculados por investigadores chilenos.

Los factores de riesgo utilizados para clasificar a los pacientes son los siguientes:

1. Hipotensión
2. Proteína C Reactiva $\geq 90 mg por litro$
3. Leucemia en recaída
4. Leucemia no linfoblastica
5. Días desde la ultima quimioterapia ≤ 7
6. Linfoma no Hodgkin
7. Neuroblastoma etapa IV
8. Plaquetas $\leq$ 50,000

Para que un niño sea estratificado como de alto riesgo tiene que tener 2 o más factores de riesgo o tener uno de los siguientes factores por si solos: hipotensión, leucemia en recaída, PCR menor a 90mg por litro, leucemia no linfoblastica, linfoma no hodgkin o neuroblastoma etapa IV. Para ser catalogado como de bajo riesgo tiene que tener solo uno de los siguientes factores de riesgo plaquetas menores a 50.000 o 7 o menos días desde la última quimioterapia.

La estratificación de riesgo tanto en adultos y niños ha resultado de utilidad haciendo que la cantidad de personas que reciben un tratamiento agresivo, haya sido reducida pero aún así hay casos catalogados de alto riesgo cuando no debiesen haber recibo un tratamiento agresivo.

Todos las estratificaciones de riesgo anteriores se basan en algún tipo de regresión logística y usualmente univariadamente, por lo que cabe preguntarse si el cálculo de estos factores a través de técnicas más modernas como los algoritmos de clasificación de machine learning permitan identificar el riesgo de complicaciones en casos de neutropenia febril de mejor manera,

[1]Las drogas antineoplásicas son el conjunto de medicamentos que evitan el crecimiento, desarrollo y proliferación de células tumorales malignas.

es decir, prediciendo los casos realmente riesgosos (sensibilidad, poder predictivo positivo) y ser capaz de identificar a los casos con bajo riesgo (especificidad, poder predictivo negativo)

1.2. Justificación y descripción del proyecto

Que existan diversas investigaciones buscando mejorar la estratificación de riesgo de complicaciones dan cuenta de la importancia de entregar un mejor tratamiento para los pacientes, debido a que se sabe que un tratamiento más racional de antibióticos repercute directamente en la calidad de vida de los pacientes de cáncer, en particular en los de los niños.

Es de particular importancia el caso de los niños por la carga emocional y social que tiene el tratamiento del cáncer para la familia y el impacto que tiene en ella posibles internaciones de larga estadía en hospitales. Esta aparte de ser emocional y social también incide directamente en la calidad de vida de los cuidadores del niño ya que impacta en la capacidad de tener un trabajo y vida más estable. Es por lo anterior que una mejor clasificación permitirá a los cuidadores del niño, evitar la preocupación de tener al niño en el hospital.

Además, si existe una sobre estratificación de riesgo significa un gasto extra para el sistema de salud que podría ser redestinado a otras necesidades que pudiera tener el sistema en ese momento. También se tiene que considerar que la implementación de un módulo de visualización de datos basado en machine learning tiene un costo menor y es de más rápida implementación que las formas alternativas para mejorar el tratamiento. El detalle de estas soluciones alternativas se muestra en el siguiente apartado.

1.2.1. Análisis de las alternativas de solución actual

El tratamiento actual de la neutropenia febril ha sido complementada con otras investigaciones que no tienen como foco la estratificación de riesgo sino que mejorar el tratamiento y prevención de esta. Algunas de estas lineas investigativas son planteadas por [21] y se resumen a continuación, primera con las lineas preventivas y a continuación del tratamiento en sí.

1. **Prevención**
   (a) **Identificación de los factores de riesgo de desarrollar una infección en pacientes con cáncer.** Estudios encontraron que hay factores genéticos que influyen en la duración de los episodios, que la intensidad de la quimioterapia también afecta la cantidad de pacientes infectados y un ultimo estudio encontró que un tratamiento profiláctico[2] antifúngico aumentaba ligeramente la probabilidad de sufrir una infección.
   (b) **Mejoramiento de la prognosis[3] de la neutropenia febril.** En los recientes años nuevas pruebas clínicas han permitido predecir de mejor manera los casos de

[2]Medidas de cuidado para prevenir infecciones posibles
[3]Conocimiento anticipado del algún suceso

bajo riesgo pero aún no se pueden identificar los factores de alto riesgo.

2. **Tratamiento:**
   (a) **Identificación de la etiología[4] de los episodios.** Cada uno de los episodios de neutropenia febril tiene un patrón diferente de los patógenos que la acompañan, por lo que saber este patrón ayudaría a entregar un mejor tratamiento.
   (b) **Nuevas opciones de tratamientos iniciales antibacteriales.** El enfoque empírico de antibióticos de amplio espectro de tratamiento para neutropenia febril ha logrado reducir en gran medida las complicaciones derivadas de esta. Aún así, la combinación de antibióticos aún es controversial por lo que el diseño de un nuevo coctel de antibióticos de amplio espectro podría mejorar más el tratamiento de esta condición. No obstante, hay autores que plantean que ampliar la cobertura antibacterial aún más tiene un valor limitado para tratar esta condición [22].
   (c) **Terapia Oral.** Es una de las alternativas con mayor interés últimamente ya que se ha observado que un tratamiento oral ambulatorio puede ser igual de efectivo que un tratamiento intravenoso e intrahospitalario.
   (d) **Necesidad de la aplicación de medicamentos más específicos para casos de fiebre persistente.** Una práctica común para el tratamiento de fiebre persistente es el ajuste terapéutico de las medicaciones para los pacientes, a pesar de que no hay pruebas de su efectividad. Lo anterior podría generar una resistencia antimicrobial, lo cual podría tener implicaciones clínicas mayores en el futuro.

Como se puede observar, existen formas alternativas para mejorar el tratamiento y resultado clínico de un episodio de neutropenia febril. Desde una mejor identificación de la causa de la enfermedad pasando por medicamentos más potentes con diferentes vías de administración hasta poder predecir de mejor manera el desarrollo del paciente. En general estás alternativas requieren una inversión alta y años de investigación, siendo soluciones en el largo plazo. No obstante, investigaciones actuales muestran que una mejor clasificación de los pacientes riesgosos puede ser una solución en el corto plazo para los pacientes ingresados por neutropenia febril [23]

1.2.2. Uso de antibióticos

El tratamiento usual de la neutropenia febril utiliza antibióticos de amplio espectro para evitar complicaciones derivadas de esta. Este tratamiento mejoró la supervivencia de los pacientes ingresados por esta condición, ya que una gran proporción de las complicaciones son derivadas de infecciones bacterianas [13]. Sin embargo, dicho tratamiento tiene como efecto colateral un posible uso indiscriminado de antibióticos, siendo muchas veces innecesarios.

La Organización Mundial de la Salud declaró durante el 2018 que *"la resistencia a los antibióticos es hoy una de las mayores amenazas para la salud mundial, la seguridad alimentaria y el desarrollo"* ya que la resistencia a los antibióticos puede afectar a cualquier persona, puesto que no depende del paciente en sí, sino que de las bacterias. Otro factor importante es que la resistencia a los antibióticos prolongan las estancias hospitalarias, los

[4]Causa de la enfermedad

costos médicos y aumenta la mortalidad general. La OMS recomienda a los profesionales de la salud *"prescribir y dispensar antibióticos solo cuando sean necesarios, de conformidad con las directrices en vigor"*, por lo que mejorar estas directrices podría evitar las consecuencias del uso indiscriminado de antibióticos [24].

1.2.3. Características de la solución que se necesita

Cualquier solución para el problema de la neutropenia febril tiene que tener las siguientes 3 características:

1. **Rapidez:** Se ha estudiado que la aplicación de antibióticos durante las primeras horas de internación en caso de que la etiología sea bacteriana mejorara de manera importante la supervivencia de los pacientes.
2. **Costo-Beneficio:** Dado que el 80 % de los pacientes pediátricos ingresados por neutropenia febril son atendidos por el sistema público, el uso racional de los recursos y de la manera más costo efectiva posible es una de las prioridades para cualquier solución.
3. **Precisión:** Se espera que cualquier solución sea lo menos arbitraria posible y que sea basada en reglas y hechos, por lo que el registro correcto de los datos es vital.

La solución actual no es rápida ni lo suficientemente precisa ya que la regla existente la tiene que computar una enfermera, muchas veces atareada y cansada haciendo propenso un error que implique una mal asignación del riesgo. A su vez, la regla actual al sobre-estimar el riesgo utiliza más recursos de los que son necesarios haciendo que sea una solución no costo efectiva.

Una posible solución es una plataforma informática en donde se puedan ingresar datos, procesarlos y entregar una clasificación de riesgo puesto que cumple los 3 criterios anteriormente mencionados. Primero, es rápida ya que la medición del riesgo se podría hacer apenas se ingresen los datos sin la necesidad de un cálculo manual por parte del usuario de la plataforma.

La segunda ventaja es que es costo-eficiente ya que una mejor estratificación de riesgo permitiría reducir los costos del tratamiento de neutropenia febril sin que la inversión sea alta. Por ejemplo, la plataforma podría funcionar desde cualquier teléfono inteligente que se conecte a un servidor dedicado para la aplicación, elementos que tiene un costo muy inferior al desarrollo de un nuevo antibiótico o nuevos exámenes médicos para tener una prognosis más adecuada.

Por ultimo, una plataforma informática podría reducir los posibles errores del personal del hospital al calcular el riesgo de cada paciente. Esto se ve al comparar la estratificación de neutropenia febril de alto riesgo y bajo riesgo (NFAR y NFBR respectivamente) declarada por el PINDA en la Base 2009-2011 y por la calculada en esta memoria basandosé en los valores de cada paciente y las reglas de clasificación de riesgo utilizadas. En la tabla 1.1 se muestra la comparación de lo declarado por el PINDA frente a lo calculado.

% Estratificación	Declarado	Calculado
NFAR	455	433
NFBR	166	188

Tabla 1.1: Comparación estratificación de riesgo calculada versus declarada, basada en las tablas entregadas por el PINDA

Hay una diferencia de 3,54 % (22 casos) entre lo declarado y calculado que puede deberse a 2 motivos. El primero puede ser que al hacer la evaluación al ingreso, el médico decide que el paciente tiene un riesgo mayor al que entregan las reglas definidas, en tal caso, las reglas tiene que ser redefinidas, en particular en los casos limites. Por ejemplo, para la regla de los 7 días desde la ultima quimioterapia, esta fija un umbral único, sin considerar que podría darse el caso de que el estado de salud de un niño que tiene 6 días desde su ultima quimioterapia y otro de 8 días sea el mismo, haciendo que su estratificación sea diferente.

La otra razón es que al calcular la estratificación de riesgo de manera manual se podrían generar errores. Un ejemplo es que la regla de PCR ≥ 90 sea entendida al revés, error que se observó múltiples veces en la base de datos. En tal caso la estratificación estaría mal hecha, lo que podría repercutir en el tratamiento del niño. Este error se podría resolver con la plataforma informática que evita que un humano haga el cálculo, evitando este posible error.

No obstante las ventajas de la implementación de esta solución, se podría dar el caso de que el PINDA no se adapte bien a la incorporación de una nueva herramienta tecnológica. Es por esto que la implementación de esta herramienta tiene que ser acompañada por capacitaciones que permitan al personal acostumbrarse a la herramienta en cuestión.

1.2.4. Beneficios de la implementación de un nuevo modelo

Si se comprueba que la hipótesis de esta memoria es correcta, la implementación de este modelo podría entregar varios beneficios a los pacientes. Entre los posibles beneficios estaría la reducción de la cantidad de antibióticos entregados disminuyendo la probabilidad de generar resistencia, toxicidad y destrucción de bacterias inofensivas. Otro beneficio posible es la reducción del tiempo de hospitalización, ya que una hospitalización extensa aumenta el riesgo para el paciente de contraer alguna infección dentro del hospital. Estos beneficios repercuten directamente en la calidad de vida del paciente evitando el deterioro de su salud y ayudándolo a tener un desarrollo psico-emocional más normal.

Una clasificación de alto riesgo implica un tratamiento intrahospitalario largo, en cambio una clasificación de bajo riesgo implica una hospitalización más corta pudiendo culminar el tratamiento de manera oral. de acuerdo a Santolaya (2004) en le caso de alto riesgo el promedio fue de 5,3 días de cama y en el de bajo riesgo 3,8 días de cama.

En el año 2014 Orme et al. [25] realizó un estudio comparativo de la calidad de vida de los pacientes y sus padres, entre los tratados dentro y fuera del hospital y encontró que la calidad de vida de los padres de los pacientes extra hospitalarios era mejor en varios dominios como

en la mantención y cuidado del hogar, tiempo con la pareja y el tiempo con otros hijos y que no había diferencia significativa entre la satisfacción con el tratamiento entre los grupos. En el caso de los pacientes, el tiempo de sueño y apetito promedio era mayor para los pacientes extra hospitalarios sin mostrar eventos adversos. Lo anterior sugiere que un tratamiento extra hospitalario de neutropenia febril puede ser una opción segura y factible para los pacientes mejorando su calidad de vida.

En cuanto a los beneficios económicos de una mejor estratificación de riesgo se tiene la reducción de los costos de cada uno de los episodios, reduciendo la carga financiera para el sistema de salud. Por ejemplo, en Estados Unidos se estimó que el costo promedio de un episodio de neutropenia febril entre el 1995 y 2000 era de 19.110 dolares (8.376 mediana)[26]. Otro estudio estimó que el gasto durante el 2012 en hospitalizaciones relacionadas con neutropenia febril en Estados Unidos fue de 2,3 billones de dolares en adultos y 479 millones de dolares en niños [27]. Cifras importantes que indican que una mejor eficiencia en la estratificación de riesgo podría reducir el gasto en salud de manera significativa.

En 2004 Santolaya et al. [28] comparó el costo y resultado entre los episodios estratificados como de bajo riesgo con los de alto riesgo. Se encontró que los episodio de bajo riesgo significan 638 dolares promedio, en cambio, los de alto riesgo significan 903 dolares promedio, es decir un costo 41,53 % superior y que la efectividad en el tratamiento era similar. Otro estudio comparativo del 2010 realizado en Canadá encontró que un tratamiento ambulatorio era casi 5,3 veces menos costoso que un tratamiento intrahospitalario y con efectividad similar (0,663 extra hospitalario y 0,649 intrahospitalario, medido en costo efectividad)[29]. Lo anterior muestra que un tratamiento intrahospitalario es significativamente más caro y que no se puede justificar en base a la seguridad y eficiencia del tratamiento intrahospitalario.

1.2.5. Hipótesis de Investigación

Dado los antecedentes generales anteriormente mencionados, se plantea la siguiente hipótesis de investigación:

- La construcción de un modelo de clasificación, que utilice datos al ingreso de los pacientes, basado en algoritmos de machine learning, aumentará la especificidad y sensibilidad en la predicción de infecciones bacterianas invasoras en episodios de neutropenia febril, que permitirá mejorar el tratamiento entregado a los pacientes reduciendo los costos directos e indirectos para la sociedad.

El objetivo de validar esta hipótesis es contribuir a los futuros trabajos que apunten a la creación de nuevos modelos de estratificación de riesgo de complicaciones en casos de neutropenia febril basados en machine learning. Se espera que estos modelos sean creados a partir de datos obtenidos de manera más rigurosa y considerando diferentes centros médicos a lo largo del país. Incluso podría ser un precedente a nivel mundial para seguir investigando esta materia.

En caso de que la hipótesis se compruebe, se propone la creación de un módulo de estratificación final, validado a nivel de usuario e institución, realizando las capacitaciones

correspondientes de la plataforma. La implementación de este módulo es relevante porque ayudaría al equipo médico en la recomendación de tratamiento al paciente, haciendo que el tratamiento sea más racional, reduciendo los costos para el sistema de salud y mejorando la calidad de vida del paciente.

1.3. Objetivos

1.3.1. Objetivos Generales

"Desarrollar un modelo de estratificación de riesgo de IBI para la optimización del tratamiento de neutropenia febril, basado en técnicas de ML"

1.3.2. Objetivos Específicos

1. Determinar las puntuaciones de riesgo de IBI en pacientes pediátricos ingresados por neutropenia febril usadas en el mundo y que pueden ser aplicadas en la base de datos del PINDA, a partir de un análisis del estado del arte en esta materia.
2. Construir una base de datos consolidada que permita la aplicación de los puntajes determinados en el objetivo anterior y los algoritmos de machine learning a usar.
3. Obtener el desempeño de los puntajes de riesgo determinados, a través de su aplicación en la base de datos creada anteriormente.
4. Calcular el desempeño de los modelos a través de la aplicación de estos en la base de datos creadas
5. Elegir el mejor modelo posible, considerando el desempeño obtenido para cada uno.
6. Crear una maqueta de plataforma informática que permita ingresar los datos del paciente y que esta le entregue un resultado dado el mejor modelo encontrado.
7. Estimar el beneficio económico y social de la implementación de esta solución.

1.4. Apoyos institucionales

Esta memoria es parte de una investigación colaborativa entre el *Web Intelligence Center* (WIC) y el Programa Infantil Nacional de Drogas Antineoplásicas (PINDA). Dicha investigación es financiada gracias a un FONDEF ganado el 2018 por esta alianza investigativa que tiene como objetivo la creación de una plataforma informática de estratificación de riesgo en el año 1 y en el año 2 la validación de esta plataforma informática.

El *Programa Infantil Nacional de Drogas Antineoplásicas* (PINDA) nace en 1988 a partir de un esfuerzo del estado para hacerse cargo tanto económicamente como administrativa de los canceres infantiles en Chile. Este programa garantiza la atención gratuita al 100 % de los pacientes beneficiarios del sistema de salud pública. Asimismo, se recibirá el apoyo constante

de la Dr. Maria Elena Santolaya, investigadora chilena con múltiples publicaciones sobre la materia.

El *Web Intelligence Center* (WIC) es una organización dedicada a la investigación de punta enfocada en el área de *Data Science*, con vasta experiencia en diferentes soluciones y proyectos tecnológicos. El centro entregará permanente guía y apoyo en el desarrollo esta memoria.

1.5. Metodología

Para probar la hipótesis de investigación se propone la siguiente metodología inspirado en el marco conceptual del Descubrimiento de Conocimiento en Base de Datos (KDD por sus siglas en inglés). La metodología propuesta esta compuesta por 7 puntos, mostrados a continuación.

1.5.1. Investigación de puntuaciones de riesgo en casos de neutropenia febril

A nivel mundial se han desarrollado múltiples estratificaciones de riesgo para neutropenia febril tanto a nivel pediátrico como adulto. De estás estratificaciones se seleccionaran algunas y se usaran como punto de referencia para el trabajo que se realizará en esta memoria. Teniendo en cuenta esto, se investigará la literatura existente para identificar y seleccionar las estratificaciones de riesgo de neutropenia febril en niños con más potencial. La selección se basará en el rendimiento del modelo investigado y la cantidad de variables que comparten con los bases de datos entregadas por el PINDA.

El producto final de esta metodología serán entre 4 y 6 puntuaciones de riesgo con más potencial que se utilizaran y aplicaran más adelante en la metodología para crear la línea base para la comparación de los modelos.

1.5.2. Implementación y evaluación de los modelos hoy existentes

En este paso de la metodología se aplicaran los modelos de estratificación de riesgo encontrados en el paso anterior y se le aplicaran a la base de datos entregada por el PINDA. El resultado de esta aplicación servirá como una línea base del modelo por crear.

Cabe destacar que los modelos seleccionados para crear la línea base son caso dependiente, es decir los resultados varían de acuerdo al conjunto de datos que se eligen, por lo que la comparación es meramente referencial y busca fijar una norma de rendimiento. Por consiguiente será imposible concluir si el modelo es válido o no a partir de esta comparación. Para poder validar el modelo se recomienda hacer un estudio en dos partes, la primera sería el modelo

creado en esta memoria y en la segunda la aplicación del modelo a un nuevo conjunto de datos.

El producto de este paso serán las métricas de rendimiento como sensibilidad, especificidad, poder predictivo positivo, poder predictivo negativo, precisión y área bajo la curva (AUC) para cada uno de los modelos investigados.

1.5.3. Procesamiento y creación de la base final de datos

Un paso fundamental para realizar un buen trabajo es asegurarse que la calidad de los datos que recibirán los modelos como *input* sea la mejor posible, por consiguiente, se llevara a acabo la integración de los datos entregados por el PINDA a través de 2 planillas de calculo en una sola y el pre-procesamiento de las bases de datos que incluirá el formateo adecuado de los datos, el tratamiento de los datos faltantes (eliminado variables o imputando valores usando KNN, promedio o moda dependiendo del caso), posibles errores de escritura de los datos entre otros.

También se ilustrara el comportamiento de los datos en gráficos para poder identificar situaciones anómalas como outliers, variables mal registradas o el desbalanceo de la base de datos. Por ultimo, con los procedimientos anteriores ya realizados se utilizaran diferentes filtros (Chi-cuadrado, Kolgomorov-Smirnoff, Varianza, Correlación) para seleccionar las variables relevantes del problema.

Además de lo anterior, se tomarán como insumo los resultados de la sección anterior para crear diferentes conjuntos de variables, que a priori serán las variables utilizada en cada uno de los modelos por separado, todas las variables utilizadas en los modelos juntas y una selección de variables usando métodos para esta tarea, a los cuales se le aplicaran los diferentes modelos de clasificación.

El producto final de este paso de la metodología será una base de datos limpia, con las variables que se utilizarán y bien formateadas, de forma que la aplicación de los modelos sea lo más sencilla posible.

1.5.4. Investigación y selección de los modelos a aplicar

Existen múltiples algoritmos de clasificación que permitirán estratificar los casos de neutropenia febril. El caso ideal sería la aplicación de todos pero hay ciertas restricciones de tiempo e interpretabilidad de los modelos por lo que no es posible la aplicación de todos.

Dentro de las restricciones esta que los modelos que son cajas negras, es decir que sabemos el output, pero no sabemos como llego a esos resultados, son de baja interpretabilidad, no obstante, en la práctica se ve que estos modelos tienen mejor rendimiento que los otros modelos. Aún así, la aplicación de estos modelos en salud es difícil ya que las personas a cargo de tomar la decisión final necesitan saber las razones de la estratificación.

Independiente de las restricción de interpretabilidad del modelo, se aplicaran algunos modelos de caja negra para fines académicos y de comparación. El producto de este apartado es una lista de 4 o 5 modelos de machine learning que serán aplicados en la siguiente sección.

1.5.5. Creación del modelo final de machine learning

Esta parte de la metodología es la principal de esta memoria ya que aquí es construirán los modelos basados en la base de datos creada anteriormente. Con estos modelos se calculará el rendimiento de cada uno a través de diferentes métricas y con estás se compararan entre ellos y con la línea base.

Además de la aplicación de los modelos se utilizará la técnica de validación cruzada (*cross validation* en inglés) con el fin de evitar un sobre-ajuste o sub-ajuste a la base de datos debido a una selección particular de los datos.

El producto de este apartado es una tabla con los rendimientos de cada uno de los modelos, con sus correspondientes métricas y sus las variables relevantes de cada uno de los modelos.

1.5.6. Evaluación de la interpretabilidad y rendimiento de los los modelos

Ya con los modelos funcionando se comparará su rendimiento con la línea base calculada previamente y se concluirá acerca del rendimiento de cada uno de los modelos.

Además se presentarán los modelos al equipo médico y las variables que predicen el riesgo y si le hacen sentido, si el modelo tiene una interpretabilidad aceptable que permita al equipo médico pueda confiar. El juicio del equipo médico será agregado a la tabla del rendimiento de cada modelo y se ponderada cada una de las variables para concluir.

1.5.7. Prototipo del módulo de visualización de riesgo

Luego de concluir acerca de cual es el modelo con mejor rendimiento e interpretabilidad, se procederá a crear una plataforma informática que recibirá las variables de ingreso de los pacientes y calculará la estratificación de riesgo en base al algoritmo y entregara el resultado de manera visual. Cabe señalar que independiente del resultado de la hipótesis de investigación, el modelo de visualización de riesgo se implementará igual, ya que una de las justificaciones de la implementación de este eran los posibles errores al calcular la regla dentro del equipo médico, errores que la plataforma podría evitar.

1.5.8. Evaluación del impacto económico y social de la solución

En esta sección se pretende estimar primero el beneficio en términos de recursos materiales como camas, medicamentos y otros, y recursos humanos como enfermeras y doctores. Lo anterior es conocido como costo-beneficio, ya que estas métricas pueden ser traducidas en dinero, lo cual hace fácil su comparación con otras soluciones. También se buscará evaluar el beneficio de esta solución en términos del costo emocional y social para la familia y cuidadores y además de las condiciones de salud que se me mejoraran gracias a esta solución (vidas salvadas, mejor tratamiento por ejemplo). Esta evaluación es denominada costo-efectividad ya que mide los efectos del mejor tratamiento en la salud de los pacientes. Es una métrica más compleja de comparar entre diferentes soluciones[30].

1.6. Alcances

Esta memoria tiene como alcance y foco la validación de la hipótesis propuesta. En el caso de que se valide la hipótesis, se espera obtener un modelo de estratificación bien justificado y relevante al problema en cuestión. Además del modelo de estratificación de riesgo, se implementará un módulo de visualización del resultado que ayude al equipo médico a entender las razones de la estratificación.

También se realizará una estimación general de los beneficios económicos y sociales de la implementación de esta solución en la red PINDA pero no se calculará los costos de la implementación del modelo, es decir los costos en hardware, software y capacitaciones para el uso de la plataforma. Esta estimación será basando fuertemente en supuestos que pueden o no cumplirse debido a que responden a proyecciones que consideran el mismo escenario actual, es decir, que no incluyen eventualidades.

Una limitación de la memoria es que los datos son solo de la Región Metropolitana por lo que los resultados no son extrapolables a los demás centros de salud del país. Además, estas bases de datos consideran solo los datos de la ficha clínica del paciente al ingreso, las cuales tienen una basta cantidad de variables pero que no contemplan toda la realidad del paciente. Por ejemplo, no existen los datos previos al ingreso del paciente por neutropenia febril, datos que podrían ayudar a mejorar la estratificación de riesgo de complicaciones.

Otra alcance es que esta herramienta no pretende ser utilizada como reemplazo al diagnostico del equipo médico sino que como apoyo a este, es decir, el equipo médico siempre tendrá la decisión final en el tratamiento asignado a cada paciente. También su buscará validar el modelo a través de la opinión y retroalimentación que entreguen los médicos. De igual manera, tampoco se pretende que este modelo reemplace a los indicadores de riesgo actuales ya que estos fueron validados a través de ensayos clínicos de manera metódica y rigurosa. De todas maneras, se deja propuesto la validación de este modelo en ensayos clínicos futuros.

El módulo de visualización tiene como fin ser un prototipo que sirva de prueba para el ingreso de nuevos datos y la entrega de la estratificación de riesgo. No pretende ser un producto final sino que ser un aporte para la futura implementación de un módulo final de

visualización a nivel país, entregando un piso mínimo de funcionamiento y dando lineamientos de lo que sirve y de lo que no. Todo lo anterior a nivel de maqueta y de manera local.

