



INAOE

**Detección automática del decaimiento de
habilidades motrices en la marcha de adultos
mayores**

German Cuaya Simbro, Angélica Muñoz Meléndez,
Eduardo F. Morales Manzanares

Reporte Técnico No. CCC-10-003
16 de marzo de 2010

© 2010
Coordinación de Ciencias Computacionales
INAOE

Luis Enrique Erro 1
Sta. Ma. Tonantzintla,
72840, Puebla, México.



Detección automática del decaimiento de habilidades motrices en la marcha de adultos mayores

M. C. German Cuaya Simbro, Dra. Angélica Muñoz Meléndez,
Dr. Eduardo F. Morales Manzanares

Coordinación de Ciencias Computacionales,
Instituto Nacional de Astrofísica, Óptica y Electrónica,
Luis Enrique Erro 1, Sta. Ma. Tonantzintla,
72840, Puebla, México

germancs@inaoep.mx, munoz@inaoep.mx, emorales@inaoep.mx

Resumen. La marcha de personas cambia con el tiempo, estos cambios están asociados a una degradación normal o bien a una degradación anómala que aumenta el riesgo de presentar una caída. El análisis de los datos obtenidos en un estudio de la marcha es importante para prevenir los efectos de una degradación anómala. El análisis así como la interpretación de grandes colecciones de datos tomados en distintos tiempos para explicar o entender el comportamiento de dichos datos plantea retos de investigación, pues se requiere el desarrollo de modelos que permitan conocer la evolución de variables en el tiempo en general para detectar la ocurrencia de eventos específicos, en este caso concreto anomalías en la marcha que pudieran representar algún tipo de riesgo a futuro, como por ejemplo, sufrir una caída. El reto de este trabajo es la construcción automática de dicho modelo.

Palabras clave: Análisis de la Marcha, Redes Bayesianas, Modelo de Regresión de Cox.

1. Introducción

Según datos del Consejo Nacional de Población (CONAPO) en México, el número de adultos mayores en el país se habrá cuadruplicado en un lapso de 50 años. De acuerdo a los datos del Instituto Nacional de Rehabilitación (INR), el riesgo de caídas en adultos mayores (personas mayores de 65 años) es mayor que en otros sectores de la sociedad. Para este sector de la sociedad una caída puede ocasionar serias lesiones o incluso la muerte. De hecho, registros estadísticos del INR muestran que las caídas en personas adultas mayores son la principal causa de muerte en este sector.

“El análisis de la marcha es el proceso donde información cuantitativa es colectada para ayudar en el entendimiento de las causas de las anomalías en la marcha de una persona y la toma de decisión del tratamiento a seguir” [13]. Dicha información es obtenida con ayuda de diversos instrumentos biomecánicos, como sistemas de cámaras, sensores electromagnéticos, etc. Algo esencial en este proceso es la interpretación de esos datos por un equipo de expertos experimentado e interdisciplinario con un conocimiento considerable en marcha normal y patológica [21].

Los métodos y tecnologías existentes para el análisis de la marcha son numerosos, y permiten la obtención de los parámetros cuantitativos característicos de un patrón de marcha de manera objetiva. Los tipos de datos obtenidos en dicho análisis pueden ser espacio-temporales como: tiempo de soporte, cadencia, largo de paso, velocidad; datos

cinemáticos como: ángulos articulares, rotación de la pelvis, ángulo de progresión del pie; datos cinéticos como: fuerza de reacción anteo-posterior, fuerza de reacción vertical.

Para la obtención de los parámetros de interés se tiene una diversidad de equipos especializados. Los laboratorios más sencillos pueden obtener los parámetros espacio-temporales mediante la utilización de cintas métricas, cronómetros y un análisis visual sistemático, o con un sistema conocido como *gaitrite* [50]. Para los análisis cinemáticos y cinéticos se cuenta con medidas de dinamometría, acelerometría, ultrasonido, goniometría digital, sistemas de análisis en dos y tres dimensiones, entre otros. Actualmente uno de los métodos más utilizados para el análisis de la marcha es la combinación de plataformas dinamométricas para la cinética, con técnicas de videogrametría para la cinemática, comúnmente en combinación con sistemas de registro de electromiografía (EMG) dinámica. Para esta última se utilizan sistemas telemétricos multicanal que permiten al paciente realizar su marcha libremente, de modo que se puede registrar simultáneamente la acción de varios grupos musculares para ser posteriormente contrastada con la información cinética y cinemática en todas las fases del ciclo de la marcha.

Como hemos mencionado nuestro problema se centra en el modelado de la marcha. Para ello es necesario el análisis de datos temporales asociados a los registros de los parámetros de la marcha tomadas en distinto tiempo. Hay tres objetivos principales en el análisis de datos temporales: predicción, modelado y caracterización [22]. El objetivo de la predicción es pronosticar de forma precisa la evolución a corto o largo plazo de los datos analizados. El objetivo del modelado, por su parte, es captar las características del comportamiento de los datos a largo plazo. Por último el objetivo de la caracterización de los datos es determinar algunas propiedades de estos, *e. g.* relevancia para la inferencia acerca del riesgo, relaciones existentes entre los parámetros y la naturaleza de las observaciones.

La mayoría de los trabajos revisados en análisis de la marcha se concentran en la adquisición de los datos como cadencia, tamaño de paso, velocidad, aceleración, para que un experto pueda revisarlos e interpretarlos y así dar un diagnóstico acerca de los problemas detectados en la marcha de una persona [12], [45]. Actualmente, la mayor parte de esa tecnología es costosa y se basa en algoritmos médicos de evaluación de la marcha que proporcionan una calificación del modo de andar de una persona o FAP, *Functional Ambulation Profile* [38].

Por otro lado, se han reportado trabajos en donde la marcha ha sido utilizada como una característica biométrica a efectos de identificación de anomalías de estabilidad, demencia senil y mal de Parkinson. Dicha identificación se ha hecho con base en el análisis de los datos de la marcha basado en hipótesis médicas, las cuales son muy restrictivas pues se limitan al análisis de parámetros específicos de la marcha, cadencia, base de sustentación, [13], [29], con el fin de proporcionar un diagnóstico acerca de la marcha de la persona. Cabe mencionar que hasta donde sabemos, no existen trabajos que analicen los datos obtenidos por un sistema de análisis de la marcha para proporcionar información a los expertos del dominio, con objeto de identificar los cambios en la marcha relacionados con la degradación patológica asociados con la pérdida de estabilidad, y por consecuencia con un aumento en el riesgo de caídas en personas adultas mayores.

Los trabajos de análisis de la marcha dentro del área de ciencias computacionales están principalmente orientados al procesamiento de imágenes provenientes de videos de personas caminando, para identificación de personas [43], [44], para la clasificación y/o la determinación del tipo de paso [1], [36], y para la detección de caídas con base en los datos adquiridos del seguimiento de personas [25]. Esto es, no se realiza un análisis de parámetros obtenidos en un análisis de la marcha para proporcionar información acerca de la calidad de la marcha de una persona. Nuestro trabajo se distingue de los anteriores pues lo que se quiere hacer es un análisis de los datos de la marcha para el desarrollo de un modelo que permita representar e identificar la degradación de habilidades motrices y evaluar el riesgo de presentar una caída.

Los trabajos que reportan procesamiento de los datos de la marcha lo hacen desde el punto de vista médico, quienes con base en su experiencia proponen factores que consideran influyen en la degradación de la marcha, y se apoyan en la medida del FAP para corroborar sus hipótesis, [13], [29]. El análisis de los datos de la marcha ofrece información a los expertos en el dominio para que proporcionen un diagnóstico acerca de la marcha de una persona. Dicha información permitirá determinar qué parámetros son los que muestran alguna degradación patológica así como la velocidad con la que están cambiando, y aún más las posibles repercusiones que esos cambios tienen a futuro sobre otros parámetros de la marcha, o incluso sobre el riesgo de presentar un evento como puede ser una caída. No es posible determinar todo lo anterior únicamente con base en la medida del FAP, pues esta medida es una calificación global de la marcha en un instante del tiempo.

Por lo anterior, es un reto y es necesario desarrollar técnicas que determinen cuáles parámetros de la marcha son relevantes para permitir la detección automática de comportamientos anormales, y aún más proporcionar información acerca del riesgo de padecer algún evento como son las caídas, con base en los cambios detectados en aquellos parámetros relevantes, identificados por las técnicas utilizadas o incluso utilizando información proporcionada por los expertos en el dominio. Esta información podría ofrecer un pre-diagnóstico o diagnóstico alternativo a los expertos médicos.

Técnicas de análisis de sobrevivencia permiten estudiar el efecto de un conjunto de variables predictivas sobre la función de riesgo, esto es, permiten estimar el riesgo de que un individuo presente el evento de estudio, teniendo en cuenta sus variables predictivas. El evento es un suceso de interés en el individuo objeto de estudio que se caracteriza por un cambio cualitativo brusco, por ejemplo, muerte, recurrencia de una enfermedad, o para abordar nuestro problema, caída. Esta técnica tiene la limitante de que únicamente trabaja con variables en su mayoría binarias o positivas crecientes, lo cual es necesario para interpretar los resultados pues sólo se permite evaluar el efecto de que una variable esté presente o no, pero no se permite considerar más información sobre una variable, e. g. qué tipo de cambio tuvo dicha variable.

Existen alternativas de modelado por parte de la Inteligencia Artificial (IA), como son Lógica Difusa, Redes Neuronales, Redes Bayesianas, por mencionar algunas, las cuales ofrecen distintas maneras de modelar datos temporales, así como representar las relaciones entre estos datos, y que permiten inferir algún suceso de interés con base en la información

de los datos representados o modelados. Existen dos retos importantes para el uso de estas técnicas en dominios médicos, el primero es que la construcción de dicho modelo depende de información acerca de los parámetros relevantes, así como de las relaciones entre ellos, la cual no siempre es conocida y por lo cual es necesario el uso del conocimiento de los expertos en dominio así como alternativas que permitan enriquecer dicho conocimiento. El segundo reto es que en dominios en donde se tienen pocos datos con respecto al número de variables, todas las técnicas mencionadas ven afectada su eficacia.

Es por eso que un análisis de los datos de la marcha así como el uso del conocimiento de los expertos en el dominio que permitan diseñar un modelo que capture la degradación de la marcha asociada al aumento en el riesgo de caída de una persona adulta mayor ayudaría al experto en el dominio a la toma de decisiones acerca del tratamiento a seguir para que dicha persona no sufra una caída y evitar así las consecuencias de dicha caída. El modelo debe permitir la representación y evaluación del cambio de los parámetros de la marcha en el tiempo, dicha evaluación debe proporcionar información acerca del aumento del riesgo de caída de acuerdo a los cambios detectados. Es importante señalar que los datos de la marcha, en general, y muy en concreto los datos que se tienen de personas adultas mayores, son incompletos, imprecisos y contradictorios, lo cual se conoce como datos inciertos.

El estudio de la representación de un sistema dinámico por medio de modelos gráficos es un tema de investigación actual. Varias aplicaciones destinadas a hacer predicción de eventos futuros a partir del análisis de datos temporales han sido desarrolladas para efectos de previsión en dominios médicos. Así, el objetivo de este trabajo es el diseño de un modelo de la marcha el cual permita evaluar el riesgo de caída de personas adultas mayores y el reto aunado a este objetivo es el aprendizaje automático de dicho modelo a partir de los datos. En específico se quiere realizar una predicción a 6 meses y 1 año del riesgo de caída de una persona adulta mayor. Se propone el uso de una EBN para el modelado de los datos y se tiene la problemática de construir dicho modelo de manera automática, esto es determinar las variables relevantes así como sus relaciones, los intervalos de los nodos temporales y el cálculo de las probabilidades conjuntas, que hasta donde sabemos es un tema investigación en ciencias computacionales.

El resto de este documento está organizado de la siguiente manera, en la sección 2 se hace una revisión de la técnica más representativa de análisis de sobrevivencia y de la técnica computacional seleccionada para el modelado de los datos de la marcha. En la sección 3 se presenta la revisión y discusión del estado del arte. La sección 4 describe la problemática atacada en este trabajo de tesis. En la sección 5 se presentan las preguntas de investigación así como los objetivos de este trabajo y las contribuciones del mismo. En la sección 6 se presenta la metodología a seguir para alcanzar los objetivos del trabajo. En la sección 7 se muestran los resultados preliminares que se tienen hasta el momento. En la sección 8 se hace una discusión de los resultados y conclusiones del trabajo, y en la sección 9 se expone el trabajo por realizar. Finalmente en la sección 10 se presenta el cronograma de actividades.

2. Marco teórico

2.1 Modelo de Regresión de Cox

Cuando hablamos de análisis de sobrevivencia nos referimos al análisis del tiempo de seguimiento (T) de cada individuo o caso de estudio hasta que ocurre un evento determinado, por ejemplo, la muerte. Dentro del análisis de sobrevivencia los métodos más utilizados son dos: Kaplan Meier [4] y el Modelo de Regresión de Cox [30], [32], [37], [55]. Estos métodos se centran en analizar el tiempo que demora en ocurrir el evento, no si éste ocurre o no. Por ello es que las fórmulas utilizadas se encargan de modelar el tiempo [46].

Cualquier estudio que implique seguimiento de pacientes tiene una duración finita, que se define en la etapa de planeación del estudio. Puede ocurrir que algunos pacientes en observación no hayan presentado el evento de estudio al momento de cerrar dicho estudio. Como no sabremos el tiempo que vivirá una de dichos pacientes sin presentar el evento de estudio, se podría pensar en no considerar la información de aquellos pacientes que no presentaron el evento de interés. Lo cual no es adecuado porque al descartarlos perderíamos la información acerca del tiempo vivido por cada uno de esos pacientes el cual es un dato muy útil que debe ser incorporado en nuestro análisis. A ese tipo de casos se les conoce como *censuras* o *casos censurados*.

Por ejemplo, en la figura 1, los casos A y C corresponden a pacientes o “unidades de observación” que presentaron el evento de estudio, como se observa el desenlace es el mismo para todos los pacientes, aunque los tiempos vividos sin presentar el evento durante el tiempo de observación sean muy distintos; en cambio B y D no presentan el evento dentro del tiempo de estudio, es decir, son casos censurados.

Uno de los mayores problemas con el análisis de tiempos es que, habitualmente, no todos los individuos que están siendo observados son ingresados al estudio y seguimiento en forma simultánea. Además, con frecuencia se pierden individuos del seguimiento por otras razones, distintas a la condición de presentar el evento de estudio. Es importante mantener en mente que los seguimientos no son perfectos y que en un estudio de fenómenos de este tipo que implica seguimiento médico muchas veces se depende de la disponibilidad del paciente.

