CIMAT

Centro de Investigación en Matemáticas, A.C.

ÁRBOLES DE DECISIÓN Y SU APLICACIÓN EN EL SÍNDROME METABÓLICO

T E S I N A

Que para obtener el grado de

Maestro en Computo Estadístico

Presenta

Arnulfo González Cantú

Director de Tesina:

Dr. Rodrigo Macías Páez

Monterrey, N.L. 29 de Enero del 2019

CIMAT

Centro de Investigación en Matemáticas, A.C.

ÁRBOLES DE DECISIÓN Y SU APLICACIÓN EN EL SÍNDROME METABÓLICO

T E S I N A

Que para obtener el grado de
Maestro en Computo Estadístico

Presenta
Arnulfo González Cantú

Director de Tesina:
Dr. Rodrigo Macías Páez

Autorización de la versión

Monterrey, N.L. 29 de Enero del 2019

Índice general

Índice de figuras

Capítulo 1

Introducción

Se presenta este trabajo como requisito parcial para obtener el grado de maestría en computo estadístico. La base de este proyecto es la utilización de los algoritmos de árboles de decisión en problemas médicos. Para tener una visión más amplia de esta técnica de aprendizaje maquina se realizó una revisión de la literatura, así como también se revisó la literatura existente de los diferentes problemas médicos a solucionar.

Esta revisión de la literatura incluye una descripción de las diferentes etapas en la evolución de los árboles de decisión. Los problemas médicos a resolver incluyeron las complicaciones del síndrome metabólico los cuales se incluyen pero no limitan a hígado graso no alcohólico, diabetes mellitus tipo 2 y retinopatía diabética. En todos los casos, el problema principal fue la clasificación. Se remarca que en el anexo a este trabajo encontrará una información mas detallada del conocimiento de estos problemas.

Dentro de los algoritmos de árboles de decisión usados para resolver los problemas medicos se incluyeron el C4.5, C5.0, árbol de decisión con mirada hacia adelante (Look Ahead Decision Tree) y un árbol de decisión sensible a costos. Además de se utilizó como herramienta de reducción de variables el algoritmo de regresión logística con paso hacia adelante.

Los resultados fueron presentados por complicación y por modelo detectado. Posterior a ello, se describen las contribuciones del trabajo y la conclusión. En caso de que el lector requiera mas información respecto a las técnicas utilizadas, se agrega un anexo donde también encontrará información médica.

1.1. Justificación y objetivos del trabajo

La justificación a este trabajo nace de la necesidad de tener herramientas que demuestren ser efectivas en solucionar los problemas en la investigación médica y que puedan ser llevadas a la práctica clínica diaria, ya que aunque es necesaria la investigación de nuevo conocimiento y caracterización de las enfermedades, es prioritario generar mayores recursos para una correcta toma de decisiones.

Esta idea de formación de recursos es apoyada por la alta incidencia de el síndrome metabólico y sus complicaciones. El gasto público desbordante es otra fuente principal que justifica el esfuerzo de este trabajo.

Por lo antes descrito, el presente trabajo cuenta con los siguientes objetivos

1. Clasificar las complicaciones del síndrome metabólico mediante el uso del árbol de decisión con variables bioquímicas y metabolómicas.

2. Uso de árbol de decisión para clasificación por su facilidad de interpretación.

3. Mejorar el algoritmo de decisión paraautomatizar el diagnostico de las complicaciones del síndrome metabólico.

4. Seleccionar las características mas importantes mediante el método de paso hacia delante, Neighborhood components analysis, y análisis de factores, reduciendo así las dimensiones.

5. Generar meta características para incrementar la eficiencia del árbol de decisión para clasificar las complicaciones del SM.

1.2. Contribución del trabajo

- Proponemos el algoritmo Generalized Cost Sensitive Look Ahead Decision Tree (GCSLADT) que puede automatizar el diagnostico de complicaciones del síndrome metabólico.
- GCSLADT supera otras variantes de árbol de decisión y criterio de síndrome metabólico para NAFLD.
- GCSLADT puede incorporar aprendizaje semi supervisado, aprendizaje de transferencia y aprendizaje profundo que funciona mejor para el diagnostico de complicación de síndrome metabólico.
- Active, Online, reinforcement e incremental learning pueden ser incorporados en la estructura del algoritmo GCSLADT.
- Incorporación de meta características ayuda a dar mayor exactitud de árbol de decisión para clasificar complicaciones del síndrome metabólico.
- La función de costo clase sensible también incrementa la exactitud de árbol de decisión para clasificar NAFLD porque los datos no son balanceados.
- Las técnicas de elegir características (NCA,FA, Forward selection) reduce la complejidad computacional del árbol de decisión por disminuir la dimensión sin reducir la exactitud.
- Se confirma la posibilidad de poder clasificar a los pacientes con diabetes, NAFLD y retinopatía con y metabolómicas.

Capítulo 2

Árboles de decisión.

2.1. Conceptos básicos

El árbol de decisión es una técnica simple y muy útil de aprendizaje máquina. Puede ser usada para regresión o clasificación; el manejo de la entropía y ganancia de información pueden explicar la relación entre las variables, explica fácilmente los datos y es fácil de seguir [58]. La manera en que el árbol de decision puede clasificar es parecida a la forma en la que los medicos toman decisiones en la vida real, y los límites dados por el árbol de decision pueden ser usados en la practica diaria.

El árbol de decisión T consiste de un gráfico dirigido con N nodos y hojas (E) que satisfacen unas propiedades en particular como son: tener solo una raíz (un nodo sin ramas que entren en él), una única vía de la raíz a cada nodo y no existen vías circulares, entre otras [7]. Cada árbol puede ser visto como una forma de separar los datos mediante cada nodo, que contiene una regla de valores de los atributos que guían a alcanzar un nodo hoja, el cual contiene la información responsable de la predicción.

Existen muchas formas de construir un árbol de decisión, la ganancia de información es uno de ellos y es basado en la entropía de Shannon ($H(\cdot)$). La entropía es una medida de incertidumbre de los datos y es definida como:

$$\text{Entropía} = -\sum_{i=1}^{m} p_i log_2(p_i) \tag{2.1}$$

Manejando solo proporciones (p_i) de variables, la entropía es fácil de obtener.

La variable seleccionada a particionar es la que tiene la mayor ganancia de información

$\Delta\phi$ definida como:

$$\Delta\phi = \text{Entropía}_b - \text{Entropía}_a \tag{2.2}$$

Así que, la ganancia de información $\Delta\phi$ es el residual de la entropía después (Entropy_b) de la partición de la variable. Este procedimiento termina cuando el subconjunto de datos es lo mas puro posible (llevando a una perfecta separación entre las clases) y la máxima reducción de entropía es alcanzada. Puede observarse que la ganancia de información puede ser sesgada hacía variables con muchos valores. Este problema puede ser resuelto con el "gain ratio", que es una normalización de la ganancia de información por la entropía, esto es:

$$\text{Gain ratio} = \frac{\text{information gain}}{\text{information content}} \tag{2.3}$$

El contenido de información es definido como $-f_i log_2 f_i$, y f_i para $i = 1, ..., d$ variables,también es la proporción del valor en la variable.

El proceso de entrenamiento sigue un proceso recursivo para maximizar la reducción de la entropia [73], y se resuelve como sigue:

Consideremos un problema de clasificación binario, con $\{\text{x,y}\}_{i=1}^{n} x \in \mathbb{R}^k, y \in \{1, 2\}$

1. Empieza desde la raíz, considera un conjunto grande de candidatos a particionar (k, β) que cubren todos los posibles k y provee suficientes subdivisiones para cada x_k. β es un valor de corte de partisión.

2. Para cada candidato (k, β), se particiona un conjunto de entrenamiento $D = \{(\mathbf{x}, y)\}$ en dos sub conjuntos:

$$D_l(k, \beta) = \{(\mathbf{x}, y) \mid x_k \leq \beta\} \tag{2.4}$$

$$D_r(k, \beta) = D \backslash D_l(k, \beta) \tag{2.5}$$

3. Encuentra el candidato (k, β) que maximize la reducción de la entropía $G(k, \beta)$:

$$(k^*, \beta^*) = argmax_{(k,\beta)} \text{Entropy}(k, \beta), \tag{2.6}$$

4. Usar (k^*, β^*) como la caracteristica indicador y el límite para el nodo de la partición actual, y repite los pasos anteriores para el sub-árbol izquierdo con $D_l(k^*, \beta^*)$ y el

derecho con $D_r(k^*, \beta^*)$.

5. Si la profundidad alcanza el máximo tamaño (o entropía) del conjunto de datos particionados $\tilde{D}$ del nodo actual es suficientemente pequeño, entonces ese nodo es una hoja, y la probabilidad de este nodo es:

$$P_T = \frac{\mid \{\mathbf{x}, y) \in \tilde{D} \mid y = 1\} \mid}{\mid \tilde{D} \mid} \tag{2.7}$$

$\mid \cdot \mid$ denota la cardinalidad del conjunto de datos. El nodo hoja es una cantidad $P_T(x)$ que indica la probabilidad de clasificación es 1.

2.2. Evolución de los árboles de decisión

El algoritmo de árboles de decisión es uno de los mas populares en la actualidad. Desde su aparición en 1963 con el algoritmo AID descrito por Mogran y Sonquist , este algoritmo ha tenido multiples mejoras a través del tiempo. Estas mejoras han sido de tal magnitud y cantidad que ha sido necesario dividir su evolución en generaciones. Estas pueden ser vistas en la figura 2.1 .

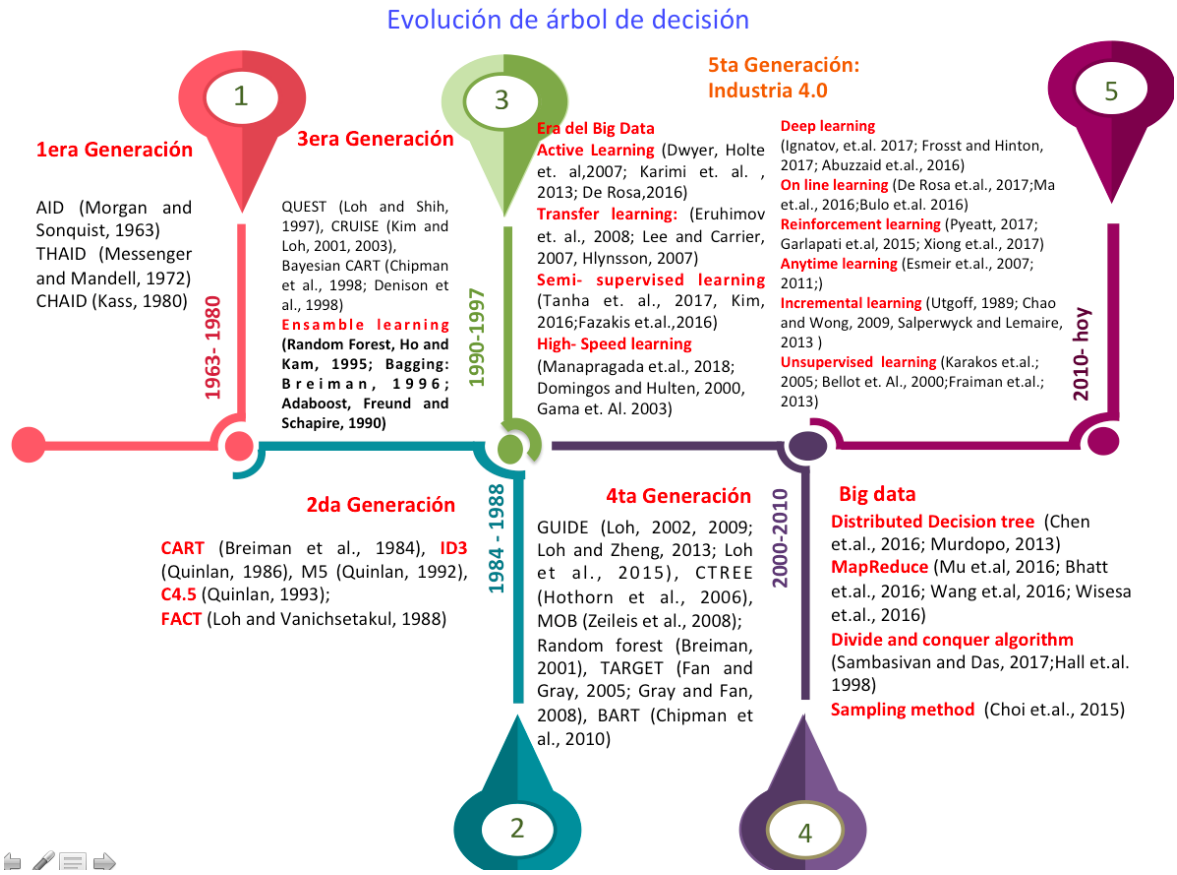

Figura 2.1: Evolución del algoritmo de árbol de decisión.

2.2.1. Primera Generación

Incluye la descripción del algoritmo AID [46], THAID [42], y el CHAID [29]. Esta generación se inició con árboles de decisión para variables continuas (regresión), ajusta un modelo constante paso a paso por una división recursiva de los datos en dos subgrupos

(nodos), con divisiones de la forma "$X \leq c$" o "$X \in A$".

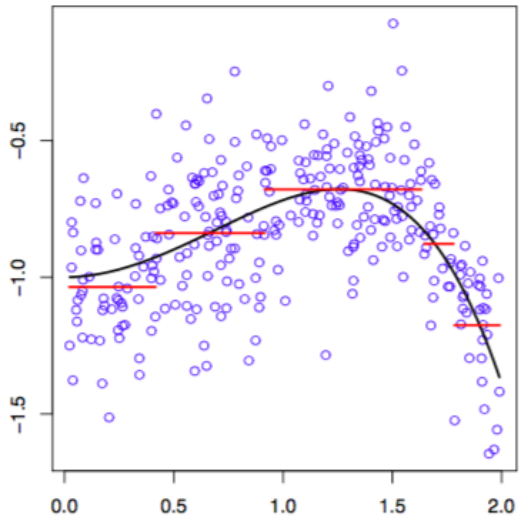

Figura 2.2: Ajuste de primera generación

Para esto, se definió el termino impureza (de cada nodo), **impureza** $\phi(t) = \sum_{i \in t}(y_i - \hat{y})$, un ejemplo de este ajuste se ve en la figura 2.2. Aunque existen otros algoritmos que también pueden clasificar (THAID) o que mejoran la velocidad (CHAID).

2.2.2. Segunda Generación

En esta generación aparece el algoritmo **C**lassification **A**nd **R**egression **T**rees (CART) descrito por Breiman et. al. [10]. Este algoritmo CART, usa la busqueda greedy utilizada en AID y THAID con otras adiciones:

1. Los árboles generados son podados en lugar de tener reglas de paro.
2. Los árboles son seleccionados por validación cruzada.
3. Se puede agregar un costo para las clasificaciones erroneas o para clases desbalanciadas.
4. Se manejan los valores perdidos por particiones surrogadas.
5. Se utilizan scores de importancia de las variables usadas para detectar el enmascaramiento.

	ID3	C4.5	CART
Criterio de Partición	Ganancia de información	Razón de Ganancia	Towing Criteria
Atributo	Categórico	Categorico y Numérico	Categórico y Numérico
Valores Perdidos	No maneja	Maneja	Maneja
Poda	No	Basado en error	Costo de Complejidad
Outlier	No maneja	No maneja	Maneja

Tabla 2.1: Características de los algoritmos de la 2da generación

6. Particiones lineales $\sum_i a_i x_i \leq c$ se obtienen al azar (RPART), es una implementación de CART en R.

Además Quinlan inicia su prolífica descripción de diferentes árboles de decisión con el algoritmo ID3 [53], M5 [55] y C4.5 [54]. En la tabla 2.1 se pueden observar las características que distinguen los diferentes algoritmos. En la figura 2.3 se puede ver como se clasifica una nueva observación, mediante las reglas obtenidas de los datos de entrenamiento.

El algoritmo ID3 recibió este nombre por que fue el tercero en procedimientos de identificación de series. Fue realizado con la intención de ser usado para datos nominales (no ordenados). Si el problema involucra variables con valor real, ellos son primero convertidos en intervalos, cada intervalo es tratado de forma no ordenada nominal. Cada split tiene un factor de rama de Bj, donde B, es el numero de atributos discretos de bins de la variable j escogida para la partición. En la practica, rara vez los datos son binarios así que la impureza de la razón de ganancia debe ser usada. Estos árboles tienen sus números de niveles igual a el número de variables ingresadas. El algoritmo continua hasta que todos los nodos son puros o no hay más variables para particionar. No hay poda en las presentaciones estándar del algoritmo.

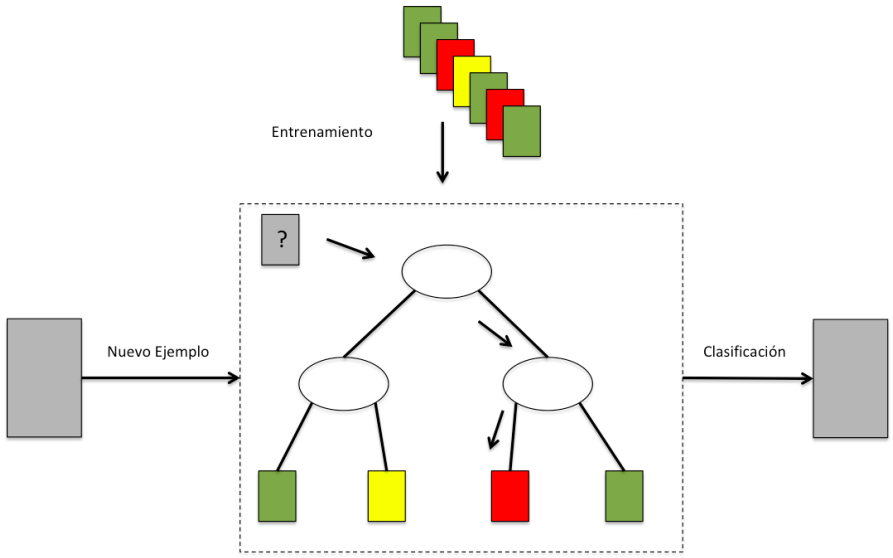

Figura 2.3: Clasificación mediante algoritmos de la 2da generación

2.2.3. Tercera Generación

En 1997, la aparición del algoritmo **Q**uick, **U**nbiased and**E**fficient **S**tatistical **T**ree(QUEST) descrito por Loh y Shih [35], inicia la tercera generación de algoritmos. Este algoritmo es el primero sin tener sesgo de selección. Se enlista algunas de sus principales características:

1. Usa ANOVA y tablas de contingencia con pruebas χ^2 para la selección de variables.
2. Mezcla clases en dos superclases para tener selecciones binarias.
3. Usa el análisis discriminante cuadratico para encontrar el punto de partición.
4. Usa imputaciones de datos mediante el promedio de los nodos.
5. Poda con el método CART.

Estas características son similiares a las encontradas en el algoritmo **C**lassification **R**ule with **U**nbiased **I**nteraction **S**election and **E**stimation, CRUISE (desarrollado por Hyunjoon kim y Wei-Yin Loh [30] [31]), en el sentido de que utilizan la prueba de χ^2 para la partición de variables y utiliza el análisis descriminante lineal para encontrar los puntos en donde partir. Sin embargo tiene otras que lo diferencían:

- Particiona cada nodo en tantos subnodos como el número de clases en la variable respuesta.

- Tiene un sesgo despreciable en la selección de variables.

- Tiene múltiples formas de lidiar con valores perdidos.

- Puede detectar interacciones locales entre pares de variables predictoras.