Otro alcance relevante de este trabajo es la muestra que se utilizara y que es proveniente de la red PINDA. Es por lo anterior que el modelo planteado podría no ser extrapolable al resto del país. Independiente de lo anterior, aun entregaría una base importante para la creación de un modelo más global y que incorporase los casos de neutropenia febril en todo el país.

Por ultimo, este trabajo no pretende entregar soluciones a posibles problemas organizacionales que tenga el proceso de atención de neutropenia febril, reiterando que solo se entregará la estratificación de riesgo de complicaciones y que la aplicación del tratamiento seguirá dependiendo del equipo médico. Tampoco se hará cargo de la posible mal implementación del modelo a falta de capacitaciones o de resistencia al cambio dentro de la red PINDA.

1.7. Resultados esperados

El resultado esperado es la creación de un prototipo de módulo de visualización de la estratificación de riesgo de complicaciones en pacientes con neutropenia febril basado en un modelo de predicción de riesgo seleccionado por el memorista. Este modelo pretende incorporar más variables de las que hoy se considera para la estratificación de riesgo. Junto con el modelo se espera entregar la justificación y relevancia de la implementación de un modelo así.

En particular, se esperan los siguientes resultados esperados:

1. Estado del arte en la estratificación de riesgo de NF, que permita determinar los posibles factores de riesgo relevantes además de mostrar que la aplicación de un modelo basado en ML sería el siguiente paso lógico para una mejor estratificación de riesgo de complicaciones.
2. Diseño y construcción de los modelos de estratificación de riesgo de complicaciones basados en algoritmos de machine learning utilizando la base de datos entregada por el PINDA. En particular, se pretende la escritura del código en *Python* que permita realizar la estratificación de riesgo además de su correspondiente documentación y justificación.
3. Comparación y validación de los modelos creados a partir de la investigación del estado del arte con la línea base. Se compararan los indicadores y métricas de desempeño relevantes para cada uno de los modelos para luego ordenarlos según desempeño e interpretabilidad del modelo. Con esta comparación se pretende validar y justificar la implementación de alguno de los modelos de machine learning si se encontrase que tiene mejores indicadores que el modelo base, justificando porque son mejores indicadores.
4. Módulo de visualización de estratificación de riesgo. Se pretende crear una herramienta prototipo que permita observar el funcionamiento del modelo en una etapa inicial, es decir, que el módulo de visualización reciba datos de *input*, el modelo calcule el riesgo y el módulo de visualización muestre como *output* la estratificación y las razones de la

estratificación.

5. Cuantificación de los beneficios económicos y sociales de la implementación de esta solución a nivel de la red PINDA en la Región Metropolitana. Se espera obtener una estimación de los beneficios directos e indirectos de esta, justificando el proceso de estimación de estos beneficios.

1.8. Estructura del informe

El presente informe se estructura en 6 capítulos diferentes que muestran el desarrollo lógico de la memora.

1. El primer capítulo introduce la memoria, describiendo el proyecto a realizar fijando objetivos, la problemática a resolver, la hipótesis de investigación, el alcance y beneficios y resultados esperados de la memoria.
2. El segundo capítulo es marco conceptual de la memoria, detallando los conceptos claves (principalmente los conceptos médicos y del proceso KDD) que serán necesarios para resolver la memoria.
3. El tercer capítulo describe todo el trabajo realizado para desarrollar el modelo de predicción utilizando la metodología KDD, desde la selección hasta el descubrimiento de los patrones.
4. El cuarto capítulo describe el trabajo realizado para crear el módulo de visualización basado en el *framework* MVC (Modelo Vista Controlador) y la implementación del modelo de predicción en este.
5. El quinto capítulo estima los beneficios económicos y sociales de la implementación de este modelo en la red PINDA de la Región Metropolitana.
6. El sexto capítulo resume todo el trabajo realizado y muestra las principales conclusiones del trabajo, asimismo, propone líneas de trabajo para futuras investigaciones.

Capítulo 2

Marco Teórico

2.1. Herramientas tecnológicas y datos utilizados

2.1.1. Datos utilizados

Los datos utilizados son los episodios recolectados por un equipo de investigadores del PINDA, quienes se adjudicaron 2 concursos FONDECYT, uno el 2009 con código 1090194 [31] y otro el 2015 con código 1120800 [23]. En ambos FONDECYT se registraron variables demográficas y de salud de los niños que ingresaron entre el 2009-2016 por neutropenia febril a la red PINDA de la región metropolitana. La cantidad de registros entre ambas bases llega a más de 1620 casos.

2.1.2. Python

Python es un lenguaje de programación con semántica dinámica. Gracias a sus características es una atractiva herramienta para el desarrollo de aplicaciones de manera rápida como también lo es porque permite conectar las diferentes componentes de un sistema.

Es un lenguaje con una sintaxis fácil de aprender y que se enfatiza en la legibilidad del código, lo que facilita y reduce el costo del mantenimiento de este. Además de lo anterior Python busca ser un código que soporte el uso de módulos y paquetes y que impulsa la reutilizabilidad del código. Por ultimo, es un lenguaje que tiene disponible su código fuente, lo que lo hace ser un código abierto [32].

2.1.3. Django

Django es un framework web de alto nivel basado en Python que incentiva el desarrollo rápido y limpio, con un diseño pragmático. Fue construido por desarrolladores con experiencia

y que se hace cargo de los problemas del desarrollo web y hace que el foco del desarrollo sea escribir la aplicación sin la necesidad de crear códigos que ya habían sido escritos. [33]. Algunas de las características de Django son:

1. **Velocidad.** Esta diseñado para que los desarrolladores pueden pasar del concepto hasta la culminación de este lo más rápido posible
2. **Seguridad.** Django se preocupa bastante de la seguridad y ayuda a los desarrolladores a evitar los errores comunes de seguridad.
3. **Escalable.** Tiene la habilidad de ser escalable de manera rápida y flexible, al incorporar de manera fuerte la modularidad del código.

2.1.4. Jupyter Notebook

Jupyter Notebook es un ambiente computacional interactivo basado en la web, que permite crear documentos Jupyter, que soportan la ejecución de docenas de lenguajes de programación. Además es un software de código abierto y su nombre deriva de los 3 lenguajes centrales que utiliza, *Julia, Python y R*, además hace alusión a los diarios del descubrimiento de las lunas de Jupiter por Galileo. Se utilizará este entorno de programación ya que es flexible y permite la portabilidad del código en combinación con Git [34, 35]

2.1.5. Git

Git es un software de control de versiones diseñado por Linus Torvalds, pensando en la eficiencia y la confiabilidad en el mantenimiento de versiones de diversas aplicaciones. Su propósito es llevar registro de los cambios realizados en archivos o documentos de un computadora y coordinar el trabajo que varias personas realizan sobre archivos compartidos [36, 37].

2.2. Conceptos médicos relevantes

2.2.1. Cáncer

El cáncer es un conjunto de enfermedades similares, pero en cada una de estas enfermedades las células del cuerpo se comienzan a dividir sin control y se diseminan a los tejidos circundantes. Este tipo de enfermedad se puede desarrollar en cualquier parte del cuerpo, ya que el control de la cantidad de células en el cuerpo es una parte normal del humano. Este proceso de control de células se llama apoptosis.

El problema sucede cuando este proceso natural del cuerpo falla y las células se dividen sin ningún control, las células viejas y enfermas siguen viviendo y nuevas células se forman cuando no son necesarias. La acumulación de estas células innecesarias se transforma en

tumores sólidos en la mayoría de los canceres, aunque hay otros como los canceres a la sangre, como la leucemia, que no suele formar tumores sólidos.

Los canceres también se pueden dividir en benignos y malignos, los benignos son los que se quedan en un solo lugar del cuerpo y que no desprenden células cancerosas que viajen a través del sistema circulatorio o linfáticos evitando la propagación del cáncer por todo el cuerpo. Los canceres malignos son los que si desprenden células cancerosas que viajan por el cuerpo y propagan la enfermedad.

2.2.2. Tipos de cáncer

Hay diferentes tipos de cáncer que dependen del tipo de célula dañada. A continuación, se muestran los tipos de cáncer que hay:

1. **Carcinoma:** son el tipo más común de cáncer y se forman en las células epiteliales, las cuales cubren superficies internas y externas del cuerpo.
2. **Sarcoma:** Son los canceres que se forman en los huesos y tejidos blandos.
3. **Leucemia:** Son los canceres que se forman a partir de los tejidos creados por la médula ósea. Los glóbulos blancos anormales se acumulan y evitan que los glóbulos blancos normales puedan hacer su trabajo, como control de hemorragias u oxigenación de la sangre.
4. **Linfomas:** Son los canceres que comienzan en los linfocitos (células T o células B), que son los glóbulos blancos que combaten las enfermedades.
5. **Mieloma múltiple:** Cáncer que comienza en las células plasmáticas, otro tipo de célula inmunitaria, las células plasmáticas anormales se acumulan en la medula ósea y forman tumores óseos por todo el cuerpo.
6. **Melanoma:** Es el cáncer que comienza en los melanocitos, células especializadas en producir melanina (pigmento que le da color a la piel)
7. **Tumores cerebrales y de medula espinal:** son tumores que se llaman de acuerdo con el tipo de células en que se forman.

2.2.3. Tratamientos para el cáncer

Hay diversos tipos de tratamiento para el cáncer, que usualmente se combinan para obtener el mejor resultado, entre ellas están

1. **Cirugía:** Este tratamiento eliminar el cáncer que se manifiesta en forma de tumores sólidos, es decir, es de carácter local y no es aplicable a canceres que se han expandido por el resto del cuerpo o cáncer a la sangre como la leucemia. Hay diferentes niveles de extirpación del tumor, está la extirpación completa del tumor y con ello el cáncer. La reducción del tumor, en la cual se elimina una parte del tumor ya que la eliminación completo podría afectar el funcionamiento de los órganos. También está la eliminación

de tumores que están generando dolor. Uno de los principales riesgo es la infección posterior a la cirugía que es tratada con antibióticos.

2. **Radioterapia:** Este tratamiento utiliza la radiación localizada en la zona del cáncer haciendo que el ADN de las células cancerígenas se vea destruido, produciendo que estás células mueran en el corto y mediano plazo. Una de las complicaciones es que la cantidad de radiación que puede recibir una parte determinada del cuerpo es limitada y además la radiación podría dañar a las células sanas que están cerca del tumor.
3. **Quimioterapia:** La quimioterapia es un tipo de tratamiento en el cual se suministran dosis de ciertos químicos que hacen que las células cancerígenas mueran o su crecimiento sea más lento. Se puede usar antes de una cirugía o radioterapia para reducir el tamaño del tumor, después de estos procedimientos para eliminar las células cancerígenas que quedan en el sistema o solo para reducir los síntomas del cáncer. La complicación más grande es que el suministro de la quimioterapia puede tener efectos secundarios en las personas y reducir la calidad de vida de las personas.
4. **Inmunoterapia:** Este tratamiento se basa en el fortalecimiento del sistema inmunitario de las personas. Al mejorar el sistema inmune del paciente se logra que el cáncer no se expanda tan fácilmente como lo hace cuando el sistema inmune está más débil. Las diferentes técnicas que hay para fortalecer el sistema están los inhibidores de punto de control, Transferencia adoptiva celular, Anticuerpo monoclonales y vacunas de tratamiento.
5. **Terapia Dirigida y medicina de precisión:** es la aplicación de ciertos medicamentos que actúan sobre los procesos de crecimiento, división y diseminación de las células cancerígenas. Estos medicamentos están asociados directamente con la genética de las células haciendo que existan casos en donde no haya medicación que sirva. Un riesgo es que las células cancerígenas puedan desarrollar resistencia a las medicamentos y el tratamiento se vuelvan ineficiente.
6. **Terapia Hormonal:** es un tratamiento que hace lento o detiene el crecimiento de las células cancerígenas a través de la limitación de las hormonas que hacen crecer estas células.
7. **Trasplante de células madre:** Este tipo de tratamiento ayudan a la restauración de las células madre que forman la sangre en individuos cuyas células madre fueron destruidas por la radioterapia o quimioterapia. Este tratamiento no afecta directamente al cáncer, pero si mejoran los efectos secundarios del tratamiento del cáncer, haciendo más eficiente y rápida la posible recuperación del paciente.

2.2.4. Neutropenia Febril

Una de las complicaciones más serias para los pacientes con cáncer es la neutropenia febril, condición que junta dos conceptos. Neutropenia es definida como un recuento anormal de neutrófilos, en específico, si el recuento absoluto de neutrófilos (`RAN`) es menor a 500 céls/mm^3 o de 1000céls/mm^3 si se predice una caída a 500céls/mm^3 en las siguientes 24 o 48 horas de la primera medición. Y febril es si es que el paciente es ingresado con un registro único de temperatura axilar mayor a 38,5 grados y dos mediciones mayores a 38 grados con una separación de 1 hora.

La condición anterior es peligrosa para este tipo de paciente por 3 razones, la primera es que al tener el sistema inmune comprometido por el tratamiento recibido (quimioterapia, radioterapia) son susceptibles a infecciones bacteriales, fúngicas o virales. La segunda es que si el tratamiento de NF es muy largo puede impactar en el tratamiento de base, que es el cáncer, retrasando la recuperación del paciente. La tercera razón es que en el mismo tratamiento de la NF el paciente puede llegar a desarrollar complicaciones. Una de estás complicaciones es la sepsis, que es un reacción masiva del sistema inmune frente a una infección, que libera químicos sin control haciendo que el cuerpo humano eventualmente deje de funcionar.

Por lo anterior es menester tener una actitud pro activa frente a la amenaza de cualquiera de estos problemas. En particular, la investigación acerca de los factores de riesgo de complicaciones llevada a cabo por diferentes autores a entregado diferentes puntajes de riesgo que han permitido estratificar de mejor manera a los pacientes. Esta estratificación a permitido entregar un tratamiento mucho más racional a los pacientes, haciendo que el proceso de recuperación sea más eficiente.

2.2.5. Tratamiento para la NF

El tratamiento actual para neutropenia febril se basa en la aproximación de los años 90' basado en ün tratamiento anti-microbiano de amplio espectro y bactericida ya que las infecciones en este tipo de hospederos progresan rápidamente y pueden ocasionar la muerte"[38]. Luego del ingreso, el tratamiento empírico debe basarse en las características epidemiológicas del hospital ingresado y del riesgo que presenta el paciente.

Para el caso de los pacientes categorizados como de alto riesgo se tienen que hospitalizar y recibir los anti-microbianos por vía intravenosa (IV), posterior a dicha hospitalización el tratamiento elegido depende de las características de cada caso. En varios ensayos médicos se han comparado diferentes estrategias de tratamiento con eficacia semejante: mono terapia, terapia combinada con aminoglucósidos, y la combinación de las anteriores con o sin terapia anti-cocaceas grampositivas.

En los casos de bajo riesgo se recomiendo seguir un tratamiento similar al de alto riesgo pero re-evaluarse a las 24 horas de haber comenzado el tratamiento parenteral (suministración de cualquier medicamento o solución que nos sea por la vía gastrointestinal, como intravenosa o subcutánea a través de agujas [39]). En caso de persistir los criterios de bajo riesgo se podría rotar el tratamiento a vía oral (cefiximama, acetil-cefuroxima o coproflxacina) y completar el tratamiento en esa modalidad, o mantener la terapia ambulatoria con antimicrobianos IV de utilización cada 24 horas.

La implementación de alguna de estás terapias de tipo ambulatorio dependerá de si el hospital a cargo del paciente tiene una capacidad de respuesta las 24 horas al día, con personal entrenado capaz de estar alerta a cualquier signo clínico que implique una nueva consulta. De todas maneras, estudios prospectivos y aleatorizados realizados en Latinoamérica han mostrado que estás terapias secuenciales tiene eficacia similar a la modalidad de manejo hospitalario, con mejor evolución psico emocional y menores costos para los pacientes y los servicios de salud. En la imagen 2.1 se muestra el procedimiento estándar para los casos de

neutropenia febril.

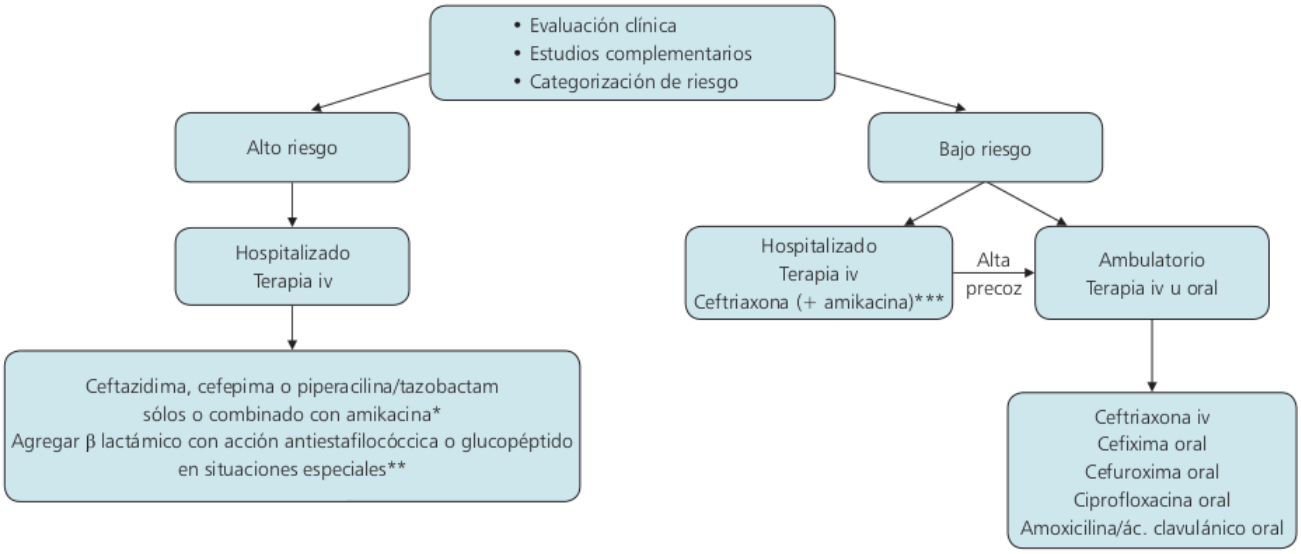

Figura 2.1: Tratamiento de neutropenia febril, Paganini 2011

2.2.6. Complicaciones

Dependiendo del estudio para la estratificación de riesgo se tienen diferentes tipos de resultados (*outcomes*) para las complicaciones, las más comunes son las siguientes [1]:

1. **Infección Bacterial Invasiva (IBI)**: Se define así si cumple al menos uno de los siguientes criterios: (1) ocurrencia de bacteriemia (bacterias en la sangre) (2) Cultivo positivo de alguna bacteria en un sitio usualmente estéril (por ejemplo: catéter venoso central) [13]
2. **Infección Bacterial Severa (IBS)**: Se define como muerte por una infección bacteriana, un cultivo positivo de bacterias en fluidos usualmente estériles (como la sangre), neumonía comprobada por radiología, diagnostico clínico inequívoco de infección bacteriana o PCR mayor a 150mg/l[40].
3. **Complicación infecciosa severa (ICS)**: se define como la presencia de sepsis y/o shock y/o bacteriemia o fungemia en una muestra de sangre y/o muerte por una infección durante el episodio de neutropenia febril[2].
4. **Evento adverso** que incluye muerte, complicaciones que necesiten unidad de cuidados intensivos y complicaciones que potencialmente arriesguen la vida del paciente que sean juzgado por un doctor tratante resultado de una infección, MDI o neumonía confirmada por radiología.[41]
5. **bacteriemia** definida como un cultivo positivo, independiente del patógeno encontrado, usando un sistema de cultivo automático. [**?**]

2.2.7. Cáncer y neutropenia febril en niños

Según datos del Informe RENCI 2007-2011 se estima que la incidencia del cáncer en menores de 15 años es de 128,2 por millón de habitantes o 480 casos al año. El informe identifica que hay 3 grupos de canceres que concentran la mayor cantidad de canceres según la clasificación ICCC-3, leucemias (40,1 %), tumores del sistema nervioso central (15,9 %) y linfomas y tumores reticuloendoteliales (9,9 %). Cabe destacar que el cáncer es la segunda causa de muerte para los niños entre 5 y 14 años después de los accidentes según datos del INE, pese a eso, la sobre vida de los pacientes infantiles de cáncer es alta, el 71,4 % de los pacientes con cáncer tiene una sobre vida mayor a 5 años. Las causas específicas del cáncer infantil aún no han podido ser descifradas por los investigadores, aun así, se sabe que aproximadamente el 10 % de los casos es mutaciones germinales, es decir, genes heredados desde los padres. Para el restante porcentaje se ha buscado encontrar los factores ambientales que causan el cáncer, pero al ser una enfermedad poco frecuente y que es difícil saber a qué factores estuvo expuesto el niño durante sus primeros años. Se piensa que podría no haber razones ambientales para la mayoría de los cánceres en niños.

2.3. Puntajes de predicción de riesgo

La investigación de puntajes de riesgos de complicaciones en adulto ha sido prolífica, desde las primeras reglas propuestas por Talcott, pasando por los recientes puntajes del Centro Multinacional de Soporte al Cancer (MASCC) y el puntaje CISNE (*Clinical Index of Stable Febrile Neutropenia*) que busca identificar a los pacientes riesgosos que aparentemente están estables.

Para el caso de los niños también han habido muchos estudios que plantean diferentes indicadores en 2017, Das et al. realiza un estudio que busca validar indicadores de estratificación de riesgo aplicándolos a una población de la India obtenidos entre julio de 2013 y septiembre de 2014 [1]. En este estudio se compararon 7 modelos de diferentes partes del mundo, incluyendo el modelo propuesto por el autor. Los modelos utilizan diversos factores y predicen diferentes resultados, en la tabla 7.1 de los anexos se ven los factores y los posibles resultados. De acuerdo a encontrado por el autor, solo 3 modelos pueden ser medidos por los indicadores de sensibilidad y especificidad y se muestran en la tabla 2.1:

Modelo	**Sensibilidad**	**Especificidad**
Das et al	86,2 %	63,3 %
Santolaya et al	82,7 %	53,1 %
Agyeman et al	98,3 %	2,8 %

Tabla 2.1: Tabla de sensibilidad y especificidad para los modelos de Das et al [1]

Como se puede ver no hay un consenso con respecto a la mejor estratificación de riesgo en el mundo. Esto se podría deber a la población que utiliza cada estudio, haciendo que los resultados estén ajustados solo a esa población. También podría ser que dado que los posibles resultados a predecir son diferentes, los factores también lo son. No obstante, algunos

de los factores son compartidos entre los diferentes estudios pero cambian el umbral en que se consideran como factor de riesgo. Como por ejemplo el conteo de plaquetas o recuento absoluto de neutrófilos, lo que da indicios de que variables utilizar para el modelo final.

Para el caso chileno, se llegó a un consenso para el tratamiento racional de neutropenia febril que identifica 8 factores de riesgo para declarar a una neutropenia febril como de alto riesgo, 3 más que los mencionados en el estudio de [1]. Los factores de riesgo son los siguientes:

1. PCR 90mg/L. La proteína C Reactiva, es una proteína creada en el hígado y es una respuesta del cuerpo frente a una inflamación.
2. Hipotensión. Presión arterial baja.
3. Leucemia en recaída. Es cuando la leucemia vuelve luego de ser tratada.
4. Leucemia no linfoblástica. Tipo de cáncer que afecta a las plaquetas.
5. Linfoma no Hodking. Tipo de cáncer que afecta los glóbulos blancos entre otras cosas
6. Neuroblastoma etapa IV. Estadio más avanzado de esta enfermedad
7. 7 días último ciclo quimioterapia-inicio Fiebre
8. Plaquetas $50.000/mm^3$

Considerado ya estos factores, se procede a hacer una estratificación de riesgos alta o baja de acuerdo a ciertos criterios. Si el paciente tiene uno de los 6 primeros factores o los dos últimos juntos se estratifica como Neutropenia Febril Alto Riesgo, en caso contrario se estratifica como Neutropenia Febril Bajo Riesgo. Esta es la línea base o *benchmark* principal a comparar.

2.4. Descubrimiento de Conocimiento en Base de Datos

KDD es un marco teórico que busca extraer conocimiento a través de fuentes de datos estructurados (Bases relacionales, XML) o datos no estructurados (textos, fotos, documentos). Esta tarea ha tomado relevancia los últimos años ya que el volumen de datos ha aumentado drásticamente. Se estima que en el 2020, por cada persona en el mundo se generen 1,7mb por minuto, es decir que en un segundo se generan aproximadamente 200 terabytes de información [42, 43].

El KDD busca hacerse cargo de este problema al incorporar algoritmos computacionales que permiten convertir estos datos brutos en conocimiento que a su vez ayuden a tomar decisiones que agreguen valor a las empresa o instituciones donde se aplique. Este proceso está compuesto por varios pasos, entre 5 y 9 pasos dependiendo del autor. En esta memoria se utilizará el proceso de KDD planteado en [43] y son los siguientes:

1. **Aprender el dominio del problema:** Tener conocimientos previos del dominio a tratar y los objetivos de este.
2. **Crear un conjunto de datos objetivos:** Seleccionar los conjuntos de datos relevantes, subconjunto de variables o muestra de datos en donde el descubrimiento será realizado.
3. **Limpieza de los datos y pre-procesamiento:** En este paso se remueve el ruido o

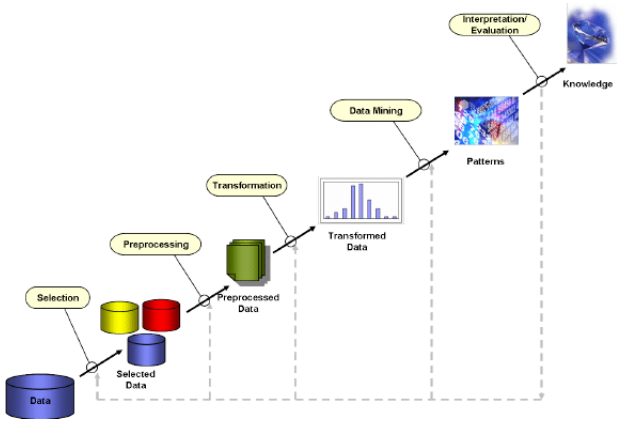

Figura 2.2: Ciclo de vida de un proyecto basado en KDD, fuente IBM [5]

los valores raros en caso de ser necesario, se deciden las estrategias para manejar los valores faltantes entre otras tareas.