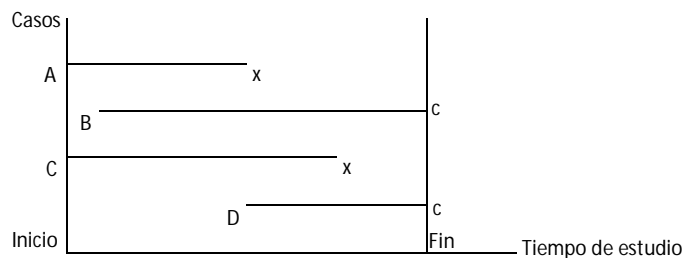


Figura 1. Ejemplo tipo de casos, donde x indica que el caso correspondiente presentó el evento de interés y la c que es un caso censurado.

Cuando se realiza un análisis de sobrevivencia generalmente lo que se desea saber es cómo influye una serie de características en una variable de estudio. En el ejemplo mostrado anteriormente, probablemente interese mucho más determinar qué variables están

influyendo en la ocurrencia precoz de un evento, más que saber el tiempo exacto en el que ocurre cada uno de dichos eventos. Para aproximarse a este problema se debe modelar la variable respuesta, que en este caso es el tiempo en el que ocurren los eventos. El método de Kaplan Meier [4] es el más simple que permite realizar pruebas de significancia estadística para evaluar la relevancia de una variable para determinar un evento de interés.

Kaplan Meier es un método por máxima verosimilitud y no paramétrico, es decir, no supone ninguna función de probabilidad, esto es, se basa en maximizar la función de verosimilitud de la muestra. Una muestra aleatoria de tamaño n , extraída de una población, estará formada por k ($k \leq n$) tiempos $t_1 < t_2 < \dots < t_k$ en los que se observan eventos. En cada tiempo t_i existen n_i "individuos en riesgo" (elementos de la muestra para los que el evento puede ocurrir) y se observan d_i eventos. Además en el intervalo $[t_i, t_{i+1})$ se producen m_i pérdidas.

Por ejemplo, verificar si la ocurrencia de un evento es distinta entre hombres y mujeres, de acuerdo al sexo. Sin embargo, no permite modelar la variable respuesta (que es el tiempo) mediante variables predictivas, como se hace en regresión. Así, Kaplan Meier permite comparar grupos, no cuantificar la influencia de diferentes variables asociadas al evento de estudio o variables predictivas.

Una forma para realizar predicción con base en el estudio de variables predictivas es usar modelos de regresión. Estos modelos se basan en suponer un comportamiento conocido de la variable respuesta (lineal, exponencial, etc.), generar un modelo aproximado y cuantificar la influencia de las variables predictivas. Esto mismo se puede aplicar al modelado de sobrevivencia, ya que existen funciones matemáticas que se aproximan a las curvas de sobrevivencia y se pueden utilizar para ver la influencia de variables predictivas. Algunos ejemplos de modelos matemáticos utilizados para estos análisis se muestran en la figura 2.

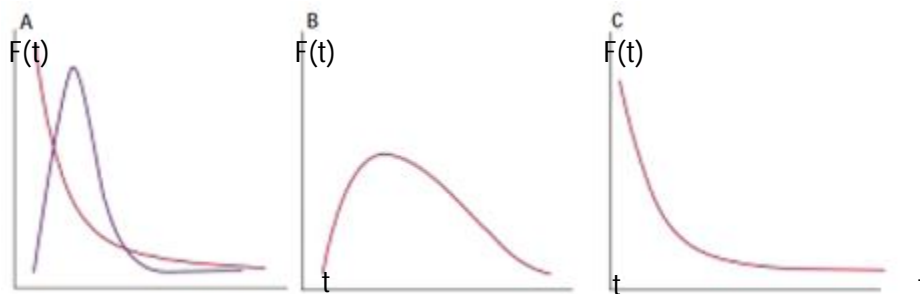


Figura 2. Algunos modelos matemáticos utilizados para análisis de sobrevivencia. A. Curvas de distribución de Weibull, B. Distribución Gamma, C. Distribución Exponencial.

Cabe mencionar que, en la práctica, no siempre los datos se aproximan a una curva conocida. El método más utilizado para resolver este problema es el modelo de regresión de Cox, ya que tiene la gran ventaja de que no se basa en modelar una curva de sobrevivencia predeterminada. De hecho, este modelo no tiene curva de sobrevivencia predefinida, pero sí permite ver la influencia de variables predictivas en la respuesta.

El Modelo de Riesgo Proporcional de Cox está definido por la fórmula 1 [20]:

$$H(t) = H_0(t) \times e^{X_1\beta_1 + X_2\beta_2 + \dots + X_k\beta_k} \quad (1)$$

Donde:

$H(t)$: función de riesgo (en inglés *hazard ratio*) de que el evento ocurra al tiempo t .

$H_0(t)$: función de riesgo mínima al tiempo t

X_k : variable predictora k

β_k : constante asociada a la variable k

En el modelo de Cox se define una función de riesgo mínima o base, cada vez que hay un evento. Éste es el riesgo de supervivencia independiente de las variables predictivas – denominado $H_0(t)$ – y define un perfil de riesgo según cada variable predictiva involucrada representada como $\beta_k x_k$. El riesgo inducido por la presencia de cada variable predictiva se ve afectado por el riesgo base $H_0(t)$.

Finalmente para determinar en qué proporción se modifica el riesgo de caída en presencia de una de las variables asociadas a éste, así como el tiempo asociado a la ocurrencia de dicho riesgo se aplica la función de supervivencia de Cox, la cual se muestra en la fórmula 2.

$$S(t) = e^{-H_0(t) * e^{M(X)}} \quad (2)$$

Donde:

$S(t)$: función de supervivencia al tiempo t .

$M(X) = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$, representación lineal de las variables predictivas.

$H_0(t)$: función de riesgo mínima al tiempo t

X_k : variable predictiva k

β_k : constante asociada a la variable k

El modelo de regresión de Cox tiene diversos supuestos que deben cumplirse para que la interpretación de los datos sea válida. En principio este método permite conocer la influencia de la presencia o ausencia de alguna variable predictiva, así dicha variable sólo puede ser binaria o a lo más creciente en el tiempo. El modelo de regresión de Cox realiza una prueba de significancia estadística entre la presencia de una variable y el aumento en el riesgo del evento de interés, para determinar la relevancia de dicha variable. El modelo de regresión de Cox, al igual que otras técnicas de modelado, muestra poca eficacia cuando se tienen pocos datos y muchas variables. Debido a que es necesario un buen número de casos para obtener estadísticas válidas. Además este modelo supone independencia entre las variables predictivas.

2.2 Redes Bayesianas

Formalmente, las redes bayesianas son gráficos acíclicos dirigidos cuyos nodos representan variables y los arcos que unen a estos nodos representan dependencias condicionales entre

las variables. Los nodos pueden representar cualquier tipo de variable, ya sea un parámetro medible (o medido), una variable latente o una hipótesis.

Si existe un arco que une un nodo A con otro nodo B, A es denominado un padre de B, y B es llamado un “hijo” de A. El conjunto de nodos padre de un nodo X_i se denota como $padres(X_i)$. Un gráfico acíclico dirigido es una red bayesiana relativa a un conjunto de variables si la distribución conjunta de los valores del nodo puede ser escrita como el producto de las distribuciones locales de cada nodo y sus padres, según se expresa en la fórmula 3.

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^n P(X_i | padres(X_i)) \quad (3)$$

Si el nodo X_i no tiene padres, su distribución local de probabilidad se toma como *incondicional*, en otro caso es *condicional*. Si el valor de un nodo es observable - y por tanto etiquetado como observado dicho nodo es un nodo de *evidencia*.

Las redes bayesianas son una alternativa para minería de datos, las cuales tienen varias ventajas:

- Permiten aprender sobre relaciones de dependencia y causalidad.
- Permiten combinar conocimiento con datos.
- Evitan el sobre-ajuste de los datos.
- Pueden manejar bases de datos incompletos.

La obtención de una red bayesiana a partir de datos es un proceso de aprendizaje, el cual consiste en dos aspectos:

1. *Aprendizaje paramétrico*: dada una estructura, se busca obtener las probabilidades *a priori* y condicionales requeridas.
2. *Aprendizaje estructural*: se busca obtener la estructura de la red Bayesiana, es decir, las relaciones de dependencia e independencia entre las variables involucradas.

Aprendizaje Paramétrico

El aprendizaje paramétrico consiste en encontrar los parámetros asociados a una estructura dada de una red bayesiana. Dichos parámetros consisten en las probabilidades *a priori* de los nodos raíz y las probabilidades condicionales de las demás variables, dados sus padres.

Si se conocen todas las variables consideradas en el fenómeno modelado, es fácil obtener las probabilidades requeridas. Las probabilidades previas corresponden a las marginales de los nodos raíz, y las condicionales se obtienen de las conjuntas de cada nodo con su(s) padre(s).

Para que se actualicen las probabilidades con cada caso observado, éstas se pueden representar como razones enteras, y actualizarse con cada observación.

Aprendizaje Estructural

Las técnicas automáticas para aprendizaje estructural de redes bayesianas consisten de dos aspectos principales:

1. Una medida para evaluar qué tan *buena* es cada estructura respecto a los datos.
2. Un método de búsqueda que genere diferentes estructuras hasta encontrar la *óptima*, de acuerdo a la medida seleccionada.

Las técnicas de aprendizaje estructural dependen del tipo de estructura de red: árboles, poliarboles y redes multiconectadas. Otra alternativa es combinar conocimiento subjetivo del experto, el cual se obtiene por la experiencia de éste, con técnicas de aprendizaje. Para ello se parte de la estructura dada por el experto, la cual se valida y mejora utilizando datos estadísticos.

El encontrar la estructura óptima de una red bayesiana es difícil, ya que el espacio de búsqueda es muy grande. Por ejemplo, hay más de 10^{40} diferentes estructuras para 10 variables. Por esto, se utilizan estrategias de búsqueda heurísticas que encuentran una solución aceptable pero, generalmente, no óptima. Otra alternativa es combinar los métodos automáticos con conocimiento de expertos.

3. Estado del arte

La mayor parte de los trabajos que hacen referencia al análisis de la marcha están enfocados a la obtención de datos propios de la marcha humana, largo de paso, cadencia, velocidad, etc. Los datos adquiridos son presentados a los expertos para que ellos los interpreten e identifiquen anomalías con base en su experiencia [12], [45].

Existe tecnología para la adquisición de datos de la marcha que además proporciona una evaluación global acerca de la calidad de la marcha de una persona, tal evaluación se basa en una medida conocida como FAP, *Functional Ambulation Profile*, [38]. La limitación de este enfoque es que al médico sólo se le proporciona un valor global de la calidad de la marcha, pero no se le da información acerca de qué parámetros de la marcha son los que están afectando dicha evaluación. Esto dificulta conocer la causa exacta de lo que ocasiona que una persona evaluada tenga problemas de la marcha, y por consiguiente no permite determinar con precisión qué tipo de tratamiento es el más adecuado para corregir la deficiencia en la marcha ni identificar los cambios asociados a la degradación de la marcha, pues en cada captura de la marcha sólo se obtiene una “fotografía” de la marcha actual de un individuo.

En dominios médicos generalmente se hace análisis de sobrevivencia para determinar el riesgo de presentar un evento de interés en el estudio de un grupo de personas con algún tipo de enfermedad. Una característica fundamental del análisis de sobrevivencia es que permite utilizar la información de datos que contienen información parcial sobre el suceso estudiado, conocidos como casos censurados [46]. El objetivo principal de este análisis es proporcionar información para la toma de decisión del tratamiento a seguir para una persona con problemas cuyas manifestaciones cuantificables no son fáciles de identificar.

Cuando se hace un análisis de sobrevivencia se hace una representación lineal de las variables predictivas $M(X)$, es decir el modelado de los datos es una función lineal. El modelo de regresión de Cox [30], [32], [37], [55] es una técnica de análisis de

sobrevivencia ampliamente utilizada en dominios médicos. Este tipo de métodos se utiliza principalmente para evaluar el efecto de algún tipo de tratamiento en una enfermedad. La mayoría de las variables analizadas con *Cox* son binarias o crecientes positivas; el modelo utiliza una función lineal para representar la ausencia o presencia de una variable, lo anterior restringe el uso de otro tipo de variables para no simplificar el modelado de los datos. Así, en problemas en donde se requiere el análisis y representación de datos temporales continuos e inciertos, así como las relaciones entre dichos datos, es necesario otro tipo de representación, pero es de interés conocer la eficacia del modelo de *Cox* en nuestro problema.

En el área de ciencias computacionales para el diagnóstico y predicción en dominios como la medicina y la industria se requiere de modelos que permitan representar adecuadamente el conocimiento de carácter temporal y manejar la incertidumbre en los datos analizados, además es necesario considerar la evolución de los datos en el tiempo y estos datos son en muchas ocasiones contradictorios, incompletos, o tienen algún tipo de ruido. Así, existen diversas alternativas para tal fin las cuales discutiremos a continuación.

Los algoritmos de regresión, algoritmos basados en redes neuronales, en redes probabilistas y en máquinas de soporte vectorial se han comparado y aplicado con éxito para clasificar datos biomédicos tales como, resultados de pruebas de sensibilidad al campo visual en la detección de glaucoma [10], imágenes digitales de lesiones de piel pigmentada [14], y señales acústicas del llanto de bebés [2], por mencionar sólo algunos casos. No hay una técnica de modelado mejor que otra *per se*. El desempeño de dichas técnicas de modelado suele depender de diversos factores, entre los que se cuentan, por ejemplo, el tipo y la representación de los datos.

La Lógica Difusa (FL) ha sido utilizada para diagnosticar la presencia de cojera en vacas [9] y los mapas cognitivos difusos se han aplicado para la predicción de enfermedades infecciosas [41]. En los trabajos revisados que utilizan Lógica Difusa no se modela la información temporal relacionada con los datos procesados para hacer la predicción, generalmente este tipo de trabajos se enfoca en la representación del conocimiento del experto en el dominio. Una de las carencias del uso de FL en dominios médicos es que esta técnica se aplica sobre un pequeño número de parámetros de los datos analizados (generalmente dados por los expertos en el dominio), además de que esta técnica llega a verse limitada para representar la incertidumbre, así como las dependencias y evolución de los datos en el tiempo.

Las Redes Neuronales (NN) han sido utilizadas para el pronóstico de infartos [7], Dorsalgia [8], estimación de la probabilidad de sobrevivir después de sufrir algún trauma [35] y variantes como el uso de FL con NN para el diagnóstico de cáncer de ovario [33]. Por otra parte, una combinación de NN con algoritmos genéticos se ha aplicado para la predicción de la contaminación del aire [31]. Las Redes Neuronales son quizá una de las herramientas más utilizadas para apoyo al diagnóstico de enfermedades en medicina, pero una de sus principales limitaciones es que de acuerdo al número de variables que se quiere analizar aumenta el número de casos requeridos, para tener un modelo efectivo. Al igual que los modelos basados en FL, un modelo obtenido con una NN tiene limitaciones para capturar

relaciones temporales y la incertidumbre de los datos. Existen otro tipo de NN conocidas como redes neuronales recurrentes, las cuales permiten el modelado de datos temporales pero cuya desventaja es que utilizan estructuras más complejas para el manejo del tiempo [18].

