UNIVERSIDAD CENTRAL "MARTA ABREU" DE LAS VILLAS
VERITATE SOLA NOBIS IMPONETUR VIRILIS TOGA. 1948

Facultad de Matemática, Física y Computación
Licenciatura en Matemática

TRABAJO DE DIPLOMA

ÍNDICE DE ALTO RIESGO CARDIOVASCULAR PARA EL MUNICIPIO DE SANTA CLARA

Diplomante: Arasay Gómez Rodríguez

Tutores: Lic. Jorge Luis Morales Martínez
Dra. Gladys Casas Cardoso

"Año 50 de la Revolución"
Santa Clara
2008

CON SU ENTRAÑABLE TRANSPARENCIA

Resumen

El presente trabajo estudia los factores de alto riesgo cardiovascular para el municipio Santa Clara. Se obtienen tres modelos matemáticos del índice de riesgo que responden a las características de la población de dicho municipio. Se explica detalladamente como se pueden aplicar técnicas de estadística multivariada para obtener dichos modelos. Los métodos más importantes que se emplean son: el análisis discriminante, la regresión logística y los árboles de decisión.

Además se efectúa una comparación entre los modelos para determinar cuál de ellos es el mejor, así como se realiza una validación de los resultados obtenidos en la aplicación de las técnicas estadísticas antes mencionadas. Todo lo anterior se realizó con la ayuda del paquete estadístico SPSS.

Introducción

El cambiante mundo moderno está sustentado por un conjunto de ciencias empleadas por el hombre para controlar y perfeccionar los procesos. La Estadística es una de estas ciencias. En los últimos años se han desarrollado varios métodos que se ocupan de los modelos matemáticos en general, métodos que han sido automatizados gracias al desarrollo de la informática, por lo que resultan de gran utilidad práctica para solucionar problemas presentes en la sociedad.

La tecnología informática disponible hoy en día, casi inimaginable hace solo dos décadas, ha hecho posibles avances extraordinarios en el análisis de datos en el área de la medicina, la bioinformática y la producción entre otras áreas. Este impacto es más evidente en la relativa facilidad con la que los ordenadores pueden procesar enormes cantidades de datos complejos. Casi cualquier problema se puede analizar fácilmente hoy en día por un gran número de programas estadísticos disponibles en ordenadores personales. Además, los efectos del progreso tecnológico han extendido aun más la capacidad de manipular datos, liberando a los investigadores de las restricciones del pasado y permitiéndoles así abordar investigaciones más sustantivas y ensayar sus modelos teóricos. Gran parte de esta creciente comprensión y pericia en el análisis de datos ha venido a través del estudio y desarrollo de la Estadística y en particular de la de la inferencia estadística.

Las condiciones económico-sociales que han caracterizado la situación del país, con las consecuentes limitaciones de los recursos materiales y financieros, los cambios ocurridos en los perfiles de riesgo, morbilidad y mortalidad en los últimos años con mayor relevancia en las enfermedades transmisibles y el deterioro ambiental y sus implicaciones en la salud de la población, especialmente en las zonas urbanas y la emergencia de antiguos problemas como la hipertensión arterial y las enfermedades cardiovasculares, han puesto de manifiesto la necesidad de perfeccionar los sistemas de vigilancia epidemiológica a cada nivel de la organización de salud, los sistemas de detección activa y su capacidad de respuesta.

La epidemiología, en su concepción más simple, puede definirse como la ciencia que estudia las epidemias. Otra definición comúnmente aceptada es que la epidemiología no

es más que:"el estudio de la ocurrencia de las enfermedades", es una definición que se basa en los métodos y propósitos que se le atribuyen a esta rama de la ciencia.

Estos métodos y propósitos pueden resumirse en lo fundamental del modo siguiente:

- Descripción del estado de salud de la población mediante indicadores e identificación de tendencias.
- Predicción de magnitud y forma que puede asumir ese estado.
- Explicación de los mecanismos de transmisión de los procesos morbosos e identificación de sus causas.
- Control de las enfermedades en el sentido de prevenir casos nuevos, erradicar los existentes y disminuir sus efectos con particular orientación hacia la prolongación de la vida.

Se define la actual epidemiología como estudio y análisis de los factores de riesgo que influyen en la aparición, presencia, frecuencia y distribución de cualquier enfermedad en una comunidad humana, para averiguar sus causas y difusión y conseguir la disminución o desaparición de aquella. Es decir, se ocupa desde el punto de vista preventivo de los fenómenos de la masa en las enfermedades transmisibles y no transmisibles.

Actualmente, dicha epidemiología dedica su especial y mayor atención a los modernos jinetes de la Apocalipsis: las enfermedades de la civilización (no transmisibles) que, amenazan cada día más la vida y el alma del hombre actual, y están representadas por:

- Las enfermedades cardiovasculares degenerativas.
- Las enfermedades mentales, alcoholismo y dependencia a drogas.
- Los intentos suicidas.
- Los diferentes tipos de cáncer.

Los accidentes y toda clase de muertes violentas, que hoy se estudian epidemiológicamente a fin de establecer qué aspectos de la existencia humana han traído estos problemas morbosos, característicos de nuestra época.

En nuestro país la vigilancia epidemiológica estuvo integrada fundamentalmente por el sistema de Vigilancia de Enfermedades Transmisibles, con sus subsistemas de

vigilancia específicos (Enfermedades Respiratoria Agudas, Enfermedades Diarreicas Agudas, Hepatitis, Meningoencefalitis Meningocócica, Parálisis Flácida, Tuberculosis, SIDA, Lepra, Enfermedades transmisibles por elementos y otros) y el Sistema de Vigilancia Alimentaria y Nutricional. Existen otros sistemas con diferentes alcances y desarrollo como son: la vigilancia toxicológica, la vigilancia de la calidad del agua, la red de monitoreo de la calidad del aire, la vigilancia químico y microbiológica de los alimentos y la vigilancia relacionada con el control de vectores, entre otros.

Una adecuada práctica de salud pública requiere que las decisiones tengan una base científica. La vigilancia en salud pública es un componente esencial y necesario para el desarrollo de los servicios de salud y ha sido definida como: el seguimiento, recolección sistemática, análisis e interpretación de datos sobre eventos de salud o condicionantes relacionadas para ser utilizadas en la planificación y evolución de programas de salud pública, incluyendo como elemento básico diseminación de dicha información a los que necesiten conocerla.

Los métodos estadísticos se aplican siempre que grupos de individuos, cuyas características o comportamientos no son predecibles, necesitan ser descritos o comparados.

Las investigaciones epidemiológicas y las de salud pública en general trabajan con fenómenos variables, por lo que es requisito fundamental que el sanitario mida lo más rigurosamente posible dichos acontecimientos y a la vez trate de minimizar la intervención de los factores externos al suceso (instrumentos de medida, subjetividad, etc.), que pueden sesgar la fiabilidad de los resultados. Para conseguir este objetivo, el procedimiento estadístico es imprescindible en la planificación, gestión y control de las actividades dentro del sector sanitario, así como en la investigación epidemiológica.

La incorporación de la matemática moderna a los métodos estadísticos ha dado lugar a la aparición de una metodología de aplicación universal en el campo de la investigación o estado de todo fenómeno cuyo acontecimiento se caracteriza por la variabilidad. Obviamente el estudio y la caracterización de este tipo de fenómeno son parte esencial de la salud pública en general y de la epidemiología en particular.

La estadística, por tanto, se ha transformado en una herramienta de uso diario e imprescindible de epidemiólogos y personal sanitario dedicado a la medicina preventiva y salud pública.

Para muchas personas, estadísticas no son otra cosa que acumulaciones de cifras que se registran en anuarios, publicaciones oficiales, cursos etc. Para muchos, el término estadísticas hace referencia al número de personas alcanzadas por determinada enfermedad o al número de muertos por determinada causa, y, para otras estadísticas es sinónimo de acumulación de casos o cifras.

Por el contrario cuando hoy hablamos de estadística queremos referirnos al conjunto de métodos utilizados para recoger, elaborar, analizar y caracterizar un conjunto de datos, de forma que podamos obtener ciertos conocimientos de estos.

Los datos reunidos después de una investigación epidemiológica, los casos o cifras obtenidos por una encuesta a una determinada población etc. por si mismo no sirven para corroborar la hipótesis de investigación, ni para comparar las observaciones obtenidas con otras circunstancias de lugar y tiempo, ni para predecir la tendencia de los hechos observados. Los datos deben ser procesados y analizados de forma sistemática para detectar así las tendencias y patrones de las relaciones, eliminando, o llevando al mínimo, las parcialidades y subjetividad del investigador.

De manera que la vigilancia en salud pública resulta esencial en el proceso de prevención y control de enfermedades y factores de riesgo y en la promoción de la salud; es una herramienta vital en la ubicación de los recursos del sistema de salud y en la evaluación de la eficiencia de los programas de prevención y control. Se convierte así en un elemento importante de la función de evaluación, especialmente en la medición del impacto, y es esencial para el desarrollo de políticas apropiadas y para el aseguramiento de la disponibilidad de servicios.

La vigilancia epidemiológica y los programas de prevención y control en salud dependen en buena medida del conocimiento de los factores de riesgo asociados a las enfermedades. Para ello, los epidemiólogos han realizado tradicionalmente estudios de casos-controles o estudios de cohorte y aplicando técnicas de contingencia con la Metodología de Mantel-Haenszel (Mantel, 1950). Este análisis se complementa

necesariamente con estudios multivariados que se realizan clásicamente con técnicas de análisis discriminante (SPSS 10) o regresión logística (Hall, 2004); pero la primera tiene restricciones sobre la distribución de las variables predictivas y la segunda supera esto pero exige una muy buena definición de las variables (muchas veces ordinales o incluso nominales) y de sus formas óptimas de codificación.

Aplicando las técnicas estadísticas mencionadas anteriormente se llega a obtener un índice de riesgo que está determinado por los factores más importantes. Los estudios que se han llevado a cabo en el mundo hasta hoy revelan que el desarrollo de enfermedades cardiovasculares está íntimamente relacionado con el estilo de vida y los factores de riesgo asociados, asimismo la intervención en la modificación del estilo de vida y en el control de los factores de riesgo pueden posponer, o al menos retrasar, la aparición de la enfermedad cardiovascular ya sea antes o después de producirse el evento clínico.

Problema: ¿Es posible obtener un índice de alto riesgo cardiovascular para la ciudad de Santa Clara utilizando técnicas estadísticas sobre una muestra de 849 personas?

Con la realización de este trabajo se pretende:

Objetivo general: Determinar un índice de "alto riesgo cardiovascular" y las variables más importantes que permitan predecir ese riesgo cardiovascular en la población de Santa Clara, aplicando técnicas estadísticas.