4. **Reducción de datos y proyección:** Encontrar las características útiles que representen los datos, dependiendo del objetivo de la tarea, usando reducción de dimensionalidad o métodos de transformación entre otras.
5. **Elección de la función del *Data Mining*:** Aquí se decide el objetivo del modelo, por ejemplo: clasificación, regresión o clustering.
6. **Elección del algoritmo de *Data Mining*:** En este paso se seleccionan los métodos que serán usados para buscar patrones en los datos, decidiendo que modelos y parámetros podrían ser apropiados (los modelos para datos categóricos son diferentes a los modelos con datos continuos) y que estos métodos estén alineados con el propósito del minado de datos, si es el entendimiento de lo que sucede en los datos o en la capacidad que tiene el modelo para predecir determinado resultado.
7. **Minado de datos:** Este paso es la aplicación de las técnicas elegidas y la búsqueda de los patrones de interés que podrían aparecer en los datos.
8. **Interpretación:** Incluye la interpretación de los patrones descubiertos y la posibilidad de regresar a pasos previo, removiendo patrones redundates o irrelevantes. Además, en esta etapa se trasladan los patrones que resultaron útiles a términos que puedan ser fácilmente entendidos por los usuarios.
9. **Usar el conocimiento descubierto:** En esta etapa se incorpora el conocimiento descubierto al sistema, tomando acciones en base a ese conocimiento o simplemente reportando los resultados a las partes interesadas. También permite chequear y resolver posibles problemas con las creencias previas.

Cabe destacar que el KDD es un proceso iterativo, en donde no se espera un avance lineal que no vuelva a los pasos anteriores sino que, en caso de ser necesario, se retroceda al paso anterior para dar cuenta de las necesidades generadas en el paso nuevo. En la figura 2.2 se puede ver el proceso tradicional de KDD.

Un marco teórico alternativo al uso tradicional del KDD es el CRISP-DM *Cross Industry Standard Process for Data Mining* que tiene 6 etapas [44] que son bastante similar al KDD y muestran en el punto 7.2 de los anexos.

2.5. Modelos ML

2.5.1. Logistic Regression

Una regresión logística es un tipo de regresión que busca predecir el resultado de una variable dependiente y categórica, que toma los valores 1 o 0 dependiendo si el suceso ocurre o no) en función de variables independientes que podrían ayudar a predecir el resultado. En otras palabras una regresión logística predice lo siguiente:

$$p_{\mathrm{i}} = \frac{1}{1 + \mathrm{e}xp^{-(\beta_0 + \beta_1 x_{1,\mathrm{i}} + \beta_2 x_{2,\mathrm{i}} + \ldots + \beta_k x_{k,\mathrm{i}})}} \tag{2.1}$$

Donde $\mathrm{p_i}$ es la probabilidad del suceso i dada las variables dependientes $\mathrm{x_i}$. Los valores *beta* son los parámetros que toma el modelo para ajustar las probabilidades. Usualmente si la probabilidad de un suceso es mayor al umbral 0,5 se considera 1 y en caso contrario 0. No obstante, dicho umbral puede ser movido de acuerdo a las necesidades particulares del problema.

Las regresiones logísticas han sido los métodos estadísticos utilizados clásicamente en las investigaciones de neutroprenia febril, que busca predecir los casos que se complicarán (asignándole el valor 1) y los casos que no se complicarán (0).

2.5.2. Support Vector Machine (SVM)

Estos métodos están relacionados con problemas de clasificación y regresión. Dado un conjunto de ejemplos de entrenamiento (de muestras) se pueden etiquetar las clases y entrenar un SVM para construir un modelo que prediga la clase de una nueva muestra. Intuitivamente, una SVM es un modelo que representa a los puntos de muestra en el espacio, separando las clases en 2 espacios lo más amplios posibles mediante un hiper plano de separación definido como el vector entre los 2 puntos. El vector más cercano a las 2 clases se llama vector de soporte, entonces una nueva muestra es ubicada en el hiper-plano y se clasifica en una de las clases más cercanas.

SVM es uno de los modelos con más potencial ya que el resultado final de este modelo es una segmentación o clasificación de los datos en hiper planos, que es justamente lo que se busca en esta memoria.

2.5.3. Clasificador Bayesiano Ingenuo

Es un tipo de algoritmo de machine learning basado en el Teorema de Bayes[45]. Se le suele llamar *Ingenuo*, porque sus supuesto de funcionamiento se suelen resumir en la independencia de las variables predictoras, es decir, que estás no tienen ninguna relación entre ellas, lo cual difícilmente se cumple, aún así el clasificador Naive Bayes ha dado buenos resultados,

compitiendo con modelos más sofisticados de predicción [46, 47].

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.2}$$

La base de este algoritmo es la ecuación 2.2, la cual calcula la probabilidad de que un evento anterior B, pueda ser capaz de predecir un evento posterior A, si es que B sucede de nuevo. En esta memoria se busca predecir una infección bacteriana invasora (A) en presencia de diferentes variables predictoras (B) (temperatura de ingreso, profilaxis, episodios previos entre otros). Para obtener la probabilidad del suceso, es necesario asumir que las probabilidades son independiente lo cual es difícilmente cierto.

En esta memoria se utilizará y comparará el rendimiento del modelo Gaussiano de Bayes para variables continuas, el modelo Multinomial de Bayes para variables categóricas y un modelo que es el resultado de la multiplicación de ambos modelos al considerar los resultados como independientes.

2.5.4. Arboles de decisión

Los árboles de decisión son modelos de predicción basados en lógicas preestablecidas, que buscan clasificar conjuntos de datos en diferentes categorías. Cabe mencionar que también existen los *arboles de regresión* en donde, en vez de predecir una categoría predicen un valor continuo.

Un árbol de decisión comienza con un nodo inicial en donde se elige una variable y una regla que se aplica a esa variable para dividir el conjunto de datos. Lo que se busca de esta regla es que los datos queden divididos de buena manera. Para definir si la división fue buena hay diferentes métricas que permiten identificar el rendimiento de dicha división, como lo son la entropía (reducción de esta), el Gini índex o el Gain Ratio.

Algunas de las características de un árbol de decisión son los mencionados nodos en donde se divide el conjunto de datos con una regla seleccionada, luego de dicha división, cada conjunto deriva en una rama o *branch* que puede culminar en un nodo, en donde una nueva regla se fijara o con una hoja o *leaf* que es el resultado final de la división de los nodos.

Entre los arboles de decisión podemos encontrar algoritmos como CART (*Classification and Regression Trees*)

2.5.5. Random Forest

Los Bosques Aleatorios son algoritmos de machine learning ensamblados [48]. Este tipo de métodos se basan en la aplicación de diferentes algoritmos de predicción diferentes para generar una decisión más robusta. En este caso, los Bosques Aleatorios utilizan como algoritmo base arboles de decisión. Cabe recordar que los arboles de decisión buscan maximizar la diferencia entre grupos y al mismo tiempo minimizar la diferencia dentro de los grupos, dicha

diferencia y similitud es medido a través de diferentes indicadores como el de Gini o Ganancia de Información. Dichos arboles de decisión son aplicados al conjunto de datos y utiliza el resultado de cada uno para votar y definir a la clase a la que pertenece la observación. En la práctica se ha observado que esta metodología genera buenos resultados clasificando las observaciones.

Una de los fundamentos que hace que los bosques aleatorios tengan buenos resultados es que al crear diferentes modelos no correlaciones, el error en cierta dirección de cada uno de ellos se va compensando con el error en otra dirección de los otros modelos. Esto hace que la predicción del conjunto de arboles sea mejor que el de un árbol en particular.

Los bosques aleatorios tienen dos mecanismos para asegurar la aleatoriedad del proceso. El primero es tomar n observaciones, siendo n menor que la cantidad total de observaciones e ir tomando n diferentes registros en cada iteración. La segunda es tomar k variables, siendo k menor que las variables totales e ir variándolas en cada iteración. En cada uno de estos casos se crean arboles de decisión y se clasifica a la observación en alguna de las clases definidas. Luego, se toma el resultado de cada uno de estos arboles y se realiza una "votación" para identificar la mejor clasificación.

Uno de los problemas con los arboles de decisión es poder determinar el impacto que tiene cada variable en la predicción realiza por el modelo. Para esto existen algunas técnicas que permiten identificar la importancia de cada variable. Una de ellas, extrae una variable e identifica la variación del error del modelo. La idea es que si una variable importante es extraída y el error aumenta significativamente, quiere decir que dicha variable era importante. Otra técnica es calcular el índice Gini del modelo e ir extrayendo variables y observando la variación del índice. La idea es que si la variable extraída hace que el índice aumente quiere decir que la clasificación se hace más impura, disminuyendo el rendimiento del modelo.

2.5.6. Gradient Boost Machine

Los algoritmos llamados *Gradient Boosting Machine* son un conjuntos de algoritmos que combinan muchos algoritmos débiles de predicción, usualmente arboles de decisión, para crear un modelo de predicción más fuerte [49]. Este tipo de algoritmos están siendo más utilizado dado su buen rendimiento en la tarea de clasificar conjunto de datos complejos.

La idea general detrás de estos algoritmos es tomar una hipótesis débil e ir haciendo pequeños ajustes o *aprendizajes* para hacer que esta hipótesis más fuerte. Lo hace al clasificar el conjunto de datos de entrenamiento de manera iterativa, dejando a los datos bien clasificados separados de los defectuosamente clasificados y aplicando los algoritmos al conjunto de datos defectuosamente clasificados para ajustar el modelo.

Algoritmos con mecánicas similares son los Bosting Adaptativo (AdaBoost), en donde se comienza con múltiples arboles de decisión, algoritmos débiles, que tienen solo una separación. Cada observación es medida por el algoritmo y a aquellas que son más difíciles de clasificar se le asigna un mayor peso y a su vez se le agregan más algoritmos débiles. Los algoritmos AdaBoost predicen la clasificación a través de un sistema de votación" efectuado

por los algoritmos débiles, en donde se asigna la clasificación más votada a cada una de las observaciones del conjunto de validación.

Los algoritmos de Boosting con gradiente usan está mecánica pero su objetivo es la disminución de la función de pérdida, es decir, la disminución del error en la clasificación. Además, la diferencia de los algoritmos adaptativos, los de gradiente, fijan el peso de los algoritmos débiles previos sin cambiarlos como sucede con los adaptativos y por ultimo, son algoritmos más completos ya que también incorporan clasificaciones con clases múltiples e incluso pueden predecir regresiones.

2.5.7. Redes neuronales

Son un modelo computacional basado principalmente en el comportamiento de las neuronas del cerebro, tratando de imitar dicho comportamiento. Estos sistemas trata de aprender en base a ejemplos y sin conocimiento ni reglas preconcebido acerca de lo que se va a aprender.

Las neuronas están conectadas en diferentes capas y cada capa entrega un peso que es multiplicado por el valor de entrada de la neurona. Estos pesos pueden hacer que la función activadora de las neuronas adyacentes sea activada o inhibida. Usualmente el objetivo de una ANN es reducir la función perdida modificando los parámetros (pesos) para predecir de mejor manera posible el resultado requerido.

La aplicación de este algoritmo es meramente académico, ya que la implementación de una estratificación basada en redes neuronales es complicado en un contexto médico. Lo anterior se debe a que el médico necesita saber las razones por la cual se esta estratificando a los pacientes. Los factores de riesgo no pueden quedar ocultos en la çaja negra"que implica las redes neuronales. De todas maneras, existen técnicas que podrían ayudar a determinar la importancia de cada una de las variables de manera aproximada.

2.6. Conceptos relevantes de machine learning y diseño de aplicaciones Web

2.6.1. Matriz de confusión y métricas de desempeño asociadas

1. **Matriz de confusión:** Una matriz de confusión es una de las formas de mostrar el desempeño de un modelo de predicción. Se basa en comparar los resultados obtenidos en la realidad con los predichos por el modelo. En caso de que la condición a predecir se manifieste en la realidad el resultado se define como positivo y en caso contrario como negativo. Al hacer lo anterior se pueden crear 4 tipos diferentes de resultados:
   (a) **Verdaderos Positivos:** Son aquellos valores que fueron predichos positivos y en la realidad fueron efectivamente positivos.
   (b) **Falsos Positivos:** Son aquellos valores que fueron predichos positivos pero que en la realidad fueron negativos.

(c) **Verdaderos Negativos:** Son aquellos valores que fueron predichos como negativos y en la realidad fueron negativos.

(d) **Falso Negativo:** Son aquellos valores que fueron predichos como negativos pero que en la realidad fueron positivos.

En la tabla 2.2 se puede ver la forma general de una matriz de confusión con sus correspondientes clases posibles.

		Resultado Real	
		Positivo	Negativo
Resultado Predicho	Positivo	Verdadero Positivo (VP)	Falso Positivo (FP)
	Negativo	Falso Negativo (FN)	Verdadero Negativo (VN)

Tabla 2.2: Ejemplo matriz confusión.

2. **Sensibilidad o Recall:**

   Sensibilidad es el indicador que mide la proporción de verdaderos positivos, del conjunto de todos los casos que terminaron complicándose (verdaderos positivos y falsos negativos). En este caso se busca que la sensibilidad sea lo más alta posible, es decir que la cantidad de personas que el modelo predice que se van a complicar al ingreso sea lo más similar posible a la cantidad real.

   Se busca que este indicador sea lo más alto posible ya que cada caso mal identificado, es decir, cada caso que se predijo como de bajo riesgo y que luego se complica es demasiado peligroso para el paciente. Este sin duda es el indicador más importante de cualquier modelo que trate con la salud de las personas y es la restricción a mantener. En otras palabras, se quiere mejorar los otros indicadores del nuevo modelo sin disminuir la sensibilidad del actual modelo.

$$Sensibilidad = \frac{VerdaderosPositivos(VP)}{VerdaderosPositivos(VP) + FalsosNegativos(FN)} \tag{2.3}$$

3. **Especificidad:** Especificidad es el indicador que busca identificar los casos que se predijeron como de bajo riesgo (verdaderos negativos) y efectivamente evolucionaron hacia ser casos de bajo riesgo.

   Se mide como la fracción de verdaderos negativos de la muestra del total de casos negativos. Este indicadores es de carácter secundario, pero de igual forma tiene una importancia fundamental en el tratamiento de los pacientes ya que predecir que un paciente será grave cuando evoluciona favorable es exponerlo a un tratamiento demasiado agresivo, que interrumpe la vida normal del paciente e incluso podría reducir la eficacia del tratamiento contra el cáncer.

$$Especificidad = \frac{VerdaderosNegativos(VN)}{VerdaderosNegativos(VN) + FalsosPositivos(FP)} \tag{2.4}$$

4. **Exactitud:** Exactitud es el indicador global de que tan preciso es el modelo, es decir cuántas veces el modelo acierta a los verdaderos positivos y los verdaderos negativos.

Este indicador no es relevante dado que no nos importa saber la cantidad de complicaciones sino que quienes se van a complicar.

$$Exactitud = \frac{VerdaderosPositivos(VP) + VerdaderosNegativos(VN)}{Todosloscasos} \tag{2.5}$$

5. **Área bajo la Curva (AUC):** El Área Bajo la Curva o AUC, también llamado Característica Operativa del Receptor o ROC por sus siglas en inglés, es un indicador que busca medir el rendimiento global de un modelo de clasificación. Los modelos de clasificación fijan umbrales para definir si una observación pertenece o no a una clase. Lo que hace la curva AUC es graficar los diferentes umbrales de clasificación, asignando cada una de las observaciones a una clase para luego crear una matriz de confusión y calcular la sensibilidad o recall y la especificidad, cuyo valor es puesto en el gráfico.

   La lógica indica que en un modelo de clasificación la sensibilidad y la especificidad van en sentidos contrarios casi siempre, entonces entre más bajo el umbral mejor sensibilidad pero peor especificidad y entre más alto el umbral de clasificación peor sensibilidad y mayor especificidad. La curva AUC indica cual es el mejor modelo que dado cierto umbral tiene mejor rendimiento en términos de sensibilidad y especificidad. En la imagen 2.3 se puede ver una curva AUC con sus correspondientes características.

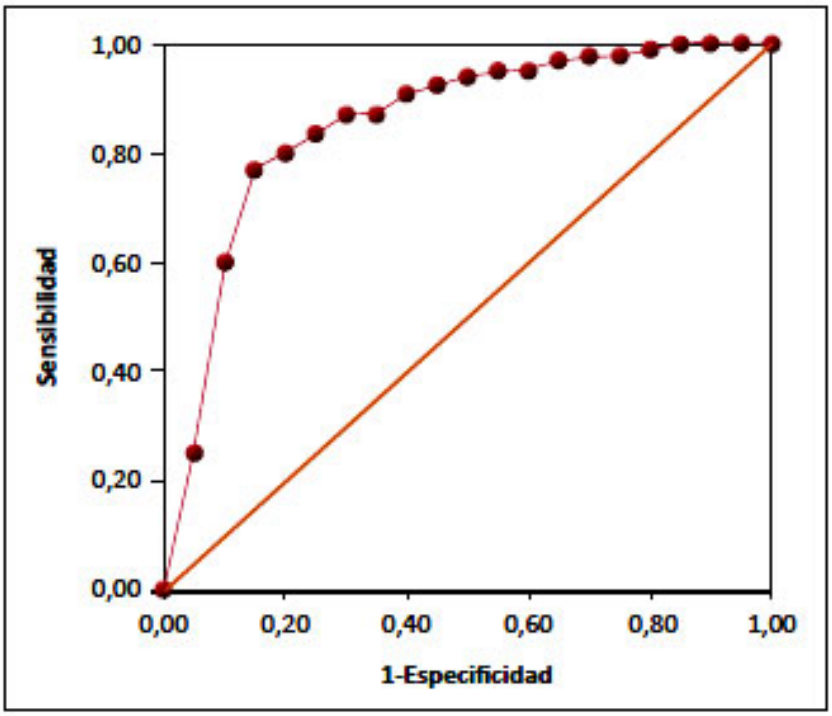

Figura 2.3: Ejemplo Curva AUC

2.6.2. Validación cruzada de K conjuntos

La validación cruzada de K conjuntos o K-Fold Cross Validation en inglés, es una técnica en estadística que permite eliminar el sesgo que se produce al tomar un conjunto en particular

de datos. Para eliminar el sesgo se recalcula el modelo utilizando diferentes conjuntos de datos de prueba y validación, promediando el valor de cada una de las medidas de evaluación de cada uno de los conjuntos, obteniendo el resultado final de la calidad del modelo.

Otra de los problemas que elimina la validación cruzada es el posible sobre-ajuste del modelo sobre un conjunto de datos dado. Por ejemplo, se podría tomar cierta cantidad de datos (80 %) para entrenar el modelo y el restante 20 % para validar el modelo, y el resultado podría ser bueno, pero al momento de calcular el modelo con datos nuevos, el modelo tuvo resultados más pobres que con el conjunto de entrenamiento y validación. En este caso se podría hablar de sobre-ajuste y la aplicación de la técnica de validación cruzada podría eliminarlo ya que prueba K conjuntos de datos diferentes, cambiando los datos de entrenamiento y validación para cada iteración, evitando que el modelo se sobre-ajuste a un conjunto en particular

2.6.3. Modelo Vista Controlador

El marco Modelo Vista Controlador es un patrón de arquitectura de software, la cual diferencia las 3 lógicas de la aplicación: lógica de la base de datos (modelo), la lógica del negocio (controlador) y la lógica de presentación (vista) [50] .

La parte Modelo se encarga de la estructura y mantención de los datos de la aplicación, el controlador es la parte que se encarga de los algoritmos dentro de la aplicación y la vista es la encargada de como se muestra los datos de salida y como se solicitan los datos de entrada.

El principal beneficio de este patrón es que permite un bajo acoplamiento entre las diferentes lógicas y la máxima cohesión dentro de ellas, haciendo que la programación sea más fácil al ser modular y escalable.

2.6.4. Métodos de selección de variables

Una de las cosas más relevantes dentro del proceso de creación de un algoritmo de machine learning es la selección de variables, asunto, que a veces es mencionado coloquialmente como "basura entra, basura sale". En otras palabras, la calidad de las variables de entrada define la calidad de las variables de salida. El problema de la selección de variables también es conocido como *Selección de Subconjuntos de Características* o FSS por sus siglas en ingles [51]. Este problema surge motivado por 3 situaciones:

1. La no monotocidad del éxito de un clasificador. Se ha demostrado empíricamente y matemática para algunos modelos que la agregación de nuevas variables a un modelo mejore el éxito del predictor. No por más variables mejor desempeño.
2. Existencia de variables irrelevantes. Se dice que una variable es irrelevante si es que conocer el valor de esta no aporta nada relevante para el predictor
3. Variables redundantes. Pueden existir variables que su valor puede ser determinado a partir de otras variables por lo que no agrega valor predictivo al modelo.

Otra de las razones para realizar un selección de variables es para tener el modelo más sencillo posible, basándose en el principio de la Navaja de Occam, que si hay dos explicaciones posibles a un fenómeno, usualmente la que tiene menos suposiciones (variables en este caso) es la correcta. Para seleccionar las variables se han creado diversas técnicas que se pueden agrupar en 3 grandes grupos: métodos de filtro, métodos envolventes y métodos ensamblados.

Métodos de filtro

1. **Baja Varianza:** Si los valores de una variable tiene varianza baja, es decir, la mayoría de los valores son iguales podría indicar que la variable no tienen valor predictivo, por lo que se podría sacar del conjunto de datos. Para este método se determina cierto umbral en donde las variables son aceptadas y cuales son descartadas.
2. **Alta Correlación:** Otra técnica es observar la correlación entre las variables, entre aquellas variables que tienen alta correlación se podría dejar una de las variables, ya que ambas tienen el mismo poder predictivo. La elección de variable sacada depende de otras características como cantidad de nulos que tiene y del entendimiento del negocio. Para esta técnica también se fija un umbral en donde las variables que lo pasan son analizadas y se procede o no a su eliminación.

   Para medir la correlación entre las variables se utilizaran dos métodos, el test de Pearson para medir la linealidad de la correlación entre las variables y el test de Spearman para medir la monoticidad de la correlación entre las variables.
3. **Chi cuadrado:** El test Chi Cuadrado es un test que mide la dependencia entre una variable y otra, observando si ambas vienen de la misma distribución o no. Este test sirve para analizar el comportamiento de las variables independientes con respecto a la dependiente, entre más similares sean las distribuciones, mayor poder predictivo tendrá.
4. **Annova:** ANOVA de un factor (también llamada ANOVA unifactorial o one-way ANOVA en inglés) es una técnica estadística que señala si dos variables (una independiente y otra dependiente) están relacionadas en base a si las medias de la variable dependiente son diferentes en las categorías o grupos de la variable independiente. Es decir, señala si las medias entre dos o más grupos son similares o diferentes [52]. También se considera a ANNOVA como una extensión del t-test ya que este solo compara 2 grupos y ANNOVA puede comparar 2 o más grupos de variables.

Métodos envolventes

Los métodos envolventes o *wrapper* en inglés, son aquellos métodos que utilizan diferentes subconjuntos de variables y es evaluado por un modelo de clasificación inducido. Estos métodos pueden ser inductivos, es decir parte desde una variable y van aumentando la cantidad de variables hasta llegar al mejor modelo. Y también pueden ser deductivos, es decir, toman todas las variables y van eliminando la que peor desempeño tiene.

Métodos embebidos

Los métodos embebidos o insertados son métodos incluidos dentro del algoritmo de aprendizaje. El más típico son los arboles de decisión que van haciendo conjuntos cada vez más pequeños de variables. Otros ejemplos es LASSO con la penalización L1 y Ridge con penalización en L2, que son penalizaciones cuando se esta construyendo el modelo. En la práctica, los dos últimos modelos hacen que el peso de muchos variables sea 0 o casi 0.

2.6.5. Balanceo de datos

Uno de los problemas que tienen los algoritmos de clasificación es cuando una clase esta sobre representada en una muestra, haciendo que la identificación de la clase minoritaria sea más difícil. Lo anterior sucede porque el algoritmo se ocupa más de clasificar adecuadamente la clase mayoritaria en detrimento de la clase minoritaria. Esto supone un gran problema cuando la clase minoritaria es una clase vulnerable, por ejemplo, si se quiere predecir quienes son las personas que desarrollaran cáncer en una población determinada, los casos de cáncer son muy bajos en comparación al total, por lo que el clasificador será incapaz de entregar los parámetros que determinaran el desarrollo del cáncer, sirviendo de poco para la predicción del cáncer.

Hay diferentes técnicas para hacerse cargo de este problema pero se pueden agrupar en 3 grandes grupos [53]:

1. Las técnicas que agregan muestras a la clase minoritaria u *oversampling*. Dentro de estás técnicas esta la *técnica de sobre muestreo sintético de la clase minoritaria* o SMOTE por sus siglas en inglés que genera nuevos registros de la clase minoritaria interpolando los valores de los registros minoritarios más cercanos. Otra técnica es el *resampling* que duplica registros al azar de la clase minoritaria.
2. Las técnicas que elimina registros de la clase mayoritaria. Dentro de estás esta el submuestreo aleatorio que elimina muestras al azar. Esta Tomek Links [54] que elimina muestras redundantes dentro de la clase mayoritaria o que estén muy cerca de la clase minoritaria. Un ultimo ejemplo es Wilson Editing [55] o ENN (Editing Nearest Neighbor) que elimina aquellas muestras donde la mayoría de sus vecinos corresponde a la otra clase.
3. Boosting que consiste en asociar pesos a cada una de las muestras e ir modificando dichos pesos en cada iteración. El objetivo de cada iteración es reducir la función de error de cada una de las iteraciones. Dentro de estás técnicas esta AdaBoost que es la aplicación directa del boosting. Esta SMOTEboost que en vez de hacer re-muestreo utiliza SMOTE y RUSboost que en cada iteración utiliza sub-muestreo aleatorio para reducir el tamaño de la muestra y aumentar el rendimiento del clasificador.

Capítulo 3

Desarrollo del modelo final de clasificación

3.1. Aplicación de los modelos de estratificación existentes

Considerando la existencia de múltiples modelos de estratificación de riesgo de neutropenia febril, primero se utilizaran dichos modelos para crear una línea base que sirva para comparar con el modelo final basado en algoritmos de machine learning. Los modelos encontrados en la literatura [1] serán aplicados a los episodios de neutropenia febril entregados por el PINDA entre 2009 y 2016. De igual manera, se pretende utilizar los resultados de los modelos de la literatura para estimar cuales son las variables más relevantes de predicción de riesgo en un episodio de neutropenia febril.

3.1.1. Modelo propuesto por Rondinelli et al.

Rondinelli et al. [2] realizaron un estudio para identificar los factores de riesgo que pudieran predecir alguna de las siguientes complicaciones infecciosas severas: *bacteriemia, fungemia, sepsis, shock séptico y muerte producto de la infección*. Para hacerlo revisaron las fichas clínicas de los pacientes menores de 18 años que desarrollaron su primer episodio de neutropenia febril durante enero del 2000 y diciembre del 2003 del Departamento de pediatría del hospital del Cáncer Camargo en San Paulo, Brasil.