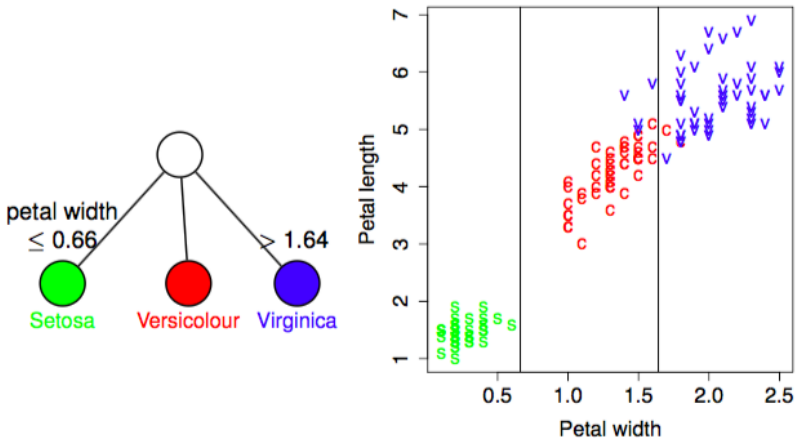

Figura 2.4: Ejemplo de el árbol de decisión CRUISE

Un ejemplo de este último algoritmo se muestra en la figura 2.4.

2.2.4. Cuarta Generación

Este mismo autor Loh, publica el algoritmo GUIDE (cuarta generación) que continua basandose en pruebas de significancia en el paso de dividir un nodo. Utiliza la prueba de χ^2. El pseudocódigo de este algoritmo se describen a continuación [34]:

Pseudocódigo del algoritmo GUIDE

1. Inicia en el nodo raíz.
2. Para cada variable ordenada X, convierte a una no ordenada variable Xál agrupar sus valores en el nodo, dentro de un pequeño numero de intervalos. Si X es no ordenada, X'= X.
3. Realiza una prueba de χ^2 de independencia para cada X'variable vs. Y en los datos del nodo y computa su probabilidad de significancia.
4. Escoge una variable X* asociada con el X'que tiene la probabilidad de significancia mas pequeña.
5. Encuentra la partición del conjunto $\{X \in S^*\}$ que minimiza la suma de los indices Gini, y usalo para dividir el nodo en dos nodos mas pequeños.
6. Si el criterio de paro es alcanzado, termina, De otra forma, aplica los pasos 2 - 5 para alcanzar un nodo hijo.
7. Poda el árbol con el metodo CART.

Este algoritmo puede dividir combinaciones de dos variables a la vez, y trata a los valores perdidos como una categoría separada. Una comparación entre los últimos algoritmos descritos se pueden ver en la tabla 2.2.

2.2.5. Quinta Generación

En la quinta generación se encuentra ya la era del big Data y la Industria 4.0. En ella se incluyen una gran cantidad de tipos de aprendizaje (learning) que tratan de superar los retos del Big Data. Se mencionará brevemente el significado de los diferentes aprendizajes de esta generación (tratando de no alejarse mucho del objetivo de esta tesis).

Aprendizaje activo (Active Learning)

Para describir el aprendizaje activo de una forma más comprensible, se debe contrastar con el aprendizaje pasivo (passive learning); que es el aprendizaje estándar, bien estudiado establecido en estadística y aprendizaje maquina). En el aprendizaje pasivo (ocasionalmente referido como aprendizaje supervisado), la meta es obtener un buen predictor de los

Característica	CRUISE	GUIDE	QUEST
Partición no sesgada	$\surd$	$\surd$	$\surd$
Tipo de partición	u, l	u, l	u, l
Ramas/Partición	≥ 2	2	2
Pruebas Interacción	$\surd$	$\surd$	
Poda	$\surd$	$\surd$	$\surd$
Costos (Usuario)	$\surd$	$\surd$	$\surd$
Previo (Usuario)	$\surd$	$\surd$	$\surd$
Ranking Variable		$\surd$	
Modelo de nodo	c, d	c, k, n	c
Bagging y Ensembles		$\surd$	
Valores perdidos	i,s	m	i

Tabla 2.2: Comparación de Métodos de clasificación con árboles. Una marca indica presencia de la característica

datos etiquetados. En el modelo de aprendizaje activo es un poco diferente a esto, ya que inicialmente los datos no tienen una etiqueta, el objetivo en este aprendizaje es el mismo que en el aprendizaje pasivo; sin embargo, en este tipo de aprendizaje esta permitido buscar una etiqueta respuesta para cualquier dato ingresado en los predictores [25].

Aprendizaje de transferencia (Transfer learning)

El aprendizaje de transferencia se refiere al proceso de aprender u obtener habilidades en el contexto de que una tarea determinada, y que puede ser usado en otro problema similar (pero no idéntico) para aprender nuevas tareas eficientemente [75].

El algoritmo de transferencia para árboles de decisión, aprende una nueva tarea u objetivo, de un modelo de decisión parcial, inducido por el ID3, que captura el conocimiento previo, de una tarea previa [53]. El algoritmo de árboles de decisión de transferencia, consta de dos partes: 1) consise en identificar atributos que no ocurren en el árbol de la tarea fuente y determina el orden en el que los nuevos atributos deben ser considerados; 2) consiste en aplicar transformaciones de la tarea fuente, para colocar los nuevos atributos en los lugares correctos, en conjunto con la etiqueta asociada.

Aprendizaje semi-supervisado (Semi-supervised learning)

Los métodos semi-supervisados no usan solamente los datos etiquetados, sino que también usan los datos no etiquetados; esto con el motivo de combinar la información de los

datos no etiquetados con los etiquetados para mejorar el desempeño de la clasificación [69]. En este caso, el algoritmo base de aprendizaje es el árbol de decisión y puede ser modificado al combinar el algoritmo con por ej. el clasificador ingenuo de Bayes (Naive Bayes Tree Classifier).

2.3. Algoritmo C4.5

El algoritmo C4.5, el sucesor y refinado ID3, es el mas popular en cuanto a métodos de çlasificación"basados en árboles. En el, las variables con valor real, son tratadas de la misma forma que en CART. Con particiones multi-vía, el algoritmo usa fundamentos heurísticos para la poda del árbol obtenido. Este algoritmo tiene prevé la poda basada en las reglas derivadas del mismo árbol. Esto es mediante la vía desde la raíz al nodo final, si existen reglas redundantes, estas son eliminadas [56].

La construcción básica del algoritmo C4.5 es la siguiente:

- Los nodos raíz es el primer nodo del árbol. En él se consideran todas los atributos, y selecciona los atributos que son más importantes.
- La información de la muestra es pasada a los nodos subsecuentes, llamados “nodos de rama” que eventualmente terminan en los nodos hoja que dan las decisiones.
- Las reglas son generadas mediante una vía que conecta el nodo raíz a un nodo hoja.

C4.5 usa los valores de probabilidad para tratar a los datos perdidos, en lugar de asignar valores comunes de algún atributo. Aunque este algoritmo es de uso extendido tiene algunas limitantes [40]:

- Ramas vacías: En caso que un nodo tenga 0 valores o valores cercanos a 0, no ayuda a construir reglas, sino que solo hace el árbol mas complejo y grande.
- Ramas sin significado: Las variables discretas pueden ayudar a formar un árbol de decisión pero no ayudan en la tarea de clasificación llevando a tener un árbol mas grande y al sobre ajuste.

2.4. Algoritmo C5.0

El algoritmo C5.0 es un algoritmo relativamente nuevo basado en su antecesor C4.5 (desarrollado por Quinlan) introduce nuevas tecnologías que incluyen el boosting y un árbol de decisión sensible a una función de costos. [51] Este algoritmo es una extensión de C4.5 que a su vez es una extensión del ID3, este algoritmo es uno que puede aplicar en el concepto del big data ya que es mejor que el C4.5 en cuanto a la velocidad, memoria y eficiencia.[12]

Un modelo C5.0 esta basado en la teoría de la información y trabaja separando el conjunto de datos en multiples sub muestras.[21]

Similar al algoritmo de Adaboost, el boosting en C5.0 es una mejora importante. Esta basado en el cálculo de un peso, el cual se incrementa con la influencia en la muestra. El peso es ajustado en cada iteración, con cada nueva muestra. El hecho de enfocarse a las muestras con peor clasificación dada por el árbol de decisión anterior hace que estas muestras tengan un mayor peso. Este método de hacer árboles de decisión es muy robusto para manejar los datos faltantes y grandes cantidades de "inputs" al modelo.

2.5. Árbol de decision con vista hacia adelante (Look Ahead Decisión Tree)

Aunque algoritmos de árboles de decisión descritos anteriormente son muy populares (como C4.5), este tipo de algoritmos basados en un enfoque glotón ("greedy"), tienen algunos inconvenientes, por ejemplo las particiones tempranas (o nodos) o anteriores pueden afectar los nodos subsecuentes, llevando a: 1) parada temprana del árbol (por encontrar un óptimo local), 2) puede afectar la solución final [11].

Una de las posibles alternativas es en utilización de un paso hacia adelante para escoger mejores particiones, poniendo atención la efectividad posterior (esto es suprime el efecto del horizonte). Aunque se piensa que puede mejorar la efectividad del algoritmo, algunos autores muestran la complejidad del árbol resultante y que tal vez pudiera afectar la exactitud [18] [17][47].

2.6. Árbol de Decisión Sensible a Costos (Cost Sensitive Decision Tree)

La clasificación en el contexto del aprendizaje maquina, maneja el problema de predecir la clase y_i del conjunto de ejemplos S, dado sus k variables. El objetivo es construir una función $f(D)$ que prediga las c_i clases de cada ejemplo usando las variables X_i. Esto pensando que los diferentes errores de clasificación tienen el mismo costo. Los métodos que usan diferentes costos de error de clasificación se conocen como clasificadores sensibles al costo; los diferentes tipos de algoritmos se presentan en la figura 2.5 [6] .

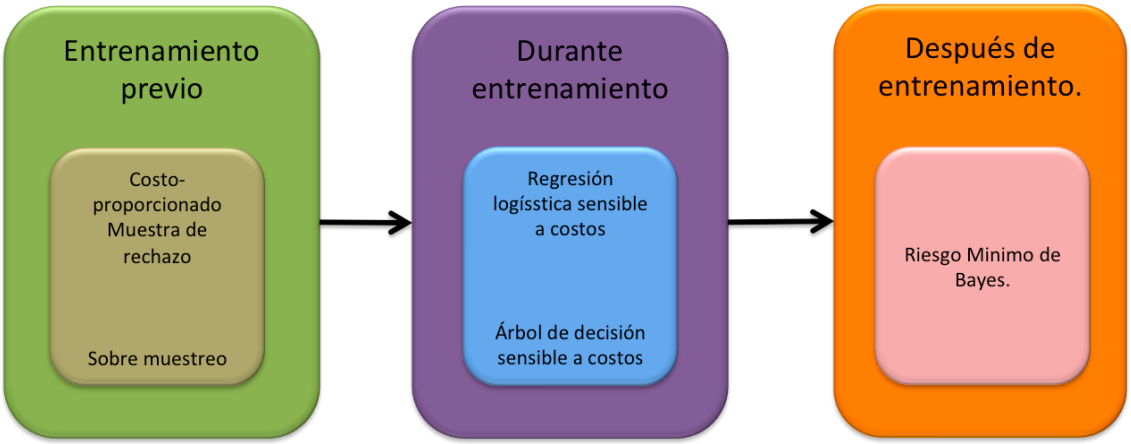

Figura 2.5: Algoritmos Sensibles a Costos

En el modelo propuesto por Correa Bahnsen [6], se usa un criterio de partición diferente durante la construcción del árbol de decisión. El propuso un modelo sensible a costos llamado Clasificador de Mínimo de Bayes. El riesgo que acompaña cada decisión es calculada. En un ejemplo de una clasificación binaria, el riesgo en la predicción de un ejemplo i como negativo es:

$$R(c_i = 0|X_i) = C_{TN_i}(1 - \hat{p}) + C_{FN_i} \cdot \hat{p}$$

y

$$R(c_i = 1|X_i) = C_{TP_i} \cdot \hat{p} + C_{FP_i}(1 - \hat{p})$$

es el riesgo cuando la predicción es positiva, donde $\hat{p}_i$ es la probabilidad estimada positiva

para el ejemplo i. Subsecuentemente, si:

$$R(c_i = 0|X_i) \leq R(c_i = 1|X_i)$$

entonces el ejemplo i es clasificado como negativo. Esto significa que el riesgo asociado con una decisión c_i es menor que el riesgo asociado con el clasificar a i como positivo. Sin embargo, cuando se usa para una clasificación binaria como base para la toma de decisiones, existe una necesidad por la probabilidad que no solo separe bien entre positivo y negativo, pero que también evalúe la probabilidad real de un evento.

Capítulo 3

Algoritmo Utilizado: Árbol de Decisión Sensible a Costos y con Mirada Adelante Generalizado

Después de la revisión de literatura médica y de los diferentes algoritmos existentes de árboles de decisión, se tienen en mente diferentes problemas a la hora de realizar el análisis de este tipo de datos. Y estos incluyen:

1. Datos desbalanceados. Los grupos de las diferentes etiquetas no son iguales.

2. Posible error al seleccionar de variables. Siendo los datos de alta dimensionalidad, se puede presentar la ocasión de que la selección de determinada variable pueda hacer crecer un árbol con una tendencia a la mala clasificación de pacientes.

Por estos problemas se decide proponer un algoritmo que incluya la solución (al menos parcial) de estos problemas: el Algoritmo Cost Sensitive Look Ahead Decision Tree.

En este algoritmo se utiliza el criterio de información basado en la reducción de la entropía, siempre tomando en cuenta primeramente a la variable que maximiza esta reducción $G(k, l)$.

Los elementos implicados en el algoritmo son:

Donde $G(k,\tau) = E_{antes} - E_{despues}$.

$(k^*, \tau^*) = \text{argmax}_{k,\tau} G(k,\tau)$

E denota la entropía, que es una medida de incertidumbre en los datos.

E_{antes} y $E_{despues}$, denotan la entropía antes y después de la partición.

$E_{antes} = E(D) = \sum_{i=1}^{m} f(w_i, D) p(c_i, D) log_2(p(c_i, D))$

$E_{despues} = \frac{|D_l(k,\tau)|}{|D|} E(D_l(k,\tau)) + \frac{|D_r(k,\tau)|}{|D|} E(D_r(k,l))$

Donde D denota conjunto de entrenamiento, $C = \{c_1, c_2, \cdots, c_m\}$, denota el indice nivel de clase,

es matriz de características, y pertenece al nivel de clase de cada muestra.

$D = \{(\mathbf{v}, y)\}$, $\mathbf{v} = \{v_1, v_2, \cdots, v_M\}$

Tuple (k,τ) divide el conjunto de entrenamiento $D = \{(\mathbf{v}, y)\}$ en dos subconjuntos :

$$D_l(k,\tau) = \{(\mathbf{v}, y) | v_k \leq \tau$$

$$D_r(k,\tau) = D \backslash D_l(k,\tau)$$

$E(D_l(k,\tau)) = \sum_{i=1}^{m} f(w_i, D_l) p(c_i, D_l) log_2(p(c_i, D_l))$

$E(D_r(k,\tau)) = \sum_{i=1}^{m} f(w_i, D_r) p(c_i, D_r) log_2(p(c_i, D_r))$

A esta forma de particionar los datos, se le puede sumar la posibilidad de agregar un peso a eventos de errores de clasificación. Esto es hacer al algoritmo sensible al costo, se puede ver en la siguiente tabla de contingencia que los errores que deben ser pesados (y disminuidos) son los falsos positivos y falsos negativos.

	Actual Positivo	Actual Negativo
Predicho positivo	C_{TP_i}	C_{FP_i}
Predicho Negativ	C_{FN_i}	C_{TN_i}

Tabla 3.1: Matriz Costos de clasificación[5]

Árbol de decisión sensible a costos

$f(w_i, D) = \frac{\sum_{i=1, c \neq i}^{m} |D(c_i)|}{|D|}$

$|D(c_i)| = \#$ de observaciones en D pertenecientes a clase c_i. $f(w_i, D) = 1 \forall i$ para árboles de decisión no sensibles a costo

3.1. Forma de evaluar la efectividad de los árboles de decisión

Una de las características mas importantes que debe contar el modelo de clasificación es el de la exactitud (**accuracy**). Este define el porcentaje sujetos que son correctamente clasificados.

Otra de los características utilizadas en las pruebas medicas es la sensibilidad (**sensitivity**) que es la habilidad de una prueba de dar un resultado positivo en casos verdaderos de enfermedad. La especificidad (**specificity**) es la habilidad de dar un resultado negativo en caso de que la enfermedad este ausente. Ambas son expresadas en proporción. Para calcular ambas características de las pruebas se determinan los casos verdaderos y falsos positivos; esto es: cuantos de los sujetos son clasificados correcta e incorrectamente. Podemos observar la formula de ambas a continuación:

$$\text{Sensibilidad } = \frac{TP}{TP + FN} \tag{3.1}$$

$$\text{Especificidad } = \frac{TN}{TN + FP} \tag{3.2}$$

donde TP son los verdaderos positivos, y FN son los falsos negativos.

Otros términos que se agregaron a la evaluación del modelo son las probabilidades post test. Estas probabilidades son por así decirlo, la inversa de la sensibilidad y la especificidad. Y se refieren a la probabilidad de que un paciente tenga una enfermedad ya que tiene la prueba positiva (**valor predictivo positivo**) o de que no la tenga si tiene una prueba negativa (**valor predictivo negativo**). Las formulas de estas probabilidades se muestran a continuación[27]:

$$\text{Valor predictivo positivo } = \frac{TP}{TP + FP} \tag{3.3}$$

$$\text{Valor predictivo negativo } = \frac{TN}{TN + FN} \tag{3.4}$$

Se remarca que los árboles obtenidos serán validados mediante una validación cruzada de 10 carpetas y mediante el método de bootstrap.

Se hace mención también que en la sección de anexo, se tendrá la oportunidad de revisar un poco sobre el tema de regresión logística y la selección hacia adelante de variables

(Forward selection), así también se revisará un poco sobre la técnica de bootstrap.

Capítulo 4

Aplicación de los árboles de decisión en el síndrome metabólico

4.1. Origen de los datos

En esta sección se detalla con mas profundidad la obtención de las bases de datos utilizadas en este trabajo. Los resultados serán discutidos en el texto y se mostrarán en las tablas. Se debe comentar que de las bases de datos obtenidas en los diferentes estudios, se calcularon las metas características para cada una de las bases de datos. Estas meta variables, refieren a nuevas variables que son construidas de las bases de datos mediante combinaciones lineales de las variables. En las tablas donde se refiera a los datos obtenidos mediante meta características se les agrega un símbolo de (+) cuando los datos son analizados en crudo más las meta características.

4.1.1. Hígado graso no alcohólico y retinopatía.

La base de datos para el el estudio de el Hígado graso no alcohólico y retinopatía se obtuvo de un proyecto llevado acabo en el Hospital General de México. Se diseñó un estudio transversal, observacional, comparativo y prolectivo en el que completamos cuatro grupos de pacientes:

1. Pacientes con EHNA y obesidad (IMC>30),

2. Pacientes con EHNA sin obesidad (IMC<25),

3. Pacientes sin EHNA con obesidad

4. Pacientes sin EHNA sin obesidad

Estos grupos son desbalanceados (tamaños de muestra diferentes), y con los siguientes criterios de: Inclusión:

- Edad: 18-45 años
- Hombres y mujeres por igual
- IMC para los obesos: >30; para los no obesos: 20 -25

Exclusión:

- Pacientes con hábito de fumar que tengan índice tabáquico>1.
- Ingestión de alcohol >10g a la semana.
- Pacientes con historia clínica de Diabetes, Hipertensión, Insuficiencia renal crónica, cáncer enfermedad inflamatoria o infecciosa aguda o crónica.
- Pacientes que estuvieran tomando medicamentos hepatotóxicos.
- Pacientes que no son conocidos por alguna de las condiciones del inciso anterior pero que durante el examen físico o los resultados de laboratorios fueran diagnosticadas.
- Pacientes con alguna patología ocular que impida visualización de fondo de ojo o presencia de alguna otra retinopatía asociada.
- Pacientes embarazadas.
- Pacientes que no acepten participar.
- Pacientes que hayan aceptado participar que no acudan a las citas programadas para medición de variables.

Variables y obtención de las mismas.