Este objetivo general se puede desglosar en los siguientes objetivos específicos:

Objetivos específicos:

1. Obtener un índice de alto riesgo cardiovascular aplicando análisis discriminante, regresión logística y árboles de decisión.
2. Determinar cuál o cuáles de los modelos de índices obtenidos es el mejor, mediante la aplicación de las curvas ROC.
3. Determinar cuáles de los factores de riesgo son los más importantes en la predicción del alto riesgo cardiovascular.

Para dar respuesta a estos objetivos nos podemos plantear las siguientes preguntas de investigación:

Preguntas de investigación:

1. ¿Cómo deben aplicarse el análisis discriminante, la regresión logística y los árboles de decisión para obtener un modelo de predicción de alto riesgo cardiovascular?
2. ¿Serán igualmente buenos clasificadores los obtenidos mediante el análisis discriminante, la regresión logística y los árboles de decisión?
3. ¿Realmente todas las variables que participan en el análisis son decisivas en la predicción del alto riego cardiovascular?

Después de elaborar el marco teórico, que aparece en esencia en el primer capítulo de esta tesis, podemos formular la siguiente:

Hipótesis de investigación:

Al aplicar métodos estadísticos de clasificación, como son el análisis discriminante, la regresión logística y los árboles de decisión; se pueden obtener buenos resultados en la determinación y predicción de alto riesgo cardiovascular.

El trabajo que se presenta a continuación está conformado por tres capítulos. El primero es una revisión bibliográfica sobre el tema del riesgo cardiovascular y los estudios sobre la construcción de índices que se han realizado recientemente en el mundo. El segundo capítulo resume los métodos univariados que se utilizaron para caracterizar la población en estudio, como son las tablas de contingencia y pruebas no paramétricas. Asimismo resume las técnicas multivariadas que se emplearon para la obtención de los índices de alto riesgo cardiovascular, estas son: análisis discriminante, regresión logística y árboles de decisión. También en este capítulo se presenta de forma breve la teoría de las curvas ROC, que permite comparar los modelos obtenidos. El tercer capítulo muestra los resultados obtenidos mediante los tres métodos estadísticos, así como su comparación mediante las curvas ROC. El trabajo culmina con la presentación de las conclusiones y de algunas recomendaciones para trabajos futuros.

Capítulo 1. Índice de riesgo cardiovascular

1.1 Riesgo cardiovascular

Los datos no tienen significado por sí mismos, sino en relación a un modelo conceptual del fenómeno que los produce. Las manzanas venían cayéndose de los árboles mucho antes de que Newton estableciese las ecuaciones del movimiento de caída libre de los cuerpos; la circulación de la sangre se comportaba de igual forma antes y después de Harvey; las estrellas y planetas se movían en el espacio de igual manera en los tiempos de Ptolomeo, de Galileo, de Copérnico, de Kepler, del mismo Newton, o de Einstein. Los datos relativos al movimiento de los planetas recogidos en las diferentes épocas, salvo por diferencias en la capacidad de la instrumentación existente en el momento, eran similares; lo que fue evolucionando a lo largo del tiempo es el modelo conceptual establecido por cada científico para explicar el fenómeno.

Aunque en la historia de la ciencia, desde un punto de vista formal, durante mucho tiempo solo existieron modelos de tipo determinista, en los que el fenómeno se expresaba mediante leyes matemáticas perfectamente formuladas, de tal manera que conocidas las variables que intervenían en el modelo y su valor, el resultado quedaba completamente determinado. Solo más recientemente se empezó a plantear modelos de tipo probabilístico, en los que conocidas las variables únicamente se calculaba la probabilidad de aparición de un resultado, y en los que se introducía por tanto un margen de incertidumbre. Realmente este tipo de modelos probabilísticos ha venido siendo utilizado desde siempre, sin una formulación matemática precisa por el ser humano, quién en su toma de decisiones se ha basado en la estimación de la probabilidad de que algo ocurra en base a lo que ha observado que ocurrió con anterioridad en situaciones similares.

La construcción de modelos de riesgo de aparición de un suceso es de gran importancia en medicina, tanto para intentar conocer las variables que influyen en que se presente ese suceso, como para analizar el mecanismo que lo produce y para predecir su aparición. En el primer caso, el conocimiento de las variables que influyen permitirá establecer medidas preventivas o terapéuticas, y en el segundo mediante el modelo se puede efectuar cálculos relacionados con la aparición del suceso, por ejemplo para

determinar las necesidades de recursos. Precisamente la teoría matemática para el cálculo de modelos de riesgo tiene su origen probablemente en este último aspecto, y más concretamente en el campo de la ingeniería, donde la demanda creciente de equipos que funcionen cada vez mejor y a un menor coste lleva aparejada la necesidad de disminuir las probabilidades de fallo de éstos.

Cualquier construcción matemática, por sencilla que ésta sea, constituye un modelo y como tal una simplificación de la realidad. Así en términos de supervivencia es habitual utilizar la mediana como dato resumen. Evidentemente en este caso se trata de una simplificación tremenda y probablemente un mejor modelo de la realidad lo constituya una estimación de la supervivencia a lo largo del tiempo, quizás obtenida mediante el método de Kaplan-Meier, lo cual no obstante sigue siendo una gran simplificación, ya que en ese modelo para el cálculo de la supervivencia solo interviene el tiempo y ninguna característica del paciente, las cuales sin ninguna duda pueden influir decisivamente en el resultado, por lo que el modelo se podrá mejorar incluyendo en el mismo el efecto de variables que se cree pueden afectar a la probabilidad de aparición del evento. No se debiera olvidar nunca que, a diferencia de los modelos deterministas propios de las leyes físicas, los modelos biológicos son en su gran mayoría modelos probabilísticos, sujetos a incertidumbre, que además se trata de simplificaciones de la realidad y que efectúan cálculos generales para valores promedio, mientras que la práctica clínica se ejerce sobre pacientes concretos con sus características individuales.

Los modelos matemáticos constituyen sin ninguna duda valiosísimas herramientas para el conocimiento, interpretación y en su caso modificación de los fenómenos, pero casi siempre se trata de modelos transitorios, sujetos a verificación y perfeccionamiento, y como todo en el mundo de la ciencia solo pueden ser aceptados con una cierta dosis de escepticismo y con una mentalidad crítica.

El riesgo cardiovascular absoluto (Villar-Álvarez F, 2005), entendido como la probabilidad de sufrir un evento cardiovascular en un tiempo determinado, es una herramienta recomendada por las guías clínicas actuales (De Backer G, 2003) y cada vez más utilizado por los clínicos en el abordaje terapéutico de los distintos factores de riesgo cardiovascular.

Dentro del concepto de riesgo cardiovascular se incluye la probabilidad de padecer las enfermedades arterioscleróticas más importantes: cardiopatía isquémica, enfermedad cerebrovascular y arteriopatía periférica. Sin embargo, existen múltiples escalas que utilizan diferentes metodologías y diferentes variables para la estimación del riesgo cardiovascular, como es el riesgo de morbimortalidad coronaria -estudios Framingham y Regicor (Anderson KM, 1991, Marrugat J, 2003)-, riesgo de mortalidad cardio y cerebrovascular -Proyecto Evaluación sistemática del riesgo coronario (Conroy RM, 2003)-, riesgo de morbimortalidad cardio y cerebrovascular - Sociedad Europea de Hipertensión (Mancia G, 2007)-, así como escalas especificas para diabéticos (Stevens RJ, 2001). Los riesgos estimados por las diferentes escalas no siempre coinciden y pueden originar cierta confusión y dificultar su aplicabilidad práctica, como se ha podido observar en diferentes estudios (Maiques A, 2004, Buitrago F, 2007).

La evolución en el tiempo del riesgo cardiovascular estimado puede servir para valorar la efectividad de las diferentes intervenciones terapéuticas que se realizan en personas con seguimiento habitual en las consultas de atención primaria. Sin embargo, su envejecimiento progresivo, junto con el aumento del riesgo cardiovascular que conlleva (Bowman TS, 2006, Wang W, 2006) por el importante peso que tiene la edad en todas las escalas de riesgo, puede estar enmascarando el efecto real de las intervenciones realizadas (Coca A, 2006, Baena-Díez JM, 2006). Por esta razón, habitualmente, en el seguimiento a largo plazo que se realiza en el ámbito de la atención primaria de pacientes crónicos, es difícil conseguir mantener la reducción inicial alcanzada con las intervenciones terapéuticas realizadas.

El riesgo relativo (RR) o razón respecto al bajo riesgo, entendido como la razón del riesgo coronario absoluto de cada sujeto y el riesgo de un sujeto de la misma edad y sexo con riesgo coronario bajo (Grundy SM, 1999) podría ser un instrumento útil para analizar la evolución del riesgo a medio y largo plazo, al ser independiente de la edad y del sexo de los pacientes. No obstante, las indicaciones terapéuticas basadas en los grandes ensayos clínicos se han realizado en función del riesgo cardiovascular absoluto.

Aunque existen numerosas publicaciones en las que se estima el riesgo cardiovascular con diferentes escalas y metodologías, en nuestro medio no se han encontrado trabajos en los que se valorara la evolución en el tiempo del riesgo coronario absoluto y relativo estimado y la influencia que puede tener el paso del tiempo en la modificación de dicha estimación.

1.2 Modelos de riesgo cardiovascular

Puesto que las enfermedades cardiovasculares constituyen una de las principales causas de mortalidad y morbilidad en los países desarrollados, es lógico que sea de gran interés el desarrollo de modelos de predicción del riesgo de padecer este tipo enfermedades, tanto para intentar conocer los posibles mecanismos que afectan al aumento del riesgo, como para poder intervenir precozmente, mediante campañas preventivas, o en su momento con tratamientos terapéuticos. Precisamente uno de los factores de riesgo que se asocian con la probabilidad de desarrollar una enfermedad cardiovascular es la presencia de hipertensión.

Para el cálculo de la probabilidad de aparición de un suceso dicotómico (enfermedad si, no) el modelo matemático más habitual se basa en la utilización de la regresión logística, que produce una ecuación en la que conocidos los valores de los diferentes factores de riesgo se puede evaluar la probabilidad de aparición de la enfermedad. Resulta evidente que en muchos procesos dicha probabilidad depende del tiempo de exposición, aumentando a medida que éste transcurre, por lo que o bien el tiempo interviene en la ecuación como factor de riesgo, o bien se utiliza un modelo específico en el que se tenga en cuenta esta característica, calculando ahora la probabilidad de que el suceso ocurra en un momento de tiempo determinado. Esto es precisamente lo que se hace en los modelos probabilísticos de supervivencia, siendo el método más conocido el denominado modelo de riesgos proporcionales o modelo de Cox (Cox, 1972).