Los Modelos Ocultos de Markov (HMM) han sido combinados con distintas técnicas para mejorar la estructura de un HMM, como el uso del algoritmo *Successive State Splitting* (SSS) para la predicción de inmunidad ante determinadas bacterias [39]. Los Modelos Ocultos de Markov permiten hacer un mejor modelado de la información temporal en los datos, a diferencia de FL o NN, pero al igual que en las redes neuronales para generar un modelo efectivo basado en HMM es necesario tener un número grande de observaciones. Una característica importante en este tipo de trabajos es que se tiene evidencia de que la construcción del modelo puede ser apoyada con técnicas adicionales para mejorar la estructura de éste.

El Aprendizaje Computacional con Máquinas de Soporte Vectorial (SVM) también se ha aplicado para análisis de datos médicos [48], [49], específicamente las SVM han sido aplicadas para la predicción de hipertensión arterial y combinadas con algoritmos de optimización para diseño de un control predictivo [47]. Las Máquinas de Soporte Vectorial inicialmente se usaron para problemas de clasificación binaria, pero se ha extendido su uso a problemas de agrupamiento, clasificación multiclase y regresión simple. El uso de SVM para predicción en dominios médicos ha mostrado las siguientes limitaciones, el problema de predicción del riesgo generalmente es reducido a un problema de clasificación binaria: hay o no riesgo, y en general el uso de SVM no mejora considerablemente la predicción con respecto a las técnicas de análisis de sobrevivencia, debido a que no hay un adecuado manejo de la información de casos censurados.

Una de las técnicas más utilizadas que permite el modelado de relaciones temporales así como la incertidumbre en los datos y que permiten hacer razonamiento temporal bajo dicha incertidumbre son las Redes Bayesianas (BN), las cual han se han aplicado para la predicción del brote de bacterias en la naturaleza [24], predicción de precipitaciones [3], modelado de efectos en tratamientos médicos [52] y variantes como las Redes Bayesianas Dinámicas (DBN) para el diagnóstico de pulmonía asociada al uso de aparatos médicos [11]; otra variante son las redes de eventos para predicción de fallas en ambientes industriales [15], [16], [17].

Al igual que los HMM, las Redes Bayesianas son una buena alternativa para la representación de conocimiento temporal, en específico las Redes Bayesianas Dinámicas, DBN. Este tipo de BN ha mostrado un mejor desempeño que los HMM cuando se tiene un número pequeño de observaciones [40]. Las DBNs son apropiadas para las tareas de monitoreo, ya que representan de forma explícita el estado del sistema en cada momento. Ahora bien, la manera en que se hace inferencia en este tipo de BN permite hacer inferencia no sólo de un evento de interés sino de varios, incluso sin conocer toda la información para la inferencia del evento, pero esto depende la estructura del modelo. La naturaleza de nuestro problema puede requerir además otro tipo de mecanismos de inferencia, que permitan obtener una función de riesgo de presentar un evento en un tiempo determinado.

Una variante de BNs son las Redes Bayesianas de Eventos (EBN) en las cuales cada valor de una variable representa el tiempo en el cual un determinado evento puede ocurrir. Así para dominios en los que se quiere dar un diagnóstico o predicción de fallas en ambientes dinámicos, como dominios médicos, las EBNs son más apropiadas que las DBNs, debido a que las fallas pueden ser fácilmente representadas por medio de eventos [15]. Esto es así pues el representar y conocer el tiempo en que ocurre un evento es fundamental para determinar el posible desenlace de éste, pues los cambios instantáneos pueden iniciar un conjunto de cambios intermedios que generan cambios subsecuentes, asociados a diferente tipo de eventos, quizá similares. Por ejemplo en nuestro problema, de acuerdo a qué tan rápido está cambiando un parámetro determinado de la marcha se puede establecer una asociación a diferentes causas, degradación normal, degradación patológica o incluso mejora en la degradación de la marcha.

Por lo anterior consideramos que las EBNs son adecuadas para el modelado de la información temporal de la marcha de personas, pues se quiere una representación de cambios (eventos) en los parámetros de ésta, que indiquen una degradación de la marcha y por lo tanto un aumento significativo del riesgo de caída, o que sólo indique un cambio asociado al deterioro normal de la marcha y por tanto un cambio poco significativo en el riesgo de caída. Para esto se quiere una representación de las variables de la marcha así como su interrelación con otras variables, además de la evolución de dichas variables en el tiempo, en específico en qué momento y circunstancias ocurre, para determinar el efecto de dichos cambios y proporcionar una medida de riesgo de un evento de interés, caídas.

Identificamos dos categorías de este tipo de redes: Redes Bayesianas de Nodos Temporales (TNBN, por sus siglas en inglés) [6] o las Redes de Eventos Probabilísticos en Tiempo Discreto (NPEDT, por sus siglas en inglés) [15], [16]. Ambos tipos de EBNs han sido utilizados en ambientes médicos e industriales, mostrando ser herramientas factibles de aplicar en dichos dominios, pues en dichas EBNs, cada valor que puede tener una variable de la red representa el momento en que un determinado evento puede ocurrir. Ambos tipos de EBNs tienen un inconveniente: son construidas manualmente con ayuda de los expertos en el dominio, esto es, el experto proporciona las variables, las relaciones y el tipo de efecto que tienen dichas variables en la predicción de un evento futuro. Nosotros proponemos el uso de técnicas de selección de variables para determinar cuáles son las variables relevantes, así como el uso de técnicas de análisis de sobrevivencia las cuales proporcionen la relevancia de un conjunto de variables para predecir un evento futuro para determinar con base en esa información las relaciones entre las variables del modelo, y que además proporcionen un mecanismo para determinar los intervalos de ocurrencia de los cambios.

Las técnicas de Inteligencia Artificial (IA) como Lógica Difusa [41], Redes Neuronales [33], Modelos Ocultos Markov [39], Redes Bayesianas [11] y Maquinas de Soporte Vectorial [47], han sido utilizadas para modelar y representar la información temporal en los datos, el conocimiento de los expertos, la incertidumbre en los datos y la falta de información. El resultado ha producido una serie de herramientas opcionales a las técnicas de análisis de sobrevivencia para apoyo a la predicción en dominios médicos. Aunque dichas técnicas han mostrado buenos resultados en dominios médicos, del orden de 70 a

80% de efectividad, no tenemos conocimiento del uso y factibilidad de dichas técnicas con datos temporales de la marcha. Además, no todas las técnicas mencionadas hacen una predicción a futuro sino que terminan haciendo una clasificación, es decir no proporcionan una medida cuantificable del riesgo de presentar un evento a lo largo de un periodo de tiempo. La mayoría de dichas técnicas descarta la información de los casos censurados.

Los trabajos mencionados que realizan un análisis de datos temporales para hacer una predicción en dominios médicos no realizan selección alguna de atributos, generalmente trabajan con atributos proporcionados por expertos o citados en la literatura correspondiente. Debido a la complejidad de procesamiento y almacenamiento de los datos, cuando se tiene un gran número de factores predictivos a analizar en grupos pequeños de pacientes, es deseable seleccionar aquellos factores predictivos más representativos para hacer la predicción del evento de interés.

Desde el punto de vista médico esto es importante pues permite validar las hipótesis de los médicos acerca de los parámetros que ellos han determinado como relevantes, y aún más esta selección permite proporcionar información adicional a los médicos para la toma de decisiones. Desde el punto de vista computacional la selección de parámetros es fundamental pues para la mayoría de las técnicas presentadas aquí, se sabe que la inclusión en el análisis de datos poco relevantes, o las dependencias mal consideradas deterioran el desempeño de dichas técnicas. Dicha selección también permitiría el desarrollo de dispositivos móviles enfocados a la captura exclusiva de dichos parámetros.

Así, un reto en este tipo de trabajos es determinar cuáles son aquellos atributos temporales y las relaciones entre ellos que dan más información para determinar el riesgo de experimentar un evento de estudio. Existen en la literatura diversas técnicas para realizar selección de atributos sobre datos estáticos, entre las técnicas más conocidas: *wrappers* [27] o filtros [54]. Y por otro lado se han reportado también trabajos acerca de selección de atributos sobre datos temporales [34], [40], [51]. Estos trabajos han sido aplicados para mejorar la estructura de un HMM o BN y en este trabajo se utilizará la selección de atributos para mejorar la estructura de una EBN, con distinto tipo de datos, e. g. datos censurados.

La selección de variables es importante debido a que en nuestro problema tenemos pocos datos con respecto al número de variables y seleccionando las más relevantes se espera diseñar un modelo eficaz. Así, se utilizarán técnicas de selección de variables como las mencionadas anteriormente. También consideramos el uso de técnicas de *oversampling* o incluso de ser necesario una combinación de ambas, sobre lo cual también existe muy poco trabajo reportado en la literatura.

4. Problemática

Hasta donde sabemos, la toma de decisión sobre el tratamiento a seguir para una persona con problemas de estabilidad es realizada por personas expertas en el dominio, y esta decisión es hecha con base en la experiencia de los expertos apoyada con los datos obtenidos por los instrumentos utilizados para realizar el análisis clínico de la marcha de una persona, como son cámaras, sensores electromagnéticos, etc.

Hasta donde se sabe, la identificación de los parámetros relevantes y las relaciones entre ellos que permitan detectar la pérdida de estabilidad, la cual implique un aumento en la probabilidad de sufrir una caída en personas adultas no están definidos. Es por eso que se propone el uso de técnicas computacionales para abordar la problemática del modelado y determinación del tipo de riesgo de caída de un individuo, de acuerdo a los cambios detectados en la marcha de éste.

En esta investigación nos centraremos en la predicción. La predicción es el proceso por el cual el comportamiento futuro de un sistema es estimado con base en la caracterización de éste. La tarea de predicción es compleja y difícil debido a múltiples razones, tales como: alta sensibilidad a las condiciones iniciales en el sistema modelado, dificultad en la determinación de las tendencias o en el reconocimiento de patrones en presencia de ruido en los datos, para ello se propone el uso de modelos que permitan representar datos temporales e inciertos.

Para el análisis del decaimiento de habilidades motrices es necesaria una **representación adecuada** de la dependencia del tiempo con respecto a cambios en los parámetros asociados a ser factores de riesgo de caída. Dado que los cambios pueden ocurrir en distintos tiempos para diferentes personas, la inferencia del riesgo debería de ser hecha en términos de los cambios que han ocurrido en el pasado o que podrían ocurrir en el futuro. Por ejemplo, un cambio puede ocurrir más rápido en una persona que en otra y la información acerca de la ocurrencia del cambio o la rapidez con la que está cambiando un parámetro ayudarían a determinar si el riesgo de caída está asociado a una degradación patológica o a una degradación normal de la marcha. Así, la inferencia del riesgo no debería de estar sujeta a un determinado tiempo, sino al tiempo en el que el cambio ha ocurrido.

Dado que los datos en dominios médicos como el nuestro son afectados por diferentes fuentes de incertidumbre, e. g., errores de medición, variabilidad de los individuos dentro del estudio, consideramos un modelo probabilístico, que permita una representación de dichos datos así como la inferencia a partir de estos, en particular una EBN.

El problema principal es que la determinación de la estructura del modelo está estrechamente ligada al conocimiento de los expertos en el dominio, pero existen dominios como el nuestro en donde no se conoce con certeza la información necesaria para la construcción del modelo, pues es necesario conocer las variables relevantes así como las relaciones entre ellas y cómo los cambios de éstas afectan al aumento o disminución de presentar un determinado evento, y finalmente saber la rapidez con la que está cambiando las variables del problema y cómo esta rapidez afecta al riesgo de presentar un evento.

Es por eso que un reto es la **construcción automática de un modelo** a partir de técnicas de análisis de los datos que nos permitan complementar así como corroborar el conocimiento de los expertos en el dominio, para determinar cuáles son variables más significativas, las relaciones temporales entre ellas, y para determinar la rapidez de cambio y el efecto que tiene en un evento futuro de interés.

El problema principal para la construcción del modelo es que no se conocen las relaciones entre las variables, ni en qué proporción están afectando el riesgo de presentar un evento de interés. Aunado a lo anterior se tienen limitaciones por disponer en estos estudios de pocos datos y muchas variables, lo cual afecta la construcción de un modelo a partir de dicha información. Así, se tiene el problema de determinar cuáles son las variables más significativas, cómo están relacionadas entre ellas y cómo afectan la aparición de futuros cambios relacionados con la degradación de la marcha patológica.

Después de lo expuesto antes, optamos por el uso de EBNs pues este tipo de red nos permite modelar e inferir con respecto a la ocurrencia de eventos, y aun más nos permite saber qué tan rápido puede cambiar una variable y cómo la rapidez de ese cambio afecta a otras variables, y finalmente determinar qué consecuencias y en qué tiempo aproximado tendrán lugar, de acuerdo a los eventos registrados y a los tiempos en los que ocurrieron estos. De este modo podemos saber si casos específicos presentan una degradación patológica de la marcha o una degradación normal de la marcha.

5. Preguntas, objetivos y contribuciones

Los objetivos de este trabajo de investigación se enfocan en dar respuesta a las siguientes preguntas de investigación.

- ¿Cómo detectar y representar los cambios en la marcha humana los cuales den indicio del decaimiento de habilidades motrices relacionado con pérdida de estabilidad, y en específico con el aumento de la probabilidad de sufrir una caída, a partir de datos biomédicos obtenidos con instrumentos para análisis de la marcha?
- ¿Cómo identificar los atributos relevantes, así como las relaciones entre ellos, del conjunto de los datos biomédicos obtenidos con instrumentos para análisis de marcha que indican el decaimiento de las habilidades mencionadas?
- ¿Cómo modelar el cambio de la probabilidad de riesgo de caída, a partir del análisis de las tendencias de las variables en el tiempo?
- ¿Cómo construir un modelo de la marcha de manera automática a partir de los datos?

5.1 Objetivo general

- El objetivo general de este trabajo es detectar y representar el decaimiento de habilidades motrices en la marcha asociadas a la pérdida de estabilidad en adultos mayores, así como determinar la probabilidad de riesgo de caída de una persona a lo largo de un periodo de tiempo.

5.2 Objetivos específicos

- Representación de las variables de la marcha de personas adultas mayores.
- Representación de la evolución en el tiempo y las relaciones entre las variables de la marcha de personas adultas mayores.
- Aplicación de técnicas que permitan trabajar con pocos datos y muchas variables.
- Construcción automática de un modelo que represente datos de la marcha humana y permita detectar cambios tempranos asociados con la pérdida de estabilidad, los

cuales incrementen el riesgo de sufrir una caída, a partir de datos temporales biomédicos proporcionados por los expertos del dominio.