Acorde al tratamiento estadístico las variables pueden modificarse como predictoras (independientes) o de respuesta (dependientes). En general las podemos clasificar en los siguientes grupos:

Variables independientes (ver anexo1): talla, peso, índice de masa corporal, obesidad, cintura, porcentaje de grasa corporal, presión arterial, curva de tolerancia a la glucosa, índice de sensibilidad a la insulina por Matsuda [39],colesterol total, triglicéridos, colesterol de alta densidad (c-HDL), colesterol de baja densidad (c-LDL).

Variables dependientes (ver anexo 1): Factor estimulante de colonias de granulocito-macrófago (GM-CSF), Interferón Gama (IFN-?) , Interleucina 10 (IL-10), Interleucina 2 (IL-2), Interleucina 4 (IL-4), Interleucina 6 (IL-6),Interleucina 8 (IL-8), Factor de necrosis-tumoral alfa (TNF α), Proteína C Reactiva. Diámetro arteriolar y venular de la retina, cruces arteriovenosos, relación arteriovenular, tortuosidad, alanino amino transferasa (ALT), glutamino amino transferasa (AST).

Se incluyeron a pacientes de la consulta externa del Hospital General de México (febrero - agosto 2012). Posterior a aplicar criterios de inclusión y exclusión se citaron a los pacientes un único día de acuerdo a las preferencias y conveniencias del mismo. Previo firma de consentimiento informado se realizó una breve historia clínica donde se interrogó sobre los antecedentes de tabaquismo, alcoholismo, consumo de medicamentos hepatotóxicos, hepatitis, hipertensión, diabetes, enfermedades crónicas o agudas infecciosas; historia familiar de diabetes, obesidad e hipertensión. Se tomó presión arterial, aplicando los criterios de la JNC7 para descartar hipertensión, peso, talla y cálculo de IMC, así como realización de impedancia bioeléctrica con aparato RJL, modelo Quantium IV (EUA).

La toma de muestra para análisis de parámetros bioquímicos se realizó de la siguiente manera: el personal de enfermería calificado colocó un punzocat de12-16Gen venas superficiales de antebrazo. Se tomó una muestra sanguínea basal, de la cual se obtuvo suero para la medición de: glucosa, urea, creatinina, ácido úrico, colesterol total, colesterol HDL, colesterol LDL, Triglicéridos, Proteína C reactiva, AST, ALT; los que se analizaron con un equipo Beckman automatizado . También se obtuvo suero para medición de insulina la que se analizó por método de ELISA con el equipo Multiskan Ascent V1.24 (USA).El coeficiente de variación de las replicaciones (CV %) osciló entre 1.81 y 2.94 %. Además se midió la concentración en suero de IL 2, IL4, IL6, IL8, IL10, IFN, GM-CSF, TNF? a través de ELISA múltiples con equipo Bioplex-ProTMAssays (Bio Rad, USA). El coeficiente de

variación de las repeticiones (CV %) fue el siguiente: GM-CSF: 19, IFN-γ: 15, IL-10: 8, IL-2: 12, IL-4: 11, IL-6: 16, IL8: 17, TNFα: 4 .

Además se realizó curva de tolerancia a la glucosa oral con 75 gr de glucosa anhidra, se tomaron muestras sanguíneas a los 30, 60, 90 y 120 minutos.

Para la fotografía de fondo de ojo se administró el midriático TP (tropicamida). La fotografía digital se realizó con la cámara Visucam NM/SA número 07740. Las fotos fueron tomadas con foco en papila central, temporal, nasal y se realizó una reconstrucción de 7 imágenes en una sola por computadora. Estas imágenes fueron analizadas de manera ciega, al azar por el retinólogo que mostró a través de la prueba de concepto mayor correlación clínica intraobservador (ver marco teórico). El retinólogo consideró como fondo de ojo anormal (FOa) la presencia de al menos 2/3 de los siguientes características clínicas: pérdida de la relación arteria-vena, tortuosidad y cruces arteriovenosos patológicos.

Finalmente se realizó en los pacientes ultrasonido de hígado y vías biliares con equipo Voluson Pro V de General Electric (USA), con un transductor de 3.5MHz, para determinar la presencia de EHNA de manera clínica apreciativa de acuerdo a la imagen observada por operador con base a tres parámetros [13] [52]: 1) Ecotextura: la esteatosis se observa como un incremento de la ecogenicidad en ecos muy finos y condensados, con apariencia de ?hígado brillante?. 2) Aumento en la atenuación: a mayor atenuación mayor dificultad de penetrar el hígado, lo que causa oscurecimiento posterior y pérdida de la definición del diafragma, lo que también resulta en un riñón relativamente hipoecoico. 3) Vasos hepáticos: disminuye la visualización de las venas porta y hepáticas, dando lugar a una apariencia blanda o sin características del hígado, por la compresión del parénquima lleno de grasa; estos hallazgos hacen difícil la diferenciación entre esteatosis hepática difusa y otras enfermedades parenquimatosas difusas. Además se calculó la distribución del tono gris de los pixeles en el lóbulo derecho del hígado y se comparó con la densidad en pixeles del riñón derecho, con lo que se calculó el índice:

$$\frac{\hat{x}_{pix}\text{hepático}}{\hat{x}_{pix}\text{renal}},$$

donde el numerador corresponde al promedio de grises de los pixeles registrados en el hígado; y el denominador al promedio de los mismos en el riñón correspondiente

4.1.2. Diabetes mellitus

Este estudio fue llevado a cabo en el Departamento de Endocrinología del Hospital Universitario "José E. González", a los pacientes que acudieron a la consulta externa, se les invito a participar a nuestro estudio previo consentimiento informado. Se incluyeron 124 pacientes los cuales fueron categorizados en la forma descrita en la figura 4.1

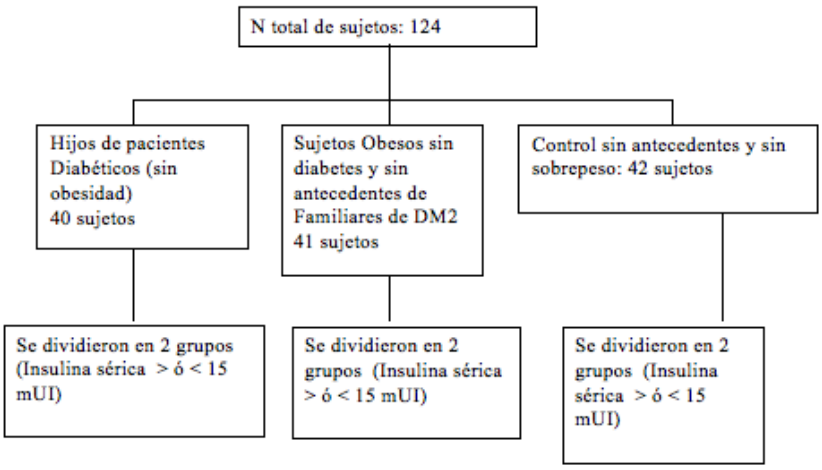

Figura 4.1: Grupos de pacientes de donde se obtuvieron los datos para la predicción de Diabetes

Para este trabajo solo se incluyeron, los grupos de sujetos normales y de los sujetos con diabetes meliitus. Y fueron reclutados bajo los siguientes criterios de inclusión.

- Sujetos sin antecedentes de diabetes mellitus, no diabéticos, sin sobrepeso (IMC <25kg/m2).

- Sujetos con Diabetes mellitus

Metodología

A los pacientes que cumplieron con los criterios de inclusión se les realizo una antropometría además de la recolección de muestra para la evaluación bioquímica:

Obtención de muestras de sangre

Previa firma de consentimiento informado se colocó una canalización y se les tomaron muestras basales que incluyen: glucosa, perfil de lípidos, insulina, acil carnitinas; aproximadamente se extrajeron 60 ml. de sangre. Posteriormente se les dará 75g de glucosa y se les mantendrá en reposo en el área de Pruebas Dinámicas de Endocrinología. En el transcurso de la prueba se realizará Curva de Tolerancia a la Glucosa de 2 horas. Al las dos horas de ingerido se repetirá la insulina, acilcarnitinas. Aproximadamente se extraerán 20 ml adicionales de sangre, para un total de 80 ml. Las muestras serán enviadas al laboratorio correspondiente para análisis. Se removerá la canalización y se citará al paciente para entrega de resultados de laboratorio y discusión de los hallazgos. Se tomarán 2 tubos de sangre, los que serán guardados en congelador para su análisis posterior.

Glucosa Enzimática. Método de Trinder-Colorimétrico. La determinación de la glucosa se realizara utilizando la glucosa oxidasa (GOD). La glucosa es oxidada por la GOD a acido glucurónico y agua oxigenada. En otra reacción secuencial interviene la peroxidasa que une al p-hidroxibenzoato con 4-aminoantipirina formando un cromógeno con absorbancia máxima de 505nm.

Acil carnitinas por espectrometría de masas en tandem. El espectrómetro de masas es un detector que identifica masas (peso) de moléculas individuales y sus fragmentos. Lo detectado se presenta en una grafica; eje de las x representa las masas y el eje de las y la cantidad de iones. El espectrómetro tiene alta sensibilidad y especificidad y capacidad de multi análisis.

Insulina por Electro quimioluminiscencia. Se realizan dos incubaciones en técnica de sándwich que duran en total 18 minutos. En la primera incubación se ponen 20° de la muestra con un anticuerpo monoclonal biotinilado específico anti-insulina y otro anticuerpo monoclonal específico marcado con quelato de rutenio. La segunda incubación incluye la adición de micropartículas de estreptavidina, dando como resultado un complejo que se fija a la base sólida por interacción entre la biotina y la estreptavidina. La mezcla se transfiere a un lector que por magnetismo separa la micro partículas. Los elementos no fijados se eliminan. Se aplica una corriente eléctrica definida y se produce una reacción de quimioluminicencia cuya emisión de luz se mide directamente con un fotomultiplicador. Los resultados se obtienen mediante curva de calibración.

Hemoglobina glucosilada. Se basa en la inhibición de la inmunoaglutinación de partículas látex. Tras introducir el cartucho en la cámara analítica del sistema DCA 2000, el resultado

aparecerá en 6 minutos. En este análisis se determinan la concentración de HbA1c, la concentración de hemoglobina total y la relación entre ambas, se reporta como porcentaje de hemoglobina A1c.

4.2. Hígado graso no alcohólico (NAFLD)

En esta sección se desarrolla la aplicación de árboles de decisión para la clasificación de pacientes con hígado graso no alcohólico (NAFLD). Se aplicaran los algoritmos descritos de C4.5, C5.0, look ahead y cost sensitive; así como la reducción de dimensión con la regresión logística con paso adelante.

4.2.1. Algoritmo C4.5 y C5.0

Para resolver el problema de clasificación de pacientes con hígado graso no alcohólico, se utilizaron todas las variables reportadas (no solo las relacionadas a las acilcarnitinas). En la figura 4.2 se puede observar como el modelo obtenido por el algoritmo C4.5 toma encuentra variables clínicas y bioquímicas básicas como la circunferencia abdominal, glucosa de ayuno y triglicéridos.

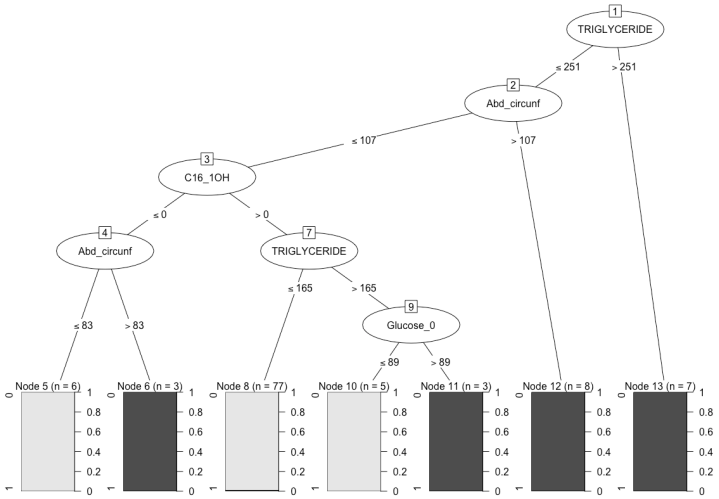

Figura 4.2: Modelo obtenido del algoritmo C4.5

El desempeño de este modelo es bastante aceptable en general, teniendo un buen balance en cuanto a sensibilidad y especificidad. La exactitud del modelo es muy alta, 97.6 %, tabla 4.1.

Validación cruzada	
Exactitud	97.6 % (IC 94.3 %, 100 %)
Sensibilidad	91.8 % (IC 81.2 %, 100 %)
Especificidad	100 %
V. Predictivo Positivo	100 %
V. Predictivo Negativo	97 % (IC 92.9 %, 100 %)

Tabla 4.1: Desempeño de modelo obtenido por C4.5

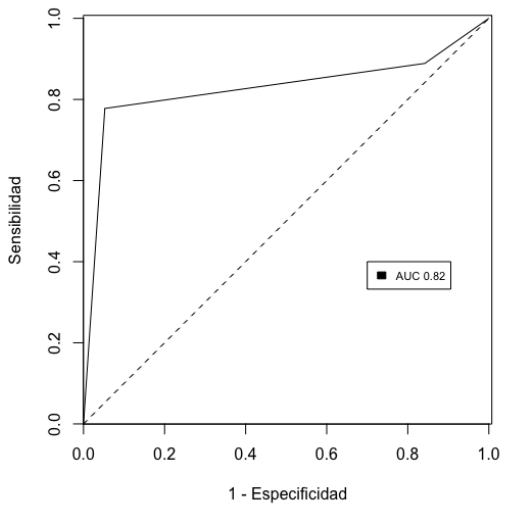

Figura 4.3: AUC del modelo obtenido por C4.5

No se muestra el modelo obtenido con C5.0 ni su tabla de desempeño porque es exactamente la misma que la obtenida por C4.5.

4.2.2. Reducción de dimensiones realizado mediante regresión logística

En la figura 4.4 se aprecia como las variables de peso, triglicéridos y acilcarnitnas de cadena larga (C14 - C18) son fuertes predictores de la presencia de hígado graso. El desempeño de este modelo es muy competitivo y sigue caracterizandose por buena especificidad, tabla 4.2.

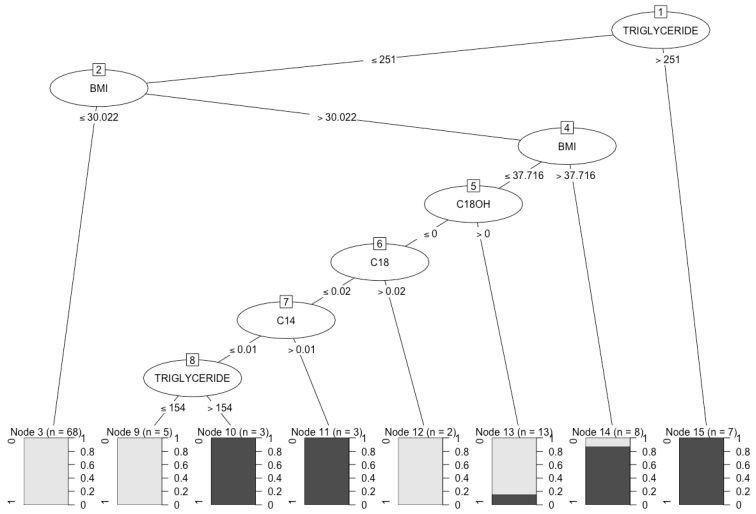

Figura 4.4: Modelo obtenido mediante C4.5 y FS/LR

Validación cruzada	
Exactitud	94.9 % (IC 78.1, 91.7)
Sensibilidad	80 % (IC 65.8 %,94.1 %)
Especificidad	87.4 % (IC 82.4 %, 92.5 %)
V. Predictivo Positivo	66.6 % (IC 54.5 %, 78.7 %)
V. Predictivo Negativo	92.2 % (IC 86.3 %, 98.1 %)

Tabla 4.2: Desempeño de algoritmo C4.5 FS/LR

4.2.3. Evaluación del desempeño mediante bootstrap

Como en la sección anterior, describimos el resultado de la evaluación de los modelos mediante bootstrap, como otra medida de evaluación de validación de los árboles obtenidos.

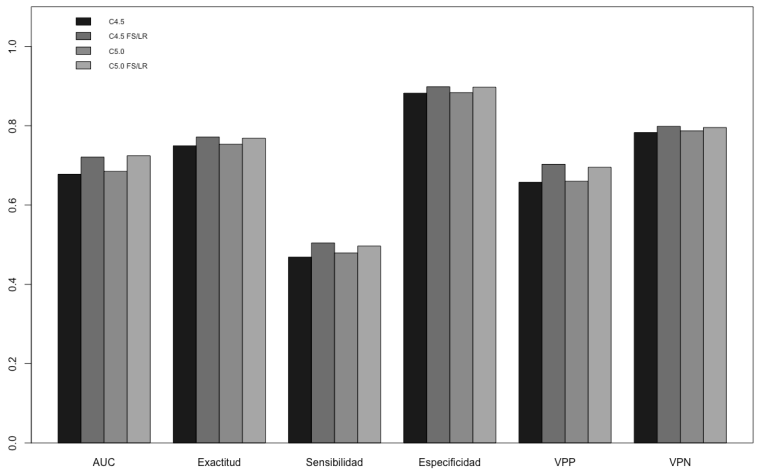

Figura 4.5: Desempeño de los algoritmos mediante bootstrap

Como se muestra en la figura 4.5, los algoritmos tienen una tendencia a ser muy específicos y en este caso C4.5 y C5.0 en conjunto con FS/LR presentan la misma especificidad ya que se obtienen los mismos modelos.

4.2.4. Aplicación de árbol de decisión sensible a costos

Como se observa en la tabla 4.3, en las primeras dos columnas se muestran los diferentes costos aplicados a el error de clasificación y en la ultima columna se muestra el cambio subsecuente de la exactitud. El árbol de decisión costo sensible puede incrementar de manera sobresaliente la exactitud en comparación a C4.5 y C5.0, pero no hay gran diferencia entre la exactitud derivada de diferentes costos.

Costo (Estimado = NAFLD/ Actual = Normal)	Costo (Estimado = Normal/ Actual = NAFLD)	Exactitud
0.25	0.75	90.6
0.75	0.25	91.24
0.33	0.67	90.73
0.67	0.33	92.45
0.5	0.5	91.24

Tabla 4.3: Resultados de clasificación de NAFLD con diferentes costos

4.2.5. Look Ahead Decision Tree.

Como se muestra en la tabla 4.4, este algoritmo no supera al anterior y de hecho empeora la clasificación (disminución de la exactitud). Aunque si se utilizan dos nodos adelante para la clasificación se mantiene con casi la misma clasificación que los algoritmos base de C4.5 y C5.0.

	Deep 1	Deep 2	Deep 3
B = 0.01	78	84.6	78.7
B = 0.03	78.9	77.4	76.7
B= 0.06	78.6	73.6	74.5
B = 0.09	78.7	80.9	79.6
B = 0.1	81.8	82.4	77.3

Tabla 4.4: Resultados para clasificar pacientes con NAFLD con diferente beta (B) y diferente profundidad (Deep) de vista hacia adelante. En las celdas de la tabla se muestra la exactitud para diferente B y la diferente profundidad del paso hacia adelante.

4.2.6. Meta características

En este segmento se describe la aportación de las meta características para la clasificación de hígado graso. Como se ve en la tabla 4.5, El algoritmo C5.0 puede mejorar su clasificación sobre todo con los datos generados de por el método de suma. El resto no muestra un gran cambio.

	C4.5	C5.0
Suma	83.9	92.8
Suma +	83.9	92.8
Resta	81.7	89.1
Resta +	81	89.8
Multiplicación	83.9	87.8
Multiplicación +	83.9	87.8

Tabla 4.5: Exactitud (%) de los algoritmos con metacaracteristicas en NAFLD

En la tabla 4.6 se observa como el algoritmo costo sensible puede mejorar la clasificación con practicamente cualquier costo y sin una predilección entre el origen de los datos.