En una primera etapa realizaron un análisis univariado de los factores de riesgo, encontrando los siguientes factores: *sexo femenino, menos de 5 años, leucemia mieloide aguda, actividad de la enfermedad de base, uso de catéter venoso central, hemoglobina menor a 7g/dL,leucocitos menores a 500 células por mm3, granulocitos menores a 500 células por mm3, monocitos menores a 100 células por mm3, plaquetas menores a 20000, temperatura mayor a 38,5C, días desde la ultima quimioterapia menor a 7 días, presencia de mucositis,*

neumonía, y ausencia de infección del tracto respiratorio superior o presencia de algún foco clínico al ingreso del paciente.

En una segunda etapa realizaron un análisis multivariado encontrando los siguientes 6 factores de riesgo: *5 o menos años, catéter venoso central, temperatura mayor a 38,5, hemoglobina menor a 7g/dL, cualquier foco clínico de infección en el primer examen y ausencia de infecciones en el tracto superior respiratorio.* A cada uno de los factores se le asignaron puntajes, que se muestran en la tabla 3.1.

Factor de riesgo	**Puntaje asignado**
5 años de edad o menos	1
Catéter venoso central (CVC)	2
Foco clínico al ingreso	4,5
Temperatura mayor a 38,5	1
Hemoglobina mayor a 7g/dL	1
Infección en el tracto respiratorio superior	2,5

Tabla 3.1: Tabla de puntajes de riesgo, elaboración propia en base a [2]

Por ultimo, los autores definieron 3 categorías de riesgo:

1. **Riesgo bajo:** menos de 5,5 puntos
2. **Riesgo intermedio:** entre 5,5 puntos y 9 puntos
3. **Riesgo alto:** más de 9 puntos

Para aplicar este modelo en la base de datos el PINDA primero se revisa si es que los factores requeridos por el modelo están en la base, donde se encuentra que 5 de los 6 factores están directos en la base de datos. El que no se encuentra es la *infección en el tracto respiratorio superior*, no obstante, se utilizará un aproximado, que es si el foco clínico de ingreso es *respiratorio alto*, se considerara como infección del tracto respiratorio alto.

Otro factor importante es que el modelo de Rondanelli et al. considera 3 categorías, bajo, intermedio y alto riesgo, a diferencia del modelo utilizado en la red PINDA que tiene solo dos categorías de riesgo, por lo que para simplificar el cálculo de la precisión del modelo se consideraran 2 umbrales para separar el nivel de riesgo. Un umbral será en los 9 puntos y la otra en 5,5, que son los cortes que los autores seleccionan para dividir entre bajo,intermedio y alto riesgo.

En tabla 3.2 se muestran los resultado del cálculo para cada uno de los umbrales de separación de riesgo, mostrando las métricas de precisión, sensibilidad, especificidad, valor predictivo positivo, valor predictivo negativo y AUC.

Umbral	Sensibilidad	Especificidad	PPV	NPV	Precisión	AUC
1	0.993	0.007	0.542	0.445	0.541	0.500
2	0.970	0.032	0.542	0.468	0.540	0.501
3	0.846	0.151	0.541	0.453	0.527	0.498
4	0.682	0.293	0.533	0.438	0.504	0.487
5	0.607	0.349	0.524	0.429	0.488	0.478
5.5	0.601	0.350	0.522	0.426	0.486	0.475
6	0.575	0.373	0.520	0.426	0.483	0.474
7	0.443	0.504	0.513	0.433	0.471	0.473
8	0.273	0.659	0.486	0.434	0.450	0.466
9	0.198	0.713	0.449	0.429	0.434	0.455
10	0.132	0.803	0.443	0.439	0.440	0.468

Tabla 3.2: Rendimiento modelo Rondanelli et al., elaboración propia

Como se puede ver, entre más alto el umbral peor desempeño tiene el modelo, lo cual es esperable. También se ve que el óptimo, medido en AUC, cuando el corte entre los de bajo riesgo y los de alto riesgo es de 2 puntos, sin embargo, este umbral clasifica a casi todos los pacientes como riesgosos teniendo una especificidad muy baja,lo cual estas bastante alejado con el objetivo de mejorar el tratamiento.

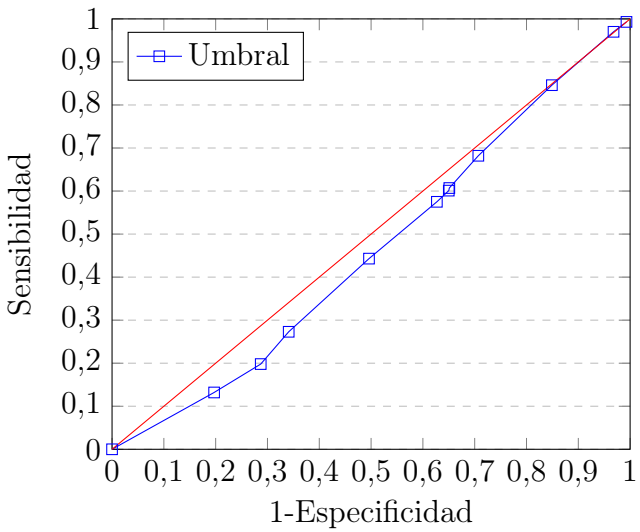

Figura 3.1: Relación entre sensibilidad y especificidad de acuerdo al umbral en el modelo de Rondanelli et al.,elaboración propia

3.1.2. Modelo de Das et al.

Das et al. [1] condujo un estudio que buscaba validar un puntaje de predicción de riesgo de complicaciones y además evaluar el rendimiento de otros puntajes publicados. La población objetivo procede del norte de India, que es un país en desarrollo con recursos limitados para la evaluación del riesgo durante julio de 2013 y septiembre de 2014, que contó con 414 episodios medidos.

Este estudio encontró que los factores relevantes para la predicción de riesgo eran *desnutrición, tiempo desde la ultima quimioterapia,presencia de focos de infección respiratorio no superior, proteína C reactiva y el recuento absoluto de neutrófilos*. En la tabla 3.3 se puede ver el puntaje asignado a cada factor de riesgo.

Factor de riesgo	**Puntaje asignado**
Desnutrición	2
7 o menos días desde la ultima quimioterapia	2
Presencia de un foco infeccioso no respiratorio alto	2
PCR $\geq 60 mg/l$	5
RAN $\leq 100 clulas/mm3$	2

Tabla 3.3: Tabla de puntajes de riesgo Das et al., elaboración propia en base a [1]

Además, se clasifica en riesgo alto y en riesgo bajo de la siguiente manera:

1. Alto riesgo: 7 o más puntos
2. Bajo riesgo: menos de 7 puntos.

4 de las 5 variables se encuentran en la base del PINDA directamente, solo el factor de riesgo *desnutrición* hay que construirlo. Lo que se hace con las indicaciones entregadas por estudio, que se basan en las tablas construidas por la Organización Mundial de la Salud [56].

En la construcción del indicador del factor de riego se encontraron más errores en el registro de los datos, en este caso con respecto a la edad. En la figura 3.2 se puede ver que hay casos extremos que se alejan de la tendencia y lo más probable es que sean errores de traspaso dela información, que sirve de antecedente para la incorporación de una plataforma que evite este tipo de errores.

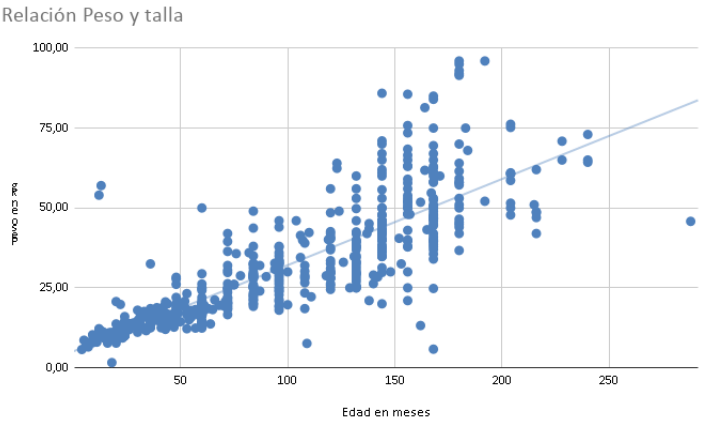

Figura 3.2: Elaboración propia en base a base del PINDA

Los resultados obtenidos se muestran en la tabla 3.4, en donde se puede apreciar un rendimiento mejor al del modelo de Rondanelli, alcanzando su mejor desempeño medido en AUC en el umbral de los 7 puntos que es justamente el umbral propuesto por los autores.

Umbral	Sensibilidad	Especificidad	PPV	NPV	Precisión	AUC
1	0.993	0.024	0.547	0.739	0.549	0.508
3	0.933	0.195	0.579	0.711	0.595	0.564
5	0.730	0.528	0.647	0.622	0.637	0.629
6	0.711	0.540	0.647	0.612	0.633	0.625
7	0.551	0.788	0.755	0.597	0.660	0.670
8	0.448	0.836	0.764	0.561	0.626	0.642
9	0.438	0.849	0.775	0.561	0.626	0.644
10	0.199	0.942	0.803	0.498	0.539	0.570
12	0.018	0.997	0.882	0.461	0.466	0.508

Tabla 3.4: Rendimiento modelo Das et al., elaboración propia

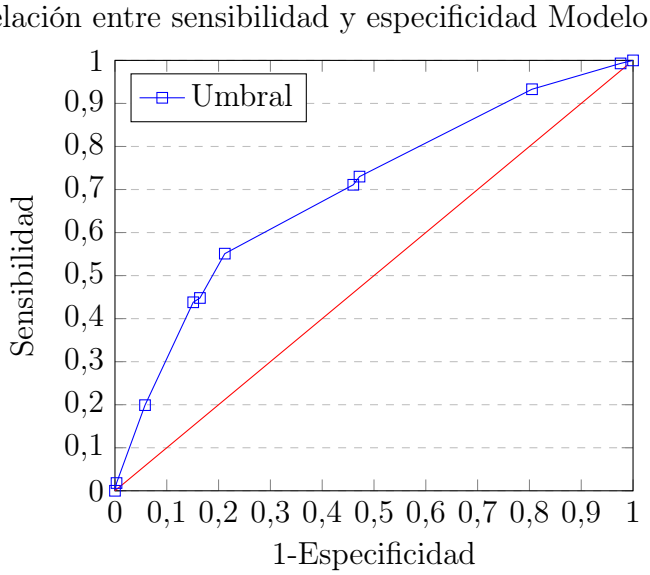

Figura 3.3: Relación entre sensibilidad y especificidad de acuerdo al umbral en el modelo Das et al,elaboración propia

3.1.3. Modelo propuesto por Agyeman et al.

El siguiente modelo a evaluar es el modelo propuesto por Agyeman et al. en 2011. En este estudio se usaron los registros de 423 episodios en 206 pacientes durante los años 2004 y 2007 del multicentro SPOG (*Swiss Pediatric Oncology Group*. Los autores buscaban identificar los factores de riesgo que permitieran predecir posibles episodios de bacteriemia en los pacientes y encontraron que *hemoglobina, plaquetas, escalofríos y necesidad de un tratamiento intrahospitalario* eran factores de riesgo para desarrollar bacteriemia [3].

La definición de riesgo para Agyeman et al. contemplaba un umbral de 3 puntos para definir riesgo alto y bajo, pero en términos prácticos significa que un paciente con un factor de riesgo ya es considerado como de alto riesgo. En la tabla 3.5 se puede ver la distribución de puntajes.

Factor de riesgo	Puntaje asignado
Escalofríos	5
Necesidad de cuidados intra-hospitalarios	3
Menos de 50.000 plaquetas por mm3	3
Hemoglobina mayor o igual a 9g/dL	3

Tabla 3.5: Tabla de puntajes de riesgo, elaboración propia en base a [3]

La aplicación sin modificaciones de este puntaje en la base del PINDA fue imposible dado que el factor de riesgo de *necesidad de tratamiento intrahospitalario* no existe en la base de datos y a priori, todo paciente es ingresado de manera intra-hospitalaria por lo que no habría una diferencia entre estos. El otro factor de riesgo no replicable son los *escalofríos* ya que sencillamente no esta en la base de datos como tal. Para hacerse cargo de este problema y aplicar el puntaje se utiliza la variable *Estado general* que sería un aproximado de los escalofríos e incluso de necesidades intra-hospitalarios.

El umbral donde se consigue el mejor AUC es en 8 puntos diferente a los 3 puntos planteados por el autor, de todas maneras, el aumento de AUC se debe al aumento en la especificidad y no en la sensibilidad, por lo que el umbral de 3 puntos sigue siendo más sensato. En la tabla 3.6 se puede ver el rendimiento que tuvo el modelo para los 5 niveles de riesgo que se podían obtener: 3, 5, 6, 8 y 11.

Umbral	Sensibilidad	Especificidad	PPV	NPV	Precisión	AUC
3	0.874	0.224	0.575	0.595	0.579	0.549
5	0.520	0.688	0.668	0.544	0.596	0.604
6	0.468	0.728	0.675	0.532	0.586	0.598
8	0.369	0.855	0.754	0.529	0.589	0.612
111	0.098	0.968	0.788	0.471	0.493	0.533

Tabla 3.6: Rendimiento modelo de Agyeman et al., elaboración propia

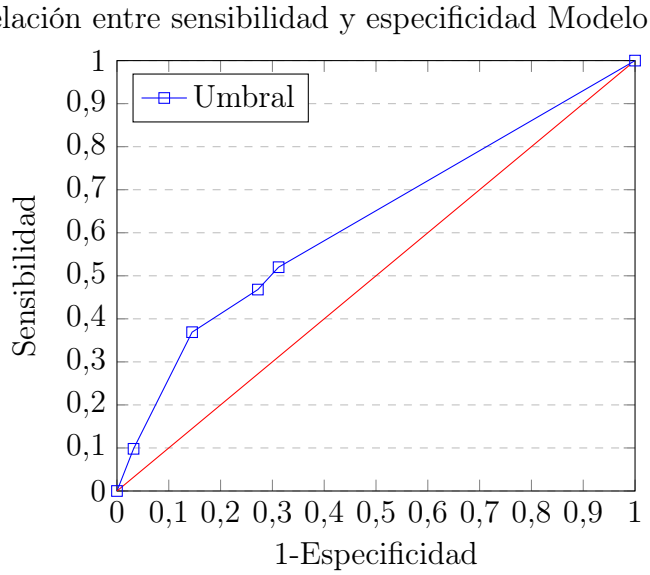

Figura 3.4: Relación entre sensibilidad y especificidad de acuerdo al umbral en el modelo Agyeman, elaboración propia

3.1.4. Modelo propuesto por Hakim et al.

Hakim et al. condujo un estudio que buscaba predecir 2 posibles resultados, Infección Bacterial o cultivo negativo por sepsis (IBD) y complicaciones clínicas [4], En esta memoria solo se aplicara al primer resultado dado que es el más similar a lo que se busca predecir en la memoria. Este estudio utilizo los registros de 390 pacientes hospitalizados por neutropenia febril durante 2 años. Asimismo, este estudio se centro fuertemente en la validación del punto de corte óptimo para estratificar los riesgos. El método que utilizaron fue la técnica de boostraping hasta encontrar el óptimo.

Los factores de riesgo y sus puntajes se muestran en la tabla 3.7. Además, los autores definen como riesgo alto a los pacientes que tienen 24 o más puntos y en caso contrario definen al paciente como de bajo riesgo.

Factor de riesgo	**Puntaje asignado**
Tipo de cáncer: Leucemia Mieloide Aguda (LMA)	20
Tipo de cáncer: Leucemia Linflobastica Aguda (ALL) y Linfomas	7
Estado general	14
Temperatura mayor o igual a 39 grados	11
Recuento absoluto de neutrófilos menor a 100 células/mm3	10

Tabla 3.7: Tabla de puntajes de riesgo Hakim, elaboración propia en base a [4]

Los factores encontrados por Hakim et al. se encuentran disponibles en la base de datos del PINDA, por lo que solo se necesitará hacer las transformaciones de puntajes correspondientes para poder aplicar este modelo. En la tabla 3.8 se puede ver el desempeño del modelo para

cada uno de los umbrales posibles. Como se puede ver, el mejor rendimiento medido en AUC se ubica en el umbral de los 14 puntos, varios puntos menos de lo que plantean los autores como corte óptimo.

Umbral	Sensibilidad	Especificidad	PPV	NPV	Precisión	AUC
7	0.870	0.263	0.584	0.628	0.593	0.566
14	0.456	0.773	0.705	0.544	0.601	0.614
17	0.387	0.832	0.733	0.533	0.590	0.610
24	0.342	0.874	0.765	0.527	0.585	0.608
25	0.069	0.976	0.773	0.468	0.483	0.522
28	0.050	0.983	0.778	0.465	0.476	0.516
41	0.000	1.000	Nula	0.456	0.456	0.500

Tabla 3.8: Rendimiento modelo de Hakim et al., elaboración propia

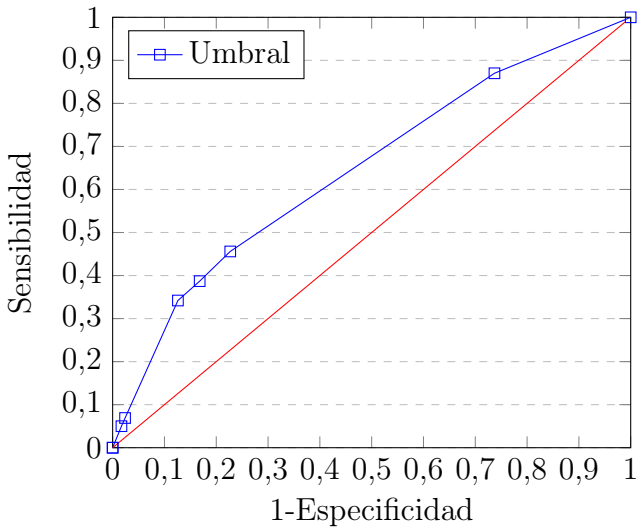

Figura 3.5: Relación entre sensibilidad y especificidad de acuerdo al umbral en el modelo Hakim, elaboración propia

3.1.5. Modelo propuesto por Santolaya et al.

El modelo propuesto por Santolaya et al. es el modelo más relevante para crear el benchmark dado que el centro del modelo es aplicado en la red PINDA de la Región Metropolitana. Esta memoria busca crear un modelo que mejora el rendimiento de ese modelo, utilizando alguna de las variables utilizadas por los modelos propuestos por otros autores.

Para crear este modelo, los autores utilizaron los 447 episodios de neutropenia febril de 257 niños registrados durante un periodo de 17 meses que fueron admitidos en alguno de los 5 hospitales de Santiago. En este estudio se realizaron análisis univariados para identificar las variables más relevantes para predecir una una infección bacteriana invasiva (IBI). Encontraron que las siguientes variables son de riesgo:

1. Proteína C reactiva menor o igual 90mg por litro
2. Recaída de leucemia

3. Hipotensión

4. Conteo de plaquetas menor o igual a 50.000

5. Quimioterapia hace 7 o menos días.

Para determinar si un episodio es de alto riesgo o bajo riesgo se aplica el siguiente criterio: de bajo riesgo si es que el paciente tenia un conteo de plaquetas igual 0 menor a 50.000 o recibió la quimioterapia hace 7 o menos días. Cualquiera de los otros factores por si solo consideraba a los pacientes como de alto riesgo. La aplicación de este modelo a los datos entregados por el PINDA es directa ya que tanto el estudio como la base depende de la misma institución.

Aplicando los criterios mencionados se encuentra un 69,7 % de sensibilidad, 62,1 % de especificidad, 68,8 % de valor predictivo positivo, 63,1 % de valor predictivo negativo, 66,2 % de precisión y un valor de AUC de 0,659.

Además de aplicar la regla propuesta por los autores del estudio, se procede a calcular el rendimiento del modelo para diferentes umbrales de riesgo medido por cantidad de factores riesgos del paciente. Los resultados se pueden ver en la tabla 3.9. El mejor resultado del modelo medido por AUC se logra con los 2 factores de riesgo, que no logra superar la regla usada por la red PINDA, teniendo solo un 0,651 de AUC.

Umbral	**Sensibilidad**	**Especificidad**	**PPV**	**NPV**	**Precisión**	**AUC**
1	0.956	0.134	0.569	0.715	0.582	0.545
2	0.664	0.638	0.687	0.613	0.652	0.651
3	0.210	0.970	0.893	0.506	0.556	0.590
4	0.013	1.000	1.000	0.458	0.462	0.507
5	0.000	1.000	Nula	0.455	0.455	0.500
Regla PINDA	0.697	0.621	0.688	0.631	0.662	0.659

Tabla 3.9: Rendimiento modelo de Santolaya et al., elaboración propia

3.1.6. Modelos no aplicados

Hay dos modelos encontrados en la literatura, la regla de eventos adversos de la SPOG y el modelo de Ammann et al. que no podrán ser aplicado debido a que al primer modelo hay 3 variables que no pueden ser extraídas directamente de la base de datos del PINDA y en el segundo una de las 4 variables utilizadas no esta presente en la base del PINDA, por lo que obviar esa variable podría implicar un sesgo en el modelo.

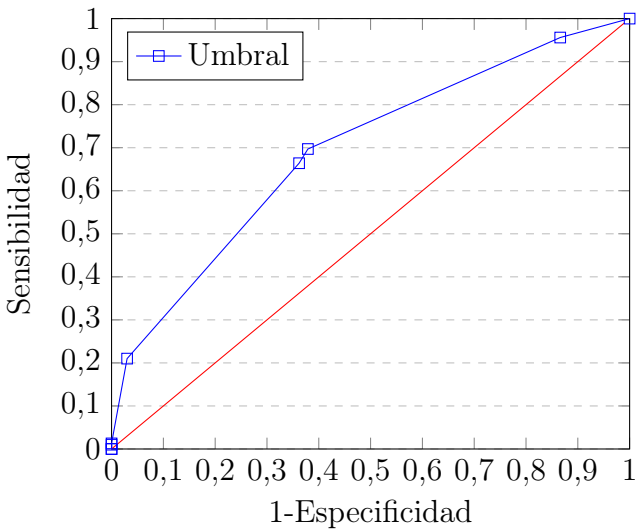

Figura 3.6: Relación entre sensibilidad y especificidad de acuerdo al umbral en el modelo Santolaya, elaboración propia

3.1.7. Resumen modelos aplicados

Variables

Las variables que se utilizaron en cada uno de los estudios comparadas se muestran en siguiente tabla 3.10, ordenados por cantidad de veces que se repite el factor de riesgo. Esta tabla entrega un indicativo de las variables más relevantes para predecir riesgo no obstante esta tabla no muestra la importancia de cada variable en cada modelo. En siguientes secciones se evaluara la importancia de cada variable utilizando métodos envolventes (*wrappers* en inglés)

Factor	**Rondinelli**	**Das**	**Agyeman**	**Hakim**	**Santolaya**	**SPOG**	**Ammann**	**Utilizado**
Hemoglobina	Sí	-	Sí	-	-	Sí	Sí	4
Plaquetas	-	-	Sí	-	Sí	Sí	-	3
Enfermedad de base	-	-	-	Sí	Sí	-	Sí	3
Foco infeccioso	Sí	Sí	-	-	-	-	Sí	3
PCR	-	Sí	-	-	Sí	-	Sí	2
Leucocitos	-	-	-	-	-	Sí	Sí	2
Catéter Venoso central	Sí	-	-	-	-	-	Sí	2
Días ultima quimio.	-	Sí	-	-	Sí	-	-	2
Estado general	-	-	Sí	Sí	-	-	-	2
Temperatura	Sí	-	-	Sí	-	-	-	2
RAN	-	Sí	-	Sí	-	-	-	2
Desnutrición	-	Sí	-	-	-	-	-	1
Edad	Sí	-	-	-	-	-	-	1
Tipo Quimioterapia	-	-	-	-	-	Sí	-	1
Hipotensión	-	-	-	-	Sí	-	-	1

Tabla 3.10: Variables utilizadas para cada modelo

Análisis de los rendimientos para cada modelo

En esta sección se compararan los modelos a través de la área bajo la curva o AUC por sus siglas en inglés. La métrica AUC es un indicador del rendimiento de los modelos que para cada umbral de clasificación compara la sensibilidad y especificidad del modelo. Este es un buen indicador ya que muestra la proporción entre la razón de verdaderos positivos con la razón entre verdaderos negativos. Con este indicador podemos saber a cuantos enfermos identifico correctamente en comparación a los sanos que catalogo como enfermos, error que se busca evitar con el desarrollo de un nuevo modelo.

Modelo	**Umbral**[1]	**Sensibilidad**	**Especificidad**	**PPV**	**NPV**	**Precisión**	**AUC**
Modelo Indio	Autor (7 puntos)	0,551	0,788	0,755	0,597	0,660	0,670
Santolaya et al.	Autor	0,697	0,621	0,688	0,631	0,662	0,659
Hakim et al.	14 puntos	0,456	0,773	0,705	0,544	0,601	0,614
Agyeman et al.	8 puntos	0,369	0,855	0,754	0,529	0,589	0,612
Rondanelli et al.	2 puntos	0,970	0,032	0,542	0,468	0,54	0,501

Tabla 3.11: Mejor rendimiento para cada modelo

El modelo que obtuvo el mejor AUC fue el modelo de Das et al. Este modelo identifico el 55,1 % de los episodios que se complicaron (Alto Riesgo) y el 77,8 % de los que no se complicaron. No obstante, la sensibilidad del modelo queda en segundo lugar, después de la sensibilidad del modelo de Santolaya. Lo anterior es muestra de que a pesar de que el modelo de Das tenga el mejor AUC no implica que se entregará el mejor tratamiento a los pacientes más críticos. En la figura 3.7 se puede ver la curva AUC graficada para cada modelo.

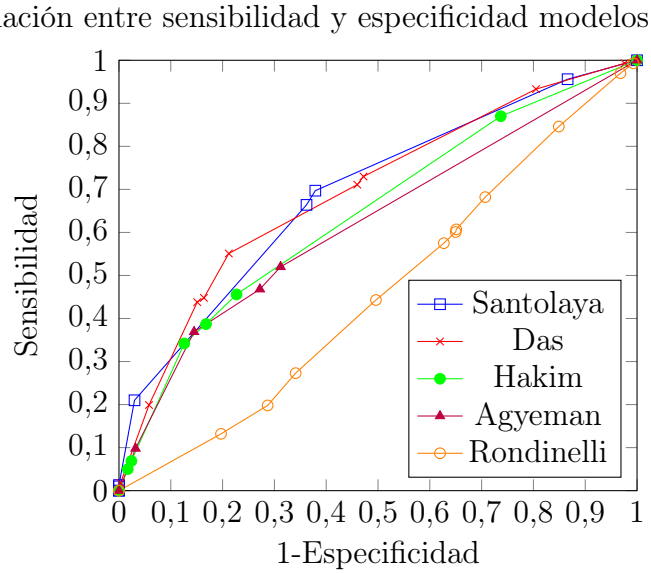

Figura 3.7: Gráfico AUC de los diferentes modelos, elaboración propia

El modelo que le sigue es el propuesto por Santolaya con un AUC de 0,659. Este modelo tuvo una precisión de 66,2 %, 0,2 puntos porcentuales menos que el modelo de Das sin embargo

tiene una mejor sensibilidad que el modelo de das sacrificando la detección de los casos con bajo riesgo. Es por lo anterior que este modelo entregaría una clasificación más segura para los pacientes más vulnerables pero también le entregaría un tratamiento innecesariamente más agresivo a más pacientes no vulnerables.