- Desarrollo de métodos que permitan la construcción automática de un modelo basado en una EBN de la marcha.

5.3 Contribuciones

- Construcción de un modelo de la marcha humana basado en una EBN que permita representar y detectar cambios asociados al incremento del riesgo de la marcha patológica de caída en adultos mayores, a partir del análisis de datos biomédicos adquiridos en distintos periodos de tiempo. Además, se desea a partir del modelo previo, proporcionar una medida de riesgo de caída asociada a una degradación patológica de la marcha de una persona adulta mayor.
- Desarrollo de métodos que permitan construir un modelo basado en una EBN a partir de pocos datos y muchas variables así como las relaciones entre ellos, para identificar tendencias de degradación patológica de la marcha en el tiempo. Técnicas de selección de atributos, técnicas de determinación de velocidad de cambio, técnicas de evaluación del efecto de los cambios y rapidez de las variables para determinar el riesgo de presentar una caída.

6. Metodología

A continuación se describen los métodos y técnicas a utilizar para alcanzar el objetivo planteado.

1. Caracterización de datos biomédicos

- 1.1. Adquirir los datos de la marcha de personas adultas mayores, los cuales serán proporcionados por el laboratorio de Análisis de Movimiento del INR de la ciudad de México. Durante el análisis de la marcha los datos involucrados con la manera de caminar de una persona son obtenidos y almacenados. Dado que en nuestro caso estamos interesados en la detección de decaimiento, es necesario tener registros de los datos de la marcha de diferentes personas en diferentes momentos.
- 1.2. Caracterizar los tipos de datos, los instrumentos con que son obtenidos, los rangos de acción, los tiempos en que han sido tomados y el tipo de personas a quienes se les ha dado seguimiento.
- 1.3. Trabajar periódicamente con los investigadores del INR para obtener información del tipo de datos que nos están proporcionando.

2. Análisis y modelado de los datos

- 2.1. Revisar y seleccionar las técnicas para el modelado de datos biomédicos que permitan representar y detectar cambios en los datos, así como técnicas que proporcionen una medida de los cambios en el riesgo de presentar un determinado evento como consecuencia de los cambios detectados con el modelado de los datos.
- 2.2. Probar las técnicas seleccionadas sobre los datos de la marcha proporcionados por el INR. Lo anterior es importante pues nos permitirá tener evidencia de la eficacia y de las limitantes de las técnicas seleccionadas al comparar la clasificación de este modelo con la clasificación de expertos, y así tener una base con la cual decir si las técnicas utilizadas son suficientes para resolver la problemática planteada.

- 2.2.1. Prueba de técnicas de análisis de sobrevivencia.
- 2.2.2. Prueba de técnicas de Inteligencia artificial.
- 2.3. Proponer un modelo que permita considerar el cambio de los datos a través del tiempo, así como la influencia que tiene dicho cambio en el riesgo de presentar un determinado evento, y permita proporcionar una medida del riesgo asociado a una degradación patológica de la marcha a lo largo de un periodo de tiempo. **En esta parte se encuentra una de las principales contribuciones de nuestro trabajo**, pues hasta donde sabemos no existen trabajos reportados que hagan una representación de la evolución de la marcha humana y que permita evaluar el riesgo de caída de una persona.
- 2.4. Proponer un método para la construcción automática del modelo anterior, esto es deseable debido a que como hemos mencionado, las diversas opciones para modelar dependen exclusivamente de la experiencia de los expertos en el dominio, pero existen dominios como el nuestro en donde los expertos no tienen seguridad de la información necesaria para la construcción de un modelo de la marcha patológica. Para ello es necesario lo siguiente:
 - 2.4.1 Determinación de atributos relevantes así como las relaciones entre ellos
 - 2.4.1.1. Utilizar el conocimiento *a priori* proporcionado por los expertos en el dominio para reducir el número de parámetros a considerar en el análisis.
 - 2.4.1.2. Utilizar técnicas de selección de atributos discutidos en la literatura especializada, donde encontramos dos métodos: los empacados (*wrappers*) y los filtros, [27], [54].
 - 2.4.1.3. Utilizar técnicas de selección de atributos relevantes de datos temporales [34], [40] y evaluar la efectividad del uso de los parámetros encontrados como relevantes con dichas técnicas de selección.
 - 2.4.1.4. Proponer técnicas de selección de atributos en datos temporales
 - 2.4.1.4.1. Comparar clasificación aplicando los puntos de la metodología 4.4.1.2 y 2.4.1.3.
 - 2.4.1.4.2. Seleccionar aquellas variables que sean identificadas que influyen en el aumento del riesgo de caída, utilizando análisis de sobrevivencia.
 - 2.4.1.4.3. Identificar cuáles son las variables que cambian la marcha de personas que tienen un decaimiento normal y descartar dichas variables como significativas para determinar un decaimiento patológico asociado al aumento de riesgo de caída.
 - 2.4.2 Determinación de la rapidez con la que ocurren los cambios
Aplicación de técnicas de análisis de sobrevivencia
 - 2.4.3 Determinación del efecto en el riesgo de los cambios en los atributos
Aplicación de técnicas de análisis de sobrevivencia.
 - 2.4.4 Determinación de los parámetros del modelo
Aplicación de técnicas de aprendizaje paramétrico
 - 2.4.5 Determinación de los intervalos
Análisis de las graficas de la función sobrevivencia

3. Identificación de atributos relevantes

La necesidad de la reducción de atributos surge debido a que se tiene una gran cantidad de atributos, de un orden mayor a 30, provenientes del análisis clínico de la marcha considerando el número de casos de estudio; alrededor de 18 actualmente. Por ello, el tratamiento de los datos para los algoritmos de minería tradicionales es costoso computacionalmente. El determinar aquellos atributos relevantes de la marcha para detección de pérdida de estabilidad ayudará a la adquisición y análisis de un subconjunto de atributos.

3.1. Utilizar técnicas de réplica de datos, esto es, aumentar el número de casos utilizando *oversampling*, para tener más casos y así evitar la degradación de tener muchos atributos contra pocos datos.

3.2. Hacer selección de atributos como alternativa al uso de *oversampling*, para tener pocos datos con pocos atributos.

3.2.1. Uso del punto 2.4.1 de la metodología.

3.3. Proponer métodos para la construcción de modelos con pocos datos y muchas variables

3.3.1. Combinación de *oversampling* con selección de atributos.

3.3.1.1. Utilizar *oversampling* sobre los datos que tenemos y posteriormente hacer selección de atributos.

3.3.1.2. Realizar una selección de atributos inicial sobre los datos y posteriormente hacer *oversampling*.

4. Evaluación

4.1. Comparar la eficacia del modelo con respecto a técnicas de análisis de sobrevivencia.

4.2. Comparar la eficacia del modelo obtenido, con respecto a modelos similares contruidos con ayuda de los expertos en el dominio.

4.3. Calcular la precisión en la clasificación del modelo con los datos iniciales, utilizando validación cruzada.

4.4. Calcular la precisión del modelo propuesto con datos de expedientes cerrados. Es decir, datos distintos a los proporcionados inicialmente.

7. Trabajo realizado y resultados preliminares

Los resultados obtenidos hasta hoy consisten en lo siguiente:

- Caracterización de una parte de los datos de la marcha de personas adultas mayores.
- Selección de las técnicas a utilizar.
- Evaluación de las técnicas seleccionadas con los datos caracterizados.
- Construcción automática de un primer modelo basado en una EBN.

7.1 Caracterización de los datos.

Actualmente se está colaborando con los investigadores del Laboratorio de Análisis del Movimiento del INR, con el fin de obtener las bases de datos del análisis de la marcha de personas adultas mayores, concretamente mujeres de entre 50 a 70 años con diagnóstico de osteoporosis, y algunas con registro de caídas previas al inicio del estudio. Esos datos se han registrado durante un estudio conducido por los investigadores del INR durante un periodo de 3 años.

A cada una de las personas a las que se les ha dado seguimiento se le realizan diversos estudios de la marcha, entre los principales está un estudio con un instrumento conocido como el sistema *gaitrite* y un estudio de estabilidad conducido con una plataforma de estabilidad estática. Un ejemplo del registro de la información obtenida en cada estudio se muestra en el Anexo A de este documento.

El INR cuenta con una base de datos de 300 expedientes, cada uno de los cuales contiene diferente número de registros de datos la marcha capturados en distintos tiempos. La falta de uniformidad en el número de registros por persona se debe a que en dicha base de datos existen datos de personas con y sin problemas de marcha, con y sin datos de caídas, a que algunas personas abandonaron el estudio, y a que otras se incorporaron en distintos momentos al estudio. Hasta el momento se nos ha proporcionado la información resumida en la tabla 1.

Tipo de expediente	No. de Expedientes
Personas con un solo registro (sin problemas de la marcha)	8
Personas con más de un registro (con problemas de la marcha)	18
Personas que han sufrido fractura	8
Expedientes totales	34

Tabla 1. Información de los expedientes proporcionados por el INR.

En el trabajo realizado hasta ahora se han utilizado los 18 expedientes con más de un registro y de estos se obtuvo el dato de aquellas personas que han sufrido una caída a lo largo del seguimiento. En promedio el tiempo del seguimiento de dichas personas es de 2.5 años, que comprenden generalmente datos de la marcha de dichas personas cada 6 meses. El resumen de los expedientes mencionados se muestra en la tabla 2.

Como hemos mencionado, sólo se tienen y se han utilizado los datos de la marcha de los diferentes registros de cada uno de los expedientes analizados. La información proporcionada en cada uno de dichos registros se resume en la tabla 3.

Información sobre expedientes	
No. de expedientes	18
Expedientes con información acerca de caída	17
Expedientes con caída	15
Expedientes sin caída	2
Información sobre las caídas	
Expedientes con reporte de 1 caída	5
Expedientes con reporte de 2 caídas	6
Expedientes con reporte de 3 caídas	2
Expedientes con reporte de 7 caídas	1
Expedientes con reporte de 8 caídas	1

Tabla 2. Información de los expedientes analizados.

No.	Nombre de parámetro	No.	Nombre de parámetro
1	Distancia (cm)	16	Long Paso Completo Izq (cm)
2	Tiempo Deambulaci3n (seg)	17	Long Paso Completo Der (cm)
3	Velocidad (cm/seg)	18	Base de Sopor Izq (cm)
4	Velocidad media normalizada	19	Base de Sopor Der (cm)
5	N3mero de pasos	20	Soporte Simple Izq (%GC)
6	Cadencia	21	Soporte Simple Der (%GC)
7	Diferencial tiempo paso (seg)	22	Soporte Doble Izq (%GC)
8	Diferencial long paso (cm)	23	Soporte Doble Der (%GC)
9	Diferencial tiempo ciclo (seg)	24	Oscilaci3n Izq (%GC)
10	Tiempo paso Izq (seg)	25	Oscilaci3n Der (%GC)
11	Tiempo paso Der (seg)	26	Estancia Izq (%GC)
12	Tiempo ciclo Izq(seg)	27	Estancia Der(%GC)
13	Tiempo ciclo Der (seg)	28	Relaci3n paso/extremidad Izq
14	Long Paso Izq (cm)	29	Relaci3n paso/extremidad Der
15	Long Paso Der (cm)	30	Toe I/O Izq (deg)
		31	Toe I/O Der (deg)

Tabla 3. Informaci3n proporcionada en un an3lisis de la marcha. La cual fue obtenida cada 6 meses.

7.2 T3cnicas utilizadas

Despu3s de la revisi3n del estado del arte encontramos que una de las herramientas m3s utilizadas para el c3lculo del riesgo de presentar un determinado evento en el domio m3dico es el an3lisis de sobrevivencia, en especifco el modelo de regresi3n de Cox [19], [28], [53], pues este tipo de t3cnicas permite la evaluaci3n de un determinado tratamiento. Por otro lado se estudi3 y utiliz3 dicha t3cnica con los datos proporcionados por el INR. Por otra parte se eligieron las Redes Bayesianas como una t3cnica computacional a aplicar, para tener un comparativo de la eficacia de cada una de estas t3cnicas. Finalmente se realiz3 un primer intento para construir una EBN de manera autom3tica.

Modelo de Regresi3n de Cox

Una de las restricciones para el uso del modelo de regresi3n de Cox es que las variables involucradas deben ser binarias, debido a esto se hizo una binarizaci3n de los par3metros para nuestro caso.

La binarizaci3n se hizo con base en el cambio que hubo de un par3metro en un a3o, *CambioPX*. Para ello se utilizaron s3lo los datos de dos registros de cada uno de los expedientes. Para ello se utiliz3 la f3rmula 4.

$$Cambio\ PX = \begin{cases} 1, & \text{si } |PXreg2 - PXreg1| > PXreg1 \times FacDec \\ 0, & \text{en otro caso} \end{cases} \quad (4)$$

Donde:

PX es el par3metro *x* de la marcha

PXreg1 es el valor del par3metro *x* en el primer registro

$PXreg2$ es el valor del parámetro x en el segundo registro
 $FacDec$ es el factor de deterioro “normal” de la marcha

En este experimento se utilizó un $FacDec = 0.05$, esto fue un $FacDec$ estimado. Como trabajo futuro y contribución de esta tesis se propone determinar cuál es el deterioro normal de la marcha en personas adultas mayores, con información de datos de personas adultas mayores con una marcha normal, que nos proporcionarían los investigadores del INR.

Una vez binarizados los datos se tiene la información necesaria para aplicar el Modelo de Riesgo Proporcional de Cox, y de igual forma dicha información fue utilizada para aplicar una BN con el mismo propósito.

Variables seleccionadas por el modelo de regresión de Cox

Dado el número de casos y variables que se tienen, el uso del modelo de regresión de Cox para obtener un conjunto de variables más representativas para la determinación del riesgo de caída no fue posible considerando todas las variables. Esto se debió a que el número de variables es considerablemente mayor al número de casos y como se menciona en [42] el número de casos depende del número de variables predictivas involucradas, para poder aplicar el modelo de regresión de Cox.

Por lo anterior, se formaron 6 grupos con 5 variables y se utilizaron todos los casos disponibles para aplicar el modelo de regresión de Cox. Así, se utilizó el modelo de regresión de Cox 6 veces distintas obteniendo en cada corrida las variables mostradas en la tabla 4. Cabe señalar que esta agrupación se hizo tomando las primeras 5 variables, y luego las segundas 5 y así sucesivamente.

. Número de Corrida	Variables encontradas
1	NUMP, DIFTP
2	Ninguna
3	SSI, SDI, SDD, OI
4	Ninguna
5	Ninguna
6	RP/EI, RP/ED

Tabla 4. Variables obtenidas en cada corrida con el modelo de regresión de Cox.