Costos	0.25 - 0.75	0.75 - 0.25	0.33 - 0.67	0.67 - 0.33	0.5 - 0.5
Suma	93.6	87.5	93.4	90	94.5
Suma +	93.7	90.3	94.8	91.2	92.7
Resta	92.1	87.9	89.2	90.2	91.2
Resta +	92.5	91.4	90.7	91.4	90.5
Multiplicación	89.4	93	90	93.6	89.7
Multiplicación +	92.3	95	89.76	93.4	90.8

Tabla 4.6: Exactitud (%) de los algoritmos con metacaracteristicas con CSDT en NAFLD

Por otro lado, el algoritmo look ahead incrementa la exactitud en combinación con las meta características originadas por resta y en Deep 2.Tabla 4.7

	Deep 1	Deep 2	Deep 3
Suma	75.2	76.4	75.9
Suma +	74.5	75	75.8
Resta	84.6	90.4	86.7
Resta +	82.3	85.4	83.1
Multiplicación	82.5	84	82.4
Multiplicación +	83.2	81	83.2

Tabla 4.7: Exactitud (%) de los algoritmos con metacaracteristicas con LADT previa selección de variables en NAFLD (todos fueron con B = 0.1)

4.3. Diabetes mellitus

En esta sección se aplican los diferentes algoritmos de árboles de decisión para la clasificación de sujetos con y sin diabetes mellitus tipo 2.

4.3.1. Algoritmo C4.5 y C5.0

Los modelos resultantes de evaluar la información metabolómica de los pacientes con diabetes mellitus tipo 2 se observan en las figuras 4.6 y 4.7 .

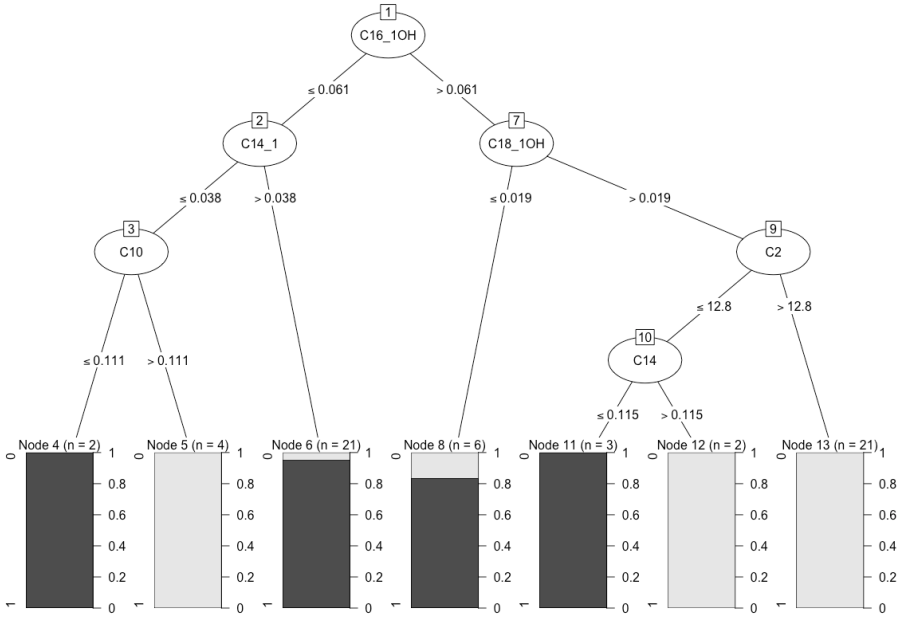

Figura 4.6: Árbol obtenido el algoritmo C4.5

Estos árboles de decisión muestran como las variables explicativas relacionadas a las acilcarnitinas de cadena larga son muy importantes en la clasificación de los pacientes con diabetes mellitus tipo 2. Ambos son muy parecidos en tamaño y en las variables tomadas en cuenta para clasificar.

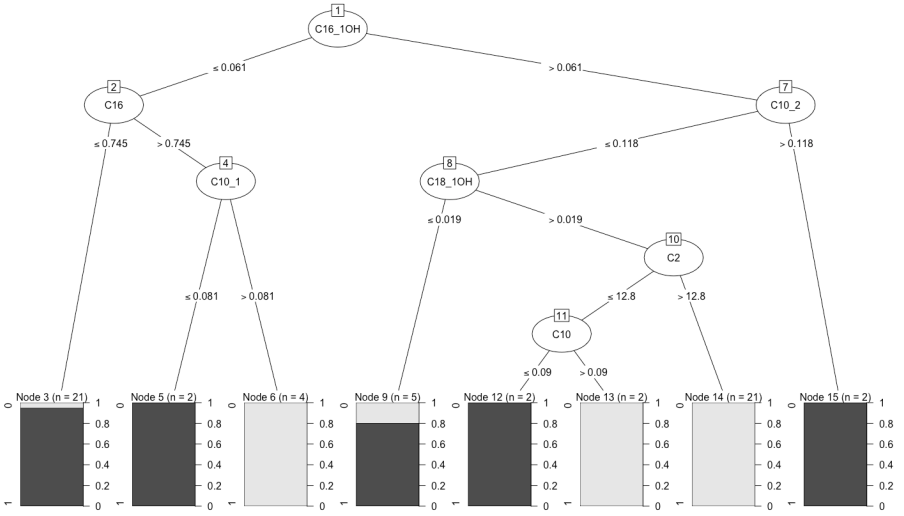

Figura 4.7: Modelo propuesto por el algoritmo C5.0

Para la comparación de ambos modelos se utilizó el análisis de área bajo la curva, la figura 4.8, donde se demuestra que el modelo mostrado por el algoritmo C5.0 es superior.

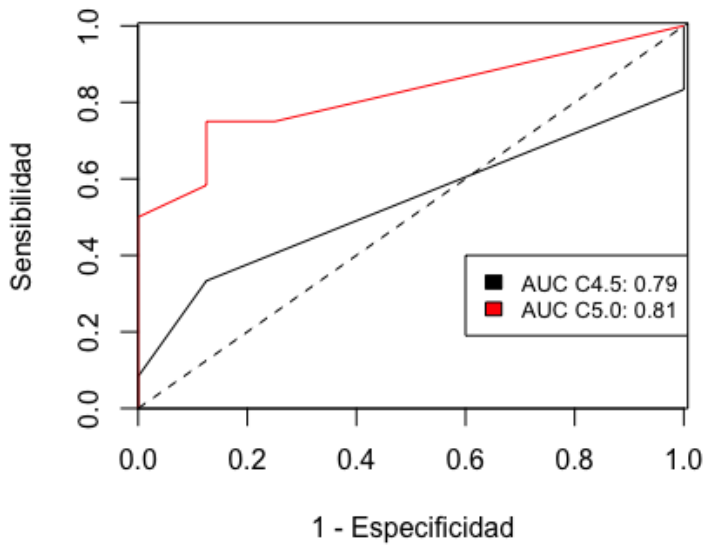

Figura 4.8: Comparación de los resultados de los algoritmos C4.5 y C5.0 por AUC

En la tabla 4.8 se muestra como el modelo propuesto por el algoritmo C4.5 se mantiene con adecuada sensibilidad, especificidad y respetable exactitud.

Validación cruzada	
Exactitud	91.2 % (IC 88.4 %, 94 %)
Sensibilidad	91.6 % (IC 88.6 %, 94.6 %)
Especificidad	90 % (IC 86.5 %, 93.4 %)
V. Predictivo Positivo	94.1 % (IC 92.1 %, 96.2 %)
V. Predictivo Negativo	90.8 % (IC 86.9 %, 94.7 %)

Tabla 4.8: Desempeño del modelo propuesto por el Algoritmo C4.5

En la tabla 4.9 , correspondiente al desempeño del modelo dado por el algoritmo c5.0 se observa como pierde algo de exactitud, sin embargo, se mantiene competitivo en cuanto a sensibilidad no así en el valor predictivo positivo y negativo. Se mantiene con adecuado valor predictivo positivo (VPP) y valor predictivo negativo (VPN).

Validación cruzada	
Exactitud	90.2 % (IC 83.1 %, 97.4 %)
Sensibilidad	90.7 % (IC 80.9 %, 100 %)
Especificidad	92.7 % (IC 87 %, 98.4 %)
V. Predictivo Positivo	91.2 % (IC 84.2 %, 98.3 %)
V. Predictivo Negativo	91.1 % (IC 80.5 %, 100 %)

Tabla 4.9: Desempeño del modelo propuesto por el Algoritmo C5.0

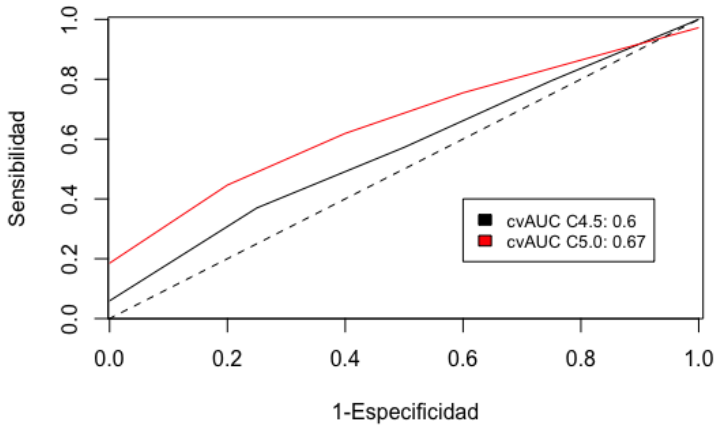

Figura 4.9: Validación cruzada de AUC de algoritmo C5.0

4.3.2. Selección de variables previamente realizado mediante regresión logística

En las tablas 4.10 y 4.11, siguientes se puede observar que utilizar una reducción de dimensiones mediante la regresión logística con pasos adelante (FS/LR) empeora el desempeño de ambos algoritmos, el algoritmo con mejor desempeño global es el C4.5. En la figura 4.10 se muestra el AUC para ambos modelos.

Vallidación cruzada	
Exactitud	81 % (IC 69.2 %, 94.6 %)
Sensibilidad	72.5 % (IC 56.3 %, 88.6 %)
Especificidad	90 % (IC 79 %, 100 %)
V. Predictivo Positivo	91.6 % (IC 82.2 %, 100 %)
V. Predictivo Negativo	78.9 % (63.9 %, 93.8 %)

Tabla 4.10: Desempeño del modelo C4.5 con FS/LR

Validación cruzada	
Exactitud	69.4 % (IC 59 %, 79.8 %)
Sensibilidad	72.2 % (IC 60 %, 84.3 %)
Especificidad	69.8 % (IC 52.8 %, 86.7 %)
V. Predictivo Positivo	77.3 % (IC 64.2 %, 90.4 %)
V. Predictivo Negativo	67.4 % (50.5 %, 84.2 %)

Tabla 4.11: Desempeño del modelo C5.0 con FS/LR

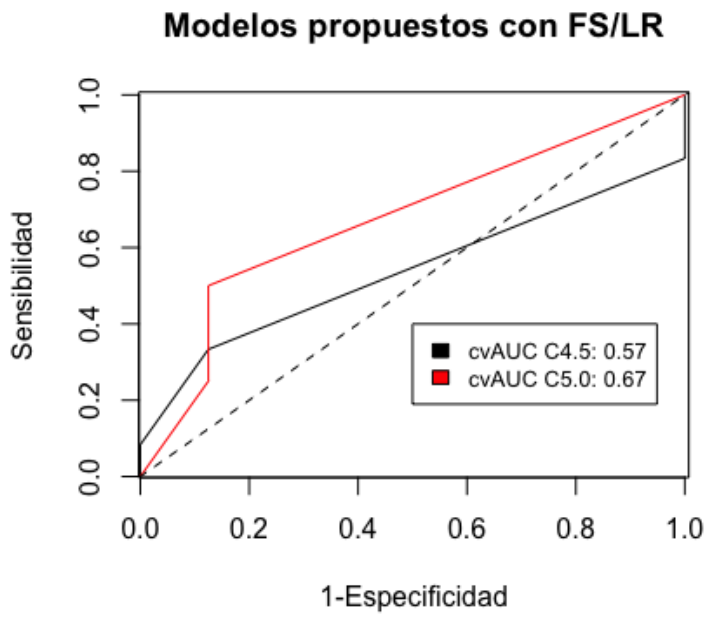

Figura 4.10: AUC de modelos dados con reducción de dimensiones con FS/LR

4.3.3. Evaluación del desempeño mediante bootstrap

En esta sección se muestra el desempeño de los algoritmos mediante un remuestreo bootstrap de los datos de entrenamiento. Este remuestreo se realizó 1000 veces y se obtuvo la media y el intervalo de confianza para cada variable.

En la figura 4.11 se observa claramente que el algoritmo con mayor área bajo la curva fue el que reunió el C5.0 FS/LR, además presento una de las mayores especificidades, aunque todos los algoritmos tienden a tener resultados muy similares.

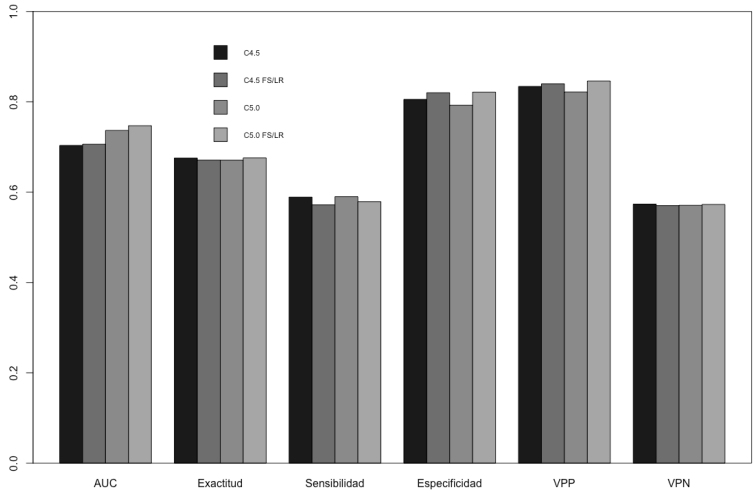

Figura 4.11: Desempeño de los algoritmos mediante bootstrap

4.3.4. Árbol de decisión sensible a costos

Se puede observar como en la tabla 4.12 se puede obtener una mejor exactitud al darle un peso distinto al 0.5. El mejor árbol fue al que se le dio 0.67 con Estimado DM / Actual = Normal), con una exactitud del 84.6 %. . En ella se muestran los diferentes costos para

el error de clasificación y los subsecuentes cambios en la exactitud del modelo.

Costo (Estimado = DM/ Actual = Normal)	Costo (Estimado = Normal/ Actual = DM)	Exactitud
0.25	0.75	81.3
0.75	0.25	84.8
0.33	0.67	83.5
0.67	0.33	84
0.5	0.5	81

Tabla 4.12: Resultados de clasificación de DM con diferentes costos

4.4. Look Ahead Decision Tree.

En la tabla 4.13, se puede observar que este algoritmo no es eficiente para clasificar a los pacientes con diabetes, ya que se mantiene con una exactitud muy cercana o inferior al 50 %. Esto con diferentes Deep y parametros B para su ajuste.

	Deep 1	Deep 2	Deep 3
B = 0.01	45.5	53.0	53.2
B = 0.03	45.5	53.2	45.7
B= 0.06	53.3	52.8	38.9
B = 0.09	53.2	53.0	45.8
B = 0.1	43.7	53.3	53.2

Tabla 4.13: Resultados para clasificar pacientes con Diabetes Mellitus con diferente beta (B) y diferente profundidad (Deep) de vista hacia adelante.

4.4.1. Meta características

Esta dificultad en la clasificación de diabetes mellitus, se demuestra de nueva cuenta en la tabla 4.14, donde se muestra que los algoritmos C4.5 y C5.0 mejoran ligeramente la exactitud de clasificación, sin embargo no alcanzan el 80 % de exactitud. La mejor exactitud la tuvo el algoritmo C5.0 con los datos originados de la Multiplicación de variables sin los datos crudos.

	C4.5	C5.0
Suma	56.9	72.2
Suma +	56.9	73.5
Resta	60.7	73.5
Resta +	62.0	68.8
Multiplicación	63.2	74.7
Multiplicación +	64.5	71

Tabla 4.14: Exactitud (%) de los algoritmos con metacaracteristicas en Diabetes mellitus

En este caso de clasificar pacientes con diabetes mellitus agregando las meta características, mediante el algoritmo costo sensible, se puede ver en la tabla 4.15 que no se mejora la exactitud, siendo muy parecido en exactitud el que da un costo (25 - 75) y las meta características originadas por multiplicación (las originadas con la resta también tienen una exactitud del 84.1 %).

Costos	0.25 - 0.75	0.75 - 0.25	0.33 - 0.67	0.67 - 0.33	0.5 - 0.5
Suma	77.2	86	77.6	81	75.9
Suma +	83.2	84.8	76.7	80.5	82.9
Resta	78.4	80.6	82.6	81	84.1
Resta +	78.4	84.4	79.3	79.3	81.6
Multiplicación	84.1	76.5	81.4	78.8	83.5
Multiplicación +	79.1	82.2	80.6	83.5	77.8

Tabla 4.15: Exactitud (%) de los algoritmos con metacaracteristicas con CSDT en Diabetes Mellitus

El algoritmo look ahead no mejora la clasificación aún con las meta características (tabla 4.16).

	Deep 1	Deep 2	Deep 3
Suma	45.8	45.5	52.8
Suma +	53.2	52.8	45.5
Resta	65.7	64.4	66.7
Resta +	63	63.5	63.2
Multiplicación	45.5	53.2	53
Multiplicación +	53.3	53.2	53.3

Tabla 4.16: Exactitud (%) de los algoritmos con metacaracteristicas con LADT previa selección de variables en Diabetes Mellitus (todos fueron con B = 0.1)

4.5. Retinopatía

En esta sección se comentan los resultados de la aplicación de los diferentes algoritmos para la clasificación de pacientes con retinopatía diabética.

4.5.1. Algoritmo C4.5 y C5.0

En las figuras 4.12 y 4.13 , se muestran los modelos obtenidos por los algoritmos C4.5 y C5.0 en la predicción de retinopatía diabética.

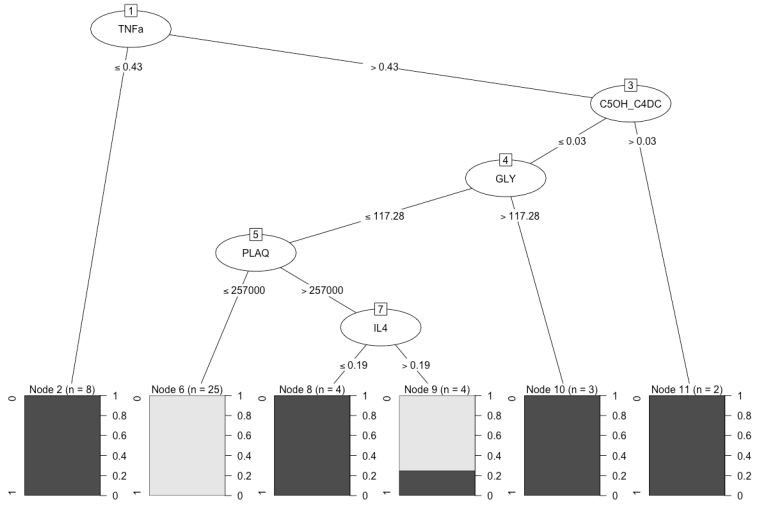

Figura 4.12: Modelo obtenido por C4.5 para la clasificación de retinopatía diabética

En este caso, se muestra (que como es de suponer por cuestiones biológicas) las variables obtenidas están relacionadas con un proceso inflamatorio (TNFa, IL4, plaquetas) y un proceso de alteración metabólica como es GLY (aminoácido, Glicina). En la tabla 4.17 se observa como este modelo se encuentra muy balanceado en su desempeño.