Como se puede ver, en la literatura se pueden encontrar múltiples formas de clasificar a los pacientes en alto y bajo riesgo. Hay modelos que consideran solo un factores de riesgo como el modelo de Santolaya y el de Agyeman en la práctica. Otros modelos como el de Das, Hakim , Agyeman, Ammann, SPOG y Rondanelli asigna cierto puntaje para cada factor de riesgo y construyen una regla a partir de ese puntaje para definir los casos de riesgo. Lo anterior indica que la aproximación usual de estos estudios no ha podido llegar a ninguna conclusión final acerca de las variables y reglas que permitan predecir el riesgo de neutropenia febril.

Teniendo en consideración lo anterior, en las siguientes secciones se probaran diferentes combinaciones de conjuntos de variables que serán aplicados en modelos de clasificación buscado generar el mejor modelo posible. Asimismo, se aplicarán técnicas de selección de características a la base del PINDA para identificar las variables que podrían llegar ser relevantes en un modelo de clasificación, sin tener necesariamente una justificación académica para el uso de esta.

3.2. Selección y creación de conjuntos de variables para los modelos

Uno de los desafíos cuando se crean modelos de machine learning es la selección de variables que se utilizaran en el modelo. Lo anterior es un desafió ya que dependiendo de la cantidad de variables, la cantidad de combinaciones posibles que se podrían crear podría llegar a las billones de combinaciones, cantidad que incluso con maquinas con alto poder de procesamiento y funcionamiento en paralelo sería imposible de calcular en un periodo corto de tiempo. Para ejemplificar la cantidad de combinaciones posibles que se podrían hacer se mostraran a continuación el conjunto inicial de variables y el cálculo de combinaciones posibles.

Luego de hacer una preselección de variables, que tiene relación con el filtrado de aquellas variables que funcionan como identificadores como por ejemplo `rut o año de ingreso` y aquellas variables que se escapan al alcance de esta memoria como las variables medidas a las 24 horas del ingreso y no al ingreso mismo, por ejemplo `Temperatura mínima o presión sistólica mínima`. También se eliminaron aquellas variables con más del 95 % de valores faltantes. El subconjunto final de variables quedo con 72 variables y es el siguiente:

1. `Abdominal`
2. `Antecedente IFI`
3. `CVC`
4. `Calcemia`
5. `Cardíaco`
6. `Coriza previa al ingreso`
7. `Creatinemia`
8. `Crépitos`
9. `Dificultad respiratoria`
10. `Edad`
11. `Enfermedad de base`
12. `Estado General`
13. `Estertores`
14. `FC`

15. FR
16. Factores de crecimiento
17. Faringe congestiva
18. Foco clínico de ingreso
19. Genito urinario
20. Glasgow
21. Glicemia
22. HCO3-
23. HG
24. Horas de fiebre previas al ingreso
25. Hospital
26. IBI
27. IL -8
28. K
29. Llene capilar (seg)
30. Mucosa ano genital
31. Mucosa oral
32. Nacionalidad
33. Neuro
34. Nitrógeno ureico
35. Nivel educacional de la madre
36. Número días desde última QMT
37. Odinofagia
38. Osteo articular
39. PAD
40. PAM
41. PAS
42. PCO2
43. PCR1
44. PO2
45. Peso
46. Piel
47. Plaquetas
48. Profilaxis
49. Protombina
50. RAM1
51. RAN 1
52. RX TX
53. Respiratorio alto
54. Respiratorio bajo
55. Rto. Absoluto leucocitos día 1
56. Rto. Absoluto linfocitos día 1
57. Saturación oxígeno
58. Servicio en que consulta
59. Sexo niño
60. Temperatura
61. Talla
62. Tejidos blandos
63. Tos previa al ingreso
64. Virus respiratorio
65. Hábil
66. Día de la semana

Además se agregaron las siguientes variables que son construcciones a partir de las anteriores:

1. BMI
2. Sobrepeso
3. Desnutrición
4. Grupo de enfermedades
5. Fin de semana
6. Diferencia FC y FR

Finalmente, el conjunto inicial de variables independientes tiene 72 variables. Con dichas 72 variables, se podrían crear 4.425125402768368e+20 combinaciones posibles si es que los conjuntos fueran de 36 elementos. Si cada combinación posible se probase en un modelo y el tiempo de ejecución del modelo fuese de 1 nano segundo (10^{-9} segundos), se necesitarían más de 14 mil años para probar todos los modelos. Es por lo anterior que es tan importante hacer la selección de variables, antes de probar los modelos.

3.2.1. Modelo utilizando variables encontradas en la literatura

Antes de realizar una selección de variables a través de métodos de filtro y envolventes, se utilizaran las variables utilizadas por los investigadores que plantearon las estratificaciones de riesgo encontradas en la revisión del estado del arte. Este modelamiento es un paso intermedio entre lo que las investigaciones han encontrado relevantes y un modelo nuevo basado en algoritmos de aprendizaje automático.

Las variables utilizadas en los modelos anteriores son las siguientes: *Hemoglobina, plaquetas, enfermedad de base, proteína reactiva C (PCR), presencia de catéter venoso central (CVC), número de días desde la ultima quimioterapia, estado general del paciente, temperatura, recuento absoluto de neutrófilos (RAN), hipotensión y desnutrición*. Esta variables serán probadas para los 7 algoritmos propuestos, evaluando su área bajo la curva (AUC) y precisión, usando exclusivamente las variable como se mencionan en las investigaciones.

Otro punto importante con respecto a estas variables es que hay tres variables encontradas en la literatura a las que se le puede hacer un tratamiento extra basado en lo expresado por la red PINDA y de la forma en que fue construida la variable. La primera es la enfermedad de base, la cual dado el conocimiento médico actual se puede agrupar en 3 grandes grupos de enfermedades.

1. **Grupo 1:** LLA, LNHB, LNH no B, LH
2. **Grupo 2:** Sarcoma PB, Neuroblastoma, Wilms, Sarcoma Ewing, Osteosarcoma, Retinoblastoma, Tumor SNC, Recaída Tumor sólido, Otro
3. **Grupo 3:** LLA recaída, LMA recaída y LNLA

Estos grupos serán utilizados en vez las enfermedades base por si solas. Las otras dos variables son variables categóricas que pasaran a ser continuas. La primera es desnutrición, la cual se transformara en *Índice Masa Corporal (IMC)* y la segunda es hipotensión, la cual se transformara en *presión sistólica*

Los resultados de 7 algoritmos aparecen en la tabla 3.12. En esta tabla aparecen las variables tal cual aparecen en la literatura, cambiando una variable a la vez y cambiando las tres variables al mismo tiempo. Cabe mencionar que tanto sensibilidad como especificidad son parámetros que dependen del umbral elegidos entre las clases, fenómeno que es capturado por la curva AUC, por lo que solo se anotara este valor.

Modelo	**AUC literatura**	**AUC IMC**	**AUC PAS**	**AUC Grupo**	**AUC Todas**
Logístico	0.783	0.785	0.785	0,746	0.785
Random Forest	0.763	0.753	0.753	0,750	0.748
Naive Bayes	0.707	0.744	0.744	0,711	0.744
Arbol de decisión	0.708	0.708	.708	0,708	0.708
Multilayer Perceptron	0.716	0.725	0.719	0,562	0.709
SVM	0.754	0.760	0.760	0,503	0.760

Tabla 3.12: Mejor rendimiento para cada modelo

Se puede ver en la tabla que los dos mejores algoritmos para la clasificación de esta condición es la clásica regresión logística y un método relativamente nuevo como *Random Forest*, siendo levemente superior este ultimo. Otro punto relevante es que la transformación de variables no mejoro significativamente los resultados del modelo pudiendo ser omitida esta transformación, sin embargo, dado que la transformación de variables categóricas a continuas permite capturar más información, se mantendrán continuas las variables de hipotensión y desnutrición. Para el caso de las enfermedades de base, dado que son muchos, se mantendrá la agrupación de las enfermedades para mejorar la eficiencia en la ejecución del modelo.

3.2.2. Métodos de filtro

Filtro por variabilidad

El primer filtro a utilizar es la variabilidad dentro de cada columna. Este filtro cuenta la cantidad de registros idénticos y los compara con los registros diferentes de la misma variable. El umbral a utilizar es del 98 %, es decir si un valor se repetía en más del 98 % de los casos, la columna a la que pertenecía el registro es eliminada.

En la tabla 3.13 se pueden ver las variables que se encontraron con la mayoría de sus registros idénticos y son las siguientes: `Neuro, Genito urinario, Osteo articular y algunos focos de infección, enfermedades, nacionalidades, virus y RX TX`. Las variables que se describen aquí fueron eliminadas de la base. No obstante, se crea una nueva variable llamada `nacionalidad` que evita eliminar las diferentes nacionalidades al agrupar las naciones diferentes a la chilena en la clase extranjero (codificada como 1) y manteniendo la clase chilena (codificado como 0).

Variables	**Registros iguales**	**Registros diferentes**	**Porcentaje**
Neuro	988	9	0,903 %
Genito urinario	982	15	1,505 %
Osteo articular	987	11	1,102 %
Foco CVC	1611	13	0,800 %
Foco Cardiaco	1623	1	0,062 %
Foco Esofagitis	1622	2	0,123 %
Foco Mucositis oral-anal	1616	8	0,493 %
Foco Osteoarticular	1622	2	0,123 %
Foco Urinario	1615	9	0,554 %
Nacionalidad 3	1619	5	0,308 %
Nacionalidad 4	1621	3	0,185 %
Nacionalidad 6	1623	1	0,062 %
Enfermedad 4	1608	16	0,985 %
Enfermedad 5	1621	3	0,185 %
Virus 3	1607	17	1,047 %
Virus 5	1610	14	0,862 %
Virus 9	1614	10	0,616 %
Virus 10	1621	3	0,185 %
Virus 12	1610	14	0,862 %
Virus 16	1616	8	0,493 %
Virus 17	1621	3	0,185 %
RX TX 3.0	1608	16	0,985 %
RX TX 4.0	1617	7	0,431 %

Tabla 3.13: Variables eliminadas a través del método de variabilidad, elaboración propia

Cabe mencionar de la tabla anterior que la suma de registros iguales con registros diferentes no es igual para todos los casos dado que existen casos con nulos en las columnas, lo que evita que sumen lo mismo en cada una de ellas.

Filtro por correlación

La siguiente forma de filtrar es a través de calcular las correlaciones de cada una de las variables. Si una variable esta muy correlacionada con otra, quiere decir que una de las variables podría ser redundante, lo que agregaría costo computacional y podría significar que el modelo se sobre ajuste a los datos entregados, evitando la generalización de este.

Para medir la correlación entre variables continuas se utilizaron 2 test de correlaciones, el test de Spearman y Pearson. El primero mide la linealidad de correlación y el otro la monoticidad en la correlación. Para considerar si una variable se sacaba o no del conjunto de datos se selecciono un umbral de 0,75 de correlación absoluta. Ambos test entregaron las mismas variables, `Talla, Peso, Presión Arterial Media y diferencia entre FC y FR`, lo cual tiene sentido. Un análisis más detallado entrega a que estas variables están correlacionados con el Índice de Masa Corporal y las presiones sistólicas y diastólicas.

Un aspecto relevante encontrado al hacer el test de correlación de Spearman es que `RAN, RAM, Leucocitos y Linfocitos` están correlacionados positivamente y estos a su vez están correlaciones inversamente con la interleucina 8.

Test Chi Cuadrado

El siguiente filtro es el *test chi cuadrado* que busca comparar el comportamiento de las distribuciones de las variables independientes con el de la variable dependiente. Si se encuentra que las variables independientes tienen una distribución independiente de la distribución de la variable dependiente, se procede a su descarte.

En particular, el test compara la frecuencia esperada de cierta variable y la compara con la frecuencia real. La hipótesis nula de este test es que ambas distribuciones son independientes, en caso de que la hipótesis nula no pueda ser rechazada, se estima que las variables son dependientes. Si las variable son independientes, las variables no se pueden utilizar dado que no ayudan a predecir el resultado, en cambio si son dependientes si podrían ayudar a determinar el resultado.

El primer paso para calcular este test es crear una tabla de contingencia que cuenta las observaciones de cada una de las variables categóricas con la variable dependiente IBI, creando su correspondiente frecuencia. Con esta tabla de contingencia se puede calcular el test, el cual entrega la significancia para cada una de las variables. Lo normal es fijar un nivel de significancia y eliminar las variables que están por sobre dicho nivel, sin embargo, en este caso, se observaran las 15 variables con el peor nivel de significancia y se evaluara su eliminación. Las variables se encuentran en la tabla 3.14

Variable	P-value
Virus 8	0,778
Miércoles	0,805
Martes	0,827
habil	0,831
Respiratorio bajo	0,838
Enfermedad 7	0,854
CVC	0,863
Virus 6	0,870
Virus 2	0,890
Nacionalidad	0,899
Enfermedad 12	0,913
Foco Respiratorio bajo	0,920
Virus 13	0,939
Hospital 3	0,956
Nivel educacional de la madre	0,976

Tabla 3.14: 10 variables con menor significancia en chi-cuadrado, basada en las bases entregadas por el PINDA

Como se ve en la tabla anterior, las variables `miércoles y martes` tienen una distribución independientes de la distribución de la variable dependiente con un alto nivel de significancia, hecho que provoco que también se observaran las otras variables de días de la semana y se encontró que también tienen distribuciones independientes, por lo que se eliminaran también los otros días de la semana. Lo mismo sucede con la variable `Hospital 3` que es uno de los hospitales de la red y se observa que el resto también tiene una distribución independiente por lo también se eliminan.

Luego de este filtro, las variables eliminadas son `Todos los días de la semana, los hospitales , nivel educacional de la madre, CVC, día hábil, respiratorio bajo, Foco respiratorio bajo, enfermedad 7 y 12, virus 2, 6 y 8, y nacionalidad.`

Filtro t-test

Con este test se verifica si es que hay diferencias estadísticas relevantes entre la distribución de un valor con respecto a determinada clase. Por ejemplo, se busca saber si es que la distribución del `IMC` entre niños con `IBI` y sin `IBI` es estadísticamente relevante. Si llega a ser relevante, la variable `IMC` podría ser utilizada como predictor de `IBI`, es decir, sería incluida en el modelo. En la tabla 3.19 se ven las variables con menor nivel de significancia a observar.

Variable	P-value
Virus 13	0,906
Presión Arterial Diastólica (PAD)	0,895
Presión Arterial Sistólica (PAS)	0,886
Frecuencia Cardiaca en Reposo (FR)	0,831
K	0,802
Protombina	0,747
Llene capilar (seg)	0,746
HCO3-	0,696
Rto. Absoluto linfocitos día 1	0,696
Profilaxis	0,674
Saturación oxígeno	0,610
Sobrepeso	0,576
Calcemia	0.576
Enfermedad 8	0,528
Virus 15	0,522

Tabla 3.15: 15 variables con menor significancia en t-test, elaboración propia

Las variables que se ven en la tabla 3.19 son las variables que con mayor seguridad tiene medias iguales entre el grupo de pacientes con `IBI` y sin `IBI`, por lo cual no aportan al poder predictivo del modelo. No obstante, la variable `Rto. Absoluto linfocitos día 1` no será eliminada ya que los médicos han indicado que podrían ayudar a predecir la evolución del paciente. Finalmente, se eliminan las siguientes variables: `Virus 13 y 15, Presión Arterial Diastólica (PAD), Presión Arterial Sistólica (PAS), Frecuencia Cardiaca en Reposo (FR), K, Protombina, Llene capilar (seg), HCO3-, Profilaxis, Saturación oxígeno, Sobrepeso, Enfermedad 8 y Calcemia`

Finalmente y luego de los procesos de selección de variables basado en filtros a través de los métodos de la variabilidad entro de las variables, correlación (tanto de Pearson como Spearman), test de Chi-Cuadrado y t-test quedaron con las siguientes 63 variables:

1. `Abdominal`
2. `Antecedente IFI`
3. `BMI`
4. `Cardíaco`
5. `Coriza previa al ingreso`
6. `Creatinemia`
7. `Crépitos`
8. `Desnutrición`
9. `Dificultad respiratoria`
10. `Edad`
11. `Enfermedades 1, 2, 3, 6, 9, 10, 11, 13, 14, 15, 16`
12. `Estado General`
13. `Estertores`
14. `FC`
15. `Factores de crecimiento`
16. `Faringe congestiva`
17. `Foco Mucositis`
18. `Foco Mucositis anal`
19. `Foco Mucositis oral`
20. `Foco Piel y tejidos blandos`
21. `Foco Respiratorio`
22. `Foco Respiratorio alto`
23. `Foco SDA`
24. `Foco Sin foco`
25. `Foco Tiflitis`
26. `Glasgow`
27. `Glicemia`
28. `HG`
29. `Horas de fiebre previas al ingreso`
30. `IL -8`

31. Mucosa ano genital
32. Mucosa oral
33. Nacionalidad 1 y 2
34. Nitrógeno ureico
35. Número días desde última QMT
36. Odinofagia
37. Osteo articular
38. PCO2
39. PCR1
40. PO2
41. Piel
42. Plaquetas
43. RAM1
44. RAN 1
45. RX TX 0, 1, 2 y 5
46. Respiratorio alto
47. Rto. Absoluto leucocitos día 1
48. Rto. Absoluto linfocitos día 1
49. Servicio 0, 1 y 2
50. Sexo niño
51. T
52. Tejidos blandos
53. Tos previa al ingreso
54. Virus 0, 1, 4, 7, 11, 14
55. betalactamicos
56. cotrimoxazol
57. finde
58. fluconazol
59. grupo 1, 2, 3
60. hipotensión
61. otro
62. voriconazol

3.2.3. Balanceo y transformación de datos

Antes de seguir con los métodos envolventes es importante mencionar dos etapas del proceso de descubrimiento de datos en base de datos que es el balanceo y la transformación de datos. Lo anterior es importante ya que en los métodos envolventes se aplican modelos estadísticos para ordenar las variables usado en cada uno de los modelos.

En la aplicación de dichos modelo podría llegar a ser necesario el balanceo de datos ya que en el proceso de eliminación de los nulos datos, la proporción entre casos IBI y no IBI podría cambiar. La proporción tolerada entre ambas clases será como máximo 60 % de la clase mayoritaria y el 40 % para la clase con minoritaria. En los casos que esta proporción se incumpla se aplicará la *técnica de sobre muestro de la clase minoritaria* o SMOTE por sus siglas en inglés, la cual crea variables sintéticas entorno a los datos que ya existen de la clase minoritaria, balanceando la muestra.

A su vez, es necesaria aplicar transformación de variables para la aplicación de ciertos modelos, en particular, en aquellos modelos que entregaran coeficientes y se necesita que puedan ser comparable entre ellos para llegar a un conclusión. Estos métodos serán aplicados dependiendo del contexto de problema y serán explicados en su respectiva sección.

3.2.4. Métodos envolventes

Los métodos envolventes son aquellos métodos en los cuales se ordenan las variables según su importancia en la predicción del fenómeno a predecir, es decir, necesitan la aplicación de un modelo para evaluar la importancia de cada variable. Por ejemplo, en un regresión logística serían los coeficientes del modelo y su peso las características que ordenarían a las variables.

Esta etapa será la continuación de la reproducción de los modelos encontrados en la

literatura, ya que, a pesar de que se cálculo el rendimiento para cada modelo, no se identifico la importancia de cada una de las variables. No obstante, dicha sección entrego cuales eran los algoritmos de mejor rendimiento y se encontró que eran las regresiones logísticas y los *arboles aleatorios* los mejores modelos para predecir el resultado. Dado que para encontrar la importancia de los coeficientes en los *arboles aleatorios* se necesitan aplicar otras técnicas de aprendizaje de maquinas, lo cual haría más engorroso el proceso. Es por lo anterior que se utilizará la regresión logística, que es un método rápido para encontrar la importancia relativa de los coeficientes.

Esta sección tendrá dos partes. En la primera se medirán la importancia de las variables para cada uno de los modelos encontrados en la literatura. En la segunda, se utilizaron los métodos inductivos y deductivos para encontrar las mejores 20 variables para predecir los episodios de IBI. Se considero este número es porque de acuerdo a la literatura ningún modelo uso más 10 variables, siendo un 20 un número razonable para comenzar el análisis. Cabe mencionar que una de las consideraciones del problema es mantener las variables que hoy utiliza la red PINDA para predecir.

Métodos envolventes en los modelos encontrados en la literatura

1. **Modelo de Rondinelli:** Este modelo utiliza las variables de *edad, presencia de catéter venoso central, Existencia de foco clínico de ingreso, temperatura, hemoglobina e infección en el tracto respiratorio superior*. A las 3 variables continuas de este modelo se le aplico una transformación MinMax, estandarizando su valor entre 0 y 1, de manera de compararlos con las otras variables categóricas del problema. El orden por importancia absoluta de las variables calculadas por la regresión logística es la siguiente:

Variable	**Valor absoluto coef.**	**Dirección del coef.**
Edad	0,902	Positivo
Temperatura	0,870	Positivo
Respiratorio alto	0,304	Negativo
Hemoglobina	0,291	Positivo
Sin foco	0,109	Negativo
Catéter Venoso Central	0,017	Negativo

Tabla 3.16: Coeficientes de la regresión logística, modelo Rondinelli, elaboración propia

   Se puede ver que entre mayor sea la edad, la temperatura y hemoglobina mayor es la probabilidad de desarrollar IBI. Las otras variables se mueven la dirección opuesta, es decir si el foco infeccioso esta en el tracto superior, hay presencia de un CVC o no hay foco infeccioso hay menos riesgo de IBI. Otro punto relevante es que la curva AUC paso de 0,501 a 0,575 al hacer la transformación MinMax en las variables continuas en vez de estratificar según los puntajes que el autor propuso.
2. **Modelo de Agyeman:** Este modelo utiliza las variables de *Escalofríos (se usara estado de salud como aproximación), hemoglobina, plaquetas y otras necesidades de cuidado*. El orden por importancia absoluta de las variables calculadas por la regresión logística es la siguiente:

Variable	Valor absoluto coef.	Dirección del coef.
Plaquetas	2,420	Negativo
Estado General	1,083	Positivo
Hemoglobina	0,186	Positivo

Tabla 3.17: Coeficientes de la regresión logística, modelo Agyeman, elaboración propia

Como se ve en la 3.17 entre más plaquetas menos probable es desarrollar IBI. En caso de que el paciente tenga un Estado General malo y un alto nivel de hemoglobina hace que sea más probable una IBI. La hemoglobina se confirma como un factor que aumenta la probabilidad de desarrollar IBI. En este caso, el rendimiento del modelo también mejoro al pasar de los puntajes del autor a una clasificación continua, desde un 0,612 a 0,669.

3. **Modelo de Hakim:** Este modelo utiliza las variables de *tipo de cáncer, mal estado o toxicidad al ingresar al hospital, fiebre en la presentación y plaquetas*. El orden por importancia absoluta de las variables calculadas por la regresión logística es la siguiente:

Variable	Valor absoluto coef.	Dirección del coef.	Grupo enfermedad
RAN	1.464	Negativo	-
Enfermedad 2	1.114	Positivo	Grupo 3
Estado General	1.093	Positivo	-
Enfermedad 13	0.927	Positivo	Grupo 3
Temperatura	0.857	Positivo	-
Enfermedad 5	0.782	Negativo	Grupo 1
Enfermedad 3	0.519	Positivo	Grupo 1
Enfermedad 10	0.501	Positivo	Grupo 2
Enfermedad 14	0.480	Positivo	Grupo 3
Enfermedad 16	0.244	Negativo	Grupo 2
Enfermedad 6	0.207	Negativo	Grupo 2
Enfermedad 12	0.186	Positivo	Grupo 2
Enfermedad 4	0.060	Negativo	Grupo 1
Enfermedad 11	0.053	Positivo	Grupo 2
Enfermedad 9	0.053	Negativo	Grupo 2
Enfermedad 1	0.047	Positivo	Grupo 1
Enfermedad 7	0.020	Negativo	Grupo 2
Enfermedad 8	0.016	Negativo	Grupo 2
Enfermedad 15	0.002	Positivo	Grupo 2

Tabla 3.18: Coeficientes de la regresión logística, modelo Hakim, elaboración propia

En la tabla 3.18 se observa el peso de cada uno de los factores del modelo. Se observa, al igual que en el modelo de Agyeman y Rondanelli, que un mal estado general y una mayor temperatura es un factor de riesgo en el desarrollo de una IBI. Además en este modelo se le suma el recuento absoluto de neutrófilos que entre mayor sea menor es el riesgo de desarrollar IBI.

Dado que este modelo considera el tipo de cáncer, es interesante observar el comportamiento de estos en la predicción de riesgo y con la agrupación que los médicos entregaron. Se observa que el grupo 3 (tipo de cáncer 2, 13 y 14) es el único grupo que tiene comportamiento similar, es decir un paciente agrupado en este grupo siempre tendrá más riesgo que uno que no. En cambio para el grupo 1 y 2, hay resultados dispares, ya que algunos tipos de canceres agregan riesgo, mientras que otros lo reducen, haciendo que el poder predictivo del clasificador no sea tan bueno.

4. **Modelo de Das:** Este modelo utiliza las variables de *desnutrición, tiempo desde la ultima quimioterapia, presencia de foco infeccioso, recuento absoluto de neutrófilos (RAN) y la proteína reactiva C (PCR).* El orden por importancia absoluta de las variables calculadas por la regresión logística es la siguiente:

Variable	**Valor absoluto coef.**	**Dirección del coef.**
PCR	6,420	Positivo
RAN	1,517	Negativo
Número días desde última QMT	1,034	Negativo
Desnutrición	0,336	Positivo
Foco infeccioso no respiratorio alto	0,227	Positivo

Tabla 3.19: Coeficientes de la regresión logística, modelo Das, elaboración propia

En este modelo aparecen dos factores nuevos que ayudan a predecir el riesgo de IBI, estos factores son la PCR y desnutrición. También se ve que el resultado del RAN esta en línea con el resultado del modelo Hakim, siendo una variable predictora de no riesgo de IBI. Otra variable que ayuda a predecir el no riesgo de IBI son los días desde la ultima quimioterapia, es decir, entre más días desde la ultima quimioterapia menos riesgo. Por ultimo si hay un foco infeccioso respiratorio no alto hay mayor riesgo de IBI.

Un punto relevante es que el factor PCR tiene un peso relativo mayor con respecto a las otras variables de los otros modelos, indicando una importancia relevante en la predicción de IBI. El rendimiento de este modelo también aumenta cuando se cambia de los puntajes propuestos por los autores a variables continuas y transformadas, pasando desde 0,670 a 0,737.