Finalmente supusimos que las variables mencionadas en la tabla 4 son las más representativas para determinar el riesgo de caída en el primer año, la tabla 5 muestra dichas variables.

Abreviación de Variable	Nombre de variable
NUMP	Número de pasos
DIFTP	Diferencial tiempo paso
SSI	Soporte Simple Izq
SDI	Soporte Doble Izq
SDD	Soporte Doble Der
OI	Toe I/O Izq
RP/EI	Relación paso/extremidad Izq
RP/ED	Relación Paso/extremidad Der

Tabla 5. Variables más representativas para determinar caída en el primer año.

Posteriormente para la construcción del modelo y prueba de su eficacia se aplicó Cox únicamente con las variables de la tabla 5.

Cabe mencionar que las variables anteriores fueron obtenidas utilizando sólo la información del primer año de cada uno de los expedientes disponibles. Es decir estas son las variables más representativas para determinar el riesgo de caída en el primer año.

Un proceso similar se siguió para determinar las variables más representativas para el segundo año, obteniendo la información de la tabla 6 y 7.

Número de Corrida	Variables encontradas
1	VEL, CAD
2	Ninguna
3	Ninguna
4	Ninguna
5	Ninguna
6	CAIDASP

Tabla 6. Variables obtenidas en cada corrida con el modelo de regresión de Cox.

Abreviación de Variable	Nombre de variable
VEL	Velocidad
CAD	Cadencia
CAIDASP	Caídas Previas

Tabla 7. Variables más representativas para determinar caída en el segundo año.

La variable CAIDASP es una variable adicional que fue agregada para indicar que el individuo ya había tenido caída registrada en el primer año. Como se observa, las variables encontradas para el segundo año son distintas a las variables del primer año. Esto es un indicativo de que los cambios de la marcha patológica se desencadenan progresivamente, y que los cambios en ciertos parámetros preceden a los cambios en otros. Esta suposición

coincide también con observaciones de los expertos del INR, que deben estudiarse con mayor profundidad.

Red Bayesiana Simple

Para la construcción del modelo de la marcha basado en una Red de Bayes Simple que permita determinar caídas se utilizaron, al igual que con el modelo de Cox, los datos del primer año. Se aplicó el algoritmo K2-learning para la construcción automática de la red a partir de los datos.

Se utilizó el software Elvira que permite la implementación de redes bayesianas. Este software permite el diseño manual de una red bayesiana o incluso la construcción automática de la red bayesiana, tanto la estructura como los parámetros, a partir de los datos (observaciones o casos).

Además, se probaron diferentes clasificadores bayesianos como son BayesNet, BayesianLogicRegresion, NaiveBayes, NaiveBayesSimple, utilizando el software weka 3.6

En nuestro caso se utilizaron, al igual que en Cox, variables binarias para la construcción de la red. De esta forma se tenía un conjunto de 17 casos con 32 variables. Esta información se introdujo a Elvira y utilizando el algoritmo K2-learning se construyó una Red Bayesiana. No se descarta el uso de otros algoritmos para la construcción de la BN en el futuro. Un ejemplo de dicha red se muestra en la figura 3.

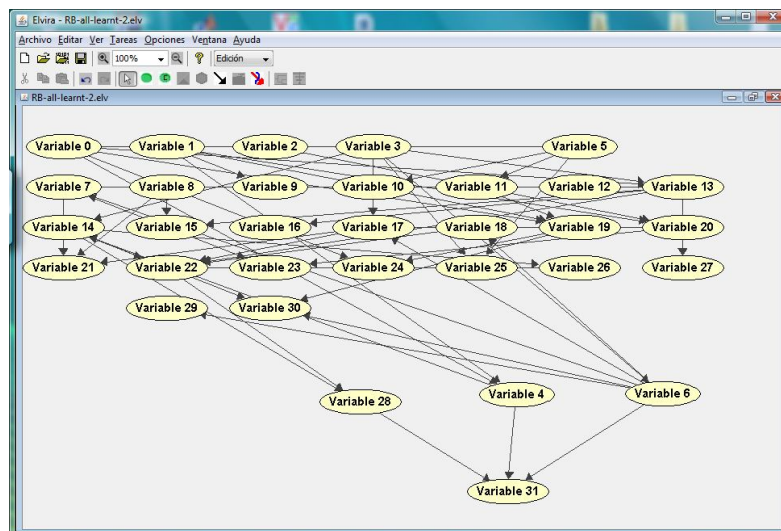


Figura 3. Ejemplo de una Red Bayesiana Simple construida con Elvira.

La tabla 8 muestra el nombre de las variables de la BN mostrada en la figura 3.

No.	Nombre de parámetro	No.	Nombre de parámetro
Variable0	Distancia (cm)	Variable16	Long Paso Completo Der (cm)
Variable1	Tiempo Deambulaci3n (seg)	Variable17	Base de Sopor Izq (cm)
Variable2	Velocidad (cm/seg)	Variable18	Base de Sopor Der (cm)
Variable3	Velocidad media normalizada	Variable19	Soporte Simple Izq (%GC)
Variable4	Número de pasos	Variable20	Soporte Simple Der (%GC)
Variable5	Cadencia	Variable21	Soporte Doble Izq (%GC)
Variable6	Diferencial tiempo paso (seg)	Variable22	Soporte Doble Der (%GC)
Variable7	Diferencial long paso (cm)	Variable23	Oscilaci3n Izq (%GC)
Variable8	Diferencial tiempo ciclo (seg)	Variable24	Oscilaci3n Der (%GC)
Variable9	Tiempo paso Izq (seg)	Variable25	Estancia Izq (%GC)
Variable10	Tiempo paso Der (seg)	Variable26	Estancia Der(%GC)
Variable11	Tiempo ciclo Izq(seg)	Variable27	Relaci3n paso/extremidad Izq
Variable12	Tiempo ciclo Der (seg)	Variable28	Relaci3n paso/extremidad Der
Variable13	Long Paso Izq (cm)	Variable29	Toe I/O Izq (deg)
Variable14	Long Paso Der (cm)	Variable30	Toe I/O Der (deg)
Variable15	Long Paso Completo Izq (cm)	Variable31	Caída

Tabla 8. Nombre de las variables de la red bayesiana de la figura 3.

Comparaci3n del modelo de regresi3n de Cox y la Red Bayesiana Simple

Dado que sólo se tienen hasta el momento 17 casos para el entrenamiento del Modelo de Riesgo Proporcional de Cox y de la Red Bayesiana, se utiliz3 *Leave One Out Cross Validation*, LOOCV, el cual es el método recomendado para este tipo de situaciones. Así se utilizaron 16 casos para entrenamiento y 1 caso para probar el modelo obtenido. De esta forma, se generaron 17 modelos de regresi3n de Cox y 17 redes bayesianas, uno por cada caso excluido.

De acuerdo al caso excluido, tanto el modelo de regresi3n de Cox como la red de Bayes incluían, excluían o conservaban variables que estaban directamente relacionadas con la inferencia de la probabilidad de caída. Así, la tabla 9 muestra la eficacia en la clasificaci3n tanto con Cox como con la Red Bayesian.

Modelo de regresi3n de Cox		Red bayesiana simple	
Eficacia	Variables más representativas para la inferencia de caída	Eficacia	Variables más representativas para la inferencia de caída
0.70588235	Diferencial tiempo paso (seg)	0.58823529	Número de pasos
	Número de pasos		Diferencial tiempo ciclo (seg)
	Soporte Doble Izq (%GC)		Diferencial tiempo paso (seg)
			Long Paso Completo Izq (cm)
			Long Paso Completo Der (cm)
			relaci3n paso/extremidad Der

Tabla 9. Resultados de los modelos utilizados.

La figura 4 muestra gráficamente el desempeño de cada uno de los modelos, y la tabla 10 muestra la eficacia de los clasificadores bayesianos utilizados.

Nombre Clasificador	Eficacia
BayesNet	0.6470
BayesianLogicRegresion	0.6470
NaiveBayes	0.5882
NaiveBayesSimple	0.5882

Tabla 10. Resultados de los clasificadores.

La eficacia fue medida de acuerdo a la clasificación que hizo cada modelo al determinar si un nuevo caso, sufrirá o no caída en un año. Los resultados obtenidos muestran mayor eficacia por parte del modelo de regresion de Cox, esto quizás se deba a que los datos están siendo utilizados del modo requerido por Cox, esto es, como datos binarios. También observamos que la BN identifica mas variables como reelevantes para determinar caída, debido a que permite considerar y modelar relaciones entre más variables.

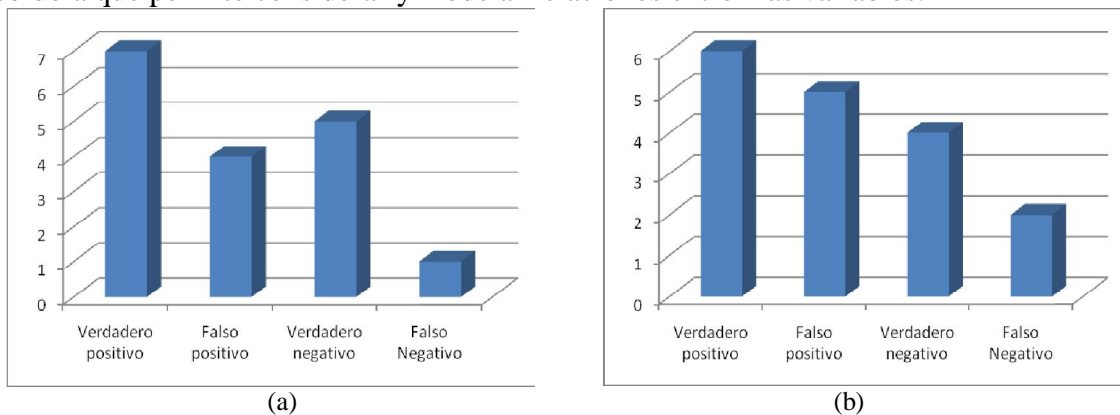


Figura 4. Resultados de clasificación aplicando: (a) Regresión de Cox y (b) Red Bayesiana

Cabe mencionar que para ambos modelos cada caso utilizado para entrenamiento contaba con 31 variables. Esta situación afecta el desempeño de ambos modelos, es por eso que como parte de la tesis se considera hacer una selección de atributos para mejorar el desempeño de cada modelo.

Un primer intento para reducir el número de variables y observar la eficacia de los modelos probados fue utilizar los parámetros que se consideran relevantes, con base en la experiencia de los expertos en el dominio, los cuales se enlistan en la tabla 11.

Variables más Representativas	
Cadencia	
Long Paso Izq (cm)	Long Paso Completo Izq (cm)
Long Paso Der (cm)	Long Paso Completo Der (cm)
Base de Sopor Izq (cm)	Base de Sopor Der (cm)

Tabla 11. Variables más relevantes proporcionadas por los expertos del dominio.

Utilizando los parámetros mostrados en la tabla 11 se obtuvieron los resultados mostrados en la figura 5 y en la tabla 12.

Modelo de regresión de Cox	Red bayesiana simple
Eficacia	Eficacia
0.82352941	0.82352941

Tabla 12. Resultado de la eficacia de clasificación de los modelos utilizados.

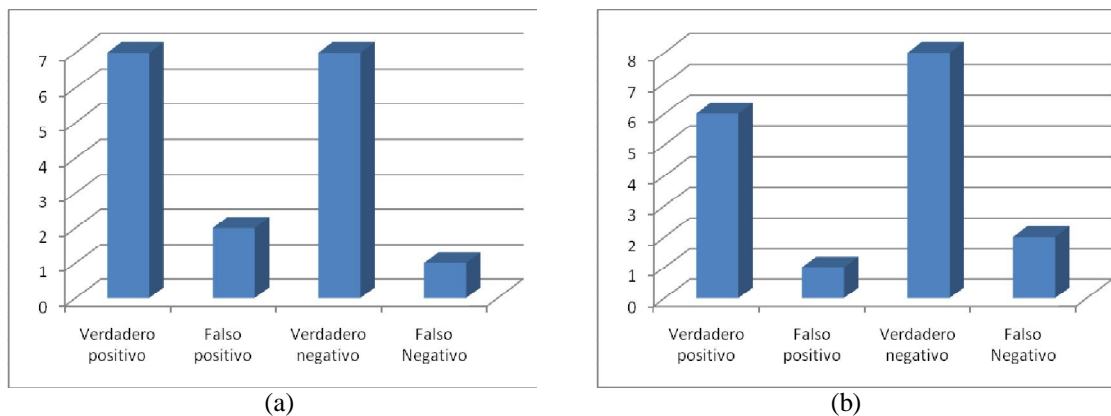


Figura 5. Resultados de clasificación aplicando: (a) Regresión de Cox y (b) Red Bayesiana.

Cabe mencionar que para el caso del modelo de regresión de Cox, el modelo proporciona una función de riesgo la cual permite tener una medida cuantificable del riesgo de caída en un periodo de tiempo, pero no proporciona el dato de si el nuevo caso presenta o no el evento. Y por otro lado, la Red Bayesiana proporciona el dato puntual de si el nuevo caso presenta o no el evento con la probabilidad asociada a este, con probabilidad mayor a 0.5, pero no permite tener un estimado de cuál es el riesgo de presentar este evento en un intervalo de tiempo.

Red Bayesiana de Eventos

Como ya se ha mencionado, una variante de las BNs que permite incorporar información temporal y además proporciona información de la ocurrencia de un evento así como el tiempo asociado a la ocurrencia de este son las EBN. Para efectos de prueba se diseñó una EBN de tipo TNBN, para evaluar la eficacia y limitantes de dicha red.

En una TNBN cada nodo representa un evento o cambio del valor de una variable y un arco corresponde a una relación causal temporal. A diferencia de las BNs comunes, cada nodo de la red contiene *nodos temporales* que representan el tiempo en que una variable cambia de estado, e inclusive permite representar el no cambio de dicha variable. Así, incluyendo más nodos temporales es posible representar más de un cambio para la misma variable, considerando además el tiempo en el que ocurre dicho cambio.

Como se ha mencionado la construcción de la EBN se hace manualmente con la ayuda de los expertos en el dominio, pues se necesita conocer qué tipo de cambios tiene una variable y el tiempo en el que ocurren dichos cambios, así como determinar de qué manera se afecta la ocurrencia de cambios posteriores en otras variables, o incluso en la misma, pero en

diferente tiempo. Y determinar así, de acuerdo a los cambios detectados y al tiempo de ocurrencia, la probabilidad de experimentar un evento a futuro.

En nuestro caso, los expertos del dominio no tienen certeza de qué variables están asociadas al aumento del riesgo de caída, ni la proporción en la que afectan dicho riesgo a lo largo de un periodo de tiempo, tampoco se conoce con exactitud las relaciones que existen entre los cambios de alguna variable de la marcha y cómo estos cambios afectan a otras variables de la marcha.