Sin embargo no es la única alternativa posible, existiendo otros posibles métodos de modelado denominados paramétricos, debido a que suponen un tipo concreto de ecuación matemática para la función de riesgo, y que aunque en la industria son muy utilizados, sin embargo no es tan normal encontrarlos en la literatura médica.

1.2.1 Estudio Framingham del corazón

Durante 50 años, el estudio de Framingham del corazón y los residentes de Framingham, Massachussets, han sido sinónimo de los notables avances logrados en la prevención de enfermedades del corazón en los Estados Unidos y en todo el mundo. El estudio es uno de los más importantes estudios epidemiológicos en los anales de la medicina americana (Wilson PWF, 1998). Si bien su contribución en la esfera del corazón de la investigación son legión, los investigadores también están utilizando nuevos datos para investigar el accidente cerebrovascular, demencia, osteoporosis, artritis, diabetes, enfermedad de los ojos, el cáncer y los patrones genéticos de las muchas enfermedades comunes.

Antes de Framingham, la mayoría de los médicos no entendían la relación, por ejemplo, entre los altos niveles séricos de colesterol y ataques cardíacos. Muchos no creían que la modificación de ciertos comportamientos podría permitir a sus pacientes evitar o corregir las causas subyacentes de la insuficiencia cardiaca y vascular.

Los investigadores querían saber qué factores biológicos y ambientales contribuían a ese rápido aumento de la discapacidad y muerte cardiovascular.

En su primer año, el estudio fue asumido por el Instituto Nacional del Corazón, ahora Instituto Nacional del Corazón, los Pulmones y la Sangre. A través de un contrato con este instituto, los investigadores de la Escuela de medicina de la universidad de Boston han desempeñado un papel importante en el estudio de Framingham del corazón.

El estudio sigue haciendo importantes contribuciones científicas mediante la mejora de sus capacidades de investigación y de la capitalización de sus recursos inherentes. Nuevas tecnologías de diagnóstico, como la ecocardiografía (una ecografía del corazón), la arteria carótida ultrasonido, la resonancia magnética del corazón y el cerebro, la tomografía computarizada del corazón y sus vasos y densitometría ósea (para la vigilancia de la osteoporosis), se han integrado en anteriores y actuales protocolos.

Como resultado de este gran estudio también tenemos la identificación de los principales factores de riesgo de las enfermedades cardiovasculares. Hoy, la gestión de

los niveles de colesterol, la presión arterial alta y la diabetes para mitigar el corazón y los accidentes cerebrovasculares y la enfermedad vascular es fundamental para una buena atención médica. De hecho, es difícil recordar un momento en que estos y otros factores de riesgo no se consideraban problemas importantes por muchos médicos.

El estudio estableció una relación entre los niveles de colesterol y el riesgo de enfermedades. Además, estableció una fuerte asociación positiva de colesterol LDL (lipoproteínas de baja densidad) con enfermedad coronaria, así como un poderoso efecto protector inverso y de los niveles de HDL (lipoproteínas de alta densidad).

Las investigaciones de la presión arterial pusieron al descubierto una serie de ideas erróneas. En general se consideró que las mujeres y las personas de edad avanzada toleran así presiones más altas. Sin embargo, los investigadores no encontraron nada que sugiera que los ancianos pueden sentirse mejor que las personas más jóvenes en un determinado grado de la hipertensión. Además, las mujeres con alta presión, al igual que los hombres, presentaban un mayor riesgo para las enfermedades del corazón.

Así como Framingham mostró el camino para la identificación de los factores de riesgo asociados con el desarrollo de enfermedad cardiovascular, también se ocupó de los elementos claves del estilo de vida Americana que contribuyen a las altas tasas de morbilidad y discapacidad.

Los investigadores del estudio descubrieron que un estilo de vida caracterizado por una alimentación deficiente, vida sedentaria, y / o aumento de peso sin límites contribuyen a la aparición de los factores de riesgo de enfermedades cardiovasculares.

Antes de Framingham, el tabaquismo no era considerado como un peligro en el desarrollo de enfermedades del corazón. El estudio demostró que los fumadores presentaban un mayor riesgo de tener un infarto de miocardio o una muerte súbita. Además, el riesgo resultó ser relacionado con el número de cigarrillos que fumaba cada día, y además se encontró que el abandono del hábito de fumar reduciría rápidamente a la mitad el riesgo en comparación con aquellos que continuaron fumando.

El estudio ha desempeñado un papel fundamental para influir en los médicos a que hagan más hincapié en la prevención, la detección y el tratamiento de las enfermedades cardiovasculares los factores de riesgo en sus primeras etapas.

Cálculo de del riesgo mediante el modelo de Framingham que utiliza el valor del colesterol total.

Las variables que intervienen son el *SEXO*, la *EDAD* en años, el *COLESTEROL* sérico en mg/dl, fracción de colesterol ligado a lipoproteínas de alta densidad *HDL*, *PRESION SISTOLICA*, *DIABETES* (No, Sí), *FUMADOR* (No, Sí).

En primer lugar hay que calcular el valor de la siguiente expresión:

Para los hombres: $L_H = b_{E1} \cdot EDAD + b_C + b_H + b_T + b_D + b_F$

Para las mujeres $L_M = b_{E1} \cdot EDAD + b_{E2} \cdot EDAD^2 + b_C + b_H + b_T + b_D + b_F$

donde los coeficientes *b* son diferentes para hombres y mujeres y los obtenemos a partir de la siguiente tabla:

Coeficientes para el modelo de Framinghan (Colesterol total)

Coeficiente	Hombres	Mujeres
b_{E1} x **Edad**	0.04826	0.33766
b_{E2} x **(Edad)^2**	0	−0.00268
b_C **Colesterol mg/dl**		
< 160	−0.65945	−0.26138
160–199	0	0
200–239	0.17692	0.20771
240–279	0.50539	0.24385
≥ 280	0.65713	0.53513
b_H **HDL–Col mg/dl**		
< 35	0.49744	0.84312
35 – 44	0.24310	0.37796
45 – 49	0	0.19785
50 – 59	−0.05107	0
≥ 60	−0.48660	−0.42951
b_T **Tensión arterial mmHg**		
PAS < 120 PAD < 80	−0.00226	−0.53363
PAS <130 PAD < 85	0	0
PAS <140 PAD < 90	0.28320	−0.06773
PAS < 160 PAD < 100	0.52168	0.26288
PAS ≥160 PAD ≥100	0.61859	0.46573
b_D **Diabetes**		
NO	0	0
SI	0.42839	0.59626
b_F **Fumador**		
NO	0	0
SI	0.52337	0.29246

Una vez calculado el valor correspondiente de *L*, se le resta la cantidad *G* (función evaluada para los valores medios de las variables en el estudio) diferente para hombres o mujeres:

GHombres = 3.0975

GMujeres = 9.92545

Exponenciamos ese valor calculado *B = exp (L−G)* y determinamos el valor de la expresión 1−SB, donde S es (función de supervivencia base a 10 años), que es diferente para hombres y mujeres:

SHombres = 0.90015

SMujeres = 0.96246

También existe la posibilidad de calcular el riesgo mediante otro modelo que utiliza el valor de LDL−col en lugar del colesterol total, siendo la mecánica de cálculo similar aunque lógicamente varían los coeficientes. El procedimiento está también descrito en el artículo (Wilson P; D'Agostino R, 1998).

Aunque existen gran número de trabajos relativos al estudio de los riesgos de enfermedad cardiovascular, el conocido como estudio de Framingham constituye un pilar básico, y en diferentes formas es ampliamente utilizado para la toma de decisiones terapéuticas en base a la estimación de riesgo proporcionada por el modelo al introducir las características de riesgo del paciente concreto. Es tan popular que incluso existen calculadoras de bolsillo que implementan el algoritmo, y también diferentes páginas Web en las que se puede efectuar dicho cálculo (Conroy RM, 2003, 3000).

1.2.2 Proyecto de Evaluación sistemática del riesgo coronario, SCORE.

Sin embargo se venía observando que el modelo de Framingham sobreestimaba en gran medida el riesgo absoluto de enfermedad cardiovascular cuando se utilizaba en países europeos, caracterizados por una baja incidencia de eventos cardiovasculares respecto al lugar de origen del estudio en Framingham Massachussets Estados Unidos, lo que podía influir al utilizar ese modelo en la decisión de tratar un exceso de pacientes en países como España o Italia, en base a una sobreestimación del riesgo real. Como es lógico esta inquietud crea la necesidad de desarrollar un modelo más adecuado para

este entorno, y así recientemente se ha publicado un trabajo correspondiente a la estimación del riesgo de desarrollar en 10 años una enfermedad cardiovascular fatal en países de Europa (Conroy RM, 2003).

El proyecto SCORE se inició para desarrollar un sistema de evaluación de riesgo cardiovascular para su uso en Europa. En él se calcula, mediante un modelo basado en la función de Weibull, el riesgo de enfermedad cardiovascular fatal en 10 años, estimándose dos ecuaciones diferentes para enfermedades coronarias y no coronarias, y se dispone de dos métodos de evaluación que se diferencian en que en uno de ellos se utiliza el colesterol y en el otro la relación colesterol/HDL, aunque según los autores no hay ventajas aparentes en la utilización de uno u otro método. Aunque se trata de un modelo similar al de Framingham los conceptos utilizados difieren sensiblemente.

Hay que tener en cuenta que se calculan por separado los riesgos de enfermedad coronaria (EC) y no coronaria (ENC), por lo que el riesgo cardiovascular total corresponderá a la suma de ambos. Además se estiman dos modelos diferentes, denominados de alto y bajo riesgo, correspondientes a países con poblaciones con alto y bajo riesgo de enfermedad cardiovascular, estando en este último grupo de bajo riesgo países como España, Italia o Bélgica.

Aquí el procedimiento de cálculo es el siguiente:

En primer lugar se calcula la probabilidad de supervivencia base para EC y ENC para la edad actual del paciente y a los 10 años, de acuerdo a la siguiente ecuación:

$$S_0(Edad) = \exp\left\{-\exp(a)\cdot(Edad-20)^p\right\}$$

$$S_0(Edad+10) = \exp\left\{-\exp(a)\cdot(Edad-10)^p\right\}$$

Los coeficientes *a* y *p* se obtienen de la siguiente tabla:

		Enf. Coronaria		Enf. No Coronaria	
		a	p	a	p
Población de riesgo bajo	Hombre	−22.1	4.71	−26.7	5.64
	Mujer	−29.8	6.36	−31.0	6.62
Población de riesgo alto	Hombre	−21.0	4.62	−25.7	5.47
	Mujer	−28.7	6.23	−30.0	6.42

Se calcula el valor de la siguiente ecuación para enfermedad coronaria y no coronaria:

$$w = b_C(Colesterol - 6) + b_T(TAS - 120) + b_F$$

donde los coeficientes se obtienen de la siguiente tabla

	Enf.Coronaria	Enf. No Coronaria
b_C Colesterol [mmol/l]	0.24	0.02
b_T TAS [mmHg]	0.018	0.022
b_F Fumador = SI	0.71	0.63

El colesterol viene dado en **mmol/l**. Para convertir el valor de mg/dl a mmol/l basta con multiplicar por **0.02586**.