Validación cruzada	
Exactitud	91 % (IC 79.3 %, 100 %)
Sensibilidad	92.5 % (IC(77.8 %, 100 %)
Especificidad	96.8 % (IC 90.7 %, 100 %)
V. Predictivo Positivo	95.8 % (IC 87.6 %, 100 %)
V. Predictivo Negativo	90.6 % (IC 72.2 %, 100 %

Tabla 4.17: Desempeño de modelo C4.5

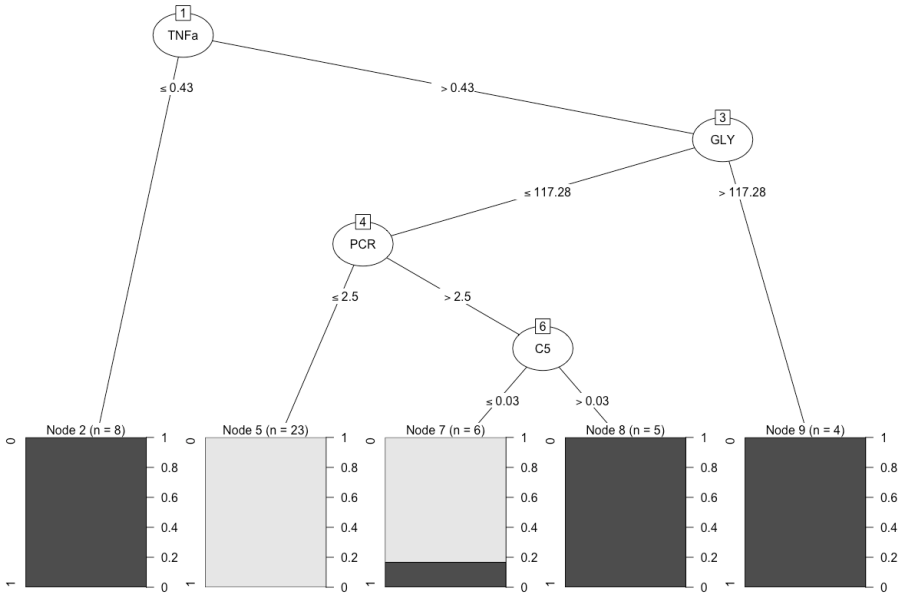

Figura 4.13: Modelo obtenido por C5.0 para la clasificación de retinopatía diabética

El modelo C5.0 también incluye una acilcarnitina de cadena corta (relacionada a problemas metabólicos) pero no carboxilada. Este cambio puede ser debido al algoritmo que trata de producir árboles de menor tamaño que su antecesor. Así también este algoritmo se mantiene con excelente sensibilidad, pero incrementa la exactitud y especificidad, tabla 4.18

Validación cruzada	
Exactitud	97.9 %(IC 93.8 %, 100 %)
Sensibilidad	93.7 % (IC 81.5 %, 100 %)
Especificidad	100 %
V. Predictivo Positivo	100 %
V. Predictivo Negativo	97.5 % (IC 92.6 %, 100 %)

Tabla 4.18: Desempeño de modelo obtenido del algoritmo C5.0

4.5.2. Selección de variables previamente realizado mediante regresión logística

En las tablas 4.19, se puede observar que la selección de variables mediante regresión logística no mejora la exactitud del modelo.

Validación cruzada	
Exactitud	68.7 % (IC 64.1 %, 73.3 %)
Sensibilidad	68.7 % (IC 57.5 %, 79.9 %)
Especificidad	72.5 % (IC 67.1 %, 77.8 %)
V. Predictivo Positivo	44.7 % (IC 35.2 %, 54.3 %)
V. Predictivo Negativo	87.5 % (IC 83 %, 91.9 %)

Tabla 4.19: Exactitud de modelos C4.5 y C5.0, mediante regresión logística con paso hacia adelante.

4.5.3. Árbol de decisión costo sensible

En la tabla 4.20 se muestra como el algoritmo costo sensible puede mejorar la exactitud de la clasificación mediante cualquier costo aplicado a la mala clasificación de los sujetos. Llegando a tener una exactitud del 88.7 % en el modelo con el costo 0.25 - 0.75.

Costo (Estimado = Retinopatía/ Actual = Normal)	Costo (Estimado = Normal/ Actual = Retinopatía)	Exactitud
0.25	0.75	88.7
0.75	0.25	87.9
0.33	0.67	89
0.67	0.33	84.5
0.5	0.5	84.4

Tabla 4.20: Resultados de clasificación de Retinopatía con diferentes costos

4.6. Look Ahead Decision Tree.

El árbol de decisión obtenido por el algoritmo de Look Ahead, con diferentes Deep, puede incrementar ligeramente la exactitud, sobre todo con Deep 2 y parámetro Beta = 0.01, (tabla 4.21). Sin embargo no llega a ser tan competitivo como el árbol obtenido con el modelo sensible a costo.

	Deep 1	Deep 2	Deep 3
B = 0.01	62.6	76.3	68.0
B = 0.03	69.6	66.3	76.6
B= 0.06	71.0	76.0	68.6
B = 0.09	67.6	72.3	68.6
B = 0.1	63.0	71.0	70.6

Tabla 4.21: Resultados para clasificar pacientes con Retinopatía con diferente beta (B) y diferente profundidad (Deep) de vista hacia adelante.

4.6.1. Meta características

En esta sección se expondrán los resultados de la aplicación de los algoritmos en la base de datos creada mediante meta características. En la tabla 4.22 se puede observar que las meta características mejoran ligeramente la exactitud de la clasificación. Sobre todo mediante el algoritmo C5.0 y con los datos obtenidos mediante multiplicación con o sin los datos crudos.

	C4.5	C5.0
Suma	58.6	73.6
Suma +	58.6	73.3
Resta	65.5	77.6
Resta +	65.5	76
Multiplicación	72.4	82.3
Multiplicación +	72.4	82.3

Tabla 4.22: Exactitud (%) de los algoritmos con meta caracteristicas en Retinopatia

En la tabla 4.23, se muestra como las meta características y el algoritmo costo sensible pueden incrementar la exactitud de la clasificación de manera importante llegando a ser superior al 90 % como en los datos originados de la resta (90.9 %) y de la multiplicación.

Costos	0.25 - 0.75	0.75 - 0.25	0.33 - 0.67	0.67 - 0.33	0.5 - 0.5
Suma	87.5	82.7	87.3	83.3	77.5
Suma +	81.8	83.1	79.9	82.7	83.6
Resta	83.6	90.9	85	86.7	85.3
Resta +	87.5	89.2	87.9	86.2	86.2
Multiplicación	87.9	90	85	85.6	82.7
Multiplicación +	84	88.7	83.3	85.6	81.8

Tabla 4.23: Exactitud (%) de los algoritmos con meta características con CSDT en Retinopatía

El algoritmo de Look Ahead, incremento de manera ligera la exactitud del modelo. El árbol con Depp 1 originado de los datos de Resta sin los datos crudos fue el que tuvo la mejor exactitud con 74.3 %. (Tabla 4.24).

	Deep 1	Deep 2	Deep 3
Suma	67.3	67.3	66.6
Suma +	66.6	66.6	67
Resta	74.3	71.6	72.6
Resta +	70.3	69.3	62.6
Multiplicación	67	67.3	67
Multiplicación +	67.3	67.6	67

Tabla 4.24: Exactitud (%) de los algoritmos con meta características con LADT previa selección de variables en Retinopatía (todos fueron con B = 0.1)

4.7. Software utilizado

En esta tesis se utilizó el paquete estadístico R v 3.4.2 , Matlab 2016 y Weka v 3.8.2 . Las librerías utilizadas en R son: pROC, DescTools, RWeka, caret, C50, corrplot, gbm, coin, psych, ROCR, OptimalCutpoints, RcmdrMisc, cvAUC.

Capítulo 5

Contribuciones, conclusión y trabajo futuro

5.1. Contribuciones al análisis de datos

El análisis de los datos utilizados en este trabajo, se caracterizan por ser de alta dimencionalidad y de gran dificultad de análisis. Los algoritmos base (por ejemplo C4.5 y C5.0) con los que se dió solución al problema de clasificación pueden ser mejorados con la previa selección de variables con una regresión logística o con otras técnicas como el dar un paso hacia adelante en su algoritmo (Look Ahead) o con la propiedad de ser sensibles a costo. Las contribuciones se resumen puntualmente de la forma siguiente:

5.1.1. Generalized Cost Sensitive Look Ahead Decision Tree (GCSLADT) supera otras variantes de árbol de decisión y criterio de síndrome metabólico para NAFLD

.

Dependiendo del problema a resolver se pueden utilizar las capacidades de este algoritmo para resolverlas, una de ellas la sensibilidad a costos cuando los datos son desbalanceados, en nuestro trabajo se presentaron dos casos así: el conjunto de pacientes con hígado graso no alcohólico y el sub grupo anidado de pacientes con retinopatía. De estos los datos con mayor desbalance fue la de NAFLD y el algoritmo presentó una exactitud que rondaba el 92 %.

En el caso del uso del paso hacia adelante (Look Ahead), se demuestra que se debe tener en consideración que la cantidad de pasos hacia adelante. Ya que el dar muchos pasos hacia adelante pudieran alterar la clasificación final del algoritmo (efecto del horizonte). También que este efecto pudiera ser alterado por el criterio de parada (beta) en la partición del árbol.

5.1.2. Active, Online, reinforcement e incremental learning pueden ser incorporados en la estructura del algoritmo GCSLADT.

Algoritmos como Active, Online, Reinforcement e Incremental Learning pueden ser incorporados al algoritmo base de GCSLADT. Con la intención de incrementar el desempeño del algoritmo inicial. Se hace hincapié a considerar el tipo de problema albergado en los datos así como el tamaño de los mismos para la decisión de usar tal o cual aprendizaje.

5.1.3. Incorporación de meta características ayuda a dar mayor exactitud de árbol de decisión para clasificar complicaciones del síndrome metabólico.

Las meta características que se utilizaron en este trabajo incluyeron las operaciones algebraicas de de suma, resta y multiplicación. Mostraron la capacidad de incrementar la exactitud en la predicción de las complicaciones del síndrome metabólico. Esta exactitud es incrementada todavía mas al agregar un "peso" a las desiciones equivocadas (Sensible a costo) del algoritmo, sobre todo para datos desbalanceados (ej. Hígado graso no alcohólico).

5.1.4. Las técnicas de elegir características (NCA,FA, Forward selection) reduce la complejidad computacional del árbol de decisión por disminuir la dimensión sin reducir la exactitud.

La selección de variables es un paso muy importante con la alta dimensionalidad de los datos. Además algoritmos como look ahead que mejoran la exactitud de los algoritmos tradicionales, tienen un limite de dimensiones (100 variables en caso de Look Ahead) haciendo muy necesario este paso.

Por otro lado, el escoger las variables mas importantes sin perdida de desempeño del algoritmo disminuye el costo computacional de la clasificación.

5.2. Contribuciones a la literatura médica

En la literatura medica, existen multiples intentos (los cuales fueron revisados en esta tesis) de clasificar adecuadamente pacientes sanos y con hígado graso no alcohólico; además en este mismo trabajo se realizó un árbol de decisión hecho por el experto, con los criterios de clasificación de síndrome metabólico, demostrando la necesidad de un algoritmo inteligente. En el uso de algoritmos como C4.5 y C5.0 se pudo demostrar que se puede mejorar la clasificación de pacientes con hígado graso no alcohólico mediante la combinación de variables químicas, clínicas y metabolómicas.

Se demostró como las alteraciones en las acilcarnitinas (principalmente de cadena larga, como C16_OH), están relacionadas con la presencia de hígado graso. El incremento en los niveles de las acilcarnitinas es asociado a esta enfermedad hepática y resistencia a la insulina. Esto ultimo dado a que la carnitina C16_1OH (y también otras como la C14_1, C18_1OH) son unas de las principales variables que pueden clasificar a los pacientes con diabetes mellitus tipo 2.

En este mismo sentido, en este trabajo se documenta por primera vez la utilidad de las acilcarnitinas (como únicas variables) para clasificar al paciente con diabetes. Por otro lado, también se logra confirmar la relación entre el trastorno metabólico expresado en estas variables metabolómicas y la presencia de retinopatía. Aunque pareciera aventurado confirmar en este momento, pero parece ser que hay bases suficientes para sostener la teoría de la retinopatía del paciente obeso.

Por otro lado, el tipo de análisis llevado a cabo originar nuevas hipótesis. El poder fundamentar que la clasificación de retinopatía por acilcarnitinas es posible, nos sugiere la realización de estudios diseñados para evaluar causalidad.

5.3. Conclusiones

Las conclusiones de este trabajo son resumidas a continuación:

- Desarrollamos Árbol de Decisión Sensible a Costos y con Mirada Adelante Generalizado (Generalized Cost Sensitive Look Ahead Decision Tree , GCSLADT) para automatizar el diagnostico de las complicaciones del síndrome metabólico.
- Demostramos el poder de meta características para clasificar las complicaciones del síndrome metabólico.

- Utilizamos decision tree para clasificar las complicaciones del síndrome metabólico por su interpretabilidad por los médicos.
- Comparamos el rendimiento de árbol de decisión con diferentes scores diagnósticos publicados en la literatura.
- El perfil bioquímico es suficiente para clasificar NAFLD con árbol de decisión automáticamente.

5.4. Trabajo futuro

Actualmente existe mucho trabajo futuro por realizar, de entre ellos se remarcan los siguientes: El algoritmo de árboles de decisión utiliza el criterio de reducción de entropía para la partición de los datos.

- Generalized Cost Sensitive Look Ahead Decision Tree (GCSLADT) funciona mejor para clasificar NAFLD, presencia de diabetes y Retinopatía, se probara en otras enfermedades.
- Queremos investigar el agregar diferentes tipos de aprendizaje (Any time, High speed, ensamble) a la estructura de GCSLADT.

Capítulo 6

Agradecimientos

Debo agradecer primeramente a Dios por darme la vida, la capacidad de entendimiento y la fortaleza para sobrellevar los problemas surgidos durante mis años de estudio.

A mi esposa Maria Elena Romero, por su incansable apoyo y comprensión.

A mis profesores Dra. Graciela González Farías, Dr. Rodrigo Macías Páez, Dr. Baidya Nath Saha, Dr. Victor Muñiz, Mtro. José Ramón Domiguez, Dr. José Jaime Hernández, Dr. Ulises Márquez. por haberme cambiado mi vida al compartir sus conocimientos.

A mis compañeros y amigos: Ariana, Edison, Nancy y Rubén por su amistad y ayuda.

A Mtro. Pavel Hipólito por su sabio consejo en el primer semestre de maestría

Finalmente agradesco a CIMAT A.C. y CONACYT por el apoyo académico y económico recibido durante los años de mi estudio.

Bibliografía

[1] Sean H Adams, Charles L Hoppel, Kerry H Lok, Ling Zhao, Scott W Wong, Paul E Minkler, Daniel H Hwang, John W Newman, y W Timothy Garvey. Plasma acylcarnitine profiles suggest incomplete long-chain fatty acid beta-oxidation and altered tricarboxylic acid cycle activity in type 2 diabetic african-american women. *J Nutr*, 139(6):1073–81, 2009. doi:10.3945/jn.108.103754.

[2] Kurt George Matthew Mayer Alberti y PZ ft Zimmet. Definition, diagnosis and classification of diabetes mellitus and its complications. part 1: diagnosis and classification of diabetes mellitus. provisional report of a who consultation. *Diabetic medicine*, 15(7):539–553, 1998.

[3] American Diabetes Association. 2. classification and diagnosis of diabetes: Standards of medical care in diabetes-2018. *Diabetes Care*, 41(Suppl 1):S13–S27, 2018. doi: 10.2337/dc18-S002.

[4] György Baffy, Elizabeth M Brunt, y Stephen H Caldwell. Hepatocellular carcinoma in non-alcoholic fatty liver disease: an emerging menace. *J Hepatol*, 56(6):1384–91, 2012. doi:10.1016/j.jhep.2011.10.027.

[5] Alejandro Correa Bahnsen, Djamila Aouada, y Bjorn Ottersten. Ensemble of example-dependent cost-sensitive decision trees. *arXiv preprint arXiv:1505.04637*, 2015.

[6] Alejandro Correa Bahnsen, Djamila Aouada, y Björn Ottersten. A novel cost-sensitive framework for customer churn predictive modeling. *Decision Analytics*, 2(1):5, 2015.

[7] Rodrigo C Barros, André CPLF de Carvalho, y Alex A Freitas. *Automatic design of decision-tree induction algorithms*. Springer, 2015.

[8] Metin Basaranoglu, Gokcen Basaranoglu, y Elisabetta Bugianesi. Carbohydrate intake and nonalcoholic fatty liver disease: fructose as a weapon of mass destruction. *Hepatobiliary surgery and nutrition*, 4(2):109, 2015.

[9] Hiram Beltrán-Sánchez, Michael O Harhay, Meera M Harhay, y Sean McElligott. Prevalence and trends of metabolic syndrome in the adult us population, 1999–2010. *Journal of the American College of Cardiology*, 62(8):697–703, 2013.

[10] L Breiman, JH Friedman, y RA Olshen. stone, cj (1984). *Classification and regression trees*, 85.

[11] Leonard A Breslow y David W Aha. Simplifying decision trees: A survey. *The Knowledge Engineering Review*, 12(1):1–40, 1997.

[12] Mr Brijain, R Patel, Mr Kushik, y K Rana. A survey on decision tree algorithm for classification. *International Journal of Engineering Development and Research*, 2014.

[13] Phunchai Charatcharoenwitthaya y Keith D Lindor. Role of radiologic modalities in the management of non-alcoholic steatohepatitis. *Clinics in liver disease*, 11(1):37–54, 2007.

[14] Michael R Chernick y Robert A LaBudde. *An introduction to bootstrap methods with applications to R*. John Wiley & Sons, 2014.

[15] Alessandro de Moura Almeida, Helma Pinchemel Cotrim, Daniel Batista Valente Barbosa, Luciana Gordilho Matteoni de Athayde, Adimeia Souza Santos, Almir Galvão Vieira Bitencourt, Luiz Antonio Rodrigues de Freitas, Adriano Rios, y Erivaldo Alves. Fatty liver disease in severe obese patients: diagnostic value of abdominal ultrasound. *World journal of gastroenterology: WJG*, 14(9):1415, 2008.

[16] Martin J Dumskyj, Jesper E Eriksen, Caroline J Doré, y Eva M Kohner. Autoregulation in the human retinal circulation: assessment using isometric exercise, laser doppler velocimetry, and computer-assisted image analysis. *Microvascular research*, 51(3):378–392, 1996.

[17] Tapio Elomaa y Tuomo Malinen. On lookahead heuristics in decision tree learning. En *International Symposium on Methodologies for Intelligent Systems*, págs. 445–453. Springer, 2003.

[18] Tapio Elomaa* y Tuomo Malinen. On look-ahead and pathology in decision tree learning. *Journal of Experimental & Theoretical Artificial Intelligence*, 17(1-2):19–33, 2005.

[19] W Gao y DECODE Study Group. Does the constellation of risk factors with and without abdominal adiposity associate with different cardiovascular mortality risk? *Int J Obes (Lond)*, 32(5):757–62, 2008. doi:10.1038/sj.ijo.0803797.

[20] David G Gardner, Dolores Shoback, y Francis S Greenspan. *Greenspan's basic & clinical endocrinology*. McGraw-Hill Medical,, 2007.

[21] Vahid Golmah. An efficient hybrid intrusion detection system based on c5. 0 and svm. *International Journal of Database Theory and Application*, 7(2):59–70, 2014.

[22] M E González Villalpando, C González Villalpando, B Arredondo Pérez, y M P Stern. Diabetic retinopathy in mexico. prevalence and clinical characteristics. *Arch Med Res*, 25(3):355–60, 1994.

[23] H Haller. Epidermiology and associated risk factors of hyperlipoproteinemia. *Zeitschrift fur die gesamte innere Medizin und ihre Grenzgebiete*, 32(8):124–128, 1977.