Dicha información es necesaria para la construcción de un modelo, y si no se tiene esa información explícita es de interés aplicar métodos que permitan la construcción de manera automática de un modelo a partir de los datos. Dado que el modelo propuesto estará basado en una EBN, **la construcción automática de dicho modelo es la contribución principal de esta tesis.**

Por otro lado, después de revisar y trabajar con el modelo de Cox observamos que cuenta con las siguientes propiedades:

1. Permite hacer una selección de variables relevantes, pero dicha selección sólo considera la presencia o ausencia de variables, y no permite considerar las relaciones entre ellas.
2. Calcula la relevancia de las variables seleccionadas con base a la ocurrencia de un evento futuro.
3. Proporciona mecanismos que permiten evaluar en qué proporción se afecta el cambio en el riesgo de presentar un evento futuro, con base en la presencia de variables específicas en distintos momentos del periodo de estudio.

Así, se propone aprovechar dichas propiedades para la construcción automática de una EBN y a continuación se muestran los primeros resultados de esos experimentos. Al igual que con el modelo de regresión de Cox y la red de Bayes simple, se usaron los datos del primer año para construir el este modelo.

El objetivo principal de la EBN es determinar qué tipo de riesgo es el que tiene una persona de sufrir una caída dada la evidencia de los valores de su marcha. Y aún más, dar un estimado del tiempo en el que podría ocurrir dicho evento así como la probabilidad de que eso ocurra.

En principio es necesario determinar un evento inicial, del cual se tenga conocimiento que es un causal de provocar cambios en eventos posteriores (cambios en los parámetros de la marcha), y que dichos cambios estén asociados al aumento en el riesgo de caída de una persona.

En nuestro caso, el evento inicial es el diagnóstico de osteoporosis, dado que es una condición necesaria para que los individuos que se les ha dado seguimiento entren al estudio. El modelo de regresión de Cox fue probado sobre este tipo de pacientes, obteniendo que para los datos del primer año las variables que más influyen en el riesgo de presentar una caída en ese año son los enumerados en la tabla 4. Es decir, el modelo de

regresión de Cox nos permitió hacer una primera selección de variables. De este modo se seleccionaron los nodos candidatos a formar parte de nuestra EBN, los cuales se muestran en la figura 6.

En principio suponemos que el diagnóstico de osteoporosis afecta a cada uno de los parámetros de la marcha elegidos como nodos candidatos de la EBN. Así, se establecieron conexiones entre todos los parámetros de la marcha y la osteoporosis. Por otro lado utilizamos de nuevo el modelo de regresión de Cox con las variables de la tabla 5, para determinar cuáles de estas están son relevantes para la determinación del riesgo de caída, y así establecer las conexiones entre los parámetros de la marcha y el evento caída de la red anterior. Descartamos aquellas variables que no afectan el riesgo de caída según la información proporcionada por el modelo de regresión de Cox, quedando la EBN conectada como se muestra en la figura 7.

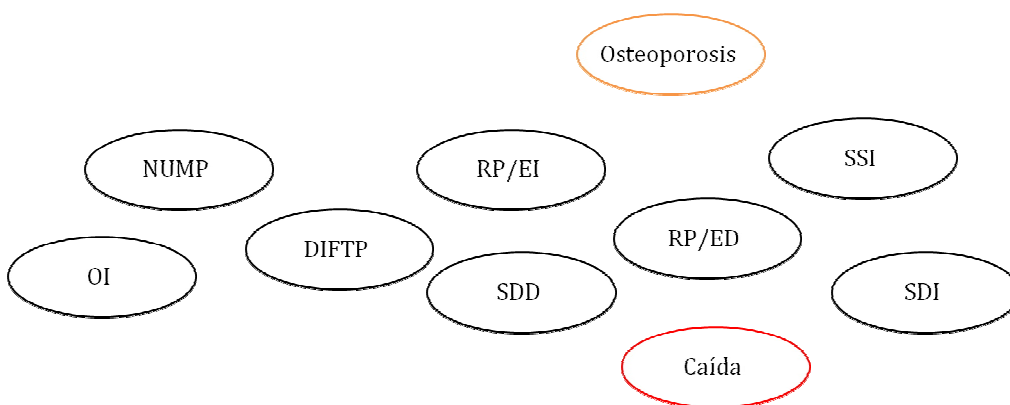


Figura 6. Nodos candidatos a ser parte de la EBN.

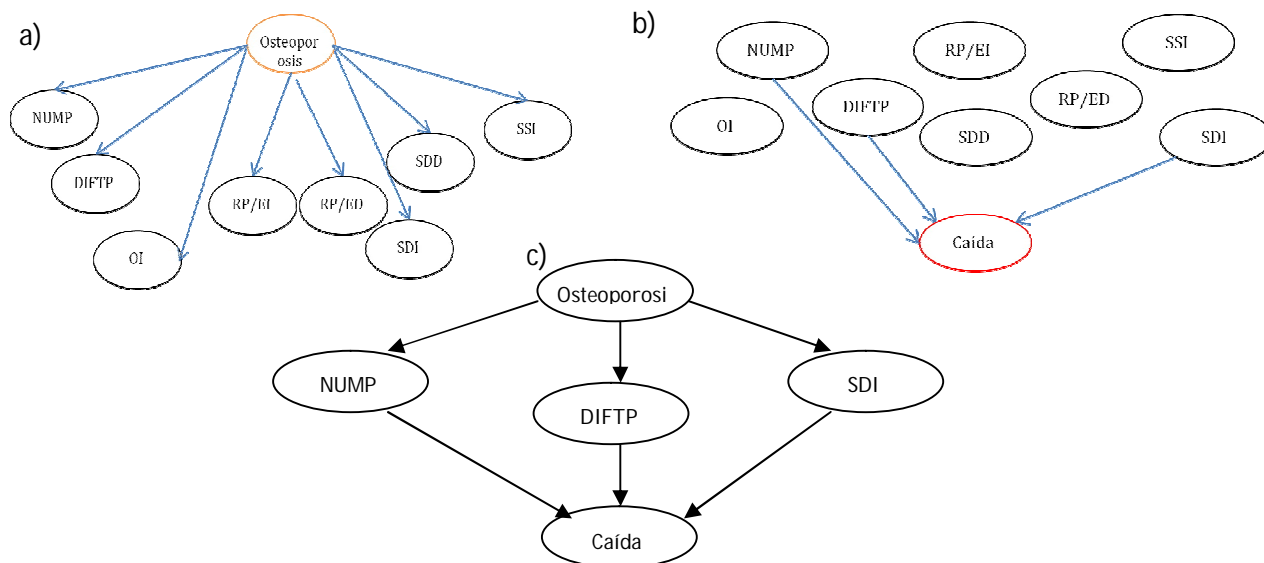


Figura 7. a) Relación entre la osteoporosis y los parámetros relevantes, b) parámetros de la marcha relacionados con la propensión a caídas, c) estructura de la EBN.

Como ya se ha mencionado, el modelo de Cox además de proporcionarnos información acerca de las variables más relevantes para determinar el riesgo de un evento, nos da información de cómo esas variables afectan el riesgo con respecto al tiempo. La salida de un análisis de datos con el modelo de regresión de Cox se muestra en la tabla 11.

De acuerdo al signo del valor β asociado a la variable predictiva, sabemos que la presencia de DIFTP y SDI, no aumenta el riesgo de caída sino lo contrario. Y determinamos que sólo NUMP es la variable que genera un aumento en el riesgo de sufrir una caída. Esto puede ser comprobado calculando la función de supervivencia para cada una de estas variables, dicho cálculo se muestra en la tabla 12.

Tiempo(Días)	$H_0(t)$ - Riesgo base	Variables predictivas	β
30	0.166	DIFTP	-2.7775
67	0.445	NUMP	3.9926
70	0.929	SDI	-2.8989
139	1.442		
177	2.182		
287	3.404		
309	4.987		
330	7.238		

Tabla 11. Información proporcionada por el modelo de regresión de Cox.

	DIFTP		NUMP		SDI	
	0	1	0	1	0	1
Días	$S(X,t)$	$S(X,t)$	$S(X,t)$	$S(X,t)$	$S(X,t)$	$S(X,t)$
30	0.84704623	0.98972894	0.84704623	0.00012385	0.84704623	0.99089777
67	0.64082428	0.97270324	0.64082428	3.3581E-11	0.64082428	0.9757857
70	0.39494846	0.94385943	0.39494846	1.3623E-22	0.39494846	0.95011445
139	0.23645438	0.91422053	0.23645438	1.1478E-34	0.23645438	0.92364193
177	0.11281567	0.87309853	0.11281567	4.3912E-52	0.11281567	0.88674945
287	0.03324004	0.80920120	0.03324004	7.5952E-81	0.03324004	0.82902482
309	0.00682611	0.73332927	0.00682611	4.185E-118	0.00682611	0.7597979
330	0.00071875	0.63752696	0.00071875	4.367E-171	0.00071875	0.67119445

Tabla 12. Valores de la función de supervivencia para cada una de las variables, donde la columnas con 0 muestran los valores de supervivencia sin la presencia de dicha variable y las columnas con 1 muestran los de supervivencia con la presencia de dicha variable.

El valor de supervivencia va de 1 a 0, donde 1 indica que no existe riesgo alguno de presentar el evento, de tal forma que un decremento en el valor de esta función indica un aumento en el riesgo de presentar el evento. Como se observa en la tabla 12, sin la presencia de la variable estudiada el valor de supervivencia es el mismo en todos los casos. Esta probabilidad de supervivencia “normal” está calculada con base en la probabilidad de riesgo base, la cual se calcula a partir de los datos de la muestra, en este caso personas con

diagnóstico de osteoporosis. Se observa que el valor de sobrevivencia es afectado drásticamente con la presencia de NUMP, esto es el riesgo de presentar una caída es mayor y aún más, la tabla 12 permite observar cómo va cambiando dicho riesgo en el tiempo.

De tal manera que se hizo una primera prueba para determinar los intervalos de tiempo para la EBN, a partir de la información proporcionada por Cox acerca del riesgo y el efecto en el tiempo de presentar un cambio en las variables involucradas en el modelo.

Dado que estamos usando una EBN tipo TNBN, se necesitan definir intervalos excluyentes. Primero se determinó un intervalo general [1-365] días y este intervalo se dividió en 3 sub-intervalos excluyentes los cuales son:

- [1-30] para NUMDP
- [31-150] para DIFTP
- [151-365] para SDI

Esta definición de intervalos fue hecha con base en lo siguiente: a los 30 días, de acuerdo a la función de sobrevivencia de NUMP, el riesgo de una caída si NUMP=1 \rightarrow riesgo =1-0.00012385, con base a esto se dio un valor “cualitativo” a esa situación, incremento moderado en el riesgo de caída, IRRC.

Para el caso de DIFTP y SDI ambas variables disminuyen el riesgo de caída, pero en diferente proporción, es decir, ambos prolongan el aumento del riesgo de caída. Para determinar los intervalos para esas variables se observó en la figura 8 y en la tabla 12, que los valores de las funciones de sobrevivencia de ambas variables eran similares en los tiempos 67, 70 y 139, y se determino a partir de esto que el intervalo [31-150] está asociado a DIFTP, y dado que la diferencia entre los valores de la función de sobrevivencia de ambas variables en los tiempos 177, 287, 309, 330 eran más marcados se determinó que el intervalo asociado a la variable SDI fuera [151-365].

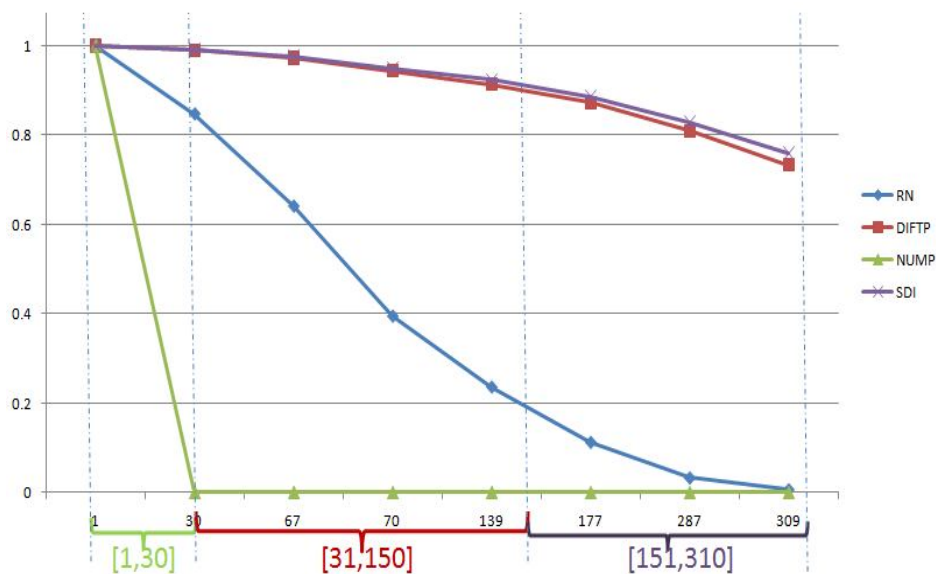


Figura 8. Gráfica de las funciones de sobrevivencia.

Ahora bien, con base en lo anterior determinamos intervalos de acuerdo a los cambios más significativos o representativos del valor de sobrevivencia de cada variable. Suponiendo que, de acuerdo al valor de sobrevivencia y el tiempo en el que ocurre, el incremento o decremento del riesgo de presentar una caída está asociado a diferente tipo de degradación de la marcha, ésta información fue utilizada para definir los nodos temporales de cada variable, los cuales se muestran en la tabla 13.

Los valores de la variable caída permiten modelar qué tipo de degradación de la marcha es, para nuestro caso, degradación patológica, degradación normal o si existe alguna terapia que está contribuyendo a mantener una marcha normal. Los valores e intervalos de tiempo para cada uno de ellos se muestran en la tabla 14. Así, la red bayesiana de eventos queda construida como se muestra en la figura 9.

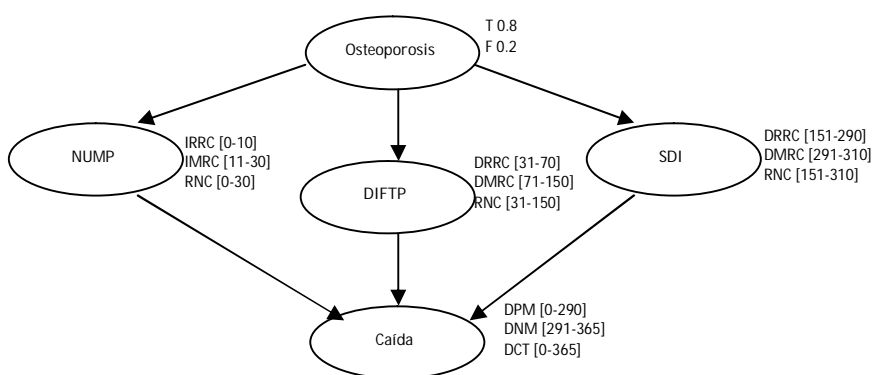


Figura 9. Red bayesiana de eventos.