El siguiente paso consiste en calcular la probabilidad de supervivencia con esos factores de riesgo a esa edad y a 10 años:

$$S(Edad) = S_0(Edad)^{\exp(w)}$$

$$S(Edad + 10) = S_0(Edad + 10)^{\exp(w)}$$

Ahora para cada tipo de enfermedad se calcula la probabilidad de supervivencia a los 10 años condicionada a la supervivencia a la edad actual:

$$S_{10}(Edad) = S(Edad + 10) / S(Edad)$$

Siendo entonces el riesgo a 10 años:

$$Riesgo_{10} = 1 - S_{10}(Edad)$$

Así se obtienen dos valores de riesgo REC_{10} para **E**nfermedad **C**oronaria, $RENC_{10}$ para **E**nfermedad **N**o **C**oronaria. El riesgo total corresponderá a la suma de ambos.

Existen otros modelos además de los dos citados, entre los que cabe quizás destacar otro trabajo anterior que incluye pacientes de España, basado en el proyecto REGICOR

(Registro Gironi del Cor) (Pérez G, 1998) donde se usan las ecuaciones de Framingham ajustadas para la población española, el ajuste consiste en sustituir la prevalencia de factores de riesgo cardiovasculares y la tasas de cardiopatías isquémicas de Framingham por las de la población española. Se usó la ecuación de Framingham que incluye el colesterol HDL debido a que la distribución poblacional del colesterol de las lipoproteínas de alta densidad indica que más del 40% de la población tiene cifras por encima de 160 mg/dL.

El poder predictivo se va desvaneciendo a medida que se intenta aplicar a poblaciones un poco diferentes o en las que factores tal vez desconocidos propios de la región o de las variaciones genéticas propias de las familias residentes en ella protegen contra el riesgo o lo empeoran. Lo único claro es que, independientemente de la población, sí parece cierto que al menos los factores de riesgo generalmente aceptados son universales, probablemente varíen en cuanto al peso específico que tengan en una población determinada

Es evidente que aunque estos modelos tienen por objeto valorar el posible riesgo de enfermedad cardiovascular de un sujeto de acuerdo a una serie de características, sin embargo presentan notables diferencias, no sólo en cuanto a las poblaciones estudiadas, sino también en cuanto a la clasificación del tipo de evento cuya probabilidad se pretende calcular, por lo que cualquiera que esté interesado en su utilización debiera como mínimo leer cuidadosamente los artículos originales que describen el modelo.

Realmente resulta curioso que un modelo como el de Framingham se haya popularizado tanto, hasta el punto de que, como ya se ha comentado, incluso existan calculadoras de bolsillo para determinarlo, similares a las que aparecieron con el cambio de moneda al euro. Por supuesto en dichas calculadoras no se explica cómo ha sido determinado el modelo, ni para qué tipo de evento cardiovascular se calcula la probabilidad, ni la sobreestimación que se obtiene para países como el nuestro del área mediterránea.

1.2.3 Proyecto "Proyección del Centro de Desarrollo Electrónico hacia la Comunidad", PROCDEC

Como parte de un proyecto de investigación conjunta entre la Universidad Central de Las Villas y la Universidad de Oviedo en España, hace algunos años se creó el proyecto titulado "Proyección del Centro de Desarrollo Electrónico hacia la Comunidad", PROCDEC, cuyo objetivo principal es desarrollar un estudio de personas supuestamente normotensas, primero en la ciudad de Santa Clara y luego en toda la nación. Especialistas en medicina integral realizan el estudio del paciente, mientras un grupo multidisciplinario realiza posteriormente el diagnóstico.

Como resultado del proyecto se desarrolla el sistema computacional Tensoft II v1.0 para automatizar el proceso de diagnóstico de esta enfermedad. Este sistema ha sido implementado en 5 policlínicos de nuestra provincia con resultados satisfactorios, sin embargo el desarrollo acelerado de las nuevas tecnologías y la necesidad de extender este estudio a otros lugares de nuestra provincia y del país, conllevó al desarrollo, en el año 2006, de una versión mejorada del sistema.

El proyecto posee datos de 849 personas supuestamente sanas, de ellos se tiene únicamente una primera observación, es decir, no existe un antes y un después de la aplicación de las primeras pruebas a las personas. No se le ha dado aún seguimiento al estudio en cuanto a la repetición de las primeras pruebas. Como no se tienen mediciones antes y después es que se puede afirmar que no se pueden aplicar los modelos de Framingham basados en regresión de Cox y análisis de supervivencia. Sin embargo se pueden obtener índices "transversales", o sea, índices en un momento del tiempo aplicando técnicas multivariadas.

En materia de riesgo en Cuba, en la ciudad de Santa Clara en el Cardiocentro "Ernesto Che Guevara ", se ha trabajado sobre la base de comparaciones, más precisamente comparaciones del riesgo coronario según las ecuaciones de Framingham-Wilson (original) y Framingham-REGICOR (calibrada española)(Alberto Morales Salinas).
Como resultado de este trabajo, para la muestra de 113 profesores de la universidad sin antecedentes personales de cardiopatía isquémica, aparentemente de bajo riesgo; se concluyó que casi un tercio de los trabajadores analizados tenían un riesgo coronario

moderado o alto según la Framingham-Wilson. He aquí la importancia del cálculo del índice de alto riesgo cardiovascular.

Conclusiones parciales

Como se ha podido ver, a nivel mundial se han realizado varios estudios sobre el cálculo de índices de riesgo para una determinada población. En este sentido pueden mencionarse: Framingham (Wilson PWF, 1998), REGICOR (Pérez G, 1998), entre otros.

Debido a la ausencia de mediciones en el tiempo, en estos momentos resulta imposible realizar el ajuste del índice de Framingham a la población de Santa Clara, utilizando el método de regresión de Cox, pero en su lugar pueden calcularse otros índices de riesgo y en particular índices que permitan cuantificar el “alto riesgo cardiovascular”.

Capítulo 2. Métodos estadísticos para el cálculo del índice de alto riesgo cardiovascular

2.1 Test Chi-Cuadrado. Tablas de contingencia

En las ciencias sociales, de la salud y el comportamiento es bastante frecuente encontrarse con variables categóricas. El sexo, la raza, el padecimiento de cierta enfermedad o de un determinado síntoma, la categoría laboral, por solo citar algunos, son ejemplos de algunas variables categóricas que podemos encontrar. Las mismas son variables sobre las cuales únicamente se pueden obtener una medida discreta con pocos valores, con orden (ordinal), o sin él (nominal).

Cuando se trabaja con variables categóricas, los datos suelen organizarse en tablas de doble entrada en las que cada entrada representa un criterio de clasificación (una variable categórica). Como resultado de esta clasificación, las frecuencias aparecen organizadas en casillas que contienen información sobre la relación existente entre ambos criterios. A estas tablas de frecuencias se les llama Tablas de Contingencia (Alvarez A, 2005).

Las Tablas de Contingencia (1994)tienen dos objetivos fundamentales:

1-Organizar la información contenida en un experimento cuando esta es de carácter bidimensional, o sea cuando está referida a dos factores (variables cualitativas).

2-Analizar si existe alguna relación de dependencia e independencia entre los niveles de las variables objeto de estudio (Vicéns Otero, 2005).

Para identificar relaciones de dependencia entre variables no pueden utilizarse solamente las Tablas de Contingencia. Prácticamente desde el surgimiento de la raza humana, el hombre se ha preocupado por conocer y entender el mundo que le rodea, descubrir las relaciones y leyes que lo rigen, para de esta manera, orientarse hacia el futuro en busca de una vida mejor. Esta es la razón por la cual estudia los diferentes fenómenos observables, buscando en ellos nexos y relaciones que permitan explicar causas y efectos. En el estudio de las dependencias entre causas y efectos, es importante analizar diferentes características involucradas en ellos. Briones (1987), ayuda en este sentido definiendo el concepto de variable como una propiedad, característica o atributo que puede darse en ciertos objetos o sujetos (García, 2007).

Para ello se debe utilizar alguna medida de asociación, acompañada de su correspondiente prueba de significación. Un ejemplo de ello es el estadístico (Chi-Cuadrado) propuesto por Pearson desde 1911. El mismo permite contrastar las hipótesis de que las dos variables utilizadas son independientes. Cuando dos criterios de clasificación son independientes, las frecuencias esperadas (m_{ij}) se estiman de la siguiente manera:

$$m_{ij} = \frac{(total\ de\ la\ \ fila\ i) \times (total\ de\ la\ columna\ j)}{N\,(total\ de\ casos)} \qquad (2.1)$$

Una vez obtenidas las frecuencias esperadas se calcula el estadístico de la siguiente manera:

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - m_{ij})^2}{m_{ij}} \qquad (2.2)$$

donde: n_{ij} representa las frecuencias que se observan en la tabla de contingencia.

El valor de χ^2 calculado se compara con el valor tabulado de una χ^2 para un nivel de confianza determinado y $(n-1)*(k-1)$ grados de libertad, donde *n* y *k* son el número de filas y columnas respectivamente. Si el valor calculado es mayor que el valor de una $X^2_{(n-1)(k-1)}$, significará que las diferencias entre las frecuencias observadas y las frecuencias teóricas o esperadas son muy elevadas y por tanto se concluirá con un determinado nivel de confianza que existe dependencia entre los factores o atributos analizados (Vicéns Otero, 2005, Alvarez A, 2005).

El estadístico Chi-Cuadrado debe cumplir ciertas exigencias. Los datos debe provenir de muestras aleatorias con distribuciones multinomiales y las frecuencias esperadas en cada celda no deben ser excesivamente pequeñas. Tradicionalmente se ha recomendado que las frecuencias esperadas deban ser mayores o iguales a 5 aunque quizás esto sea muy exigente. Lo importante es cómo evitar en general este problema. En esencia, las tablas de contingencia no pueden tener dimensiones demasiado grandes. Si se quiere eliminar frecuencias esperadas bajas, se deben reducir las dimensiones de la tabla. Para mejorar la aproximación de la distribución en una

tabla 2×2, se utiliza frecuentemente la "corrección por continuidad de Yates"(1934). Esta corrección consiste en reducir en 0.5 el valor absoluto de las diferencias $n_{ij} - m_{ij}$ del estadístico X^2 antes de elevarlas al cuadrado. Algunos autores como Conover y Mantel (1974) sugieren que, con muestras pequeñas, esta corrección permite que el estadístico X^2 se ajuste mejor a las probabilidades de la distribución X^2 pero en honor a la verdad no existe aun un consenso generalizado por lo que han desarrollado una controversia sobre los méritos o insuficiencias de esta corrección, pero pese a todo, es bastante usada.