5. **Modelo de Santolaya:** Este modelo utiliza las variables de *Recaída de leucemia (Enfermedad 2, 13 y 14), días desde la ultima quimioterapia, proteína reactiva C (PCR), hipotensión y conteo de plaquetas.* El orden por importancia absoluta de las variables calculadas por la regresión logística es la siguiente:

Variable	Valor absoluto coef.	Dirección del coef.
PCR	6,384	Positivo
Plaquetas	1,800	Negativo
Número días desde última QMT	1,436	Negativo
Enfermedad 2	1,091	Positivo
hipotensión	0,996	Positivo
Enfermedad 13	0,857	Positivo
Enfermedad 14	0,418	Positivo

Tabla 3.20: Coeficientes de la regresión logística, modelo Santolaya, elaboración propia

Desde la tabla se puede ver que el conteo de plaquetas y los días desde la ultima quimioterapia son predictores de bajo riesgo de IBI mientras que PCR, enfermedades 2, 13 y 14 e hipotensión son predictoras de riesgo de IBI. La PCR de nuevo mostró ser un predictor relevante para el desarrollo de IBI, similar para los días desde la ultima quimioterapia y plaquetas que tanto en el modelo de Das como en el modelo de Agyeman mostraron ser predictoras de bajo riesgo. Hipotensión mostró ser un nuevo predictor de IBI, sumando un nuevo factor relevante. Por ultimo, el rendimiento mejoro desde 0,659 a 0,767 cuando se utilizaron variables continuas transformadas en vez de la puntuación planteada por el autor.

Como conclusión de esta sección tenemos que en general los modelos aumentan su rendimiento luego de hacer las correspondientes transformaciones a las variables continuas y que una clasificación limpia, es decir basada en un umbral de probabilidad, entrega una mejor clasificación que la puntuación planteada por los autores de los correspondientes modelos. En la tabla 3.21 se puede ver la dirección de cada una de las variables en cada uno de los modelos, además se hace una clasificación global de la dirección de la variable.

Factor	Rondinelli	Das	Agyeman	Hakim	Santolaya	Dirección global
Hemoglobina	(+)	-	(+)	-	-	(+)
Plaquetas	-	-	(-)	-	(-)	(-)
Foco infeccioso	(+)	(+)	-	-	-	(+)
PCR	-	(+)	-	-	(+)	(+)
Catéter Venoso central	(-)	-	-	-	-	(-)
Días ultima quimio.	-	(-)	-	-	(-)	(-)
Estado general	-	-	(+)	(+)	-	(+)
Temperatura	(+)	-	-	(+)	-	(+)
RAN	-	(-)	-	(-)	-	(-)
Desnutrición	-	(+)	-	-	-	(+)
Edad	(+)	-	-	-	-	(+)
Hipotensión	-	-	-	-	(+)	(+)
Enfermedad 2	-	-	-	(+)	(+)	(+)
Enfermedad 13	-	-	-	(+)	(+)	(+)
Enfermedad 14	-	-	-	(+)	(+)	(+)

Tabla 3.21: Variables utilizadas para cada modelo y su correspondiente dirección

De la tabla 3.21 se observa que `hemoglobina` alta, un `foco infeccioso identificado`, `PCR` alta, `estado general` malo, `temperatura` alta, estar `desnutrido`, tener mayor `edad`,

`hipotensión` y alguna de las `enfermedades 2, 13 o 14` son los factores de riesgo de un neutropenia febril. En cambio, un alto nivel de `plaquetas`, presencia de `catéter venoso central`, mayor cantidad de `días desde la ultima quimioterapia` y `RAN` alto son factores que disminuyen el riesgo.

Ahora bien, la aplicación conjunta de todas las variables anteriores entrega resultado mostrado en la tabla 3.23. Se puede ver en dicha tabla que la dirección de cada uno de las variables medidas coincide con la aplicación de los modelos individuales encontrados en la literatura. Otro hallazgo interesante es que las variables se podrían dividir en 4 grupos:

1. **Grupo 1:** Esta compuesto por la proteína reactiva C, que tiene un coeficiente que esta muy por sobre los demás coeficientes, siendo por lejos el mejor predictor.
2. **Grupo 2:** Esta compuesto por las variables RAN, plaquetas, Días desde la ultima quimioterapia, enfermedad 2, Estado general, hipotensión, enfermedad 13 y temperatura. Estas variables tienen coeficientes en torno al valor 1 y a pesar de que no tienen el mismo poder predictivo que la PCR, aportan en la predicción.
3. **Grupo 3:** Tiene las variables desnutrición, enfermedad 14, hemoglobina y edad, que tiene coeficientes bajos pero aún así entregan algo de poder predictivo.
4. **Grupo 4:** Es CVC y Sin Foco y casi no aportan poder predictivo al modelo, por lo que podrían ser eliminadas sin problemas.

Variable	**Valor absoluto coef.**	**Dirección del coef.**
PCR	6.067	Positivo
RAN	1.320	Negativo
Plaquetas	1.256	Negativo
Número días desde última quimioterapia	1.136	Negativo
Enfermedad 2	1.076	Positivo
Estado General	0.943	Positivo
hipotensión	0.756	Positivo
Enfermedad 13	0.742	Positivo
Temperatura	0.734	Positivo
Desnutrición	0.299	Positivo
Enfermedad 14	0.261	Positivo
Hemoglobina	0.251	Positivo
Edad	0.248	Positivo
Catéter Venoso Central	0.067	Negativo
Sin foco	0.057	Positivo

Tabla 3.22: Variables utilizadas para cada modelo y su correspondiente dirección

Cabe destacar que la curva AUC del modelo fue de 0,786, superior a lo mostrado por el mejor modelo por separado que fue el modelo de Santolaya con un área bajo la curva de 0,767. En la siguiente sección se tomaran como base las variables correspondientes a los 3 grupos con mejores rendimientos y a dicha base se le agregaran variables en la medida que la agregación de dicha variable mejore el poder predictivo del modelo.

Métodos envolventes en todas las variables.

Los métodos envolventes como se ha mencionado en el marco teórico son aquellos métodos que utilizan algún algoritmo para ordena las variables entre menos a más importantes. Dentro de los métodos envolventes está el inductivo y el deductivo. El primero mide el rendimiento del modelo en la medida en que se le agregan variables que más mejoren el rendimiento del modelo y el segundo, toma todas las variables y elimina aquellas variables que permitan mejorar el rendimiento del modelo.

En el caso del método inductivo, se utilizará como base las variables encontradas en la literatura y esté conjunto de variables se le agregaran variables y se seleccionara aquel conjunto que maximice el AUC. Para el caso del deductivo, se tomaran todas las variables y se irá eliminando de una en una hasta que quede el modelo base de la literatura. Para ambos casos, la cantidad mínima de episodios para que el modelo sea valido será de 300 o más episodios.

Al realizar el proceso inductivo de selección de variables, es decir, tomar las variables de base encontradas en la literatura y agregando variables en la medida que se aumenta el rendimiento del modelo, se encontraron las siguientes variables que al ser agregadas aumentaban el rendimiento del modelo:

`Virus 14, Foco Mucositis, Crépitos, Foco Tiflitis, PCO2, RAM, Enfermedad 3, Glicemia, Rto. Absoluto leucocitos día 1, Tejidos blandos, FC, Foco Respiratorio, BMI, RX TX 5 y otro`.

Cuando se incorporan estás variables, quedan 304 registros con todas las variables completas y el porcentaje de registros con `IBI` es de 57.56 % por lo que aún se encuentran en balance las dos clases. Por ultimo, el rendimiento máximo de este modelo es de 0,8421 medido en AUC, bastante superior a lo mostrado por los modelos encontrados en la literatura.

Para el proceso deductivo, se comienza eliminando la cantidad de variables necesarias para que la cantidad de registros sea de al menos 300 casos. Se realiza creando subconjuntos de variables y eliminando dicho subconjunto hasta que en alguno de los procesos de eliminación quede una cantidad superior a 300. Luego de realizar dicho filtro, el subconjunto de variables a eliminar son las siguientes: `PO2, PCO2, Odinofagia, Faringe congestiva, Rto. Absoluto linfocitos día 1, Glicemia y Rto. Absoluto leucocitos día 1`.

El proceso de eliminación deductiva consiste en la creación de un modelo inicial con todas las variables de la base de datos para luego ir eliminando de manera iterativa la variable que aumente el rendimiento del modelo, hasta que dicho rendimiento deje de aumentar. El modelo final del proceso anterior mantuvo a las variables de la literatura más las siguientes 16 variables: `Sexo niño, Crépitos, fluconazol, Virus 1, RAM, Enfermedad 3, Calcemia, Tejidos blandos, FC, Virus 7, Nacionalidad 2, Foco SDA, Creatinemia, Factores de crecimiento, Abdominal, BMI, Antecedente IFI, Foco Respiratorio alto`. La base con la cual se cálculo este modelo mantuvo 462 registros y tuvo un rendimiento medido en AUC de 0,8430 mayor al mostrado en el proceso inductivo.

Ahora bien, con el proceso de filtro inductivo y deductivo, se seleccionaron aquellas varia-

bles que en la medida que se agregaban o eliminaban variables mejoraban el rendimiento del modelo. Todas las variables que pasaron este proceso y mejoraron el rendimiento de algún modelo se mantendrán y pasaran a la etapa final de selección de variables. Aquellas variables que no fueron agregadas en ninguno de los dos filtrados serán eliminadas y son las siguientes:

1. Virus 4
2. Tos previa al ingreso
3. Nacionalidad 1
4. Servicio 0
5. Virus 0
6. RX TX 0
7. Nitrógeno ureico
8. Enfermedad 15
9. Servicio 2.0
10. grupo 1.0
11. Mucosa oral
12. betalactamicos
13. Enfermedad 9.0
14. IL -8
15. voriconazol
16. Foco Mucositis oral
17. RX TX 2
18. finde
19. Rto. Absoluto linfocitos día 1
20. PO2
21. Virus 11
22. Enfermedad 16.0
23. Mucosa ano genital
24. grupo 2.0
25. Piel
26. Estertores
27. cotrimoxazol
28. Dificultad respiratoria
29. Foco Mucositis anal
30. Foco Piel y tejidos blandos
31. RX TX 1
32. Glasgow
33. Enfermedad 1.0
34. Cardíaco
35. grupo 3.0
36. Osteo articular
37. Respiratorio alto
38. Faringe congestiva
39. Foco Sin foco
40. Servicio 1.0
41. Enfermedad 11
42. Enfermedad 6
43. Odinofagia
44. Coriza previa al ingreso
45. Horas de fiebre previas al ingreso

Además de las variables anteriormente mencionadas, se agregará Nacionalidad 2 ya que no hay fundamento en la literatura para que determinada nacionalidad sea más susceptible a complicaciones en episodios de neutropenia febril y lo más probable es que la importancia de esta variable sea una casualidad más que tenga verdadero poder predictivo.

Con las variables que pasaron los procesos de filtrado y selección se trabajara en la siguiente sección en donde a través de la creación de combinaciones entre las diferentes variables se elegirá el mejor modelo para clasificar riesgo.

3.3. Desarrollo del modelo de clasificación

La aproximación inicial para el desarrollo del modelo final era la creación de diferentes subconjuntos de variables, con diferentes cardinalidades para cada uno de los subconjuntos pero a pesar de todos los procesos de selección y filtrado, la cantidad de variables a utilizar aún es alto para la creación de los dichos subconjuntos. El conjunto final de variables a elegir tiene 39 variables, que si se combinan en conjuntos de 19 o 20 variables, se podrían crear más de 68 billones de conjuntos posible, cantidad difícil de computar en la realización de esta memoria.

Como método alternativo se utilizaran los métodos inductivos y deductivos para las variables elegidas, pero estos modelos comenzarán desde 2 variables hasta las 39, sin tomar como base las variables encontradas en la literatura. Otro diferencia con la anterior aplicación de estos métodos es que a parte de las regresiones logísticas, se utilizaran bosques aleatorios para calcular el rendimiento de los modelos.

En la figura 3.8 se observa el comportamiento de la aplicación de los métodos. Se observa que los modelos logísticos tienen un bajo rendimiento hasta la variable 10 y luego comienzan tienen un comportamiento más estable entre las 10 y 25 variables, en torno a un AUC de 0,847, para luego tener una leve caída en su rendimiento cuando se agregan todas las variables. En el caso de los bosques aleatorios vemos el mismo comportamiento que en el logístico, pero su comportamiento no es tan estable y tiene un rendimiento en torno a 0,859. Es decir, los bosques aleatorios tienen un comportamiento levemente mejor a las regresiones logísticas.

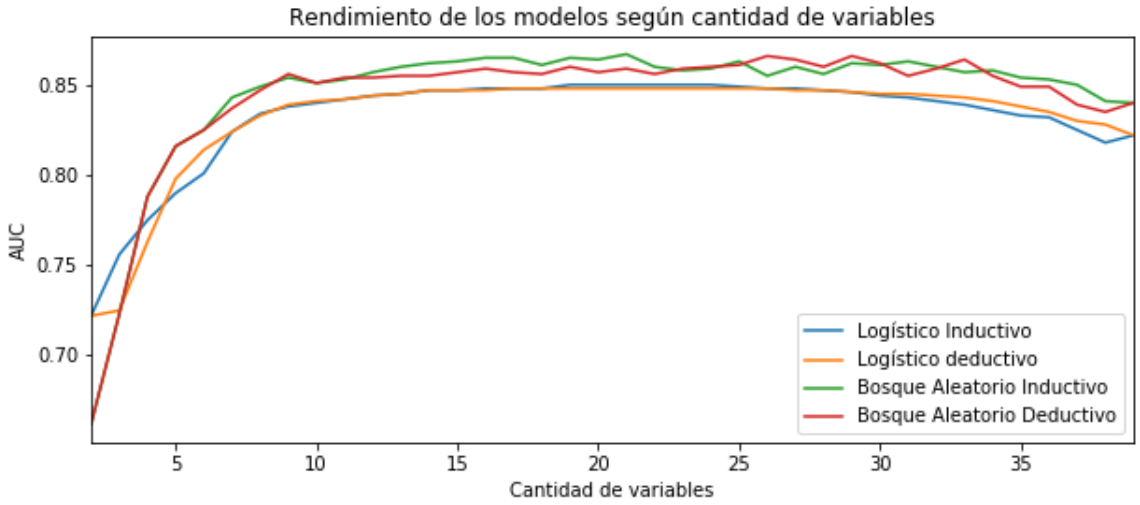

Figura 3.8: Rendimiento por cantidad de variables, elaboración propia

A continuación, se eligen los 3 mejores conjuntos de variables de cada uno de los métodos calculados (inductivo logístico y bosque aleatorio y deductivo logístico y bosque aleatorio) y se analizaran dichas variables y se propondrá el modelo. En la tabla 7.2 de los anexos se pueden ver la cantidad de variables, el rendimiento y las variables de los conjuntos con mejor desempeño.

Se observa que el rendimiento óptimo de los modelos se alcanza entre las 15 y 28 variables implicando que el mejor modelo para predecir IBI se ubica en ese rango de variables A continuación se reportara las variables que más veces aparecen en los diferentes modelos.

Variable	**Apariciones en modelos**
PCR, Enfermedad 2 y Estado General.	12
Enfermedad 3	11
RAN	10
Calcemia, Enfermedad 13.0, Factores de crecimiento, Rto. Absoluto leucocitos	9
Abdominal, RAM, Tejidos blandos	8
BMI, Foco Tiflitis, HG, hipotensión, Sexo, Virus 7	7
Crépitos, Foco Respiratorio alto, Virus 14, Plaquetas	6
Fluconazol y RX TX 5	5
Desnutrición, FC, PCO2, Edad y Creatinemia	4
Antecedente IFI, Enfermedad 14.0, Foco Mucositis, Número días desde última QMT, Glicemia, Virus 1 y temperatura.	3
Otra profilaxis, Foco Respiratorio y Foco SDA	2

Tabla 3.23: Variables utilizadas para cada modelo y su correspondiente dirección

Como se puede ver en la tabla anterior, hay variables que aparecen en todos los modelos como `PCR, enfermedades 2 y estado general` mientras que hay variables que solo aparecen 2 veces en los diferentes modelos como `Otra profilaxis, Foco Respiratorio y Foco SDA.` Para elegir cuales serán las variables finales del modelo se observaran las variables que aparezcan al menos 6 veces en los modelos. Lo anterior deja afuera las variables `fluconazol, RX TX 5, desnutrición, frecuencia cardiaca, PCO2, edad y creatinemia, antecedente IFI, enfermedad 14, foco mucositis, número días desde última QMT, glicemia, virus 1, temperatura, otra profilaxis, foco Respiratorio y foco SDA`. Cabe destacar que variables como `desnutrición, edad, enfermedad 14, número de días desde la ultima quimioterapia, foco respiratorio y temperatura` si aparecían en la literatura pero no pasaron este ultimo filtro.

Las variables que aparecieron en todos los modelos, serán utilizadas en el modelo final ya que pasaron a todos los filtros hechos y estás son `PCR, enfermedad 2 y Estado General`. La discusión ahora, se centrara en las variables que aparecieron entre 6 y 11 veces en los diferentes modelos.

Considerando que hay variables dentro de este rango de repeticiones que ya han sido utilizadas en otros modelos predictivos y hay literatura que las avalan serán agregadas a las variables ya elegidas y son las siguientes `RAN, Enfermedad 13, IMC, plaquetas, hipotensión, hemoglobina y foco respiratorio alto.`

Teniendo 10 variables de base y dejando solo 11 variables a probar (`sexo` se eliminará porque su valor predictivo podría deberse a casualidades) se pueden utilizar técnicas combinatorias para elegir al mejor modelo posible. Al crear todos los subconjuntos posibles de variables, se encontró que las variables que mejoran más el rendimiento del modelo son las siguientes: `Calcemia, Factores de crecimiento, Tejidos blandos, Crépitos, Enfermedad 3, Rto. Absoluto leucocitos día 1 y RAM1`. El conjunto de todas las variables elegidas hacen que el AUC del modelo sea 0.834, rendimiento que es menor a lo encontrado en otras combinaciones de variables, pero se estima, que este modelo al tener 17

variables y ser calculados por un método más determinista, como lo es una regresión logística, podría asegurar una mejor generalización y usabilidad del modelo del modelo.

Ahora bien, una de las restricciones de este modelo es que la sensibilidad no sea inferior a lo que plantean los modelos hoy existentes y en particular por lo indicado por Santolaya en 2001, cuyo modelo declaro tener una sensibilidad del 92 %. Dado que la sensibilidad de un modelo depende del umbral que separa la clases, se variara dicho umbral en el modelo propuesto para que la sensibilidad calculada sea del 92 %.

Para que la sensibilidad del modelo sea del 92 % se encontró que el umbral óptimo esta en un 38,03 % de probabilidad de IBI, es decir, aquellos episodios con una probabilidad mayor al 38,03 % son declarados como de alto riesgo. Al hacer este corte, la especificidad quedo en 44,5 %, valor que es menor que lo planteado por Santolaya en 2001, en donde la especificidad quedo en 76 %.

Dados estos valores se podría concluir que la hipótesis se rechaza ya que no se logro mejorar la especificidad en los casos de neutropenia febril, no obstante, cabe señalar que los valores entregados por Santolaya en 2001 corresponde a otra base de comparación por lo que esta conclusión no es válida.

Por consiguiente, hay que utilizar los valores de sensibilidad y especificidad calculados para esta población en particular, hecha en secciones anteriores. Dichos cálculos muestran que el modelo con mejor AUC tiene una sensibilidad del 55,1 % (Modelo de Das) y el siguiente en AUC tiene una sensibilidad de 69,7 % (Santolaya), por lo que al aplicar las reglas de la literatura, incluida, la regla de Santolaya se tiene una sensibilidad muy por debajo del 92 %.

Es por lo anterior que se elige un nivel de sensibilidad del 90 % para la fijación del umbral de clasificación, sensibilidad que es levemente inferior lo declarado por Santolaya, pero superior a lo encontrado con las reglas de la literatura. El umbral de clasificación en este caso sube a un 40,77 %, dejando una especificidad del 49,17 %.

Capítulo 4

Implementación del modelo de visualización

Luego de la selección de las variables que compondrán el modelo final, se construye un prototipo funcional de una aplicación web que permita recibir los datos ingresados por un usuario, luego que en el backend, el modelo haga los cálculos correspondientes y entregue una clasificación de riesgo acompañada por la probabilidad de IBI. El objetivo de esta plataforma es mostrar de manera simple y directa la utilidad y el flujo lógico del proceso de clasificación de riesgo. Además muestra como una herramienta como esta permite una clasificación más rápida y efectiva ya que no es necesaria la creación de puntajes o reglas más simples para calcular el riesgo como en los otros modelos encontrados en la literatura.

4.1. Descripción del prototipo desde la vista del usuario

El prototipo de aplicación web consiste en 2 vistas. La primera es la vista en donde el usuario ingresa los datos y la segunda es donde el usuario puede observar los resultados del modelo, que muestra la probabilidad de IBI, la clasificación del episodio. A través de este proceso se puede obtener la clasificación de manera rápida y precisa, ayudado a que el paciente sea clasificado de la mejor manera posible.

La vista de ingreso de datos es un formulario en donde aparecen cada una de las variables que calculara el modelo con los correspondientes campos en donde el usuario tendrá que ingresar el valor de la variable del paciente. Los campos son los siguientes:

1. `Rut`: Es el identificador del paciente y del episodio en sí. Con el rut y la fecha se crea una llave que permite diferenciar el episodio de otros. Se construye sin caracteres especiales (sin puntos ni guiones)
2. `Edad`: Es la edad del paciente en meses. Se utiliza para calcular el umbral de hipotensión del paciente.
3. `Peso`: Es el peso del paciente en kilos. Se utiliza para crear el Índice de Masa Corporal

(IMC) del paciente, que es calculado por la aplicación web, evitando que el usuario lo tenga que hacer. Se puede ingresar con decimales.

4. `Talla`: La altura del paciente medido en centímetros. También se utiliza para calcular el IMC.
5. `Recuento absoluto de leucocitos (RAL)`: Es el recuento absoluto de leucocitos (RAL) que presenta el paciente al ingreso. Se tiene que ingresar como entero.
6. `Proteína Reactiva C (PCR)`: Es la medición de la PCR al ingreso medida en mg/l. Se ingresa como entero.
7. `Plaquetas`: Es la cantidad de plaquetas por mm^3 al ingreso. Se ingresa como entero.
8. `Recuento absoluto de monocitos (RAM)`: Es el conteo absoluto de monocitos por mm^3 al ingreso. Se ingresa como número entero.
9. `Recuento absoluto de neutrófilos (RAN)`: Es el conteo absoluto de neutrófilos al ingreso por mm^3. Se ingresa como número entero.
10. `Hemoglobina`: Es la hemoglobina (hg) al ingreso del paciente y se ingresa puede ingresar con decimales.
11. `Presión Sistólica`: Es la presión sistólica al ingreso del paciente, medida en milímetros de Mercurio (mmHg) y se utiliza para definir si el paciente tiene hipotensión o no. Se ingresa como entero.
12. `Calcemia`: Es la cantidad de calcio que hay en la sangre al ingreso del paciente. Se puede ingresar como decimal.
13. `Estado del Paciente`: Es el estado del paciente al ingreso. Puede ser bueno, regular o malo. Si es bueno, la variable ingresa al modelo como 0, en cualquiera de los otros dos casos ingresa como 1, dado que es la forma en que el modelo fue entrenado. Es un checkbox que se despliega con cada una de las 3 opciones: **bueno, malo** o **regular**
14. `Foco respiratorio alto`: Se muestra la pregunta *¿El foco respiratorio es alto?* que tiene 2 respuestas posibles: **sí** o **no**.
15. `Factores de crecimiento`: Se despliega la pregunta *¿Uso factores de crecimiento post quimioterapia?* con dos respuestas posibles:**sí** o **no**.
16. `Tejidos blandos`: Se despliega la pregunta *¿Hay algo anormal en los Tejidos blandos?* con dos respuestas posibles: **sí** o **no**.
17. `Crépitos`: Se muestra la pregunta: *¿Hay crépitos en el paciente?* con dos respuestas posibles: **sí** o **no**.
18. `Enfermedad`: Se despliega la pregunta *¿Cuál es la enfermedad de base del paciente?* con cuatro opciones, **LNLA, LNHB, LLA en recaída** y **otros**. Las primeras 3 opciones entregan un valor de 1 al modelo en caso de ser seleccionadas y 0 en caso de que se seleccione otra.

Al final de esta vista hay un botón de enviar, el cual envía las variables al modelo, el cual calcula la probabilidad de IBI tomando cada una de estas variables y transformándolas cuando sea necesario. Además, entrega la clasificación del episodio gracias al umbral definido anteriormente en esta memoria. La figura 4.1 muestra la primera vista del usuario, mientras que la figura 4.2 muestra el resultado del proceso, indicando la clasificación, luego la probabilidad y para terminar el valor de cada una de las variables.

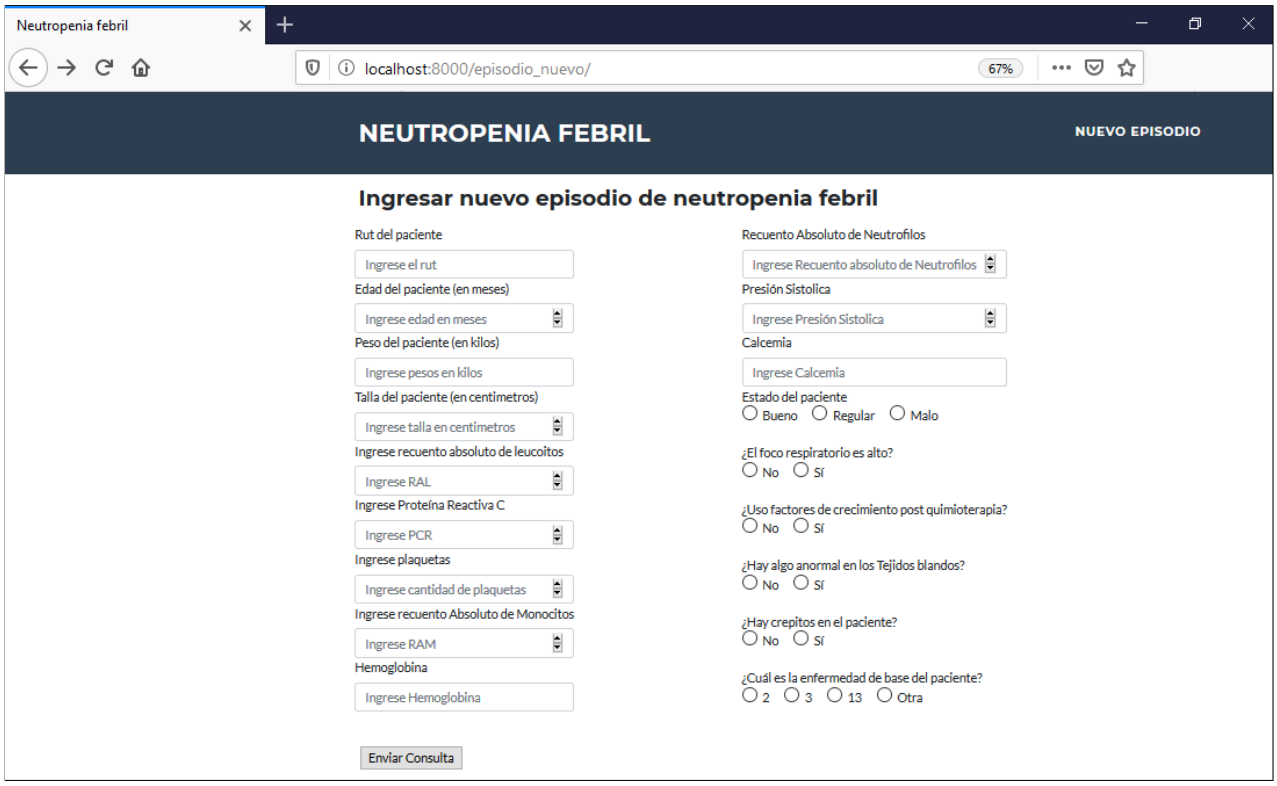

Figura 4.1: Vista del usuario del prototipo funcional al ingresar datos

4.2. Creación del modelo funcional

El prototipo funcional fue desarrollado en Django, un framework de desarrollo web basado en Python. La primera etapa del prototipo fue la creación de un proyecto llamado *neutropenia febril* y una aplicación llamada *clasificación NF*. Además se crea la carpeta *templates* la cual tiene todos los archivos *html* que permitirán la visualización de la página en los navegadores.