DIFTP			NUMP			SDI		
Nodo Temporal	Valor de la var.	Intervalo propuesto	Nodo Temporal	Valor de la var.	Intervalo propuesto	Nodo Temporal	Valor de la var.	Intervalo propuesto
Decremento rápido en el riesgo de caída (DRRC)	1	[31-70]	Incremento moderado en el riesgo de caída (IRRC)	1	[0-10]	Decremento rápido en el riesgo de caída (DRRC)	1	[151-290]
Decremento moderado en el riesgo de caída (DMRC)	1	[71-150]	Incremento rápido en el riesgo de caída (IMRC)	1	[11-30]	Decremento moderado en el riesgo de caída (DMRC)	1	[291-310]
Riesgo normal de caída (RNC)	0	[31-150]	Riesgo normal de caída (RNC)	0	[0-30]	Riesgo normal de caída (RNC)	0	[151-310]

Tabla 13. Definición de nodos temporales. Cada una de las variables cuenta con 3 nodos temporales, cada nodo temporal con un intervalo excluyente con respecto a los otros nodos temporales de la variable respectiva.

Caída		
<i>Nodo temporal</i>	<i>Valor de la var.</i>	<i>Intervalo propuesto</i>
Degradación patológica de la marcha, alto riesgo de caída (DPM)	1	[0-290]
Degradación normal de la marcha apoyada con terapia (DNM)	1	[291-365]
Degradación con terapia (DCT)	0	[0-365]

Tabla 14. Definición de nodos temporales para la variable caída.

Las probabilidades conjuntas de la EBN construida se calculan directamente de los datos. De igual manera se utilizó LOOCV para probar la eficacia de la EBN con los datos que se tiene. La eficacia de la red bayesiana de eventos se muestra en la tabla 15 y la figura 10.

Red Bayesiana de Eventos

Efectividad
0.64705882

Tabla 15. Eficacia de la EBN.

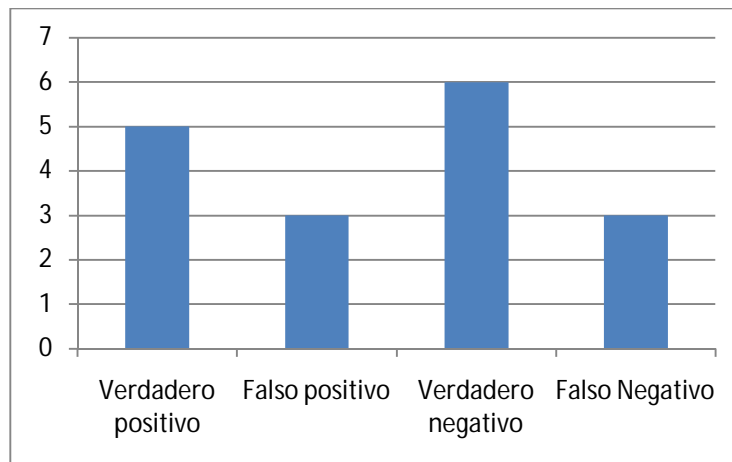


Figura 10. Resultados de clasificación con la red bayesiana de eventos.

7.3 Evaluación de las técnicas utilizadas

De acuerdo a los resultados presentados, se ha comprobado la factibilidad del uso de técnicas de IA, en específico BN, para el modelado de la marcha patológica, contrastando su eficacia con técnicas de análisis de sobrevivencia, como el modelo de regresión de Cox, señalando las ventajas y desventajas de ambas técnicas.

Se observó que al reducir el número de variables analizadas utilizando el conocimiento de los expertos en el dominio se mejoraba la eficacia de ambos modelos, debido a que se tienen pocos datos con respecto al número de variables.

Se ha propuesto un primer modelo de la marcha patológica que determina el riesgo de caída de personas adultas mayores. Este modelo se orienta hacia la principal contribución de este trabajo pues hasta donde sabemos no existe un modelo de este tipo.

Se ha mostrado la factibilidad de la construcción de manera automática de un modelo de la marcha basado en una EBN que permita determinar el riesgo de caída de una persona asociado a los cambios de los parámetros de la marcha. Cabe insistir que en esto radica una de las contribuciones de la tesis pues dichos modelos generalmente se construyen manualmente con ayuda de los expertos en el dominio.

8. Conclusiones y discusión de resultados

Hasta el momento se ha probado la factibilidad del uso de técnicas de análisis de sobrevivencia en específico el modelo de regresión de Cox, al igual que técnicas de IA, en específico BN, para modelar los datos de la marcha de personas adultas mayores. Los resultados muestran que ambas técnicas permiten obtener información útil para los expertos en el dominio aunque dicha información es proporcionada de manera separada y es obtenida de manera limitada pues se está trabajando con datos restringidos debido a los requerimientos de los modelos.

También hemos corroborado la necesidad de técnicas tales como selección de variables pues tenemos pocos datos con muchas variables. Como se ha mostrado, sólo se ha trabajado con pocos datos pero al tener más datos se espera mejorar los modelos presentados en la sección anterior.

Hasta el momento sólo se ha trabajado con datos de un año de seguimiento, es decir dos registros de datos de la marcha. Ambos modelos se usaron para determinar si un nuevo caso analizado tendría caída, con base en la información de los datos de la marcha.

Para ambos modelos se consideró únicamente si existió un cambio en algún parámetro de la marcha de un registro a otro, no se sabía si el cambio detectado estaba asociado al aumento o disminución del parámetro estudiado. Esto hace suponer que se está perdiendo información relevante para determinar las relaciones que puedan existir entre parámetros de la marcha asociados con el aumento en el riesgo de caída, es por eso que una de las contribuciones de este trabajo es incorporar en el modelo utilizado la representación de este tipo de cambios.

Una alternativa para considerar la información para el modelo de regresión de Cox es el uso de variables adicionales asociadas al aumento, decremento o ningún cambio de un parámetro, esto traería consigo un aumento considerable en el número de variables y ello implicaría un mayor número de muestras para entrenar el modelo. Otra alternativa sería utilizar representación de los cambios de los datos como variables binarias con información sobre el tipo de cambio que experimentó una variable. Por otro lado, esta situación es menos problemática para el modelo con una Red Bayesiana pues cada variable puede tener más de un valor y por lo tanto el número de nodos en el modelo no se modifica.

Uno de los objetivos de este trabajo es la representación de los cambios de la marcha asociados al aumento de riesgo de caída, que hasta el momento no se ha realizado. Para ello se aplicarán redes de eventos, las cuales permiten modelar cómo se relacionan la ocurrencia de eventos iniciales con eventos futuros. Una de las aportaciones de este trabajo es el diseño de un modelo que permita determinar los cambios en determinados parámetros de la

marcha y la relación que tienen en el cambio de otros parámetros que inicialmente no se detectaban como indicativos de sufrir una caída. También dicho modelo permitirá determinar el aumento del riesgo de acuerdo a los cambios de la marcha. Y aún más como contribución principal de este trabajo se prevé la construcción de dicho modelo a partir de los datos.

De acuerdo a los resultados obtenidos, el modelo de regresión de Cox permite obtener probabilidades de riesgo de caída de acuerdo al cambio que tienen las variables predictivas en un periodo de tiempo, aunque como se ha mencionado la limitante es que sólo se considera presencia o ausencia de dichas variables, y además no se permite modelar relaciones entre las variables. Por otro lado, las BN permiten el modelado de relaciones entre variables además de considerar no sólo variables binarias. De tal forma que ambas técnicas pueden ser tratadas como técnicas complementarias para la construcción de un modelo basado en EBN, pues una EBN permite un modelado de las relaciones entre variables no binarias y además obtener una medida del riesgo de presentar un evento. El reto en esto es que el diseño de dicho tipo de BN no se realiza de manera automática y depende en su mayoría de la experiencia de los expertos en el dominio, y esto es una desventaja en dominios en los que los expertos no tienen certeza o desconocen por completo información fundamental para el diseño de un modelo, como el de la marcha patológica, información como tipos de cambios, relaciones entre variables, efecto del cambio de una variable en el riesgo de presentar un evento de interés.

Los resultados mostrados hasta el momento dan evidencia que es posible realizar la construcción de un modelo basado en EBN que permite inferir el riesgo de caída de acuerdo a los cambios detectados en la marcha. Se descubre de manera automática a partir de los datos el modelo, y se puede recurrir a los expertos para validar y complementar este modelo.

9. Actividades por realizar

Ahora bien, de acuerdo al trabajo realizado queda por hacer, en principio el mejoramiento del modelo para determinar riesgo de caída no sólo de personas con osteoporosis detectada sino para personas adultas mayores en general, modificando la función de riesgo base calculada por el modelo de regresión de Cox. Para ello proponemos utilizar la información de personas sin osteoporosis para la construcción de dicha función de riesgo base, cuyos datos también estarán disponibles en el Laboratorio de Análisis de Movimiento del INR.

Cubrir las limitaciones del modelo de regresión de Cox, pues hasta el momento con los resultados obtenidos sólo sabemos que hubo un cambio en una variable y es deseable para complementar lo obtenido con la información de los expertos en el dominio, además incorporar la información de si ese cambio es un aumento o decremento en el valor del parámetro.

Mejorar la construcción automática de la EBN, pues como se ha comentado las variables detectadas por el modelo de regresión de Cox difieren con el conocimiento proporcionado por parte de los expertos del INR sobre las variables relevantes. Así, se propone el uso de técnicas reportadas para selección de atributos en datos temporales [34], [40], [51]. Así

como mejorar la selección de las relaciones utilizando otras opciones, como correlación entre variables, y no limitarnos exclusivamente a la información del modelo de regresión de Cox. Refinar la manera de seleccionar los intervalos de cambio, proponiendo medidas de cambio relevante con base a los valores de las funciones de supervivencia de cada variable predictiva. Probar otro tipo de redes de eventos como las NPEDT [15], [16], [17] y comparar su eficacia con la TNBN.

Utilizar técnicas distintas a la selección de variables que permitan manejar el problema de tener muchas variables y pocos datos, utilizar *oversampling* o una combinación de *oversampling* con selección de variables para trabajar con los datos que actualmente se tienen, hasta que se tenga por completo la base de datos del INR.

Evaluar la eficacia del modelo propuesto probándolo con datos nuevos así como utilizando otro tipo de datos de la marcha, como los proporcionados en un estudio de estabilidad.

10. Plan de trabajo

La figura 11 muestra el cronograma de actividades.

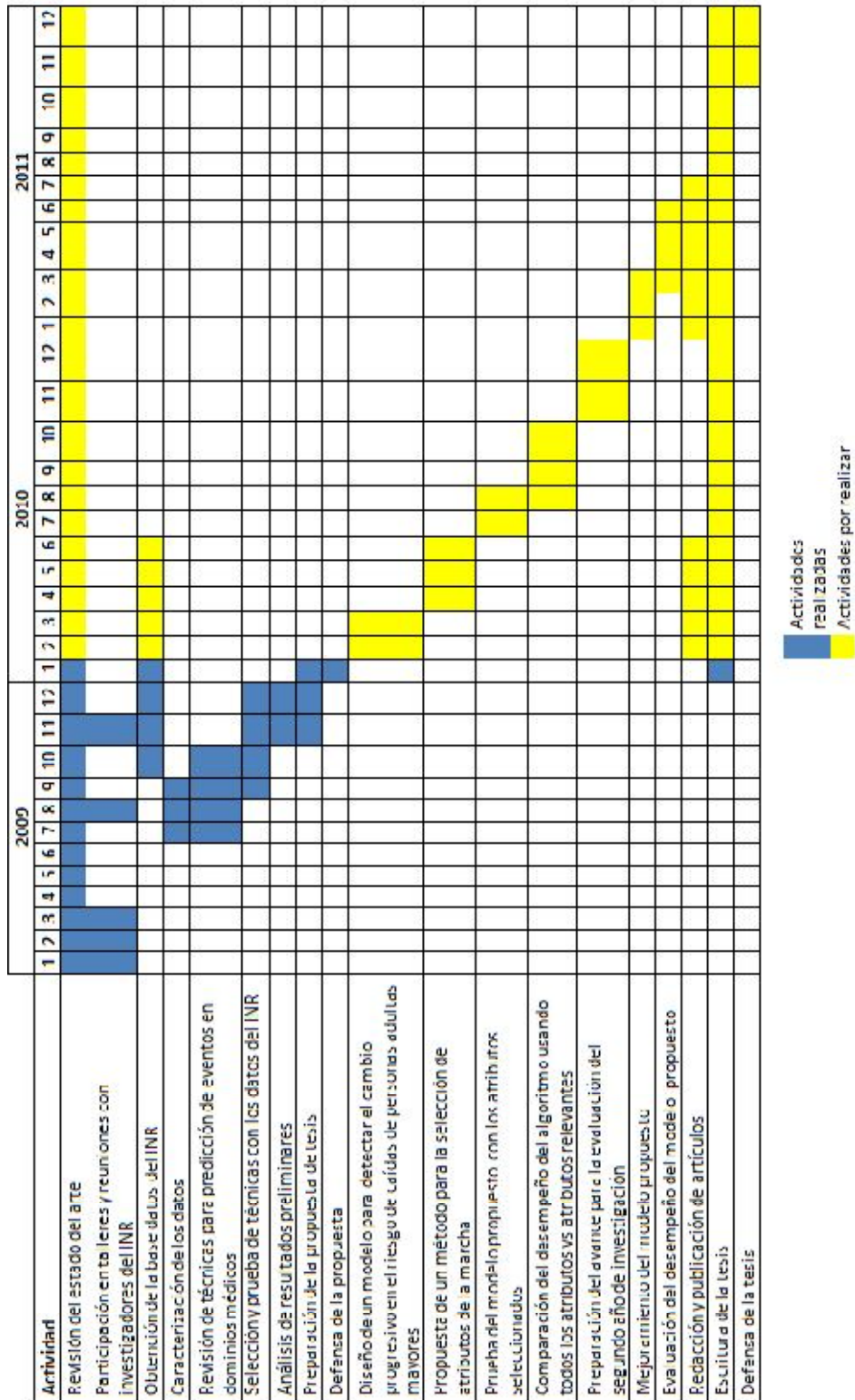


Figura 11. Cronograma de actividades.