Se puede utilizar, en lugar del Test Chi-cuadrado de Pearson, un Test basado en la distribución hipergeométrica y en la hipótesis de independencia, conocido como Test exacto de Fisher (1935). El mismo ofrece la probabilidad exacta de obtener las frecuencias de hecho obtenidas o cualquier combinación alejada de la hipótesis de independencia.

Otros estadísticos que se utilizan para el análisis de la relación entre las variables son el Chi-cuadrado de razón de verosimilitud y el Chi-cuadrado de asociación lineal.

El Chi-cuadrado de razón de verosimilitud (Fisher, 1924, Pearson., 1928) se obtiene mediante la fórmula:

$$Razon\,de\,Verosimilitud = 2\sum_i \sum_j n_{ij} \log\left(\frac{n_{ij}}{m_{ij}}\right) \qquad (2.3)$$

En este caso estamos en la presencia de un estadístico asintóticamente equivalente al Chi-Cuadrado aunque con una forma más complicada, el mismo es muy utilizado para estudiar las relaciones entre variables categóricas, particularmente en el contexto de los modelos log-lineales.

El Chi-cuadrado de asociación lineal mide además la tendencia a la dependencia lineal entre las variables de fila y columna en una tabla de contingencia y no debe utilizarse para datos nominales porque realmente se calcula multiplicando el cuadrado del coeficiente de correlación de Pearson por el número de casos menos 1. Tiene su propio

grado de libertad. En epidemiología es conocido como la prueba de Chi Cuadrado de Mantel-Haenszel. (Cristóbal J, 2005, Álvarez A, 2005)

2.1.1 Algunas medidas de asociación entre variables aleatorias discretas nominales y ordinales

En muchas investigaciones, más que discernir la dependencia de dos variables interesa la naturaleza y fortaleza de la asociación. Los indicadores que miden esto se llaman *medidas de asociación*. En general ninguna medida resume adecuadamente todos los tipos de asociación posibles. Las medidas se diferencian por su interpretación, y por la forma en que ellas pueden reflejar asociaciones perfectas o parciales. Las medidas difieren además por la forma en que ellas se afectan por otros factores, por ejemplo los totales marginales. Así, hay medidas que son muy "sensibles a los marginales", en el sentido que su valor está muy influenciado por la distribución de los totales marginales de filas y columnas. Tales medidas reflejan por tanto información sobre los marginales, además de sobre la asociación.

Una medida de asociación determinada puede tener un valor bajo para una tabla dada, pero para ello no significa que las variables objeto de estudio no estén relacionadas, sino que ellas no están relacionadas en la forma a la que es sensible dicha medida. Por ello ninguna medida individual es la mejor para todas las situaciones. Al seleccionar una, debe tenerse en cuenta el tipo de datos, la hipótesis de interés así como las propiedades de cada medida. A veces se calcula un gran número de medidas y luego se referencian aquellas que más "ayudan a respaldar" las hipótesis del investigador, como si fueran las únicas que se hubieran calculado; pero esto no es correcto a menos que haya un verdadero análisis de cuáles medidas son las que se necesita observar.

Las medidas nominales (se asumen apenas que las dos variables de la tabla están medidas nominalmente) pueden suministrar solamente alguna indicación sobre la estrechez de la asociación y no pueden indicar casi nada sobre la dirección o cualquier otra cosa de la naturaleza de la relación.

El test Chi-cuadrado en sí, no proporciona una buena medida del grado de asociación entre las dos variables; pero como está tan expandido el uso del Chi-cuadrado en la

dócima de independencia, se ha estimulado la definición de medidas de asociación basadas en el Chi-cuadrado tratando de minimizar la influencia del volumen de la muestra y de los grados de libertad, así como restringir el rango de los valores de la medida al intervalo [0-1]. Estas medidas ayudan entonces a comparar los resultados del Chi-cuadrado en tablas diferentes cuando hay variación de las dimensiones y de los volúmenes de las muestras. Sin estas correcciones, no se puede comparar con el Chi-cuadrado tales tablas. El llamado coeficiente Phi modifica el Chi-cuadrado dividiéndolo por el volumen de la muestra y extrayendo la raíz cuadrada del resultado:

$$\phi = \sqrt{\frac{\chi^2}{N}} \qquad (2.4)$$

Para tablas en las cuales una dimensión es mayor que 2, Phi no yace necesariamente entre 0 y 1, ya que el valor del Chi-cuadrado puede ser más grande que el volumen de la muestra. Por tanto Phi sólo queda estandarizado en el intervalo [0-1] en tablas en las cuales $R = 2$ ó $C = 2$ (*R* y *C* denotan siempre el número de filas y columnas).

Para obtener una medida que debe yacer siempre entre 0 y 1 Pearson sugirió el uso de:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} \qquad (2.5)$$

que es conocido como coeficiente de contingencia. El valor de este coeficiente está siempre entre 0 y 1; pero generalmente no puede llegar a ser 1. De hecho, el máximo valor de *C* depende sobre el número de filas y columnas. Por ejemplo para una tabla 4×4, el máximo valor de *C* es 0.87. Cramer introdujo la siguiente variante:

$$V = \sqrt{\frac{\chi^2}{N * (K - 1)}} \qquad (2.6)$$

donde $K = \min(\textit{número de filas}, \textit{número de columnas})$

El estadístico, conocido como V de Cramer, puede alcanzar el máximo de 1 en tablas de cualquier dimensión y además coincide con Phi si una de las dimensiones es igual a 2.

Entre las tres mencionadas se recomienda la V de Cramer. Su carácter estandarizado permite al menos comparar la "estrechez de la asociación" entre tablas diferentes; pero esta estrechez o fortaleza de la asociación que estamos comparando, no responde a ningún concepto intuitivo claro de asociación (Press, 2005).

La idea de la Reducción Proporcional en el error fue introducida por Goodman y Kruskall (1954). Con este tipo de medidas, el significado de la asociación resulta más claro. En esencia estas medidas indican relación entre:

- La magnitud o probabilidad del error al predecir los valores de una variable basándonos en el conocimiento sólo de esa variable.
- La magnitud o probabilidad del error al predecir los valores de esa variable basándola en el conocimiento de otra variable adicional.

El estadístico lambda de Goodman y Kruskal siempre toma valores entre 0 y 1. Un valor de 0 significa que la variable independiente no ayuda en la predicción de la variable dependiente. Un valor de 1 significa que la variable independiente permite pronosticar exactamente las categorías de la dependiente (la perfección ocurre solamente cuando en cada fila hay solamente una celda diferente de cero). Cuando dos variables son independientes, lambda es 0; pero un valor 0 para lambda, no significa que las variables tengan independencia estadística. Debe recordarse que no existe ninguna medida sensible a todos los tipos imaginables de asociación.

2.2 Pruebas no paramétricas

El uso de pruebas no paramétricas se ha generalizado asombrosamente desde hace ya algún tiempo (Siegel, 1970). A estas pruebas a menudo se les llama “distribuciones libres”, debido a que uno de sus principales méritos es que no suponen que las variables se ajusten a una distribución determinada. Otra de sus ventajas es su sencillez y robustez (Siegel, 1970).

Dentro de las pruebas no paramétricas, una de las más utilizadas es la U de Mann-Whitney para la comparación de dos grupos independientes. Ella es una de las alternativas de la prueba paramétrica de *t* de *Student* más usadas (Weiss, 2002).

La prueba de Mann Whitney, (denominada también test de suma de rango de Wilcoxon) trabaja con variables ordinales o en particular dicotómicas porque su esencia es ranquear los valores de las variables. Su fundamentación matemática, se resume a continuación:

Sean X_1 y X_2 variables ordinales independientes con distribución cualquiera desconocida. Supongamos que se quiere verificar la hipótesis de que sus dos distribuciones son coincidentes, en el sentido de que los rangos de los valores que aparecen en las respectivas muestras no difieren significativamente (SPSS 10).

El test se basa en el ranqueo de los datos de la muestra total (compuesta de dos grupos) y la observación de si estos valores ranqueados de un grupo y del otro se intercalan adecuadamente como una medida de que las distribuciones no difieren.

El criterio de Mann-Whitney parte de determinar el número de veces que un valor del grupo más pequeño precede a un valor del grupo más grande. Si los volúmenes de las muestras son iguales analiza las dos orientaciones y toma la menor. Mann-Whitney construye el estadístico U como el número mínimo de estas dos determinaciones. Si las distribuciones de las variables son iguales el estadístico U no debe ser demasiado grande.

Para muestras pequeñas se puede determinar la distribución del estadístico U condicionada a la hipótesis fundamental y construir un test con probabilidad exacta. Para muestras grandes, a partir de U se construye el estadígrafo:

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 * (n_1 + n_2 + 1}{12}}} \qquad (2.7)$$

Donde n_1 es el volumen de la muestra más pequeña y n_2 el de la más grande y se demuestra que Z tiene aproximadamente distribución normal normalizada si la hipótesis fundamental es cierta.

El criterio de la suma de rango de Wilcoxon consiste en calcular la suma W de los rangos para el grupo de volumen menor (o para el primer grupo, si las dos muestras

tienen igual volumen). Si la hipótesis fundamental es cierta, esta suma W debería ser aproximadamente la mitad de la suma total de los rangos en la muestra completa.

Para muestras pequeñas, la distribución de W se determina con precisión y se pude construir un test exacto. Para muestras grandes, se construye el estadístico.

$$Z = \frac{W - \frac{n_1(n_1 + n_2 + 1}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} \qquad (2.8)$$

Tiene también aproximadamente distribución normal normalizada cuando la hipótesis fundamental es cierta.

Se demuestra que ambos criterios conducen a la misma significación y por ello se habla indistintamente del Test de Rangos de Mann-Whitney o del Test de Suma de Rangos de Wilcoxon. La mayoría de los paquetes estadísticos lo conocen como el test de Mann-Whitney para distinguirlo del test de Wilcoxon de diferencias ranqueadas.

2.3 Análisis discriminante

El problema de la clasificación es uno de los primeros que aparecen en la actividad científica y constituye un proceso consustancial con casi cualquier actividad humana, de tal manera que en la resolución de problemas y en la toma de decisiones la primera parte de la tarea consiste precisamente en clasificar el problema o la situación, para después aplicar la metodología correspondiente y que en buena medida dependerá de esa clasificación. Por supuesto también es así en la medicina, ciencia en la que el diagnóstico constituye una parte primordial, siendo una fase previa para la aplicación de la terapia. En la medicina, diagnosticar es equivalente a *clasificar* a un sujeto en una patología concreta en base a los datos correspondientes de su reconocimiento, exploración y pruebas complementarias. Cuando hablamos de clasificar a un sujeto en un grupo determinado, a partir de los valores de una serie de parámetros medidos u observados, y esa clasificación tiene un cierto grado de incertidumbre, resulta razonable pensar en la utilización de una metodología probabilística, que nos permita cuantificar esa incertidumbre (SPSS 10).