[24] M Hernández-Ávila, J Rivera-Dommarco, T Shamah-Levy, L Cuevas-Nasu, LM Gómez-Acosta, EM Gaona-Pineda, et al. Encuesta nacional de salud y nutrición de medio camino 2016. *Cuernavaca, Morelos, México: Instituto Nacional de Salud Pública*, 2016.

[25] Daniel Joseph Hsu. *Algorithms for active learning*. Tesis Doctoral, UC San Diego, 2010.

[26] Kento Imajo, Takaomi Kessoku, Yasushi Honda, Wataru Tomeno, Yuji Ogawa, Hironori Mawatari, Koji Fujita, Masato Yoneda, Masataka Taguri, Hideyuki Hyogo, et al. Magnetic resonance imaging more accurately classifies steatosis and fibrosis in patients with nonalcoholic fatty liver disease than transient elastography. *Gastroenterology*, 150(3):626–637, 2016.

[27] Abhaya Indrayan y Rajeev Kumar Malhotra. *Medical biostatistics*. Chapman and Hall/CRC, 2017.

[28] Guoyu Jia, Fusheng Di, Qipeng Wang, Jinshuang Shao, Lei Gao, Lu Wang, Qiang Li, y Nali Li. Non-alcoholic fatty liver disease is a risk factor for the development of diabetic nephropathy in patients with type 2 diabetes mellitus. *PloS one*, 10(11):e0142808, 2015.

[29] Gordon V Kass. An exploratory technique for investigating large quantities of categorical data. *Applied statistics*, págs. 119–127, 1980.

[30] Hyunjoong Kim y Wei-Yin Loh. Classification trees with unbiased multiway splits. *Journal of the American Statistical Association*, 96(454):589–604, 2001.

[31] Hyunjoong Kim y Wei-Yin Loh. Classification trees with bivariate linear discriminant node models. *Journal of Computational and Graphical Statistics*, 12(3):512–530, 2003.

[32] April D Lake, Petr Novak, Petia Shipkova, Nelly Aranibar, Donald G Robertson, Michael D Reily, Lois D Lehman-McKeeman, Richard R Vaillancourt, y Nathan J Cherrington. Branched chain amino acid metabolism profiles in progressive human nonalcoholic fatty liver disease. *Amino acids*, 47(3):603–615, 2015.

[33] Alan M Laties. Central retinal artery innervation: absence of adrenergic innervation to the intraocular branches. *Archives of Ophthalmology*, 77(3):405–409, 1967.

[34] Wei-Yin Loh. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):14–23, 2011.

[35] Wei-Yin Loh y Yu-Shan Shih. Split selection methods for classification trees. *Statistica sinica*, págs. 815–840, 1997.

[36] A Lonardo, C D Byrne, S H Caldwell, H Cortez-Pinto, y G Targher. Global epidemiology of nonalcoholic fatty liver disease: Meta-analytic assessment of prevalence, incidence, and outcomes. *Hepatology*, 64(4):1388–9, 2016. doi:10.1002/hep.28584.

[37] Giulio Marchesini y Rebecca Marzocchi. Metabolic syndrome and nash. *Clinics in liver disease*, 11(1):105–117, 2007.

[38] Colin D Mathers y Dejan Loncar. Projections of global mortality and burden of disease from 2002 to 2030. *PLoS medicine*, 3(11):e442, 2006.

[39] Masafumi Matsuda y Ralph A DeFronzo. Insulin sensitivity indices obtained from oral glucose tolerance testing: comparison with the euglycemic insulin clamp. *Diabetes care*, 22(9):1462–1470, 1999.

[40] Mohammed M Mazid, ABM Shawkat Ali, Kevin S Tickle, et al. Improved c4. 5 algorithm for rule based classification. En *Proceedings of the 9th WSEAS international conference on Artificial intelligence, knowledge engineering and data bases*, págs. 296–301. World Scientific and Engineering Academy and Society (WSEAS), 2010.

[41] Shlomo Melmed. *Williams textbook of endocrinology*. Elsevier Health Sciences, 2016.

[42] Robert Messenger y Lewis Mandell. A modal search technique for predictive nominal scale multivariate analysis. *Journal of the American statistical association*, 67(340):768–772, 1972.

[43] Ivana Mikolasevic, Tajana Filipec-Kanizaj, Maja Mijic, Ivan Jakopcic, Sandra Milic, Irena Hrstic, Nikola Sobocan, Davor Stimac, y Patrizia Burra. Nonalcoholic fatty liver disease and liver transplantation-where do we stand? *World journal of gastroenterology*, 24(14):1491, 2018.

[44] Matthias Möhlig, Frank Isken, y Michael Ristow. Impaired mitochondrial activity and insulin-resistant offspring of patients with type 2 diabetes. *N Engl J Med*, 350(23):2419–21; author reply 2419–21, 2004.

[45] Luca Montesi, Chiara Caselli, Elena Centis, Chiara Nuccitelli, Simona Moscatiello, Alessandro Suppini, y Giulio Marchesini. Physical activity support or weight loss counseling for nonalcoholic fatty liver disease? *World Journal of Gastroenterology: WJG*, 20(29):10128, 2014.

[46] James N Morgan y John A Sonquist. Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association*, 58(302):415–434, 1963.

[47] Sreerama Murthy y Steven Salzberg. Lookahead and pathology in decision tree induction. En *IJCAI*, págs. 1025–1033. Citeseer, 1995.

[48] Taiji Nagaoka, Takashi Sakamoto, Fumihiko Mori, Eiichi Sato, y Akitoshi Yoshida. The effect of nitric oxide on retinal blood flow during hypoxia in cats. *Investigative ophthalmology & visual science*, 43(9):3037–3044, 2002.

[49] María Araceli Ortiz-Rodríguez, Lucía Yáñez-Velasco, Alessandra Carnevale, Sandra Romero-Hidalgo, Demetrio Bernal, Carlos Aguilar-Salinas, Rosalba Rojas, Antonio Villa, y Josep A Tur. Prevalence of metabolic syndrome among elderly mexicans. *Arch Gerontol Geriatr*, 73:288–293, 2017. doi:10.1016/j.archger.2017.09.001.

[50] Jen-Jung Pan y Michael B Fallon. Gender and racial differences in nonalcoholic fatty liver disease. *World J Hepatol*, 6(5):274–83, 2014. doi:10.4254/wjh.v6.i5.274.

[51] Su-lin Pang y Ji-zhang Gong. C5. 0 classification algorithm and application on individual credit evaluation of banks. *Systems Engineering-Theory & Practice*, 29(12):94–104, 2009.

[52] Víctor Huggo Córdova Pluma, Alejandra Correa Morales, José Luis Artigas Arroyo, María del Carmen de la Torre, Miriam Vargas González, et al. Hígado graso no alcohólico: su diagnóstico en la actualidad. tercera parte. *Medicina Interna de México*, 25(3):217–228, 2009.

[53] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[54] J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

[55] John R Quinlan et al. Learning with continuous classes. En *5th Australian joint conference on artificial intelligence*, tomo 92, págs. 343–348. Singapore, 1992.

[56] JR Quinlan. C4. 5: Programs for machine learning. morgan kaufmann, san francisco. *C4. 5: Programs for machine learning. Morgan Kaufmann, San Francisco.*, 1993.

[57] Mary E Rinella, Zurabi Lominadze, Rohit Loomba, Michael Charlton, Brent A Neuschwander-Tetri, Stephen H Caldwell, Kris Kowdley, y Stephen A Harrison. Practice patterns in nafld and nash: real life differs from published guidelines. *Therapeutic advances in gastroenterology*, 9(1):4–12, 2016.

[58] Lior Rokach y Oded Maimon. *Data mining with decision trees: theory and applications*. World scientific, 2014.

[59] Maria E Romero-Ibarguengoitia, Arturo Herrera-Rosas, Alfredo A Domínguez-Mota, Jinny T Camas-Benitez, María F Serratos-Canales, Mireya León-Hernández, Antonio González-Chávez, Eduardo López-Ortiz, Srinivas Mummidi, Ranvidranth Duggirala, y

Juan C López-Alvarenga. Nonalcoholic fatty liver disease can be predicted by retinal vascular changes in patients with obesity without hypertension or diabetes. *Eur J Gastroenterol Hepatol*, 29(8):962–967, 2017. doi:10.1097/MEG.0000000000000900.

[60] Maria Elena Romero-Ibarguengoitia, Felipe Vadillo-Ortega, Augusto Enrique Caballero, Isabel Ibarra-González, Arturo Herrera-Rosas, María Fabiola Serratos-Canales, Mireya León-Hernández, Antonio González-Chávez, Srinivas Mummidi, Ravindranath Duggirala, et al. Family history and obesity in youth, their effect on acylcarnitine/aminoacids metabolomics and non-alcoholic fatty liver disease (nafld). structural equation modeling approach. *PloS one*, 13(2):e0193138, 2018.

[61] Edward Roufail, Michelle Stringer, y Sandra Rees. Nitric oxide synthase immunoreactivity and nadph diaphorase staining are co-localised in neurons closely associated with the vasculature in rat and human retina. *Brain research*, 684(1):36–46, 1995.

[62] Johan Rung, Stéphane Cauchi, Anders Albrechtsen, Lishuang Shen, Ghislain Rocheleau, Christine Cavalcanti-Proença, François Bacot, Beverley Balkau, Alexandre Belisle, Knut Borch-Johnsen, et al. Genetic variant near irs1 is associated with type 2 diabetes, insulin resistance and hyperinsulinemia. *Nature genetics*, 41(10):1110, 2009.

[63] Susan L Samson y Alan J Garber. Metabolic syndrome. *Endocrinology and Metabolism Clinics*, 43(1):1–23, 2014.

[64] A L Sberna, B Bouillet, A Rouland, M C Brindisi, A Nguyen, T Mouillot, L Duvillard, D Denimal, R Loffroy, B Vergès, P Hillon, y J M Petit. European association for the study of the liver (easl), european association for the study of diabetes (easd) and european association for the study of obesity (easo) clinical practice recommendations for the management of non-alcoholic fatty liver disease: evaluation of their application in people with type 2 diabetes. *Diabet Med*, 35(3):368–375, 2018. doi:10.1111/dme.13565.

[65] Anjali R Shah y Thomas W Gardner. Diabetic retinopathy: research to clinical practice. *Clin Diabetes Endocrinol*, 3:9, 2017. doi:10.1186/s40842-017-0047-y.

[66] Teresa Shamah-Levy, Lucía Cuevas-Nasu, Verónica Mundo-Rosas, Carmen Morales-Ruán, Leticia Cervantes-Turrubiates, y Salvador Villalpando-Hernández. Health and nutrition status of older adults in mexico: results of a national probabilistic survey. *Salud publica de Mexico*, 50(5):383–389, 2008.

[67] P Singer. Diagnosis of primary hyperlipoproteinemias. *Zeitschrift fur die gesamte innere Medizin und ihre Grenzgebiete*, 32(9):129–33, 1977.

[68] Jay S Skyler, George L Bakris, Ezio Bonifacio, Tamara Darsow, Robert H Eckel, Leif Groop, Per-Henrik Groop, Yehuda Handelsman, Richard A Insel, Chantal Mathieu, et al. Differentiation of diabetes by pathophysiology, natural history, and prognosis. *Diabetes*, 66(2):241–255, 2017.

[69] Jafar Tanha, Maarten van Someren, y Hamideh Afsarmanesh. Semi-supervised self-training for decision tree classifiers. *International Journal of Machine Learning and Cybernetics*, 8(1):355–370, 2017.

[70] Herbert Tilg, Alexander R Moschen, y Michael Roden. Nafld and diabetes mellitus. *Nature Reviews Gastroenterology and Hepatology*, 14(1):32, 2017.

[71] Carmine Vecchione, Angelo Maffei, Salvatore Colella, Alessandra Aretini, Roberta Poulet, Giacomo Frati, Maria Teresa Gentile, Luigi Fratta, Valentina Trimarco, Bruno Trimarco, et al. Leptin effect on endothelial nitric oxide is mediated through akt–endothelial nitric oxide synthase phosphorylation pathway. *Diabetes*, 51(1):168–173, 2002.

[72] Salvador Villalpando, Vanessa de la Cruz, Rosalba Rojas, Teresa Shamah-Levy, Marco Antonio Avila, Berenice Gaona, Rosario Rebollar, y Lucia Hernández. Prevalence and distribution of type 2 diabetes mellitus in mexican adult population: a probabilistic survey. *Salud Publica Mex*, 52 Suppl 1:S19–26, 2010.

[73] Quan Wang, Yan Ou, A Agung Julius, Kim L Boyer, y Min Jun Kim. Tracking tetrahymena pyriformis cells using decision trees. En *Pattern Recognition (ICPR), 2012 21st International Conference on*, págs. 1843–1847. IEEE, 2012.

[74] T J Wolfensberger y A M Hamilton. Diabetic retinopathy–an historical review. *Semin Ophthalmol*, 16(1):2–7, 2001.

[75] Jun won Lee y Christophe Giraud-Carrier. Transfer learning in decision trees. En *Neural Networks, 2007. IJCNN 2007. International Joint Conference on*, págs. 726–731. IEEE, 2007.

[76] XD Ye, AM Laties, y RA Stone. Peptidergic innervation of the retinal vasculature and optic nerve head. *Investigative ophthalmology & visual science*, 31(9):1731–1737, 1990.

Anexo

Antecedentes del Síndrome Metabólico

Definición

Se puede definir síndrome como una agrupación de hallazgos clínicos que pueden ocurrir conjuntamente mas que lo que se pudiera deber al azar [63]. A pesar de los multiples nombres dados en el pasado, el síndrome metabólico es ahora usado universalmente. La definición fue propuesta por un grupo de trabajo de la OMS que inició en 1998 y termino en 1999 [2].

El síndrome metabólico tiene múltples criterios de clasificación dependiendo de la sociedad o instancia internacional. Sin embargo, sobre sale los criterios de clasificación de Haller que incluyen obesidad, diabetes, hiperlipoproteinemia e hígado graso [23]; y el de Singer que incluye estos mismos padecimientos más hipertensión [67], hoy en día es común entre las diferentes definiciones la persistencia de esos trastornos integrados como la presencia de obesidad, adiposidad abdominal o indicadores de resistencia a la insulina, metabolismo de la glucosa alterado, hipertensión, y dislipidemia aterogénica. Más adelante se describirá con profundidad la epidemiología del síndrome metabólico.

Epidemiología

La prevalencia reportada del Síndrome metabólico varia dependiendo de la definición usada, edad, género y estado socioeconómico. Sin embargo, de estudios publicados en la última década, se estima que cerca de una cuarta parte a un tercio de los adultos pudieran cumplir con los criterios del síndrome.

En la encuesta de nutrición y salud de Estados Unidos (**N**ational **H**ealth and **N**utrition **E**xamination **S**urvey (NHANES) en 1999 a 2010, en adultos mayores de 20 años, la prevalencia ajustada por edad fue del 25.5 % de 1999 a 2000, disminuyendo hasta 22.9 % del

2009 al 2010 [9]. En europa el estudio **D**iabetes **E**pidemilogy: **C**ollaborative **A**nalysis of **D**iagnostic **C**riteria in **E**urope (DECODE) incluyó datos de 9 estudios poblacionales realizados en Finlandia, Holanda, Reino Unido, Suecia, Polonia e Italia, usando los valores de corte de la **F**ederación **I**nternacional de **D**iabetes (IDF) el 41 % de los hombres y 38 % de las mujeres cumplen los criterios a los 46 a 71 años [19].

La prevalencia general del síndrome metabólico en México se desconoce; sin embargo, una publicación reciente documento una prevalencia de 72.9 % del síndrome en mexicanos mayores de 65 años [49].

Complicaciones

En este anexo se presentaran brevemente lo que en este trabajo llamaremos complicaciones del hígado graso, que como se verá mas adelante todas estas complicaciones tienen un origen común: un trastorno metabólico asociado a alteraciones en el metabolismo de la glucosa, ácidos grasos y la subsecuente alteración inmunológica con una presencia de inflamación de bajo grado y predisposición a presentar alteraciones de otras vías metabólicas (ej. retinopatía).

Hígado graso no alcohólico

El Hígado Graso No Alcohólico (**N**on **A**lcoholic **F**atty **L**iver **D**isease, NAFLD), es definido como el incremento en los depositos de grasa en el hígado con fenotipos histologicos - clínicos que van desde una simple esteatosis (deposito de grasa presente en >5 % de los hepatocitos) a la esteatohepatitis no alcohólica (**N**on**A**lcoholic **S**teato**H**epatitis, NASH) [43]. NASH es una presentación mas agresiva de la enfermedad e incluye una presentación histologica de esteatosis, “balonamiento” de los hepatocitos e inflamación lobular que lleva a fibrosis avanzada y, finalmente, cirrosis y carcinoma hepatocelular [64]. El carcinoma hepatocelular es la sexta causa mas común de cancer en el mundo y es predispuesto con la presencia de cirrosis, pero datos emergentes sugieren que HCC puede desarrollarse del NAFLD no cirrótico que es fuertemente asociado al síndrome metabólico [4].

NAFLD se ha convertido en la enfermedad hepática mas común en el mundo, con una prevalencia estimada del 10 al 40 %. [36]. La prevalencia de NASH es aproximadamente 3 %, pero puede estar presente en mas del 25 % de los individuos con obesidad [50].

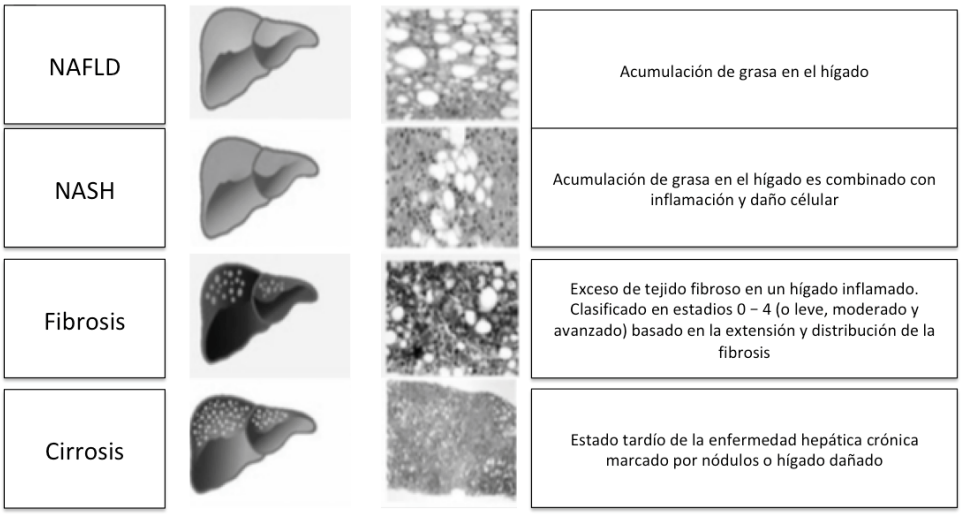

Figura 6.1: Estadíos de la enfermedad hepática crónica

Síntomas de NAFLD

Se puede considerar al hígado graso no alcohólico como una enfermedad silente debido a una perdida de síntomas en los estadios iniciales. Ocasionalmente, los pacientes con NASH pueden presentar fatiga, perdida de peso injustificada, y disconfort en el lado derecho del abdomen [57]. Si el paciente no es tratado, también tendrá los síntomas asociados a la cirrosis.

Etapas clínicas de NAFLD

Se puede observar las diferentes etapas de la enfermedad hepática crónica causada por NAFLD, figura 6.1.

Diagnóstico de NAFLD

Puede ser diagnosticado por diferentes modalidades de imágenes, incluyendo el ultrasonido abdominal, que aparentemente es el mas usado [15]. También se puede utilizar las medidas no invasivas como la tomografía computada y Resonancia magnetica; aunque no son confiables en reflejar el respecto de la histología del hígado en pacientes con NAFLD [26].