Luego de la creación de la estructura general del prototipo, se construye una estructura que maneje la lógica del prototipo. La primera parte de esta lógica incluye el cálculo de los coeficientes de la regresión logística, que es el algoritmo finalmente usado para el cálculo de la probabilidad de IBI. Este calculo incluye las 18 variables relevantes para el cálculo, a saber: `Índice masa corporal, recuento absoluto de leucocitos, monocitos y neutrófilos, proteína reactiva C, plaquetas, hemoglobina, hipotensión, estado del paciente, calcemia, tejidos blandos, crépitos, foco respiratorio alto, factores de crecimiento y enfermedades LNLA, LNHB y LLA en recaída`. Luego de la realización de este cálculo, los coeficientes del modelo quedan fijados y son utilizados para calcular la probabilidad de IBI.

Otra aspecto importante del prototipo funcional es el archivo en donde se crean los formularios. Este archivo permite la creación de las diferentes características de los campos para ingresar los datos de entrada. Por ejemplo, se crea la pregunta "¿Uso factores de crecimiento post quimioterapia?" a lo que se crea un checkbox con las respuestas "si" o "no", respuestas

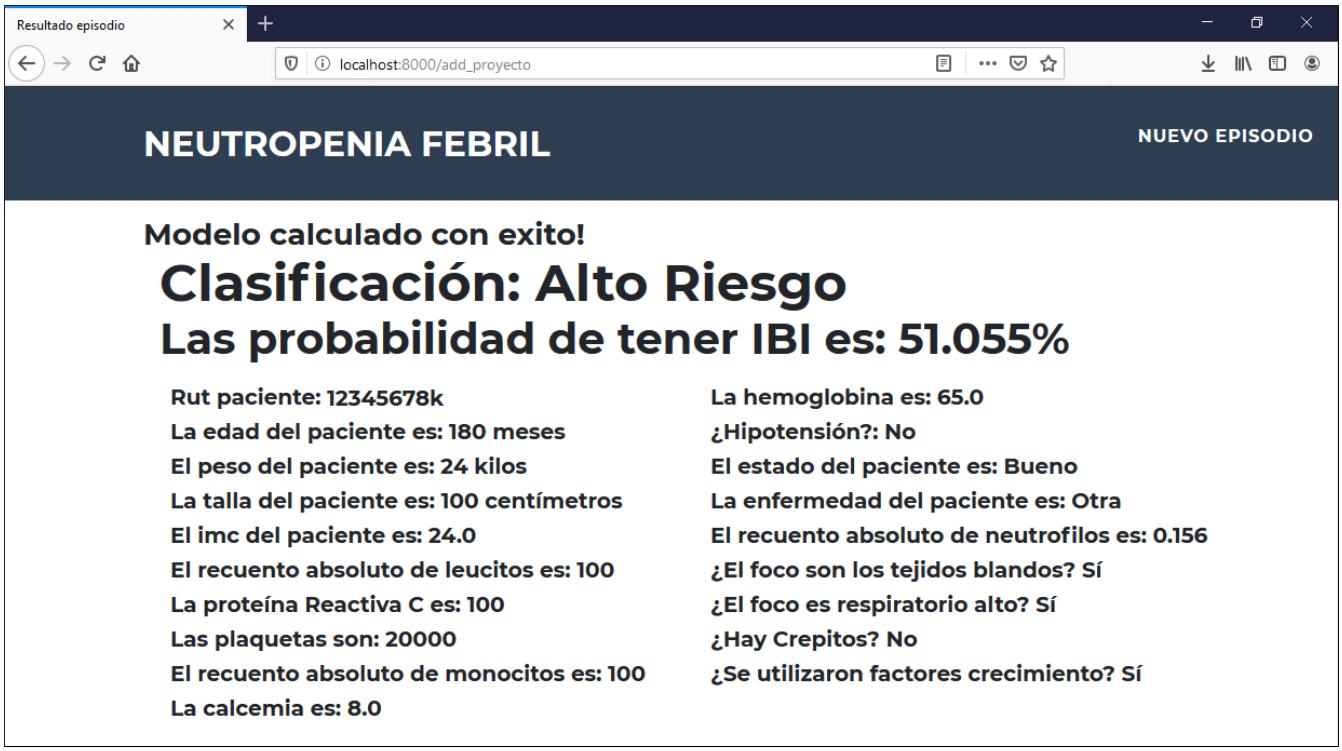

Figura 4.2: Vista del usuario del prototipo funcional al mostrar el resultado de la clasificación

que tienen asociadas los valores 1 y 0 respectivamente, valores que son ingresados al modelo posteriormente.

El siguiente componente del framework de Django, es el archivo *views*, componente en donde fueron calculados los coeficientes del modelo. Además de tener el calculo del modelo, tiene una función principal: `AddEpisodio`, función que recibe los datos ingresados en el formulario, transforma los valores continuos a valores entre 0 y 1. Posterior a esta transformación, los datos de entrada son ingresados al modelo y este entrega la probabilidad de IBI. Todas las variables que se muestran en la segunda vista del usuario, son enviadas al *template new episode*, a través de un diccionario, en donde las llaves tienen asociados los valores a mostrar y el *template* muestra dichos valores en el navegador.

Por ultimo, cabe destacar que este prototipo funcional es una maqueta del proceso lógico de la clasificación de los pacientes en alto y bajo riesgo. El objetivo de este no era ser una herramienta probada y definitoria para el uso en los hospitales ya que se reconoce que es un prototipo básico y su usabilidad inicial podría implicar que no sea utilizado por los usuarios para el cálculo de la clasificación.

En la figura 4.2 se observa el resultado final de la maqueta que es una pagina en donde se muestra la clasificación de riesgo, la probabilidad de riesgo de acuerdo al modelo y por ultimo cada uno de los campos utilizados para calcular el riesgo con su respectiva figura.

También es importante recalcar que la utilización de herramientas nuevas en organizaciones con protocolos establecidos no pasa exclusivamente por la calidad de la herramienta

en términos de usabilidad y capacidad técnica, sino que por la adopción dentro de la organización que es impulsada por agentes claves y con el acompañamiento de capacitaciones y mecanismos que aseguren que su utilización será el nuevo statu quo dentro de la organización, haciendo que el cambio no sea solo a nivel de herramienta, sino que de la cultura también.

Capítulo 5

Estimación del impacto económico del modelo

La presente sección tiene como objetivo estimar el impacto económico y social de la implementación del modelo creado en esta memoria en la red PINDA de la Región Metropolitana, y potencialmente en el país. Para la medición del impacto económico y social se estimaran 2 valores: el **ahorro económico** que genera una mejor clasificación de riesgo en los pacientes con neutropenia febril, evitando estadías más prolongadas para los pacientes reduciendo los gastos asociados a dichas estadías. El otro valor es la cantidad de **días ahorrados** para el paciente al evitar estadías más largas. Cabe mencionar que el presente capitulo es una estimación a grandes rasgos del impacto económico del modelo y no pretende ser un análisis exhaustivo de los posibles costos y beneficios de la implementación del modelo.

5.1. Impacto para la red PINDA metropolitana

Para la estimación del impacto económico y social de la implementación de este modelo hace sentido separar el impacto, primero en la Red PINDA de la RM y luego hacer una extrapolación hacia el resto del país. Lo anterior considerando que se tienen datos exclusivamente de la RM, haciendo que este calculo sea más preciso, ya que, si se hiciese solo a nivel nacional podría generar ruido en la estimación.

El primer paso para estimar el beneficio de la implementación del modelo es calcular una aproximación de los casos de neutropenia febril ingresados en la red PINDA para la posterior extrapolación en los siguientes 10 años. Para realizar lo anterior se agruparon los casos de neutropenia febril entregados por el PINDA entre 2009 y 2015 (2016 no se considero ya que constaba con solo un episodio en todo el año). Los episodios desglosados por mes y año aparecen en la tabla 5.1.

Como se en la tabla 5.1 hay meses con una cantidad baja de episodios, es por lo anterior que no se tomará el promedio anual sino que el promedio mensual sin contar aquellos meses con 3 o menos casos. Lo anterior entrega un promedio de 23,3 mensuales, es decir aproximadamente

Mes	2009	2010	2011	2012	2013	2014	2015
Enero	0	22	14	0	30	22	13
Febrero	0	16	17	0	14	13	21
Marzo	0	20	23	1	26	22	26
Abril	3	28	24	0	27	28	19
Mayo	19	15	32	0	26	29	22
Junio	24	30	26	1	23	30	25
Julio	23	30	19	11	19	38	18
Agosto	21	25	0	32	18	23	19
Septiembre	30	21	0	27	22	20	23
Octubre	23	27	0	32	21	15	25
Noviembre	27	18	0	34	21	27	30
Diciembre	25	24	0	21	23	25	22

Tabla 5.1: Episodios de neutropenia febril por mes durante 2009 al 2015, elaboración propia en base a datos del PINDA

280 casos de neutropenia febril al año. Con esta cantidad se puede proyectar los casos de neutropenia febril para los siguientes 10 años que sumarían cerca de 2.800 nuevos casos ingresados por esta condición en la RM

5.1.1. Ahorro directo por menores costos de tratamiento

Con esta estimación de casos se puede calcular el ahorro directo por costos menores del tratamiento. Cabe recordar, que el ahorro se produce al clasificar correctamente a un paciente de bajo riesgo, haciendo que su tratamiento sea menos invasivo que si se hubiese clasificado como de alto riesgo. A si mismo, el objetivo principal de este modelo es entregar el mejor tratamiento posible a los pacientes, es decir, ni sobre estimar (que aumenta los costos y disminuye la calidad de vida del paciente) ni sub estimar los riesgos (que podría implicar una complicación severa). Teniendo en cuenta lo anterior, siempre se elegirá la sobre estimación ya que esta apunta hacia el cuidado de la vida humana.

Teniendo presente lo anterior, se puede emparejar la **sobre estimación** del riesgo con un alto nivel de **sensibilidad** y un menor nivel de **especificidad** y emparejar la **sub estimación** con un menor nivel de sensibilidad y un alto nivel de especificidad, es por esto, que el impacto económico del modelo dependerá en donde se fije el umbral entre las clasificaciones de riesgo.

Como ya se ha visto en esta memoria, hay dos grupos de lineas bases relevantes, el primero son los niveles declarados por Santolaya (2001) y los segundos son los calculados en esta memoria basados en la literatura. No obstante, considerando que los niveles de Santolaya corresponden a una muestra diferente, no serán utilizados como linea base. Es decir, que como línea base se utilizara alguno de los modelos de la literatura.

Como línea base se utilizara el mejor modelo encontrado en la literatura el cual fue el planteado por Santolaya (2001), que se encuentran en la tabla 3.11. Dichos valores son 69,7 % de sensibilidad y 62,1 % de especificidad. Teniendo estos valores se procede a fijar el umbral de

clasificación en donde se separan las clases, para esto se tomará el umbral mínimo que puede tener el modelo para tener la misma sensibilidad del modelo de Santolaya (sobre 69,7 %) y sobre ese valor se moverá el umbral para determinar la ganancia o perdida de especificidad y dicha diferencia será cuantificada como los episodios bien clasificados.

Umbral	**Sensibilidad Modelo**	**Especificidad Modelo**	**Especificidad Santolaya**	**% de casos de NFBR mejor clasificados**
59,5 %	70,51 %	82,08 %	62,10 %	19,98 %
59,1 %	72,22 %	81,50 %	62,10 %	19,40 %
58,7 %	73,08 %	80,92 %	62,10 %	18,82 %
57,8 %	73,50 %	78,61 %	62,10 %	16,51 %
57,2 %	73,93 %	76,88 %	62,10 %	14,78 %
56,4 %	74,79 %	76,30 %	62,10 %	14,20 %
56,0 %	75,64 %	75,72 %	62,10 %	13,62 %
53,9 %	76,07 %	72,25 %	62,10 %	10,15 %
52,6 %	76,50 %	70,52 %	62,10 %	8,42 %
51,9 %	77,35 %	69,94 %	62,10 %	7,84 %
49,4 %	79,91 %	65,90 %	62,10 %	3,80 %
48,9 %	80,77 %	65,32 %	62,10 %	3,22 %
48,5 %	81,20 %	63,58 %	62,10 %	1,48 %
47,9 %	82,05 %	63,01 %	62,10 %	0,91 %

Tabla 5.2: Porcentaje de mejores clasificados, de acuerdo al umbral definido para la clasificación

Como se puede ver en la tabla 5.2 la cantidad de casos mejor clasificados como NFBR puede llegar hasta casi el 20 % si se considera el mismo nivel de sensibilidad que se fijo en la línea base. Por otra parte, se puede aumentar la sensibilidad desde el 69,7 % de la línea base hasta el 82,05 % y seguir teniendo más casos bien clasificados. Dado que la sensibilidad y la especificidad se tienen que balancear entre sí, se podría tener una mejor sensibilidad al 82,05 % cambiando el umbral y se tendrían más casos mejor clasificados de alto riesgo pero menos caso de bajo riesgo bien clasificados, no siendo este el objetivo de la memoria y su hipótesis de investigación.

Teniendo los porcentajes de la tabla 5.2 se puede estimar la cantidad de casos en términos absolutos que serian clasificados adecuadamente por año. De acuerdo a los datos entregados por el PINDA, el 45,56 % de los episodios de neutropenia febril no se complican, es decir que deberían estar clasificados como de bajo riesgo. Además, teniendo en cuenta que cada año ingresan 280 casos, 128 de estos deberían ser clasificados en bajo riesgo. Al nivel de especificidad del modelo de Santolaya, solo 79,5 casos serían correctamente clasificados, mientras que para el modelo propuesto en esta memoria 105 casos serian bien clasificados, significando 25 pacientes que recibirán un tratamiento más adecuado según sus necesidades.

Ahora bien, la forma de cuantificar el efecto económico y social de un mejor tratamiento como se mencionó será la cuantificación del menor costo del tratamiento y los días menos de hospitalización para el paciente. Santolaya (2004) calcula que el tratamiento de un episodio de alto riesgo de neutropenia febril cuesta 903 dolares mientras que uno de bajo riesgo cuesta 638 dolares en promedio desde el 2001 hasta el 2003 [28]. Los valores anteriores serán ajustados por inflación primero y luego por el precio del dolar.

La inflación acumulada entre enero del 2004 y enero del 2020 es de 70,38 % [57] por lo que el costo de cada episodio aumenta a 1.538 dolares para el caso de alto riesgo y 1.087 dolares para el de bajo riesgo. Una diferencia cercana a los 451 dolares, haciendo que los 25 casos mejor estratificados signifiquen un ahorro de 11.275 dolares al año.

Para calcular la cantidad de días ahorrados se utiliza la cantidad de días promedios que pasa un paciente usando camas en un determinado hospital. De acuerdo a Santolaya (2004), los días promedios para un tratamiento ambulatorio (bajo riesgo) fueron 3,8 días (1 cama de hospital y 2,8 días de cama transitorios) y para un tratamiento intrahospitalario (alto riesgo) fueron 5,3 días de cama en el hospital mismo, es decir 1,5 días de cama menos por paciente bien clasificado como de bajo riesgo. Y considerando los 25 casos mejor clasificados, hay un ahorro de 37,5 días en tiempo del paciente que podría ser utilizado en otras actividades. El tiempo anterior es sin considerar el tiempo ahorrado a los familiares y cuidadores del niño.

5.2. Impacto para Chile

Dado que ya se estimó el impacto que tendrá en la red PINDA de la Región Metropolitana la implementación de este modelo, ahora se podría hacer una estimación para los próximos 10 años del impacto que tendrá en Chile la implementación de este modelo. Considerando que el modelo se aplica a un subconjunto de la población, es necesario determinar el porcentaje de prevalencia que hubo en la Región Metropolitana ajustando por la población de esta y la cantidad de personas que están en el subconjunto buscado en todo el país. Este modelo se construyó en base a una población de pacientes pediátricos, es decir menores de 18 años por lo que el subconjunto que se buscará es el de los pacientes menores de 18 años.

Según datos del CENSO del 2017 [58] en la Región Metropolitana hay aproximadamente 1.763.997 personas menores de 18 años, es decir el 24,8 % de las personas que viven en la Región Metropolitana están en el conjunto de interés. Suponiendo 281 casos de neutropenia febril esperados al año, se ve que la prevalencia en este conjunto de la población es del 0,0159 %. Aplicando esto al resto del país, es decir a las restantes 2.747.789 de personas dentro del subconjunto, se esperarían 437 casos más de neutropenia Febril por año. Cabe señalar que estos supuestos no considera la prevalencia del cáncer en la población, aún así, se considera una buena aproximación para la cantidad estimada de casos de neutropenia febril.

Además de la estimación anual, se realizará una estimación considerando el crecimiento o decrecimiento de la población chilena (hasta los 18 años) para los próximos 10 años. Como se puede ver en la tabla 5.3, la cantidad de personas dentro del rango de edad, decrece a un ritmo promedio anual de 0,43 %, aún así, la cantidad de personas potencialmente afectadas por neutropenia febril siempre supero los 4,5 millones de personas.

5.2.1. Ahorro directo por menores costos de tratamiento

Con la estimación de los casos de neutropenia febril a nivel nacional y por año, se puede estimar a groso modo la cantidad de recursos ahorrados provenientes de una mejor clasifica-

Año	**Población**	**Tasa**	Caso NF
2021	4.729.088	-	752
2022	4.719.445	-0,20 %	750
2023	4.711.212	-0,17 %	749
2024	4.705.066	-0,13 %	748
2025	4.697.073	-0,17 %	747
2026	4.682.405	-0,31 %	745
2027	4.658.229	-0,52 %	741
2028	4.625.071	-0,71 %	735
2029	4.587.765	-0,81 %	729
2030	4.549.081	-0,84 %	723
Promedio	4.666.444	-0,43 %	742

Tabla 5.3: Evolución población Chilena hasta los 18 años de edad, incluyendo la cantidad de casos de neutropenia febril estimados.

ción de riesgo. Primero, como se considera un horizonte de tiempo de 10 años y el promedio esperado es de 742 casos por año, se estiman 7420 casos para los 10 años. Segundo, si se considera el 45,56 % como ratio de los casos que no se complican basados en la base del PINDA, se tendrían 3380 casos de bajo riesgo de los cuales la regla actual con especificidad de 62,10 % solo identificaría 2.099 casos mientras que el modelo propuesto en esta memoria con especificidad 82,8 % (al mismo nivel de sensibilidad) identificaría 2.775 casos, es de decir 676 casos más bien clasificados.

Con los 676 casos extras clasificados adecuadamente, y utilizando los costos planteados planteados por Santolaya (2004), actualizados por inflación, se tendría un ahorro monetario de 304.876 dolares y que con la tasa de cambio promedio de los últimos 12 meses (756,59 pesos) [59] significarían un ahorro de 230.598.040 pesos chilenos. Además, la cantidad de días ahorrados para todos los paciente y sus familia serian 1014 días o más de 2,77 años.

Resumiendo, clasificar un mejor caso de neutropenia febril significan para el estado un ahorro de 451 dolares y 1,5 días para el paciente además de entregarle el mejor tratamiento posible y una mejor calidad de vida tanto para el como para la familia. Al año, sería un ahorro de 23 millones de pesos y 101,4 días evitados de hospitalizaciones que no se justifican.

Que la hipótesis se haya comprobado y a qué además esta potencialmente signifique un ahorro para el estado y mejor calidad de vida para los pacientes comprueba la necesidad de implementar una solución de este tipo en el sistema público, no obstante, entendimiento de que la evaluación económica usa supuestos bastante fuertes pero que incluso en el peor de los casos esta solución ayudaría a los niños con cáncer y sus familias.

5.3. Impacto en el tratamiento general de neutropenia febril

Para la medición del impacto económico se tomo el sistema de puntuación que mejor nivel de sensibilidad entrega, que fue el propuesto por Santolaya (2001). No obstante, el rendimiento de dicho modelo es bastante bajo para estándares de salud, ya que de 10 casos de IBI declarada solo identificara 7. En cambio, el modelo propuesto en esta memoria con un mismo nivel de especificidad es capaz de identificar un caso extra de IBI declarado, en ese sentido, el análisis económico hecho anteriormente sirve solamente para indicar la factibilidad económica en términos generales, es decir, que no significara un costo extra para el estado.

Es por lo anterior, que la principal ventaja del modelo propuesto no es el ahorro monetario de su implementación sino que es la entrega de un tratamiento más racional a los pacientes, sabiendo identificar de mejor manera los casos de alto riesgo y teniendo un nivel alto de identificación de los casos de bajo riesgo, en consecuencia, se justifica el umbral de clasificación propuesto en esta memoria del 40,77 % que entrega un nivel de sensibilidad del 90 % de sensibilidad y un 49,17 % de especificidad, que podría implicar un mayor costo, pero entregando el 90 % de las veces el tratamiento más adecuado al paciente.

Cerrando este capítulo, cabe decir que en términos de un análisis costo-beneficio, esta solución claramente será un aporte al sistema de salud nacional, al reducir los costos que se incurrirían al sobre-estimar el riesgo de IBI. Del mismo modos, al hacer un análisis costo-efectividad, también se podría suponer beneficiosa esta solución, ya evitaría complicaciones derivada de la sobre-estimación del riesgo, mejorando la calidad de vida. No obstante, al ser un análisis más complejo, escapa a los alcances de esta memoria, ya que entraría en un dominio médico el definir cuanto mejora la calidad de vida y tratamiento al implementar esta solución.

Capítulo 6

Conclusiones

6.1. Conclusiones generales

Esta memoria se basa en la creación de un modelo de clasificación de riesgo de infecciones bacterianas invasoras (IBI) en pacientes pediátricos con cáncer de la red PINDA y que logre tener un rendimiento superior a los modelos clasificadores existentes sobre esta materia. Para crear el modelo, se aplicó el proceso de descubrimiento de conocimiento en base de datos (`KDD` por sus siglas en inglés) a la base de datos entregada por el PINDA que esta compuesta por diferentes indicadores biológicos que en su mayoría ya han sido estudiados por la literatura y han demostrado ser relevantes en la predicción de un episodio de IBI. Considerando todo lo anterior se plantea la siguiente hipótesis de investigación:

- La construcción de un modelo de clasificación, que utilice datos al ingreso de los pacientes, basado en algoritmos de machine learning aumentará la especificidad y sensibilidad en la predicción de infecciones bacterianas invasoras en episodios de neutropenia febril que permitirá mejorar el tratamiento entregado a los pacientes reduciendo los costos directos e indirectos para la sociedad.

Inicialmente, se crea una linea base para comparar el rendimiento del modelo y esta es la regla aplicada en la red PINDA basada en la puntuación creado por Santolaya (2001). Para realizar la comparación se utilizan las métricas de sensibilidad, especificidad y área bajo la curva (AUC por sus siglas en inglés). El modelo desarrollado en esta memoria tiene un AUC final de 0,834 en comparación al 0,659 de la linea base, siendo una diferencia importante de rendimiento entre ambos modelos. Cuando se evalúa por las otras métricas se encuentra que al mismo nivel de sensibilidad del modelo de Santolaya (69,7 %) el modelo desarrollado tiene una especificidad de 82,08 %, casi 20 puntos porcentuales más (la linea base tiene 62,1 %). Para el mismo nivel de especificidad (62,1 %) de la linea base se tienen aproximadamente 12 puntos porcentuales más de sensibilidad (82,05 %). Con lo anterior se valida la hipótesis planteada de que se puede crear un modelo con un rendimiento superior a los existentes.

Cabe destacar que de las 14 variables utilizadas para el modelo final, 4 son utilizadas de alguna manera en el modelo de Santolaya (`Proteína Reactiva C, Plaquetas, Enfermedad`

de base e hipotensión) y 9 aparecen en algún modelo propuesto por la literatura (Índice de masa corporal, recuento absoluto de neutrófilos, hemoglobina, estado del paciente y foco respiratorio alto además de las 4 de Santolaya) . Solo son 5 las variables que son completamente nuevas para las puntuaciones de riesgo existentes y son las siguientes: Recuento absoluto de leucocitos (RAL), recuento absoluto de monocitos (RAM), calcemia, factores de crecimiento, foco en tejidos bandos y crépitos.

Solo 5 de las 14 variables son completamente nuevas por lo que existe un balance adecuado entre lo que la literatura muestra acerca de las variables relevantes en la estratificación de riesgo y las descubiertas a través del método KDD. Si se hubieran encontrado variables totalmente diferentes a las de la literatura podría significar que el modelo difícilmente fuese validado por la comunidad médica e incluso que prediga mal lo que realmente sucede. O en el caso contrario, encontrar las mismas variables implicaría que el desarrollo del modelo de esta memoria habría sido innecesario.

Otro hallazgo relevante de esta memoria es que los modelos desarrollados hasta la fecha se basaban en puntajes asociados a la existencia de cierta condición, por ejemplo, un recuento absoluto de neutrófilos menor a 500 equivale a 2 puntos. Lo anterior simplifica el proceso de clasificación de los episodios, simplificación que era necesaria hacer ya que la clasificación era realizada de manera manual por el equipo médico. No obstante, la masificación de las aplicaciones web ha facilitado la implementación de modelos más complejos y precisos en la palma de la mano, y ha hecho posible que ya no sea necesaria dicha simplificación. En consecuencia, hoy es posible ingresar los datos médicos relevantes del paciente en cosa de minutos y recibir la probabilidad de IBI en solo segundos, haciendo que la clasificación sea mucho más precisa y permita un mejor tratamiento

A su vez, también se construye un prototipo funcional de la aplicación, lo cual permite realizar lo anteriormente descrito, es decir ingresar datos y recibir la probabilidad de IBI, teniendo como finalidad mostrar el proceso para que los futuros usuarios puedan entregar retroalimentaciones del uso de la aplicación. Finalmente, se estima el beneficio económico y social de la implementación de este modelo tanto en la red PINDA de la RM como en todo el país. Para hacer lo anterior se estima la reducción de costos por un tratamiento menos agresivo cuando se clasifica adecuadamente como de bajo riesgo al episodio.