Cuando los grupos están bien definidos y se trata de determinar un criterio para etiquetar cada individuo como perteneciente a alguno de los grupos, a partir de los valores de una serie limitada de parámetros. En este caso las técnicas más utilizadas se conocen con el nombre de análisis discriminante (AD), aunque como alternativa también se puede usar la regresión logística.

Uno de los problemas más simples en cuanto a metodología de clasificación es el de etiquetar a un sujeto como enfermo SI/NO (en nuestro caso sería clasificar al sujeto como de alto, moderado o bajo riesgo cardiovascular), a partir del resultado de una prueba diagnóstica. Pero en casi cualquier actividad, no solo científica, es raro que las cosas sean tan simples y que se maneje una sola variable para tomar la decisión de clasificar; lo habitual será disponer de un conjunto de variables, y entonces resulta ideal utilizarlas de forma conjunta, lo que nos conduce a un enfoque multivariado de la cuestión.

En el análisis discriminante estudiamos las técnicas de clasificación de sujetos en grupos ya definidos. Partimos de una muestra de N sujetos en los que se ha medido p variables cuantitativas independientes, que son las que se utilizarán para tomar la decisión en cuanto al grupo en el que se clasifica cada sujeto, mediante el modelo matemático estimado a partir de los datos. Dentro del análisis discriminante nos encontramos a su vez con dos enfoques diferentes, uno al que se denomina predictivo y otro explicativo. Estos dos enfoques están determinados por los objetivos que se quieren lograr con el análisis, ellos son:

- Analizar si existen diferencias entre los grupos en cuanto a su comportamiento con respecto a las variables consideradas y averiguar en que sentido se dan dichas diferencias (descriptivo o explicativo).
- Elaborar procedimientos de clasificación sistemática de individuos de origen desconocido, en uno de los grupos analizados (predictivo).

En el análisis discriminante predictivo se trata de estimar a partir de los datos unas ecuaciones que aplicadas a un nuevo sujeto, para el que se determinan los valores de las diferentes variables, pero del que se desconoce a qué grupo pertenece, nos proporcionen una regla de clasificación lo más precisa posible. Se trata pues de

formular un algoritmo por el que se pueda determinar a qué grupo pertenece una nueva observación. Este tipo de análisis para nosotros constituye la base del trabajo puesto que nuestro objetivo primero es lograr un índice general para predecir el riesgo cardiovascular en sujetos de los que se conoce un conjunto de datos que reflejan en cierta medida su estado de salud.

A diferencia del anterior, en el análisis discriminante descriptivo estamos más interesados en las variables empleadas para diferenciar los grupos, en las variables explicativas, y lo que se desea es determinar cuáles de esas variables son las que más diferencian a los grupos, cuales son importantes y cuales no a efectos de clasificar los sujetos. Este tipo de análisis es igualmente importante puesto que permite obtener cuáles de las variables que tenemos son las más importantes en la clasificación de que un sujeto tenga alto, moderado o bajo riesgo cardiovascular (SPSS 10).

El caso de dos grupos

El análisis discriminante permite diferenciar entre cualquier número de grupos. Sin embargo, por simplicidad, se comenzará con el caso de dos grupos, para ampliar posteriormente el razonamiento a k grupos.

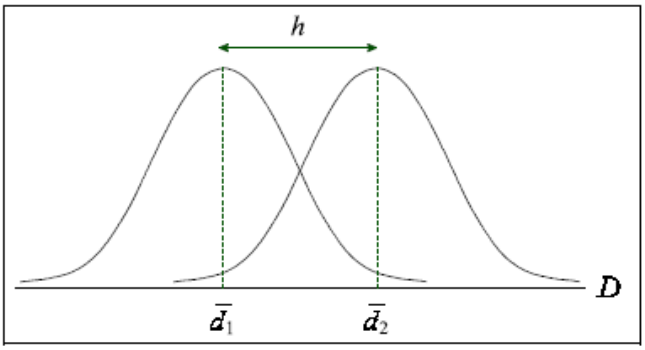
Supóngase que se tienen dos variables independientes X_1 y X_2 capaces de diferenciar lo más posible a ambos grupos. El propósito del análisis discriminante consiste en aprovechar la información contenida en dichas variables para crear una función D que sea una combinación lineal de X_1 y X_2 y que diferencie lo más posible a ambos grupos. La función discriminante tiene la forma:

$$D = b_1 X_1 + b_2 X_2$$

Donde b_1 y b_2 son las ponderaciones de las variables independientes que consiguen hacer que los sujetos de uno de los grupos obtengan puntuaciones máximas en D, y los sujetos del otro grupo puntuaciones mínimas.

Una vez hallada la función discriminante D, se puede realizar una representación gráfica de la función discriminante. Los grupos aparecen representados por sus histogramas y las proyecciones de los centroides aparecen marcadas por líneas de puntos.

Figura 2.1 Histogramas de cada grupo y centroides representados sobre la función discriminante



Sustituyendo en la función discriminante el valor de las medias del grupo 1 en las variables X_1 y X_2, se obtiene el centroide del grupo 1: $\bar{d}_1 = b_1\bar{x}_1^{(1)} + b_2\bar{x}_2^{(1)}$

De igual modo, sustituyendo las medias del grupo 2, se obtiene el centroide del grupo 2: $\bar{d}_2 = b_1\bar{x}_1^{(2)} + b_2\bar{x}_2^{(2)}$

La función D debe ser tal que la distancia h entre los dos centroides sea máxima, consiguiendo de esta forma que los grupos estén lo más distantes posible. Esta distancia puede expresarse de la siguiente manera: $h = \bar{d}_1 - \bar{d}_2$ donde $\bar{d}_1$ y $\bar{d}_2$ son las medias del grupo 1 y del grupo 2 en la función D.

De manera general se desea reducir la dimensionalidad de las p variables independientes a una sola dimensión (la de la combinación lineal D) en la que los grupos se diferencien lo más posible. Las puntuaciones de los sujetos en esa nueva dimensión (denominadas puntuaciones discriminantes) serán las que nos permitan llevar a cabo la clasificación de los sujetos.

Es importante señalar que los grupos deben diferenciarse de antemano en las variables independientes. El análisis busca diferenciar los dos grupos al máximo combinando las variables independientes pero si los grupos no difieren en las variables independientes, el análisis será infructuoso: no podrá encontrar una dimensión en la que los grupos difieran. Dicho de otro modo, si el solapamiento entre los casos de ambos grupos es excesivo, los centroides se encontrarán en la misma o parecida ubicación en el espacio

p-dimensional y, en esas condiciones, no será posible encontrar una función discriminante útil para la clasificación.

Los supuestos del análisis son los mismos que los del análisis de regresión múltiple.

Métodos para hallar las funciones discriminantes

Las variables independientes pueden incorporarse a la función discriminante utilizando dos estrategias distintas. Por defecto, el SPSS utiliza una estrategia de inclusión forzosa de variables que permite construir la función discriminante incorporando todas las variables independientes incluidas en el análisis.

Método paso a paso:

En la estrategia de inclusión por pasos, las variables independientes van siendo incorporadas paso a paso a la función discriminante tras evaluar su grado de contribución individual a la diferenciación entre los grupos. Las opciones de este apartado permiten seleccionar el estadístico que será utilizado como método de selección de variables:

1. **Lambda de Wilks**: Cada variable independiente candidata a ser incluida en el modelo se evalúa mediante un estadístico que mide el cambio que se produce en el valor de la lambda de Wilks al incorporar cada una de las variables al modelo. En cada paso se incorpora al modelo la variable que produzca el mayor cambio en la lambda de Wilks.
2. **Varianza no explicada**: Utiliza como criterio de inclusión la suma de la variación entre todos los pares de grupos no explicada por las variables ya incluidas en el modelo. Se incorpora al modelo la variable que minimiza la cantidad de varianza no explicada. La cantidad de varianza explica por el modelo, R^2, es proporcional, en una constante c, a la distancia H de Mahalanobis.
3. **Distancia de Mahalanobis**: Se incorpora en cada paso la variable que maximiza la distancia de Mahalanobis(Mahalanobis, 1936) entre los dos grupos más próximos.

4. **Menor razón F**: Se incorpora en cada paso la variable que maximiza la menor razón F para las parejas de grupos. El estadístico F utilizado es la distancia de Mahalanobis ponderada por el tamaño de los grupos.
5. **V de Rao**: El estadístico V de Rao(Rao, 1952) es una transformación de la traza de Lawley-Hotelling que es directamente proporcional a la distancia entre los grupos. Al utilizar este criterio, la variable que se incorpora al modelo es aquella que produce un mayor incremento en el valor de V.

Cualquiera que sea el método seleccionado, en la estrategia de inclusión por pasos siempre se comienza seleccionando la mejor variable independiente desde el punto de vista de la clasificación (es decir, la variable independiente en la que más se diferencian los grupos). Pero esta variable sólo se selecciona si cumple el criterio de entrada. A continuación, se selecciona la variable independiente que, cumpliendo el criterio de entrada, más contribuye a conseguir que la función discriminante diferencie a los grupos, etc. Cada vez que se incorpora una nueva variable al modelo, las variables previamente seleccionadas son evaluadas nuevamente para determinar si cumplen o no el criterio de salida. Si alguna variable de las ya seleccionadas cumple el criterio de salida, es expulsada del modelo.

El problema de la clasificación

En los apartados precedentes se ha estudiado básicamente, cómo construir o estimar la función discriminante. Si nuestro objetivo consiste en averiguar en qué difieren dos grupos, con lo visto hasta ahora es más que suficiente. Sin embargo, la mayor utilidad de una función discriminante radica en su capacidad para clasificar nuevos casos.

La clasificación de casos es algo muy distinto de la estimación de la función discriminante. De hecho, una función perfectamente estimada puede no pasar de una pobre capacidad clasificatoria.

Una vez obtenida la función discriminante podemos utilizarla, en primer lugar, para efectuar una clasificación de los mismos casos utilizados para obtener la función: esto permitirá comprobar el grado de eficacia la función desde el punto de vista de la clasificación. Si los resultados son satisfactorios, la función discriminante podrá

utilizarse, en segundo lugar, para clasificar futuros casos de los que, conociendo su puntuación en las variables independientes, se desconozca el grupo al que pertenecen.