Su relación con el síndrome metabólico

El síndrome metabólico es considerado como uno de los principales retos de salud y esta muy asociado con el NAFLD [45]. El síndrome metabolico es un grupo de anormalidades que incrementan el riesgo cardiovascular. Un estudio conducido por Marchesini y Marzocchi, concluyó en que el contenido de la grasa hepática es significativamente mayor en pacientes con síndrome metabólico comparado con los individuos libres de enfermedad, independientemente de IMC, edad o género [37].

Insulinoresistencia (IR) y diabetes mellitus tipo 2 como factores de riesgo asociados a NAFLD

La incidencia de NAFLD en adultos con DM2 alcanza al rededor de 75 % de la población general [28]. El hígado graso puede causar daño hepático e inflamación por estrés oxidativo y puede progresar a fibrosis y culminar en cirrosis [8]. La insulino resistencia y el estres oxidativo son factores centrales en la patofisiología de las anormalidades metabólicas de NAFLD [70]. Esto puede ser visto en la figura 6.2.

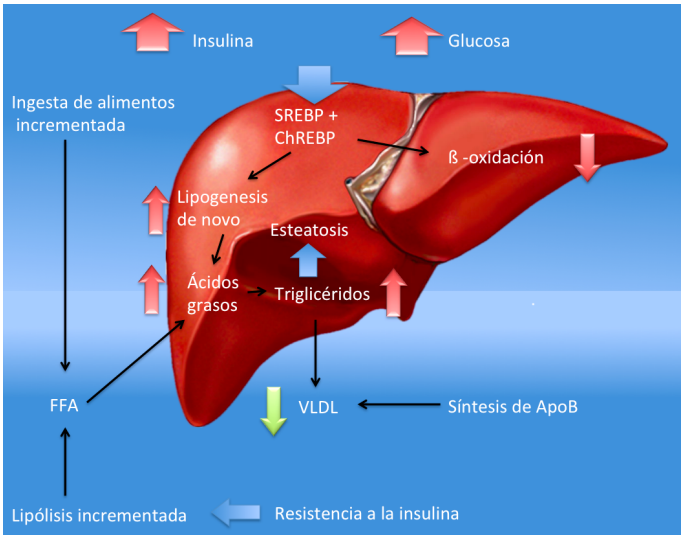

Figura 6.2: Patogénesis de la acumulación de grasa dentro de las células del hígado

Hígado graso no alcohólico y Acilcarnitinas

Recientemente Romero - Ibarguengoitia et. al. describieron un trabajo donde se relaciona el patrón de acilcarnitinas con la presencia de obesidad e hígado graso [60] . Este estudio fue analizado mediante ecuaciones estructurales. Parte de los datos trabajados en esta tesina, corresponden a observaciones publicadas en ese artículo. Aunque la visión hecha en el artículo original, remarca la historia familiar de los pacientes y su relación con obesidad. Los autores concluyen que si existe una alteración en el patrón de acilcarnitinas e inflamación en los pacientes con hígado graso. El patrón de acilcarnitinas descrito como anormal fue el incremento de las acilcarnitinas de cadena larga (sospechando una obstrucción en la vía de la beta oxidación.

Por otro lado, uno de los estudios más representativos fue el de Statish C Kalhan y colaboradores que incluyó 11 pacientes no diabéticos con esteatosis hepática y 24 pacientes con esteatohepatitis que fueron comparados con 25 controles sanos. El patrón metabolómico no pudo diferenciar entre esteatosis hepática y esteatohepatitis; sin embargo en comparación con los controles sanos las concentraciones de carnitinas libres, butirilcarnitina y metilbutiril carnitina (C3-C5) estuvo más alta (73). Otro estudio ha reportado tanto acilcarnitinas de cadena larga (C18, C18:2, C16) como corta (C4 yC3) relacionado a diferentes grados de EHNA en humanos [32].

Diabetes mellitus tipo 2

La diabetes mellitus tipo 2 es una de un conjunto de enfermedades englobadas en el termino diabetes mellitus. En este y otro tipo de diabetes, varios factores genéticos y ambientales pueden resultar en una perdida o disfunción progresiva de las células productoras de insulina (células beta), manifestandose clínicamente como hiperglucemia. Una vez que la hiperglucemia aparece, el paciente desarrolla un incremento a presentar complicaciones crónicas [68]

La diabetes puede ser diagnosticada basandose en el criterio de glucosa plasmática en ayuno (FPG, por sus siglas en íngles), a las 2 horas después de una carga de tolerancia oral de glucosa (2-h PG, 2 hrs post glucose, CTOG) o con la hemoglobina glucosilada (A1c) [3].

Criterios para el diagnostico de Diabetes Mellitus

- FPG $\geq$ 125 mg/dl (7.0 mmol/L). El ayuno definido como no ingesta calorica por al menos 8 horas*.
- 2-h PG $\geq$ 200mg/dl (11.1 mmol/L) durante la CTOG. La prueba debe ser realizada como fue descrita por la OMS, usando una carga de glucosa que contenga 75 grs o de glucosa anhídrida disuelta en agua.
- A1c $\geq$ (48 mmol/mol). La prueba deberá ser realizada en un laboratorio usando el método estandarizado a la prueba del DCCT.

El número de personas con diabetes se ha incrementado de 108 millones en 1980 a 422 millones en el 2014. La prevalencia mundial de diabetes entre adultos mayores de 18 años se ha incrementado de 4.7 % en 1980 a 8.5 % en 2014 [38]. En México la **E**ncuesta **N**acional de **Sa**lud y **Nut**rición de **M**edio **C**amino (ENSANUT MC) 2016 mostró un ligero incremento en la prevalencia de diabetes por diagnóstico médico previo (9.2 %) con respecto a la encuesta del 2012 (9.2 %). El mayor incremento se observo entre los hombres de 60 a 69 años de edad y entre las mujeres con 60 o más años de edad, aproximadamente un 30 % de esta población tiene diabetes [24].

La diabetes mellitus tipo2 y las acilcarnitinas

La diabetes mellitus tipo 2 es la forma predominante de Diabetes en todo el mundo, siendo el 90 % de todos los casos [41]. La palabra diabetes viene del griego (día= a través de=, bainein= ir, tes=gente), es decir: ?lo que va a través?, esto referido por el exceso de orina. Mellitus (del girego melli= miel) que sabe dulce o a miel (característica de la orina de estos pacientes).

El número de pacientes con diabetes esperados para el 2025 son de 300 millones de personas. Por lo que la Diabetes mellitus tipo 2 se ha convertido en uno de los problemas mundiales de salud pública. En México la prevalencia de la Diabetes mellitus tipo 2 es del 14.42 % (7.3 millones de personas) (3). Esta prevalencia a aumentado un 7 % comparando los resultados de la Encuesta Nacional de Salud 2006 y la Encuesta Nacional de Enfermedades Crónicas 2004 [72] demostrando que México no es la excepción en el problema mundial de la Diabetes mellitus.

Este incremento tan pronunciado en la prevalencia de la Diabetes Mellitus tipo 2 es

debido en parte al incremento en la prevalencia de la obesidad que en México con una prevalencia combinada de sobrepeso y obesidad en mujeres mayores a 20 años de un 71.9 % y en hombres del 66.7 % [66].

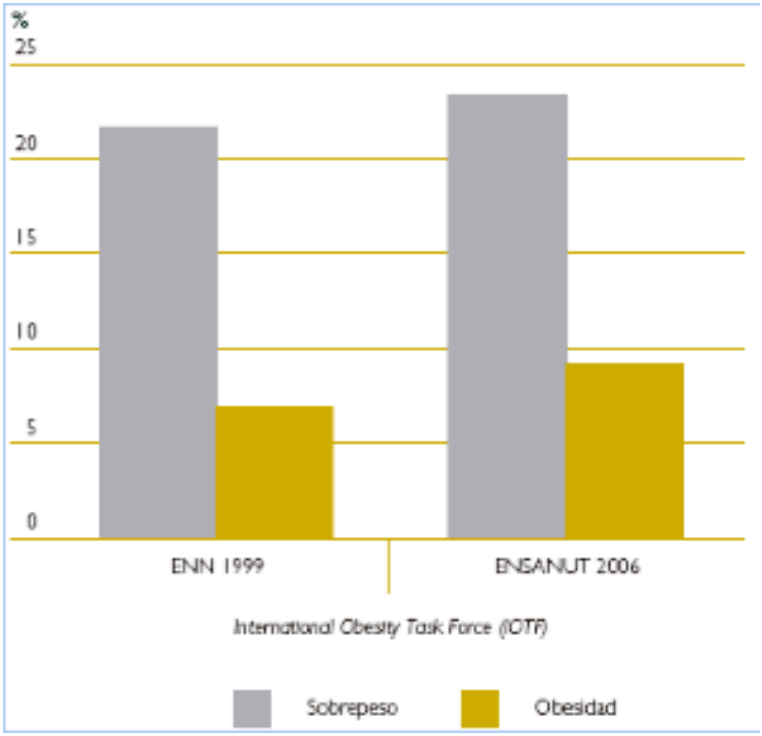

Figura 6.3: Comparación de la prevalencia de sobrepeso y obesidad entre 1999 y 2006 en mujeres de 12 a 19 años de edad de acuerdo con los criterios propuestos por el IOTF. México

Fisiopatología de la Diabetes Mellitus tipo 2

La patogénesis de la Diabetes Mellitus tipo 2 es compleja e incluyen factores genéticos y ambientales. Algunos de los genes descritos como predisponentes a Diabetes Mellitus tipo 2 son HHEX, SLC30A8, CDKAL1 y especialmente TCF7L2 el cual es asociado fuertemente a la enfermedad [62]. Existen otras alteraciones genéticas puntuales causantes de diabetes pero son descritas como diabetes monogénicas y sus fisiopatologías son distintas. En la última década Ralph Defronzo describo las múltiples alteraciones metabólicas encontradas en el metabolismo de los carbohidratos (El ominoso octeto), en donde se describe la resistencia a la insulina. La resistencia a la insulina es una de las alteraciones preclínicas de la diabetes mellitus y se define como la disminución del efecto de la insulina en los tejidos periféricos.

Teorías de la resistencia a la insulina

La resistencia a la insulina se manifiesta por una disminución del transporte de la glucosa estimulado por la insulina, por una alteración del metabolismo de la glucosa en los adipocitos y en el músculo esquelético, y por una supresión alterada de la producción hepática de glucosa. La sensibilidad a la insulina está influida por varios factores entre los que se encuentran la edad, el peso, el grupo étnico, la grasa corporal (especialmente la abdominal), la actividad física y los fármacos. La resistencia a la insulina se asocia con la progresión de la ATG y de la Diabetes mellitus tipo 2, aunque rara vez se observa una diabetes mellitus en las personas con resistencia a la insulina incluso cuando no son obesos, lo que implica la existencia de un importante componente genético en el desarrollo de la resistencia a la insulina.

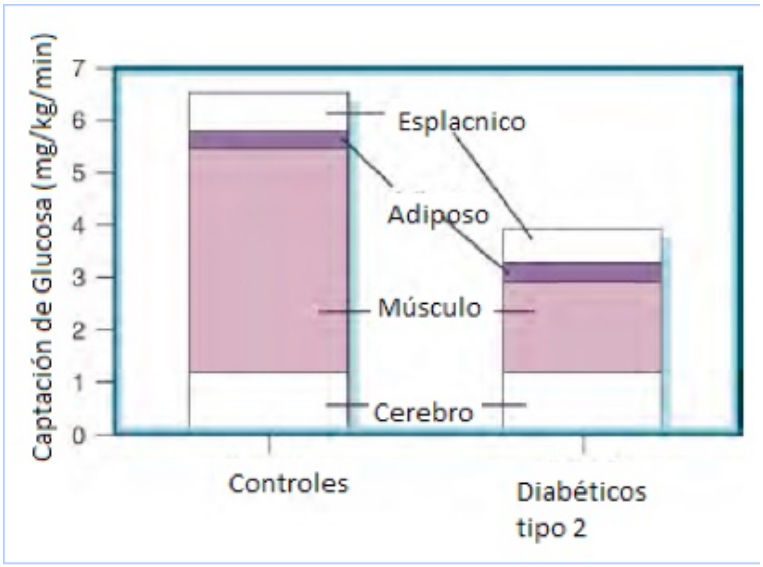

Figura 6.4: Captación de glucosa dependiente de insulina.

El principal lugar de almacenamiento de la glucosa después de una comida es el músculo esquelético, y el principal mecanismo de almacenamiento de glucosa es a través de su conversión a glucógeno. Los estudios que utilizan la técnica del clamp hiperinsulinemico euglucémico han demostrado que en las personas resistentes a insulina con o sin DM2 hay un déficit en la captación no oxidativa de la glucosa, en relación principalmente con un defecto en la síntesis de glucógeno.

Triglicéridos intramusculares.

La captación de glucosa estimulada por la insulina es inversamente proporcional a la cantidad de triglicéridos intramusculares. Se ha demostrado una importante correlación entre la concentración de triglicéridos intramusculares mediante biopsia, TC y resonancia magnética. Los familiares de primer grado de las personas con DM2 tienen un aumento de la grasa intramiocelular, y en este grupo existe también una correlación con la resistencia a la insulina.

Esta acumulación de triglicéridos intracelulares disminuye la señalización intracelular del receptor de insulina al mantener una fosforilzación parcial del sustrato del receptor de insulina, evitando la señalización normal por la vía de la MAP cinasa y terminando en la expresión génica celular; disminuye la expresión de canales GLUT 4 en la membrana celular y la captación de glucosa por la célula.

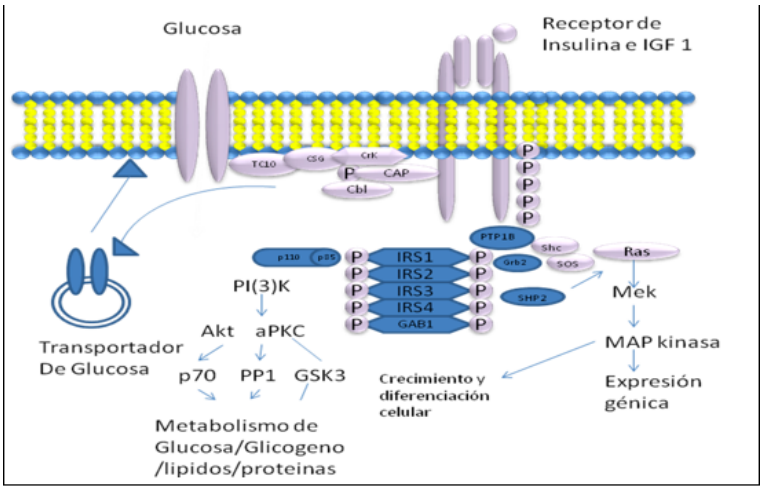

Figura 6.5: Señalización post receptor de insulina

Fisiología normal de las acil carnitinas

En el cuerpo, los ácidos grasos son degradados a acetil-CoA la cual entra al ciclo del acido cítrico. Esta degradación ocurre en la mitocondria por la beta oxidación. La oxidación de los ácidos grasos empieza con la activación del acido graso, la reacción ocurre fuera y dentro de la mitocondria.

Los ácidos grasos de cadena mediana y corta pueden entrar a la mitocondria sin dificultar pero los ácidos grasos de cadena larga deben entrar unidos a la carnitina con una unión

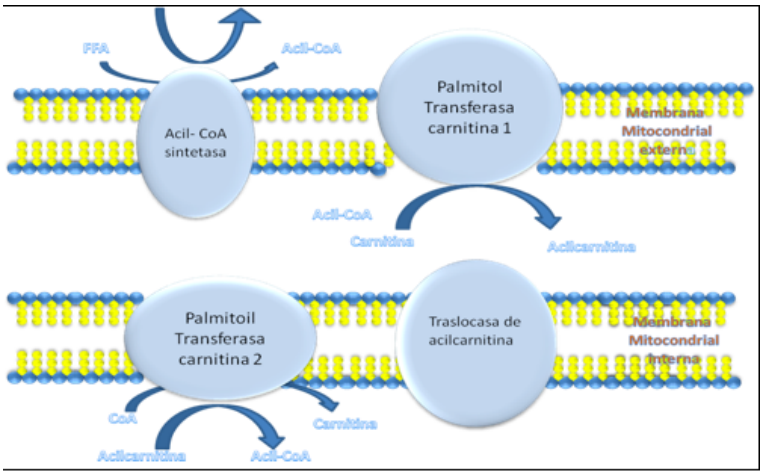

Figura 6.6: Internalización de ácidos grasos de cadena larga a la matriz mitocondrial

ester antes de que pueda entrar a través de la membrana mitocondrial. La carnitina es un b hidroxi gama trimetilamonio butirato, y es sintetizado en el cuerpo de lisina y metionina. La translocasa moviliza el ester ácido graso- carnitina dentro del espacio de la matriz en intercambio con carnitina libre. En el espacio de la matriz, el ester es hidrolizado haciendo el acido graso activado una molécula disponible para la beta oxidación y proveyendo creatina libre para intercambios posteriores.

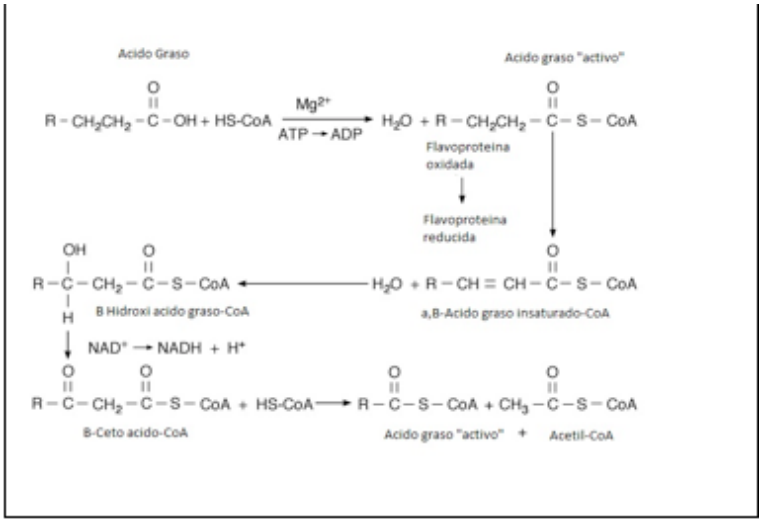

Figura 6.7: Ejemplo de la Beta oxidación de un acido graso libre.

La beta oxidación procede por la extracción de fragmentos de 2 carbonos del acido

graso. La energía generada en este proceso es muy alta. Por ejemplo, el catabolismo de 1 mol de un acido graso de 6 carbonos que pasa a través del ciclo del acido cítrico da Co2 y H2O y genera 44 mol de ATP, comparado con los 38 mol generados por el catabolismo de 1 mol de 6 carbonos de la glucosa.

Esta vía de oxidación de los ácidos grasos libres se mantiene inactiva mientras la célula es expuesta a glucosa y se activa en momentos de disminución de glucosa y a la exposición de ácidos grasos libres.

Evidencia de las alteraciones del metabolismo de las acilcarnitinas como etiología de resistencia a la insulina.

Recientemente se ha demostrado que la alteración en la beta oxidación mitocondrial es parte de la fisiopatología de la resistencia a la insulina, esto al presentar diferentes patrones de acilcarnitinas séricas, a continuación describiré algunas de las evidencias recientes.

Sean H. Adams et al. Encontró en su estudio al comparar 44 mujeres obesas con diabetes mellitus tipo 2 y 12 mujeres sin diabetes mellitus (todas africano americanas), que la relación de acilcarnitinas totales: carnitinas libres fue significativamente incrementada (150 a 170 %) en las mujeres con Diabetes mellitus tipo 2; además la concentración de ácidos grasos de cadena larga se mantuvo incrementada hasta en un 300 % en las pacientes con DM2 (p=0.004). Estos resultados son consistentes con la hipótesis de que una beta oxidación insuficiente debida en parte a la baja capacidad del ciclo del acido tricarboxílico, incrementa la acumulación de acetil CoA y genera moléculas de acil carnitina de cadena corta que activan las vías pro inflamatorias implicadas en la resistencia a la insulina [1].