11. Referencias

- [1]. Abdelkader C., Cutler R. et Davis L. (2002), View-invariant estimation of height and stride for gait recognition, *Proceedings of the International European Conference on ComputerVision Workshop Copenhagen on Biometric Authentication*, 155–167.
- [2]. Amaro E., Reyes C., Arch E. et Mandujano M. (2007), Statistical vectors of acoustic features for the automatic classification of infant cry, *International Journal of Information Acquisition*, 4, 347-355.
- [3]. Ancell A. (2009), *Predicción Probabilística de Precipitación*, Tesis Doctoral, Universidad de Cantabria.
- [4]. Armitage P., Berry G. et Matthews J. N. S. (2002), *Statistical methods in medical research*, ed Wiley, 4a edición.
- [5]. Arroyo G., Alvarez Y. et Sucar L. E. (2000), SEDRET - an intelligent system for the diagnosis and prediction of events in power plan, *Expert Systems With Applications*, 18, 2, 75-86.
- [6]. Arroyo G., Sucar L. E. (1999), A Temporal Bayesian Network for Diagnosis and Prediction, *Proceedings 15th Conference on Uncertainty in Artificial Intelligence (UAI)*, 13-20.
- [7]. Baxt W. (1990), Use of an artificial neural network for data analysis in clinical decision making: the diagnosis of acute coronary occlusion, *Neural Computation*; 2: 480-489.
- [8]. Bounds D. et Lloyd P. J. (1988), A multi-layer perceptron network for the diagnosis of low back pain, *Proceedings IEEE International Conference on Neural Networks*, 2, 481-489.
- [9]. Carvalho V., Naas I. A., Mollo M. et Massafera V. (2005), Prediction of the Occurrence of Lameness in Dairy Cows using a Fuzzy-Logic Based Expert System-Part I, *Agricultural Engineering International: the CIGR Ejournal*, 7.
- [10]. Chan K., Lee T., Sample P., Goldbaum M., Weinreb R. et Sejnowski T. (2002), Comparison of machine learning and traditional classifiers in glaucoma diagnosis, *IEEE Transactions on Biomedical Engineering*, 49, 9, 963-974.
- [11]. Charitos T., Gaag L., Visscher S., Schurink K. et Lucas P. (2009), A dynamic Bayesian network for diagnosing ventilator-associated pneumonia in ICU patients, *Expert Systems with Applications*, 36, 2, 1249-1258.
- [12]. Dabiri F., Vahdatpour A., Noshadi H., Hagopian H. et Sarrafzadeh M. (2008), Ubiquitous Personal Assistive System for Neuropathy, *The 2nd*

International Workshop on Systems and Networking Support for Healthcare and Assisted Living Environments (HealthNet).

[13]. Davis R., Ounpuu S., DeLuca P. et Romness M. (1999), Clinical Gait Analysis and Its Role in Treatment Decision-Making, *MedGenMed*, 1, 1.

[14]. Dreiseitl S., Ohno L., Kittler H., Vinterbo S., Billhardt H. et Binder M. (2001), A comparison of machine learning methods for the diagnosis of pigmented skin lesions, *Journal of Biomedical Informatics*, 34, 28-36.

[15]. Fernández S., Díez F.J., Arroyo G. et Sucar L.E. (2007), Comparison of two type of Bayesian networks: a case study, *Applied Artificial Intelligence*, 21, 3, 185 - 209.

[16]. Galán S. F., Aguado F., F. J. Díez et Mira J. (2002), NasoNet, Modeling the Spread of Nasopharyngeal Cancer with Networks of Probabilistic Events in Discrete Time, *Artificial Intelligence in Medicine*, 25, 3, 247-264.

[17]. Galán S. et Díez F. (2002), Networks of probabilistic events in discrete time, *International Journal of Approximate Reasoning*, 30, 2, 181-202.

[18]. Graves A. et Schmidhuber J. (2009), Offline handwriting recognition with multidimensional recurrent neural networks, , *Advances in Neural Information Processing Systems*, 21, 545-552.

[19]. Gresele P., Binetti B. M., Branca G., Clerici C., Ascutti S., Morelli A., Semeraro N. et Colucci M. (2008), TAFI deficiency in liver cirrhosis: Relation with plasma fibrinolysis and survival, *Thrombosis Research*, 121, 6, 763-768.

[20]. Harrel F. E. (2001), Cox proportional hazard regression model, *Model in regression modeling strategies*, 465 – 507.

[21]. Haworth J. (2008). Gait, aging and dementia, *Reviews in Clinical Gerontology*, 18, 39 – 52.

[22]. Hamilton J. D. (1994), Time series analysis, *Princeton University Press*, Princeton, NJ.

[23]. Jha S., Li W. et Seshia S. (2009), Localizing Transient Faults Using Dynamic Bayesian Networks, *High Level Design Validation and Test Workshop IEEE International*, 82-87.

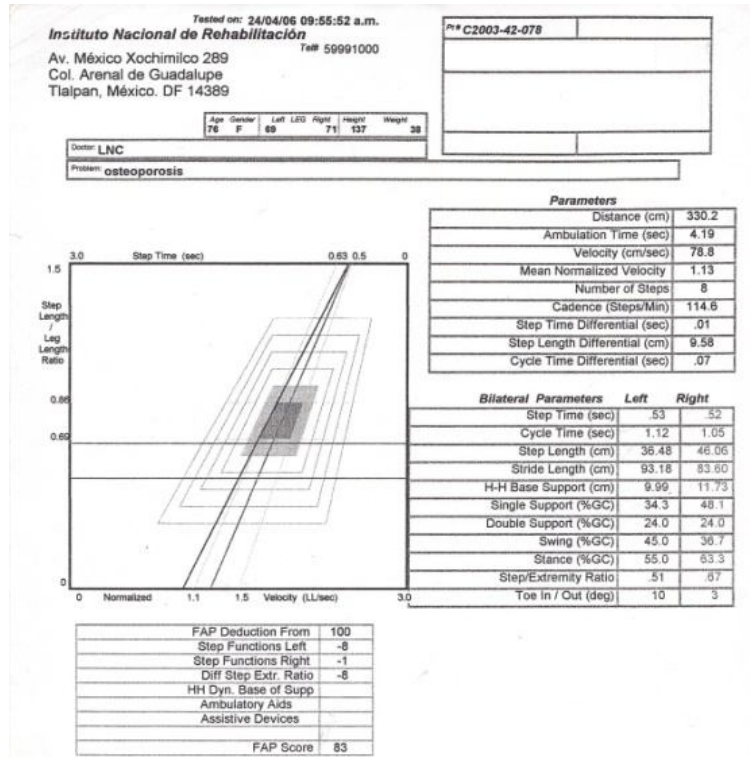
[24]. Johnson S., Fielding F., Hamilton G. et Mengersen K. (2009), An Integrated Bayesian Network approach to Lyngbya majuscula bloom initiation, *18th World IMACS / MODSIM Congress*, 4255-4261.

- [25]. Kangas M, Konttila A, Lindgren P, Winblad I et Jämsä T. (2008), Comparison of low-complexity fall detection algorithms for body attached accelerometers, *Gait Posture*, 2, 2, 8591.
- [26]. Kawahara Y., Fujimaki R., Yairi T. et Machida K. (2005), Diagnosis Method for Spacecraft Using Dynamic Bayesian Networks, *8th International Symposium on Artificial Intelligence*, 85-91.
- [27]. Kim Y., Street W. et Menczer F. (2000), Feature selection in unsupervised learning via evolutionary search, *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, 365 – 369.
- [28]. Kornacker M., Stumm J., Pott C., Dietrich S., Süßmilch S., Hensel M., Nickelsen M., Witzens-Harig M., Kneba M., Schmitz N., Ho A. D. et Dreger P. (2009), Characteristics of relapse after autologous stem-cell transplantation for follicular lymphoma: a long-term follow-up, *Annals of Oncology*, 722-728.
- [29]. Lee H. et Chou L. (2006), Detection of gait instability using the center of mass and center of pressure inclination angles, *Physical medicine and rehabilitation*, 87, 4, 569-575.
- [30]. Li H. et Luan Y. (2003), Kernel Cox regression models for linking gene expression profiles to censored survival data, *Biocomputing*, 65–76.
- [31]. Lin S. W., Sun C. H. et Chen C. H. (2008), Temporal Data Mining Using Genetic Algorithm and Neural Network: A Case Study of Air Pollutant Forecasts, *Geo-Spatial Information Science*, 7, 1, 31-38.
- [32]. Lloyd D. Fisher et Lin D. Y. (1999), Time-dependent covariates in the Cox proportional-hazards regression model, *Annual Review of Public Health*, 20, 145 - 157.
- [33]. Madu E. O., Stalbovskaya V., Hamadicharef B., Ifeachor E. C., Van Huffel S. et Timmerman D. (2005), Preoperative Ovarian Cancer Diagnosis using Neuro-Fuzzy Approach, *European Conference on Emergent Aspects on Clinical Data Analysis (EACDA 2005)*.
- [34]. Martínez M. et Sucar L. E. (2008), Learning Dynamic Naive Bayesian Classifiers, *Florida Artificial Intelligence Research Symposium (FLAIRS-21)*, 655-659.
- [35]. McGonigal M. (1994), A new technique for Survival Prediction in Trauma Care Using a Neural Network, *Proceedings World Conference on Neural Networks*, 3495-3498.

- [36]. Meyer D. (1997), Human gait classification based on hidden Markov model, *3D Image Analysis and Synthesis*, 139–146.
- [37]. Nan B., Emond M. J. et Wellner J. A. (2004), Information bounds for Cox regression models with missing data, *Annals of Statistics*, 32, 2, 723-753.
- [38]. Nelson A. J. (1974), Functional ambulation profile, *Physical Therapy*, 54, 10, 1059-1065.
- [39]. Noguchi H, Kato R, Hanai T, Matsubara Y, Honda H, Brusica V et Kobayashi T. (2002), Hidden Markov model-based prediction of antigenic peptides that interact with MHC class II molecules, *Journal of bioscience and bioengineering*, 94,3, 264-270.
- [40]. Palacios M., Brizuela C., Sucar L.E. (2008), Evolutionary Learning of Dynamic Naive Bayesian Classifiers, *Florida Artificial Intelligence Research Symposium (FLAIRS-21)*, 660-665.
- [41]. Papageorgiou E. I., Papandrianos N. I., Karagianni G., Kyriazopoulos G. et Sfyras D. (2009), Fuzzy Cognitive Map Based Approach for Assessing Pulmonary Infections, *ISMIS 2009*, 109-118.
- [42]. Peduzzi P., Concato J., Feinstein A. R. et Holford T. R. (1995), Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates, *Journal of Clinical Epidemiology*, 48, 1503-1510.
- [43]. Phillips P., Sarkar S., Robledo I., Grother P. et Kevin Bowyer (2002), The Gait Identification Challenge Problem: Data Sets and Baseline Algorithm, *International Conference on Pattern Recognition*, 1.
- [44]. Romero M., Martínez J. et Carrasco J. (2008). Gait Recognition Based on Silhouette, Contour and Classifier Ensembles. *Proceedings of 13th Iberoamerican Congress on Pattern Recognition (CIARP2008)*, Habana, Cuba. Published, 527-534.
- [45]. Saboune J. et Charpillet F. (2005), Markerless Human Motion Capture for Gait Analysis, *Reporte Técnico*.
- [46]. Salinas M. (2008), Modelos de Regresión VI. Análisis de Supervivencia, *Ciencia & Trabajo*, 10, 28, 75-78.
- [47]. Seijas C., Caralli A., Villasana S., Sáenz L. et Arteaga F. (2006), Predicción de hipertensión arterial usando maquinas de vectores de soporte, *Ingeniería UC*, 13, 3, 13-18.

- [48]. Stafford W. (2004), Support vector machine applications in computational biology, *Kernel Methods in Computational Biology*, 71-92.
- [49]. Tong S., Koller D., Kaelbling P. (2001), Support Vector Machine Active Learning with Applications to Text Classification, *Journal of Machine Learning Research*, 999-1006.
- [50]. Uden C. et Besser M. (2004), Test-retest reliability of temporal and spatial gait characteristics measured with an instrumented walkway system (GAITRite®), *BMC Musculoskeletal Disorder*, 5.
- [51]. Vidal R. et Favaro P. (2007), DynamicBoost: Boosting Time Series Generated by Dynamical Systems, *IEEE 11th International Conference on Computer Vision*, 1-6.
- [52]. Visscher S., Lucas P., Schurink C., Bonten M. (2009), Modelling treatment effects in a clinical Bayesian network using Boolean threshold functions, *Artificial Intelligence in Medicine*, 46, 251-266.
- [53]. Yee-Moon A., Wai-Kei C., Wang M., Hiu-Shuen I., Goggins W. B., Yu C., Lui S. et Zanderson J. (2007), Prognostic Value of Cardiac Troponin T Is Independent of Inflammation, Residual Renal Function, and Cardiac Hypertrophy and Dysfunction in Peritoneal Dialysis Patients, *Clinical Chemistry*, 53, 882-889.
- [54]. Yu L. et Liu H. (2007), Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution, *Proceedings of The Twentieth International Conference on Machine Learning (ICML-03)*.
- [55]. Zhou M. (2001), Understanding the Cox Regression Models with Time-Change Covariates, *The American Statistician*, 55, 2, 153-155.

Anexo A



Registro de datos de la marcha, con el sistema gaterite.

Subject Information

Number: C2003-42-004
 First Name: Irma
 Middle Name:
 Last Name:
 Date: 02/14/06
 Trial Type: Closed Base
 Trial Eyes: Eyes Open

Balance Test

General Statistics

X Avg -0.07 cm
 Y Avg -4.05 cm
 X SD 0.52 cm
 Y SD 0.42 cm
 X Range 2.62 cm
 Y Range 2.13 cm
 X Max 1.62 cm
 X Min -1.00 cm
 Y Max 0.93 cm
 Y Min -1.20 cm
 X Disp Avg 0.42 cm
 Y Disp Avg 0.35 cm
 X Skew -0.72
 Y Skew -0.14
 X Kurtosis -0.15
 Y Kurtosis -0.47

Radial Measures

Radial Disp Avg 0.59 cm
 Radial Disp SD 0.32 cm

Area Measures

Area Circular 3.22 cm sq
 Area Rectangular 5.59 cm sq
 Area Effective 1.41 cm sq

Path Length Measures

Path Length 206.44 cm
 Unit Path 2.06cm/sec
 Path/Area 64.03 1/cm

Velocity Measures

Vx Max 6.73cm/sec
 Vx Min -6.59cm/sec
 Vy Max 6.76cm/sec
 Vy Min -6.79cm/sec
 V Avg 2.06cm/sec

BOS Relative Measures

Dev CG X 0.64 cm
 Dev CG Y 2.67 cm

Ellipse - 95th Percentile

Covariance -0.12
 Area95 3.48 cm sq
 Major95 1.458 cm
 Minor95 0.759 cm
 Tan95 1.43
 Slope95 -55.10 deg

Registro de datos de estabilidad.