En el análisis discriminante predictivo se trata de estimar a partir de los datos unas ecuaciones que aplicadas a un nuevo sujeto, para el que se determinan los valores de las diferentes variables, pero del que se desconoce a qué grupo pertenece, nos proporcionen una regla de clasificación lo más precisa posible. Se trata pues de formular un algoritmo por el que se pueda determinar a qué grupo pertenece una nueva observación. Este tipo de análisis para nosotros constituye la base del trabajo puesto que nuestro objetivo primero es lograr un índice general para predecir el riesgo cardiovascular en sujetos, dado que se conozca de ellos un conjunto de datos de su estado de salud y otros.

2.4 Regresión logística

La regresión logística (RL) es uno de los instrumentos estadísticos más expresivos y versátiles de que dispone para el análisis de datos en áreas del conocimiento como la clínica y la epidemiología, su origen se remonta a la década del 60.

Su uso se universaliza y se expande desde principios de la década del 80 debido especialmente a las facilidades informáticas con que se cuenta desde entonces. En los últimos años se ha verificado una presencia muy marcada de esta técnica, tanto en literatura orientada a tratar temas metodológicos como en artículos biomédicos.

Cuando se desea conocer cómo una serie de factores influyen en una variable cualitativa o categórica dicotómica, es decir con dos posibilidades, como por ejemplo:

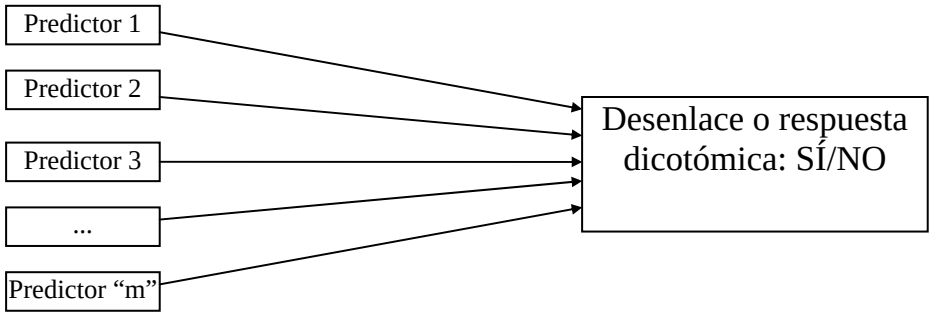
- Un paciente hospitalizado muere o no antes del alta
- Un sujeto operado se infecta o no durante cierto lapso postoperatorio.

En situaciones como la antes mencionadas, suele interesar la evaluación del efecto de uno o más antecedentes sobre el hecho de que el acontecimiento se produzca.

Se utilizará la regresión logística cuando se tenga *una variable dependiente dicotómica*, es decir se utilizará la regresión logística para resolver un problema de clasificación dicotómico. Esta situación es muy frecuente, ya que muchas veces en las investigaciones biomédicas o epidemiológicas se desean identificar los predictores de la

ocurrencia de un determinado fenómeno (que ocurra un suceso o no ocurra). Todas las variables que son candidatas a predecir la ocurrencia de ese fenómeno se utilizarían como variables independientes en un modelo de regresión logística, como muestra la figura 1.

Figura 2.2 Aplicación de la regresión logística



Lo que se procura mediante la regresión logística es en principio, expresar la probabilidad de que ocurra un hecho en cuestión como función de ciertas variables (supongamos que son m) que se presumen relevantes o excluyentes.

El caso más simple es aquel en que se incluye una sola variable independiente:

$$\ln\left(\frac{p}{1-p}\right) = a + b_1 x_1 \qquad (2.9)$$

El caso más general es el que presenta la ecuación logística siguiente:

$$\ln\left(\frac{p}{1-p}\right) = a + b_1 x_1 + b_2 x_2 + \ldots + b_m x_m \qquad (2.10)$$

donde trabajando algebraicamente podemos llegar a la siguiente expresión:

$$P = \frac{1}{1 - \exp(-a - b_1 X_1 - b_2 X_2 - \ldots - b_m X_m)} \qquad (2.11)$$

Donde $a, b_1, b_2, \ldots\ldots\ldots b_m$ son los llamados parámetros del modelo y **exp** denota la función exponencial. La expresión anterior se conoce como función logística.

Al construir el modelo, las variables explicativas (también llamadas covariables) pueden ser de cualquier naturaleza: dicotómicas, continuas o nominales. Esta flexibilidad en

cuanto a información de entrada constituye uno de sus mayores atractivos. (Morales, 2003)

Se trata de un contexto muy parecido al de la regresión múltiple, la diferencia es que ahora hemos sustituido la variable dependiente ("y") por otra expresión. Ahora la variable dependiente no tiene un sentido numérico en sí misma, sino que es el logaritmo neperiano (ln) de la probabilidad (p) de que ocurra un suceso, dividido por la probabilidad de que no ocurra, (1-p). Al cociente $\frac{P}{1-P}$ en inglés se le llama"odds"**,** que se ha querido traducir por "ventaja".

$$odds = \frac{P}{1 - P} \qquad (2.12)$$

Es más fácil calcular una odds que definirla. Se calcula una odds dividiendo el número de quienes tienen una característica por el número de quienes no la tienen. Si en un estudio hay 50 pacientes reclutados en un centro de salud y 25 que no proceden del mismo, son de un hospital; la odds de proceder del centro de salud es 2. Esto significa que hay el doble de pacientes que vienen del centro de salud que del hospital.

$$odds_{\text{Centro de Salud}} = \frac{\text{nº pacientes del Centro de Salud}}{\text{nº pacientes que no son del Centro de Salud}} = \frac{50}{25} = 2$$

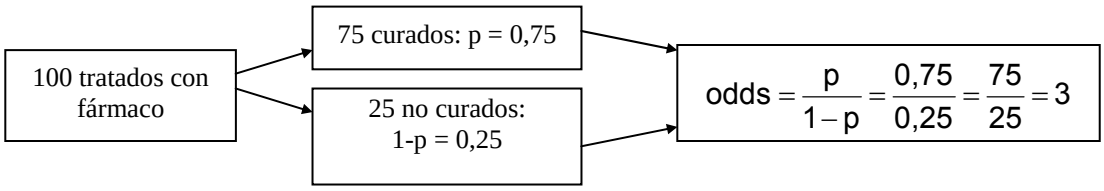
Por tanto, para calcular una odds basta con dividir el número de individuos con la característica de interés por el número de individuos que carecen de ella.

Conceptos de odds y odds ratio

El ejemplo de la figura 2 ayudará a entender los conceptos de *odds* y *odds ratio*. Supongamos que en una muestra de 100 pacientes que han recibido un fármaco se ha alcanzado éxito en 75 de ellos. Si se divide la probabilidad de curación (p = 75/100 = 0,75) por la probabilidad de no curación (1-p = 25/100 = 0,25), se obtendrá la odds de curación para ese tratamiento, que valdría 3, que es el resultado de dividir 75% entre 25% (odds = 0,75/0,25 = 3), o bien simplemente dividir 75 entre 25 ¿Cómo se interpreta una odds de 3 en el ejemplo? Se entendería que por cada paciente en que no se alcanzó el

éxito terapéutico hay 3 en que sí se logró, es decir, con ese tratamiento la probabilidad de éxito es 3 veces mayor que la de fracaso. Tienen una ventaja de 3 para curarse.

Figura 2.3 Concepto de ventaja (*odds*): 75 curaciones en 100 pacientes tratados con un fármaco.



Aunque este concepto de "odds" pueda parecer al principio extraño, se maneja con gran frecuencia en el mundo anglosajón, por ejemplo en el lenguaje de las apuestas. Supongamos que un caballo ha ganado en la última temporada un 80% de las carreras y ha perdido (no ha ganado) un 20%. La odds de ese caballo sería de 4. Cuando se escucha en una película que las apuestas van 4 a 1, se interpretaría que este caballo tiene un 80% de probabilidades de ganar.

Para transformar una odds en una proporción el proceso es a la inversa.

$$\text{Proporción} = \frac{\text{odds}}{1 + \text{odds}} \qquad (2.13)$$

Si la odds de curarse con un tratamiento (figura 2) es de 3, la proporción sería:

$$\text{Proporción} = \frac{3}{1+3} = \frac{3}{4} = 0{,}75\,(75\%)$$

Tanto las proporciones como las odds expresan lo mismo pero usando dos escalas numéricas distintas: las proporciones oscilan entre 0 y 1 y las odds entre 0 e infinito. A veces interesa pasar de una escala a otra, utilizándose para ello las expresiones que hemos visto: $odds = p/(1-p)$ y $prop = odds/(1+odds)$.

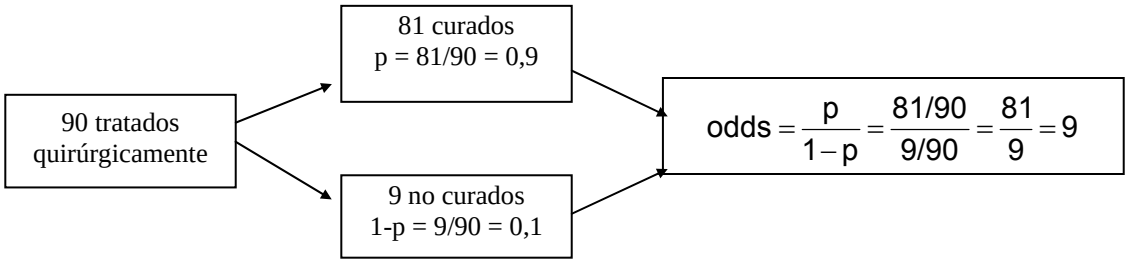
Sabiendo lo que es una *"odds"*, ahora estudiaremos lo que es una *odds ratio*. La traducción más lógica es *razón de odds o razón de ventajas*. Pero el término *odds ratio*, que es cada vez más utilizado en la literatura médica, ha recibido diversas traducciones al castellano: razón de oportunidades, razón de posibilidades, oportunidad relativa, razón de probabilidades o razón de productos cruzados, e incluso algo tan extraño como "razón de momios". Una buena opción que sirve para evitar confusiones y se ha

hecho mayoritaria es incorporar directamente el término inglés y decir siempre *odds ratio* (abreviadamente OR), lo mismo que con otros términos originalmente ingleses, pero que ya son de uso habitual en castellano (el "*stop"* de las carreteras o el "*penalti*" en el fútbol).

¿Qué es una *odds ratio*? Un cociente entre dos odds, la división de una odds por otra odds es una razón de odds u "odds ratio".

En el ejemplo anterior (figura 2), de 100 pacientes tratados médicamente con un fármaco, se curaron 75 (odds = 75/25 = 3). Supongamos ahora que otros 90 pacientes se trataron quirúrgicamente y se alcanzó el éxito terapéutico en 81 de ellos. La odds esta vez sería de 9 (odds = 81/9 = 9) como muestra la figura 3.

Figura 2.4 Odds de curación si se producen 81 éxitos entre 90 pacientes tratados quirúrgicamente.