Kitt Falk Petersen et al. en su artículo en el New England Journal of Medicine describe la alteración mitocondrial al realizar un pinzamiento hiperinsulinemico euglucémico en combinación de glucosa marcada en pacientes sanos, jóvenes, delgados e insulinoresistentes en desendientes de pacientes con diabetes mellitus tipo 2, comparados con sujetos controles insulinosensibles pareados por edad, peso y actividad física. Realizaron una resonancia magnética con espectroscopia para medir el contenido lipídico intramiocelular e intra hepático además de la evaluación de la razón de la actividad de fosforilación oxidativa mitocondrial en el músculo. Encontraron que la razón de captura de glucosa estimulada por insulina en el músculo fue aproximadamente 60 % más baja en los sujetos insulinoresistentes que en los sensibles a la insulina (p=<0.001) y fue asociado a un incremento del 80 % en el contenido lipídico intramiocelular (p=<0.005) [44].

M. Möder · A. Kießling et al. Demostraron que los pacientes con Diabetes mellitus tipo 2 presentan una alteración en el patrón urinario de las acilcarnitinas que concuerda con los cambios a nivel sérico.

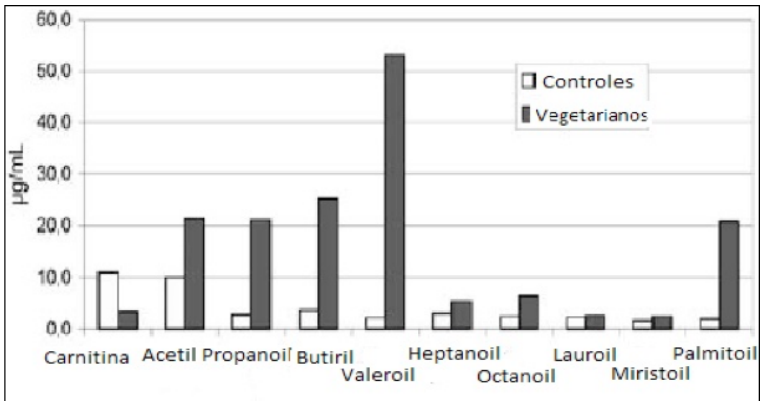

Figura 6.8: Acil carnitinas urinarias en sujetos controles

Estos pacientes tenían mayor concentración urinaria de acilcarnitinas de cadena larga (principalmente palmitoil carnitina) mediante el análisis de inyección de flujo- electroscopia de masas inonización electro espray.

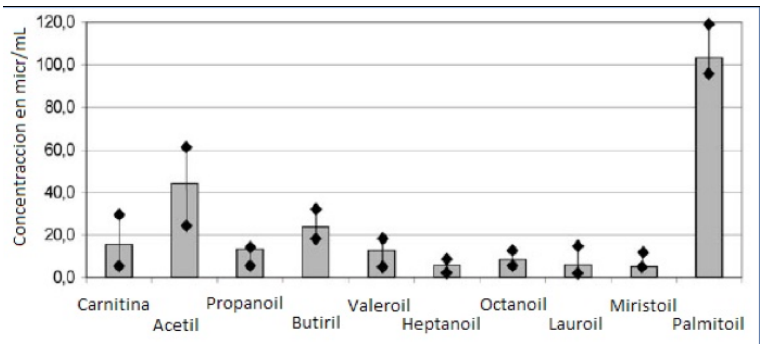

Figura 6.9: Concentraciones urinarias de acilcarnitinas en personas con DM2.

Retinopatía

El término retinopatía se refiere a la presencia de algún daño no especificado en la membrana sensible a la luz del ojo llamada retina. Una de las principales causa de retinopatía es la diabética, la cual es la principal causa de ceguera en los Estados Unidos, fue inicialmente descrita por Eduard Jaeger en 1856, pero sus relaciones causales entre los exámenes de

retina y la diabetes mellitus fueron inicialmente controversiales hasta 1875 cuando Leber confirmo los hallazgos [74].

En el 2005, 5.5 millones de personas tenían retinopatía diabética y 1.2 millones se encontraban en peligro de padecerla [65]. En México existe un estudio que evaluó la prevalencia de retinopatía diabética encontrandose que mas del 34 % de la población estudiada tenia retinopatía no proliferativa [22].

Existen dos categorías principales de retinopatía diabética: no proliferativa y prolifrativa. El edema macular diabético puede presentarse en cualquier etapa. La diferenciación entre ambas etapas de la retinopatía es la proliferación de nuevos vasos sanguíneos los cuales son altamente predisponentes a presentar sangrados y exudados de materiales lipídicos [20].

Revisión de fondo de ojo y el lecho vascular sistémico

La forma no invasiva de conocer el estado de la vasculatura sistémica es a través de la visualización de los vasos retinianos (mediante la lampara de hendidura), que provee un campo de estudio de cómo diversas patologías cambian la microcirculación humana. La tecnología actual permite hacer una medición objetiva de dichos cambios. Los vasos retinianos no tienen inervación adrenérgica que pueda iniciar cambios en el tono vascular [76],[33], se ha postulado que el diámetro vascular es dependiente de cambios miogénicos [16], así como vías que involucran la función endotelial, inflamación y autorregulación metabólica a través de elaboración de factores vasodilatadores (óxido nítrico, adenosina, prostanoides) y vasocontrictores (endotelina, angiotensina II) en respuesta a demandas metabólicas. El óxido nítrico juega un papel central en la regulación del tono vascular e inhibie la adhesión plaquetaria y leucocitaria en las células endoteliales [61], [48],[71]. Otros marcadores de inflamación como el complemento y las interleucinas, niveles elevados de proteína C, interleucina 1, 6 y TNF alfa se asocian a mayor diámetro venular independiente de la presión arterial y diabetes Mellitus (24). Recientemente se ha demostrado que la PCR puede tener efectos sobre la vasorreactividad del óxido nítrico en el endotelio retiniano arteriolar (25). La IL 10 se asocia a mejor reactividad del sistema vascular y suele encontrarse disminuida en pacientes con obesidad y los niveles de IL 17 se asocian a la mayor estimulación de citocinas IL- 6 y 8.

En base a lo descrito previamente se realizó el estudio de Romero - Ibarguengoitia [59], donde se documenta la relación de lesiones retinianas en pacientes con obesidad y sin

diabetes. En este trabajo se continua con la intención de poder clasificar estos pacientes como portadores de lesiones retinianes con las acilcarnitinas, que como se mencionó en la sección de NAFLD, tienen relación con inflamación y la presencia de obesidad.

Regresión logística.

Los métodos de regresión se han convertido en una parte principal del análisis de los datos, en lo que se refiere a la relación entre las variables respuesta y las explicativas. Se puede distinguir la regresión logística de los otros tipos de regresión, porque la variable respuesta es binaria o dicotómica. Donde la variable respuesta debe ser predicha mediante una probabilidad, y no mediante la predicción de un valor determinado continuo .

Se puede usar la cantidad $\pi(x) = E(Y|x)$ para representar la media condicional de Y dado x cuando la regresión logística es usada. La ecuación especifica a este modelo es:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \tag{6.1}$$

$\pi(x)$ se puede transformar en la transformación logit:

$$g(x) = ln[\frac{\pi(x)}{1-\pi(x)}] = \beta_0 + \beta_1 x \tag{6.2}$$

La importancia de esta transformación es que g(x) tiene propiedades deseables de un modelo de regresión lineal. El logit, g(x), es lineal en sus parámetros, puede ser continuo y su rango de $-\infty$ a $+\infty$ dependiendo del rango de x.

La cantidad de variables incluidas en el modelo deben ser k-1 variables, siendo K la cantidad de observaciones a predecir. Por lo que se debe tener cuidado al momento de incluir las variables y no saturar el modelo.

La selección paso a paso es ampliamente usado en otros tipos de regresiones como en la lineal. Este procedimiento se basa en la selección estadística de las variables "mas importantesz esta importancia es medida mediante la significancia estadística de los coeficientes de las variables. En la regresión logística se utiliza el mayor cambio de log verosimilitud en relación con un modelo que no contiene la variable.

La selección de variables hacia adelante (**Forward**,en inglés), es uno de los tres métodos disponibles para la selección automática de variables (los otros son, hacia atras (**backward**) y exhaustivo (**exhaustive**). Estos métodos pueden ser criticados por reunir variables que

clínicamente pudieran ser retiradas del modelo y se sugiere siempre agregar variables por expertise del investigador. Sin embargo, en el presente trabajo se utiliza este método ya que en la actualidad no se cuenta con una manera clínica de selección de variables.

A continuación se describe el algoritmo de selección hacia adelante.

1. Paso (0): Inicia con el ajuste del "modelo solo con interceptoz la evaluación de su log-verosimilitud. Esto es seguido de el ajuste de cada posible variable mediante regresiones logísticas univariadas. Se agregan solo las variables con menor "p valor".

2. Paso (1): Se comienza con el ajuste de la regresión logística, conteniendo ya la primera variable. Se ajusta el siguiente modelo agregando la variable con menor "p valor", y se procede a realizar el paso 2, de otra forma se detiene.

3. Paso (2): Se ajusta el modelo conteniendo las dos primeras variables. Dado que la primera variable agregada puede no ser ya significativa (en presencia de la segunda variable), se realiza una eliminación hacia atrás. Esto basado en el cambio en el p valor con o sin la primera variable.

4. Paso (3): El paso (3) es idéntico al paso (2). El programa ajusta el modelo que incluye la variable seleccionada durante el paso previo. Este paso continua hasta el paso (s).

5. Paso (S): Este paso, ocurre cuando:

   a) Todas las variables an entrado al modelo ó

   b) Todas las variables en el modelo tienen p valor, para remover las que tengan menor p valor.

Bootstrap

El “bootstrap” es una de las técnicas que ahora es parte de un abanico de pruebas estadísticas no parametricas que comunmente son llamados métodos de remuestreo. Fue definido por Efron en el año 1979 como un procedimiento de remuestreo. El objetivo del bootstrap es estimar un parámetro de los dato (media, mediana o desviación estándar). También se pueden construir intervalos de confianza [14].

El elemento básico para el boostraping es la distribución empirica. Esta distribución empirica es solo la distribución discreta que da igual peso a cada punto (ósea probabilidad

1/n). El principio del bootstrap menciona que F es la distribución de la población, y T(F) es la función que define el parámetro a obtener. Nosotros deseamos estimar un parámetro cualquiera de una muestra de n observaciones independientes e igualmente distribuidas. Entonces F_n juega el papel de F y F_n* la distribución bootstrap, tiene el papel F_n* en el proceso de remuestreo.

De esta forma nosotros calculamos la media del parámetro con el que se evalúa el modelo (ej. sensibilidad, especificidad, etc.).

Pseudocódigo ID3

En el siguiente espacio se describirá el pseudocódigo de ID3

```
ID3 (Ejemplos, Atributo, Atributos) Create: Un nodo raíz para el árbol; asigna
        todos los Ejemplos a la raíz
if  Todos los Ejemplos son positivos then
 |  return un solo nodo raíz, con etiqueta = +
end
if  Todos los Ejemplos son negativos then
 |  return un solo nodo raíz, con etiqueta = −
end
if  Los Atributos están vacios then
 |  return un solo nodo raíz, con la etiqueta = el valor mas común de Atributo en
 |   Ejemplos
end
Otherwise
 |  A ← el atributo de Atributos que mejor clasifica en Ejemplos
 |  El atributo decisión para la raíz ← A
 |  foreach Posible valor v_i de A do
 |   |  Agregar una nueva rama debajo de la raíz, correspondiente a la prueba A = v_i
 |   |  Haz que Ejemplos_vi sea el subgrupo de Ejemplos que tienen el valor vi para
 |   |   A
 |   |  if Ejemplos_vi then Esta vacío
 |   |   |  Debajo de esta nueva rama se agrega un nodo hoja con etiqueta = el
 |   |   |   valor mas común de Atributo en Ejemplos
 |   |   |  else
 |   |   |   |  Debajo esta nueva rama agregar el sub árbol
 |   |   |   |  ID3(Ejemplos, Atributo, Atributos{A})
 |   |   |  end
 |   |  end
 |  end
 |  return raíz
end
```

Pseudocódigo C4.5

Se presenta el pseudocódigo del algoritmo C4.5.

```
Input  : atributos valuados en el conjunto de datos D
Árbol = {}
if  D es "puro" u otro criterio de paro se cumple then
 |  termina
end
forall Atributos ∈D do
 |  Computar el criterio de información teórico si particionamos en a
end
a_mejor = El mejor atributo acorde a los criterios computados antes.
D_v = Inducción de los sub-datos de D basados en a_best
forall D_v do
 |  Árbol_v = C4,5(D_v) Adjunta el árbol_v a la rama correspondiente del árbol
end
return Árbol
```

Algoritmo 1: Pseudocódigo de algoritmo C4.5

Pseudocódigo del boosting C5.0

Para describir el desarrollo de este algoritmo, asumiremos que el conjunto de muestras S consiste de n muestras y un sistema de aprendizaje que construye diferentes árboles de decisión. El boosting construye árboles de decisión de las muestras, esto es, construye T árboles de decisión, y C^t es el árbol de decisión arrojado por el sistema de aprendizaje en el intento t y C^* es el árbol final que es formada al agregar los T árboles de decisión. w_i^t es el peso de la i-ésima muestra en la prueba t $(i = 1, 2, \ldots, N; t = 1, 2, \ldots, T)$. P_i^t es el factor normalizado de w_i^t y β_t es el factor que ajusta el peso. También se puede definir una función indicadora:

$$\theta^t(i) = \begin{cases} 1, \text{ la i-ésima muestra es mal clasificada} \\ 0, \text{ la i-ésima muestra es bien clasificada} \end{cases} \tag{6.3}$$

Los principales pasos para el boosting es el siguiente:

1. Inicializar las variables; ajusta un valor al numero de T (usualmente es 10). Ajusta $t = 1, w_i^1 = \frac{1}{n}$.

2. Calcula $P_i^t = w_i^t / \sum_{i=0}^{n}(w_i^t$, donde $\sum_{i=0}^{n}(P_i^y) = 1$.

3. Sea P_i^t el peso de cada muestra y construye C^t bajo esta distribución.

4. Calcula la taza de error de C^t como $\epsilon^t = \sum_{i=0}^{n}(P_i^t \theta_i^t)$.

5. Si $\epsilon^t < 0.5$, los experimentos se terminan, sea $T = t + 1$; de otra forma si $\epsilon^t = 0$, los experimentos se terminan, sea $T = t$; de otra forma si $0 < \epsilon^t < 0{,}5$, ve al paso 6.

6. Calcula $\beta^t = \epsilon^t / (1 - \epsilon^t)$.

7. Ajusta el peso acorde a la tasa de error, que es

$$w_i^t t + 1 = \begin{cases} w_i^t \beta^t, & \text{la muestra es clasificada equivocadamente.} \\ w_i^t, & \text{la muestra es clasificada correctamente.} \end{cases}$$

8. Si $t = T$, los experimentos son terminados. De otra forma, sea $t = t + 1$ y ve al paso 2 para empezar con el nuevo experimento.

Finalmente, se obtiene el boosted tree C^* sumando los votos de los árboles de decision $(C^1, C^2, \ldots, C^T)$, donde el voto para C^t vale $log(1/\beta^t)$ unidades. Esto es $C^* = \sum_{t=1}^{T}(1/\beta^t)C^t$. Esto significa cuando clasifica una muestra de prueba usando un modelo de árbol de decision, primero, se clasifica esta muestra por $C^t(1 \leq t \leq T)$, y podemos tener los T resultados. Entonces se cuentan al final los votos de cada clase acorde al peso de $C^t(1 \leq t \leq T)$ y selecciona la clase que tiene el mas alto voto como resultado final. [12].

La poda del árbol producido por el algoritmo C5.0 es hecha desde el punto de vista de la probabilidad en la tasa de mala clasificación; esto es, del intervalo de confianza. Cuando el control de la poda del árbol es llevada por este intervalo de confianza (CF): mientras mas grande el valor, menos ramas son podadas, mientras mas bajo sea el valor mas ramas son podadas. Así como en el algoritmo C4.5, el valor por default de CF es de 0.25, y asumimos que la tasa de error en la clasificación esta acorde a una distribución binomial.

Input : S_l, un conjunto de atributos asignados al Nodo l, en el árbol T; SM (Medición de partición)
Output: MejorSM (S_l (la mejor partición de atributo para el nodo l)
Árbol = {}
foreach *atributos* $a_i \in S_l$ **do**
  SM(a_i) ← Calcular la medida de partición (SM,l,i)
  MejorSM (S_l) ← arg max [SM(a_i)]
  SM_{Crit} ← Encontrar un valor critico (MejorSM (S_l))
  Iniciar un grupo de atributos potencialmente particionables $E_l \leftarrow \emptyset$
end
foreach *atributo* $a_i \in S_l$ **do**
  if *SM* (a_i) $> SM_{Crit}$ **then**
    $E_l \leftarrow a_i$
  end
end
Crea t, el arreglo para guardar la evaluación de subárboles
foreach $a_e \in E_l$ **do**
  t_e ← ConstruyeÁrbolJ48 (a_e) Exactitud (t_e) ← EvaluaÁrbol (t_e,ConjuntoValidación$_l$)
end
Encuentra el "Mejor" subárbol
MejorÁrbol $leftarrow$ **arg max Exactitud** (t_e)
MejorSM(S_l) ← **MejorÁrbol**
Regresa MejorSM (S_l)

Algoritmo 2: Pseudocódigo de Look Ahead agregado a J48

Pseudocódigo de la integración del Look Ahead al algoritmo C4.5 (J48 de Weka).

Pseudocódigo

Pseudocódigo Árbol de Decisión Sensible a Costos

En esta sección se presenta el pseudocódigo del árbol de decisión sensible a costos:

```
Input  : Datos de entrenamientos S; el conjuto de atributos C, parámetro δ
Método: ACSDT
Output : A árbol de decisión
Crea un nodo árbol;
if S es puro o C esta vacío then
 | regresa árbol tiene un nodo hoja;
end
maxQuality = 0; El máximo valor de la función heurística
/* Selecciona el atributo con el mayor valor de función heurística   */
for i = 0; i</C/; i++ do
 | Computa el máximo valor (denotado como maxValue) y el minimo valor
 |   (denotado como minValue) del atributo a_i; cp = 1/2(maxValue + minValue),
 |   paso = 1/4(maxValue-minValue);
 | Quality(a_i) = ASCP(cp,paso);
 | if Quality(a_i >maxQuality then
 |  | A=a_i;maxQuality=Quality(A);
 | end
 | else
 |  | /* Remueve atributo                                              */
 |  | if /C/>δ y Quality(a_i) <1/δ *maxQuality then
 |  |  | C=C-{a_i
 |  | end
 |  | :
 | end
end
if maxQuality=0 then
 | regresa árbol;
end
árbol=árbol ← A; tc(A) =0; /* Particiona S en dos conjuntos de datos: S_1,
   S_2.                                                                 */
Coloca el objeto con VA_{x_i} ≤
```

Algoritmo 3: Pseudocódigo de Árbol de decisión Sensible a Costos

Árbol de Decisión Sensible a Costos y con Mirada Adelante Generalizado

El pseudocódigo del algoritmo propuesto se presenta a continuación.

```
Árbol de Decisión Sensible a Costos y con Mirada Adelante Generalizado
 (GCSLADT)(D,d)
% D: conjunto de datos, d: tamaño de profundidad
Input  : atributos valuados en el conjunto de datos D
Output: Un GCSLADT
if  D es "puro" u otro criterio de paro se cumple then
 |  termina
end
forall Atributos ∈D do
 |  Computar el criterio de información teórico si particionamos en a
end
a_mejor = El mejor subconjunto de atributos de tamaño d, acorde a los criterios de
 información teórico computados antes
Árbol = Crea una rama de decisión que prueba a_mejor en la raíz
D_I = Inducción de los sub-datos de D basados en a_mejor
forall D_I do
 |  Árbol_v = (GCSLADT)(D_I, d)
 |  Adjunta el árbol_v a la rama correspondiente del árbol
end
return Árbol
```