La estimación del beneficio económico y social arrojo que la implementación de este modelo en los próximos 10 años, podría clasificar bien a más de 676 episodios extras de lo que el puntaje actual clasifica bien. Estos 676 casos extras significan un ahorro para el sistema de salud púbico de más de 230 millones de pesos y para los pacientes un ahorro de 1014 días al no tener que pasar días y noches en el hospital.

En términos generales se cumplen los objetivos de la memoria, que fue la construcción de un modelo de clasificación de riesgo de IBI en episodios de neutropenia febril en niños con cáncer, mejorando el rendimiento de los actuales modelos y permitiendo un tratamiento más racional a los pacientes ingresados por esta condición.

La primera etapa fue la investigación de modelos de clasificación existentes en el mundo, encontrado varios modelos que apuntaban hacia lo mismo, una mejora clasificación del riesgo de IBI. Estos modelos fueron aplicados a la base entregada por el PINDA para generar una

base de comparación con el modelo por crear. Además, los modelos de la literatura entregaron la base inicial para los primeros algoritmos, mostrando que los mejores algoritmos para la clasificación son las *regresiones logísticas y los arboles aleatorios (Random Forest).*

El proceso final de selección de variables permitió elegir las variables que tenían el mayor poder predictivo, haciendo que el rendimiento del modelo propuesto en esta memoria fuera superior al mejor existente, validando la hipótesis de investigación. A su vez, el prototipo funcional de visualización de la clasificación mostró que es posible generar una herramienta simple pero que incorpore las complejidades de un modelo multivariable y que a su vez entregue una mejor clasificación de riesgo.

Finalmente se comprobó que la implementación de este modelo tendrá beneficios directos a los pacientes al entregar un tratamiento más racional, mejorando su calidad de vida en términos de salud y tiempo requerido para el tratamiento, y para el sistema de salud público, que ahorrara recursos al evitar sobre-estimar el riesgo de IBI.

6.2. Trabajo Futuro

Esta sección busca ser un aporte a las futuras investigaciones acerca de clasificadores de riesgo de IBI, tal que permita a futuros interesados construir desde aquí posibles mejoras a nuevos clasificadores e incluso en la aplicación de otros dominios relacionados con la estratificación de riesgo en la salud.

Una línea de trabajo de investigación es la posible existencia de otras variables de riesgo que no han sido identificadas hasta el momento. Lo anterior es consecuencia de que la revisión bibliográfica realizada arrojo la existencia de varios clasificadores con diferentes variables cada uno, sin que uno haya mostrado un rendimiento superior a los otros. Considerando lo anterior, agregar nuevas variables al modelo de predicción podría ser útil, en particular variables que tengan que ver con lo que sucede dentro del lugar de residencia del paciente, agregando información previa al momento del ingreso del paciente.

Otra futura mejora es que los datos sean ingresados plenamente a la plataforma propuesta, evitando posibles valores perdidos y en caso de que existan, explicar la razón de la perdida. Lo anterior se debe a que a pesar de que la base de datos tenia más de 1.600 registros, la cantidad de registros se reducía al incorporar más variables (que tenían valores faltantes), haciendo que el modelo final se validará sobre un umbral de 300 registros, valor que está en el orden de cantidad de registros usados en la literatura pero que podría generar un sobre ajuste del modelo al conjunto final de datos usados, situación que se busco reducir al incorporar técnicas de muestreo aleatorio, en particular *k-folds*.

En la línea de lo anterior, hay espacio de trabajo e investigación en la implementación de esta herramienta en el espacio médico real, considerando que cualquier cambio en una organización genera resistencias dentro de esta y que la implementación de esta herramienta tendrá que ser acompañada con un plan de capacitaciones y validación de la herramienta por parte de la organización.

Por ultimo, el modelo y la plataforma han sido enfocada en los casos de neutropenia febril en niños, pero la misma metodología podría ser utilizada para ser aplicada en adultos, ampliando significativamente el impacto de un modelo de clasificación considerando que por cada niño con cáncer hay 17 adultos, es decir el impacto del modelo podría ser hasta 17 veces más grande. Existen clasificadores en el mundo aplicados a adultos, pero que han sido validados con muestras diferentes a las chilenas por lo que la aplicación de este metodología en episodios chilenos podría crear una clasificación superior en adultos y por consiguiente, un mejor tratamiento, que beneficie enormemente la calidad de vida de las personas.

Capítulo 7

Anexos

7.1. Marco teorico

7.1.1. Variables por modelos de la literatura

En la siguiente tabla podemos ver las diferentes variables y outcomes utilizados por los diferentes modelos encontrados en la literatura.

Modelo	Outcome predicho	Variables y su valoración	Estratificación de riesgo
Amman et al.	Infección Bacterial Severa (IBS)	1.Médula osea involucrada = 7 2.Otro diagnostico que pre-B ALL = -3 3.Sin signos clínicos de infección viral =4 4.PCR entre 6 y 50mg/l=3, PCR>50mg/l=6 5.TLC 0,6-10 x10^9 /litro =3 TLC <= 0,5 por 10^9/litro=6 6.Catéter venoso central = 2 7. Nivel de hemoglobina entre 71 y 100 g/l =-2 y <=70g/l = -4	***Alto riesgo:** Más de 4 puntos ***Bajo riesgo:** 4 o menos puntos
Rondinelli et al.	Complicaciones Infecciosas Severas	1. Catéter venoso central = 2 2. Edad $\leq 5aos = 1$ 3. Infecciones en sitio clínico = 4.5 4. Infecciones en el tracto respiratorio =2.5 5. Temperatura corporal $\geq 38,5C = 1$ 6. Hemoglobina $\leq 7g/\mathrm{d}L = 1$	Bajo riesgo: puntos entre 2,5 $\leq 5,5$ Medio riesgo: puntos entre $\geq= 5,5a9$ Alto riesgo puntos ≥ 9
Regla SPOG de eventos adversos	Eventos adversos	1. Quimio. más intensa que ALL = 4 2. Hemoglobina<90g/l=5 3. TLC<0,3 x10^9 /litro =3 4. Plaquetas <50000/mm3 = 3	Bajo riesgo: puntos<9 Alto riesgo: puntos>=9

Hakim et al.	Infección bacterial probada o cultivo negativo de sepsis	1. Puntaje por diagnostico de cancer: AML = 20, ALL/linfoma = 7, solids=0 2. Presentación con serio malestar/toxicidad = 14 3. Fiebre al ingreso >=39 C = 11 4. RAN<100mm3 =10	***Bajo riesgo:** <24 puntos ***Alto riesgo** $\geq 24 puntos$
Santolaya et al.	Infección Bacterial Invasiva (IBI)	1. Recaída de leucemia 2. Quimioterapia hace $\leq$ 7di*as* 3. PRC $\geq 90 mg/l$ 4. Hipotensión 5. Plaquetas $\leq 50{,}000/mm^3$	***Alto riesgo:** 2 factores de riesgo cualquiera o 1,2,3 y 4 por si solos. ***Bajo riesgo:** Ningún factor de riesgo identificado o solo tener menos de 50000 plaquetas por mm3 o menos de 7 días desde la ultima quimioterapia
Agyeman et al.	Bacteremia	1. Escalofríos observados 2. Hemoglobina >90/l 3. Plaquetas <50.000/mm3 4. Cualquier necesidad que necesite hospitalización	***Bajo riesgo:** ningún factor de riesgo ***Alto riesgo:** uno cualquiera
Indigenous model	Complicaciones	1. Mal nutrición = 2 2. Quimioterapia hace $\leq$ 7di*as* = 2 3. Presencia de un foco infeccioso = 2 4. RAN $\leq 100/mm^3 = 2$ 5. PRC>60 mg/l = 5	***Bajo riesgo** <7 puntos ***Alto riesgo** $\geq 7 puntos$

Tabla 7.1: Modelos utilizados en el mundo para estratificar riesgo, elaboración propia en base a Das 2017

7.2. CRISP-DM

Un marco teórico alternativo al uso tradicional del KDD es el CRISP-DM *Cross Industry Standard Process for Data Mining* que tiene 6 etapas [44] que son bastante similar al KDD y se listan a continuación:

1. **Comprensión del negocio y definición de las necesidades del cliente:** Es una fase inicial y se enfoca en la comprensión de los objetivos del proyecto que luego es convertido en un problema de minería de datos.
2. **Entendimiento de los datos:**Esta fase toma los datos iniciales y realiza tareas para familiarizarse con los datos, descubriendo relaciones interesantes a través de estadísticos tradicionales y visualización de los datos.
3. **Preparación de los datos. Análisis de estos y selección de características:**Incluye todas las tareas necesarias para construir el conjunto final de datos que incluyen la limpieza, tratamiento de valores faltantes, formateo de los registros entre otros.
4. **Modelado:** En esta etapa se seleccionan todas las técnicas de minería de datos que sean pertinentes al problema y se calibran los hiper parámetros para cada modelo. En general la aplicación de los modelos tiene situaciones particular que hacen necesario el re-procesamiento de los datos lo que hace volver a la fase de preparación.
5. **Evaluación y obtención de resultados:** En esta etapa ya se han hecho funcionar los modelos y se pueden evaluar en las diferentes métricas para determinar si es puesto en producción y su aplicación en el proceso de análisis de datos.
6. **Despliegue o puesta en producción:** La creación del modelo no es típicamente el final de proyecto sino que el paso previo para que los resultados sean analizados y se puedan tomar decisión en base a ellos. Dependiendo de los requisitos los resultados

pueden ser un informe, un dashboard o incluso la realización periódica del análisis de los datos.

Para el caso de CRISP-DM es clave entender que es un proceso iterativo o cíclico en donde volver pasos atrás antes de avanzar es parte de la metodología. En la figura 7.1 se puede ver el ciclo de vida de un proyecto basado en CRISP DM.

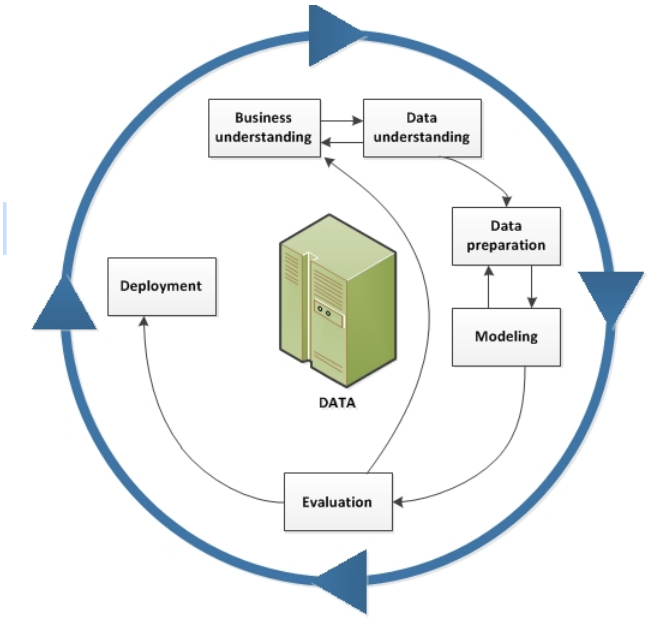

Figura 7.1: Ciclo de vida de un proyecto basado en CRISP DM, fuente IBM [6]

7.3. Mejores variables por modelos

En la tabla 7.2 se pueden ver las mejores variables para cada uno de los modelos posibles.

Método	AUC	N Var	Variables
1er conjunto bosque aleatorio inductivo	0,8663	28	PCR, Rto. Absoluto leucocitos, Calcemia, Enfermedades: 2, 3, 13 y 14, Factores de crecimiento, RAN, Desnutrición, Número días desde última QMT, Estado General, Focos: Respiratorio alto, Tiflitis y Mucositis, Antecedente IFI, BMI, Virus: 7 y 14, Crepitos, Sexo, RX TX 5, hipotensión, Abdominal, RAM, HG, Tejidos blandos, fluconazol.
2ndo conjunto bosque aleatorio inductivo	0,8660	15	PCR, Rto. Absoluto leucocitos, Calcemia, Enfermedades: 2, 3 y 14, Factores de crecimiento, RAN, Desnutrición, Número días desde última QMT, Estado General, Foco Respiratorio alto, Antecedente IFI, BMI, Virus 7.
3er conjunto bosque aleatorio inductivo	0,8657	27	PCR, Rto. Absoluto leucocitos, Calcemia, Enfermedades: 2, 3, 13 y 14, Factores de crecimiento, RAN, Desnutrición, Número días desde última QMT, Estado General, Focos: Respiratorio alto, Tiflitis y Mucositis, Antecedente IFI, BMI, Virus: 7 y 14, Crepitos, Sexo, RX TX 5, hipotensión, Abdominal, RAM, HG, Tejidos blandos,
1er conjunto bosque aleatorio deductivo	0,8654	13	Rto. Absoluto leucocitos, Enfermedades: 2 y 13, T, HG, Calcemia, fluconazol, Estado General, Virus: 7 y 14, RAN, Factores de crecimiento, PCR.
2nd conjunto bosque aleatorio deductivo	0,8615	19	Rto. Absoluto leucocitos, Creatinemia, Virus: 7 y 14, T, HG, fluconazol, RAN, Enfermedades: 2, 3 y 13, PCO2, Calcemia, PCR, Sexo, hipotensión, Estado General, Abdominal, Factores de crecimiento.
3er conjunto bosque aleatorio deductivo	0,8607	16	Rto. Absoluto leucocitos, Enfermedades: 2, 3 y 13, hipotensión, T, HG, Calcemia, fluconazol, RAN, Estado General, Virus: 7 y 14, Factores de crecimiento, PCR, Sexo.
1er conjunto logístico inductivo	0,8467	26	PCR, Estado General, PCO2, Enfermedades: 2, 3 y 13, Rto. Absoluto leucocitos, Glicemia, RAM, Plaquetas, RX TX 5, Abdominal, Foco Tiflitis , Tejidos blandos, otro, Foco Mucositis, Foco Respiratorio , Virus: 7 y 14, BMI, fluconazol, RAN, Crepitos, Desnutricion, FC, Edad.
2ndo conjunto logístico inductivo	0,8466	13	PCR, Estado General, PCO2, Enfermedad 2 y 3, Rto. Absoluto leucocitos, Glicemia, RAM, Plaquetas, RX TX 5, Abdominal, Foco Tiflitis , Tejidos blandos.
3er conjunto logístico inductivo	0,8466	14	PCR, Estado General, PCO2, Enfermedad 2 y 3, Rto. Absoluto leucocitos, Glicemia, RAM, Plaquetas, RX TX 5, Abdominal, Foco Tiflitis , Tejidos blandos, otro.
1er conjunto logístico deductivo	0,8427	20	Crepitos, Creatinemia, BMI, Enfermedad 2, 3 y 13, Virus 1, Tejidos blandos, RAN, RAM, Edad, Calcemia, PCR, Sexo, hipotensión, FC, Estado General, Foco Respiratorio alto, Plaquetas, Factores de crecimiento,
2ndo conjunto logístico deductivo	0,8423	24	Foco Tiflitis , Crepitos, Creatinemia, BMI, Enfermedad 2, 3 y 13, HG, Virus 1, Tejidos blandos, RAN, RAM, Edad, Foco SDA, Calcemia, PCR, Sexo, hipotensión, FC, Estado General, Foco Respiratorio alto, Abdominal, Plaquetas, Factores de crecimiento,
3er conjunto logístico deductivo	0,8423	25	Foco Tiflitis , Crepitos, Creatinemia, BMI, Enfermedad 2, 3 y 13,, HG, Virus 1, Tejidos blandos, RAN, RAM, Edad, Foco SDA, Calcemia, PCR, Sexo, hipotensión, FC, Estado General, Foco Respiratorio , Foco Respiratorio alto, Abdominal, Plaquetas, Factores de crecimiento,

Tabla 7.2: Resumen rendimientos modelos con métodos inductivos y deductivos

Bibliografía

[1] Anirban Das, Amita Trehan, Sapna Oberoi, and Deepak Bansal. Validation of risk stratification for children with febrile neutropenia in a pediatric oncology unit in India. *Pediatric Blood & Cancer*, 64(6), 2017.

[2] Patrícia Imperatriz Porto Rondinelli, Karina de Cássia Braga Ribeiro, and Beatriz de Camargo. A proposed score for predicting severe infection complications in children with chemotherapy-induced febrile neutropenia. *Journal of Pediatric Hematology/Oncology*, 28(10):665–670, October 2006.

[3] Philipp Agyeman, Christoph Aebi, Andreas Hirt, Felix K. Niggli, David Nadal, Arne Simon, Hulya Ozsahin, Udo Kontny, Thomas Kühne, Maja Beck Popovic, Kurt Leibundgut, Nicole Bodmer, and Roland A. Ammann. Predicting bacteremia in children with cancer and fever in chemotherapy-induced neutropenia: results of the prospective multicenter SPOG 2003 FN study. *The Pediatric Infectious Disease Journal*, 30(7):e114–119, July 2011.

[4] Hana Hakim, Patricia M. Flynn, Deo Kumar Srivastava, Katherine M. Knapp, Chenghong Li, James Okuma, and Aditya H. Gaur. Risk prediction in pediatric cancer patients with fever and neutropenia. *The Pediatric Infectious Disease Journal*, 29(1):53–59, January 2010.

[5] (PDF) Explorations of the BDI Multi-agent support for the Knowledge Discovery in Databases Process.

[6] Conceptos básicos de ayuda de CRISP-DM, October 2014.

[7] Organización mudnial de la salud. Cáncer en el mundo.

[8] Ministerio de Salud. Plan nacional del cancer 2018-2018.

[9] Diccionario de cáncer, February 2011.

[10] Neutropenia Febril: La Otra Cara de la Lucha Contra el Cáncer, June 2011.

[11] Daniel Ricardo Martínez Ávila, Horacio Santos González, and Sailyn Reyes Castilla. Neutropenia febril postquimioterapia, Instituto de Oncología y Radiobiología. *Revista Cubana de Farmacia*, 50(1):44–52, March 2016.

[12] Philip A. Pizzo, K.J. Robichaud, Fred A. Gill, and Frank G. Witebsky. Empiric antibiotic and antifungal therapy for cancer patients with prolonged fever and granulocytopenia. *The American Journal of Medicine*, 72(1):101–111, January 1982.

[13] M. E. Santolaya, A. M. Alvarez, A. Becker, J. Cofré, N. Enríquez, M. O'Ryan, E. Payá, J. Pilorget, C. Salgado, J. Tordecilla, M. Varas, M. Villarroel, T. Viviani, and M. Zubieta. Prospective, multicenter evaluation of risk factors associated with invasive bacterial infection in children with cancer, neutropenia, and fever. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 19(14):3415–3421, July 2001.

[14] Santolaya de P and María Elena. Neutropenia febril en el niño con cáncer: Conceptos actuales sobre criterios de riesgo y manejo selectivo. *Revista médica de Chile*, 129(12):1449–1454, December 2001.

[15] C. Marcela Palavecino. Toxicidad antibacterianos: farmacocinetica-farmacodinamia: prevención y manejo. *Revista Médica Clínica Las Condes*, 25(3):445–456, May 2014.

[16] MedlinePlus. Resistencia a los antibióticos.

[17] Huniades Urbina Medina. Infección nosocomial. *ARCHIVOS VENEZOLANOS DE PUERICULTURA Y PEDIATRÍA*, 64:7, 2001.

[18] C Figueras Nadal, E Roselló Mayans, and F Álvez González. Infección fúngica invasiva (IFI): actualización. page 13.

[19] Juan C. Benitez-Agudelo, Ernesto A. Barceló-Martinez, and Melissa Gelves-Ospina. Características psicológicas de los pacientes con larga estancia hospitalaria y propuesta de protocolo para su manejo clínico. *Cirugía Plástica Ibero-Latinoamericana*, 42(4):391–398, December 2016.

[20] Estadísticas | PINDA.

[21] Claudio Viscoli and Elio Castagnola. Treatment of febrile neutropenia: what is new? *Current Opinion in Infectious Diseases*, 15(4):377, August 2002.

[22] W. V. Kern. Risk assessment and risk-based therapeutic strategies in febrile neutropenia. *Current Opinion in Infectious Diseases*, 14(4):415–422, August 2001.

[23] Maria Santolaya-De Pablo. Towards a more rational approach in the comprehensive management of the child with cancer, fever and neutropenia: Filling the gaps. 2012.

[24] Organización mundial de la salud. Resistencia a los antibióticos.

[25] Lisa M. Orme, Franz E. Babl, Chris Barnes, Peter Barnett, Susan Donath, and David M. Ashley. Outpatient versus inpatient IV antibiotic management for pediatric oncology patients with low risk febrile neutropenia: A randomised trial. *Pediatric Blood & Cancer*, 61(8):1427–1433, 2014.

[26] Nicole M. Kuderer, David C. Dale, Jeffrey Crawford, Leon E. Cosler, and Gary H. Lyman. Mortality, morbidity, and cost associated with febrile neutropenia in adult cancer patients. *Cancer*, 106(10):2258–2266, 2006.

[27] Eric Tai, Gery P. Guy, Angela Dunbar, and Lisa C. Richardson. Cost of Cancer-Related Neutropenia or Fever Hospitalizations, United States, 2012. *Journal of Oncology Practice*, 13(6):e552–e561, 2017.

[28] Early hospital discharge followed by outpatient management versus continued hospitalization of children with cancer, fever, and neutropenia at low ... - PubMed - NCBI.

[29] Oliver Teuffel, Eitan Amir, Shabbir M. H. Alibhai, Joseph Beyene, and Lillian Sung. Cost-effectiveness of Outpatient Management for Febrile Neutropenia in Children With Cancer. *Pediatrics*, 127(2):e279–e286, February 2011.

[30] Víctor Zárate. Evaluaciones económicas en salud: Conceptos básicos y clasificación. *Revista médica de Chile*, 138, September 2010.

[31] Miguel Luis O'ryan Gallardo, Tamara Nieves Viviani Salgado, Juan Pablo Torres Torretti, Milena Villarroel Cickovic, Carmen Eliana Salgado MuÃoz, Carmen Luz Aviles Lohmann, Juan Aristides Tordecilla Cadiu, and Monica Cecilia Varas Palma. PERFILES TRANSCRIPCIONALES Y DIAGNOSTICO MICROBIOLOGICO MOLECULAR PARA AVANZAR EN EL DIAGNOSTICO DE SEPSIS Y EN LA IDENTIFICACION DEL MICROORGANISMO CAUSAL EN NIQOS CON CANCER, NEUTROPENIA Y FIEBRE. 2009.

[32] What is Python? Executive Summary.

[33] The Web framework for perfectionists with deadlines | Django.

[34] Project Jupyter. Library Catalog: jupyter.org.

[35] Proyecto Jupyter, April 2020. Page Version ID: 125011822.

[36] Git.

[37] Git, April 2020. Page Version ID: 125298297.

[38] Hugo Paganini, María Elena Santolaya de P, Martha Álvarez, Manuel de Jesús Araña Rosaínz, Ricardo Arteaga Bonilla, Anibal Bonilla, Miguela Caniza, Fabianne Carlesse, Pio López L, Lourdes Dueñas de Chicas, Tirza de León, José Marcó del Pont, Mario Melgar, Laura Naranjo, Carla Odio, Mónica Rodríguez, and Marcelo Scopinaro. Diagnóstico y tratamiento de la neutropenia febril en niños con cáncer: Consenso de la Sociedad Latinoamericana de Infectologia Pediátrica. *Revista chilena de infectologÃa*, 28:10–38, March 2011.

[39] Parenteral definición.

[40] Roland A. Ammann, Andreas Hirt, Annette Ridolfi Lüthy, and Christoph Aebi. Identifi-

cation of children presenting with fever in chemotherapy-induced neutropenia at low risk for severe bacterial infection. *Medical and Pediatric Oncology*, 41(5):436–443, November 2003.

[41] Roland A. Ammann, Nicole Bodmer, Andreas Hirt, Felix K. Niggli, David Nadal, Arne Simon, Hulya Ozsahin, Udo Kontny, Thomas Kühne, Maja Beck Popovic, Annette Ridolfi Lüthy, and Christoph Aebi. Predicting adverse events in children with fever and chemotherapy-induced neutropenia: the prospective multicenter SPOG 2003 FN study. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 28(12):2008–2014, April 2010.

[42] How Much Data Is Generated Per Minute? The Answer Will Blow Your Mind Away [Infographic].

[43] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34, November 1996.

[44] CRISP-DM: La metodología para poner orden en los proyectos, August 2016.

[45] Thomas Bayes and null Price. LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions of the Royal Society of London*, 53:370–418, January 1763.

[46] I Rish. An empirical study of the naive Bayes classiïer. page 6.

[47] David D. Lewis. Naive (Bayes) at forty: The independence assumption in information retrieval. In Claire Nédellec and Caroline Rouveirol, editors, *Machine Learning: ECML-98*, Lecture Notes in Computer Science, pages 4–15. Springer Berlin Heidelberg, 1998.

[48] Random Forest, el poder del Ensamble | Aprende Machine Learning.

[49] Gradient Boosting Classifiers in Python with Scikit-Learn, July 2019.

[50] 5.2. El patrón de diseñoo MTV (El libro de Django 1.0).

[51] Tema 10. Selección de Variables, language = en, author = Moujahid, Abdelmalik and Inza, Inaki and Larranaga, Pedro, pages = 5, file = Moujahid et al. - Tema 10. Selección de Variables.pdf:/home/carlosvega/Zotero/storage/BMIHPZZD/Moujahid et al. - Tema 10. Selección de Variables.pdf:application/pdf.

[52] Julian Cardenas. Qué es ANOVA de un factor y cómo analizarla fácilmente, November 2015.

[53] J Moreno, Daniel Rodriguez, M. Sicilia, José Riquelme, and Y Ruiz. SMOTE-I: mejora del algoritmo SMOTE para balanceo de clases minoritarias. January 2009.

[54] Two Modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics*,

SMC-6(11):769–772, November 1976.

[55] Francisco Azuaje, Ian Witten, and Frank E. Witten IH, Frank E: Data Mining: Practical Machine Learning Tools and Techniques. *Biomedical Engineering Online - BIOMED ENG ONLINE*, 5:1–2, January 2006.

[56] WHO | WHO Global Database on Child Growth and Malnutrition.

[57] Índice de Precios al Consumidor. Library Catalog: www.ine.cl.

[58] CENSO 2017 - Piramide poblacional · Instituto Nacional de Estadísticas - Gobierno de Chile.

[59] Sii | Servicio de Impuestos Internos.