La odds ratio (OR) se obtiene al dividir la odds de un tratamiento por la odds de otro:

$$OR = \frac{\text{Odds}_{\text{T. QUIRÚRGICO .}}}{\text{Odds}_{\text{FÁRMACO}}} = \frac{9}{3} = 3$$

Se obtiene una OR = 3 para el éxito terapéutico del tratamiento quirúrgico respecto al tratamiento con el fármaco. Una OR, por tanto, es el cociente o razón entre dos odds y carece de unidades de medida.

Para poder interpretar una OR es necesario siempre tener en cuenta cuál es el factor o variable predictora que se estudia y cuál es el resultado o desenlace. Aquí el factor es el tratamiento y la respuesta o desenlace es el éxito terapéutico. La OR no tiene interpretación absoluta, siempre es relativa. Una OR de 3 se interpreta como una ventaja 3 veces superior de una de las categorías (la categoría quirúrgica en el factor tratamiento)

relativamente a la otra categoría (fármaco) para alcanzar el desenlace o resultado (éxito terapéutico).

El valor nulo para la OR es el 1. Una OR = 1 implica que las dos categorías comparadas son iguales. El valor mínimo posible es 0 y el máximo teóricamente posible es infinito. Una OR inferior a la unidad se interpreta como que el desenlace es menos frecuente en la categoría o grupo que se ha elegido como de interés con respecto al otro grupo o categoría de referencia. La odds del grupo de interés se debe colocar siempre en el numerador y la de referencia en el denominador.

Generalizando, podría escribirse una tabla como la que se muestra a continuación.

Tabla 2.1 Disposición de una tabla para el cálculo de una odds ratio

	Respuesta	
Factor	Si	No
Categoría A	a	b
Categoría B	c	d

En esta disposición de la tabla, la odds ratio se calcula por el producto cruzado

$$\mathrm{OR} = \frac{\mathrm{ad}}{\mathrm{bc}} \qquad (2.14)$$

De todos modos, al manejar una OR se presenta una aparente incongruencia con nuestro modo habitual de pensar. ¿Hasta qué punto es verdad que el tratamiento quirúrgico es 3 veces mejor que el farmacológico? Nuestro modo habitual de razonar es que si el tratamiento quirúrgico ha curado al 90% y el farmacológico sólo al 75%, diremos que existe una razón de probabilidades de curarse con valor 1,2:

$$\frac{90\%}{75\%} = \frac{0{,}9}{0{,}75} = 1{,}2$$

En epidemiología este cociente, que surge de dividir proporciones (p_A/p_B) se conoce como “riesgo relativo” o “razón de riesgos” (RR).

$$RR = \frac{p_A}{p_B} \qquad (2.15)$$

Pero la odds ratio (OR) sólo se aproxima al riesgo relativo (RR) cuando el suceso es raro y ocurre en menos del 10% de los sujetos (p< 0,1), por lo que su interpretación debe matizarse en función de lo frecuente que sea el suceso que se usa como respuesta o variable dependiente.

Se ha hecho esta larga introducción sobre la odds ratio porque es el estimador que más fácilmente puede obtenerse e interpretarse en un análisis de regresión logística.

Odds ratio en la regresión logística

Volviendo a la regresión logística, podría escribirse también su ecuación:

$$\ln(\text{odds}) = a + b_1 x_1 + b_2 x_2 + ... + b_m x_m \qquad (2.16)$$

A toda la expresión de la variable dependiente ln (p/1-p) se le llama logit (p). Por consiguiente:

$$\text{logit}\,(p) = \ln\,(\text{odds}) = \ln\left(\frac{p}{1-p}\right) \qquad (2.17)$$

La transformación logarítmica es necesaria para adaptarse a un fenómeno como la probabilidad cuyos límites teóricos son tan estrechos como 0 y 1. En cambio, los límites teóricos de ln (odds) van desde $-\infty$ hasta $+\infty$.

Como sucede con la regresión lineal, también cuando se ajusta un modelo de regresión logística, el ordenador también devuelve coeficientes b_i para cada una de las variables independientes x_i que pueden considerarse predictores del suceso considerado como respuesta o variable dependiente (y = logit(p)).

El estimador habitual de asociación entre variables que se obtiene directamente de la regresión logística es la odds ratio (OR). Esto hace a la regresión logística un procedimiento muy útil para construir modelos matemáticos de factores predictivos, ya que sus resultados son interpretables como odds ratios. La regresión logística es muy utilizada, cada vez más, tanto en epidemiología de factores de riesgo como en epidemiología clínica.

En una regresión logística, al igual que en la regresión lineal múltiple, es posible introducir variables independientes (x_i) categóricas o dicotómicas en los modelos.

También es posible incluir como variables independientes variables cualitativas con varias categorías como estado civil (soltero, casado, viudo, etc.). Pero ello, como se ha visto requeriría la creación de tantas variables artificiales (dummy) como categorías, menos una, que se reserva como estrato de referencia. La regresión logística se emplea habitualmente en uno de los diseños epidemiológicos más utilizados: los estudios de casos y controles.

2.5 Árboles de decisión

En un estudio real existen frecuentemente múltiples variables (predictivas o independientes) que pueden tener asociación con una variable dependiente. La presentación de muchas tablas de contingencia, no siempre refleja las asociaciones esenciales, y usualmente se convierte en un listado inútil de tablas que desinforman en lugar de orientar, aún cuando se utilice la V de Cramer para ordenar la fortaleza de las asociaciones. Un estudio multivariado trata de enfocar el efecto posible de todas las variables conjuntamente incluyendo sus posibles correlaciones; pero puede ser particularmente interesante, si se considera además las posibilidades de la interacción entre las variables predictivas sobre la variable dependiente. Cuando el número de variables crece, el conjunto de las posibles interacciones crece en demasía, resulta entonces prácticamente imposible analizarlas y por ello adquiere especial interés una técnica de detección automática de interacciones fundamentales. (R.)

Estos procedimientos de clasificación (Breiman, 1984) permiten crear un modelo basado en un árbol que clasifica los casos en grupos o valores de una variable dependiente (en nuestro ejemplo, pacientes con alto riesgo o no) basados sobre los valores de un conjunto de variables independientes, o predictoras, (en nuestro caso, los factores de riesgo). Los árboles de decisión suministran herramientas de validación para el análisis exploratorio de datos y la clasificación, pueden ser usados para:

1. **Segmentación**: Identificar las personas de la población que probablemente sean miembros de un grupo particular (segmentar la población en subgrupos acorde a las variables independientes y en los cuales predomina la pertenencia a uno u otro grupo de la variable dependiente)

2. **Estratificación**: Asignar casos a una o varias categorías (en particular Alto riesgo cardiovascular, riesgo moderado o. Bajo riesgo cardiovascular)
3. **Predicción**: Crear reglas y usarlas para eventos futuros, tales como la verosimilitud de que un nuevo paciente pueda pertenecer a uno u otro grupo (por Ej. Predecir alto , moderado o bajo riesgo de padecer cardiopatías en un futuro)
4. **Reducción de datos y examen de variables**: Seleccionar un subconjunto de predictores a partir de un conjunto grande de variables para usar en un modelo formal paramétrico (por ejemplo ¿cuáles son los factores de riesgo más importantes?)
5. **Identificación de interacciones**: Identificar relaciones que permitan determinar los únicos subgrupos de variables para usar en la construcción de los modelos paramétricos (¿cuáles factores interactúan produciendo modificaciones de riesgo?
6. **Unión de categorías y discretización de variables**. Recodificar las categorías grupales (si fuera el caso) y las variables continuas, ordinales y nominales con mínima pérdida de información (¿cómo recodificar las variables independientes para obtener la mejor predicción?)

Existen varios algoritmos de construcción de árboles de decisión que se distinguen por los criterios con los cuáles se despliega el árbol. En la literatura de Estadística y de Inteligencia Artificial es bien conocido el clásico algoritmo ID-3 incorporado en el sistema C4.5 (Quinlan). Por ejemplo: CHAID (viene del inglés *Chi-square Automatic Interaction Detector*), es un algoritmo que en cada paso selecciona la variable independiente (predictor) que tiene una interacción más fuerte con la variable dependiente, sobre la base de la mayor significación de un test Chi-cuadrado. *Exhaustive* CHAID es una modificación de CHAID, que en cada paso examina todas las formas posibles de desplegar cada predictor. Por su parte, CART o CRT, que proviene del inglés *Classification and Regresion Tree* despliega los datos en segmentos que son lo más homogéneos posible respecto a la variable dependiente, y QUEST**,** que viene del inglés *Quick, Unbiased, Efficient Statistical Tree*, es un método que es rápido y evita el sesgo que aparece en otros métodos cuando hay predictores con muchas categorías.

El análisis de CHAID surge realmente como una técnica de segmentación. Es muy útil en todos aquellos problemas en que se quiera subdividir una población a partir de una variable dependiente y las posibles variables predictivas que cambien esencialmente los valores de la variable dependiente en cada una de las subpoblaciones o segmentos. Ejemplos típicos asociados con su origen son los problemas de estudio de mercado. En estos casos la variable dependiente puede ser la aceptación o no de un producto y las variables predictivas un conjunto de características psico o socio económicas de la población que pueden influir en esta aceptación o no. La técnica de CHAID es capaz de segmentar la población en grupos de acuerdo con determinados valores de esas variables y sus interacciones que distinguen de forma óptima, diferencias esenciales en el comportamiento de la variable dependiente (García, 2007).

Desde esta formulación inicial, se concibió la posibilidad de aplicación a diversas investigaciones como pudiera ser en la salud. La más típica de ellas, es precisamente en epidemiología, en el estudio de los factores de riesgo asociados a una enfermedad (en nuestro caso riesgo cardiovascular). En tal caso, la variable dependiente puede ser simplemente la variable que distingue un grupo de enfermos y sanos (en nuestro caso pacientes con elevado riesgo cardiovascular o no) y las variables predictivas los posibles factores de riesgo (Zamora Rodríguez, 1997).

Más que segmentar la población en este caso la técnica de CHAID se usa en este caso para:

- Conocer cuáles, entre decenas de variables (posibles factores de riesgo) pueden ser eliminadas.
- Comprender el orden de importancia de los factores de riesgo en la caracterización de la enfermedad y en particular ayudar a detectar posibles factores confusores o modificadores de riesgo.
- Entender cómo ciertos factores de riesgo interactúan con otros.
- Conocer que efectos interactivos incluir en un análisis discriminante o de regresión logística de casos-controles respecto a factores de riesgo.
- Buscar entre cientos de tablas de contingencia y seleccionar aquellas que son más significativas estadísticamente.

- Simplificar las cross tabulaciones combinando categorías de variables predictoras que no difieren significativamente.

El análisis de CHAID tiene otras aplicaciones importantes en salud. En particular permite:

- Construir una escala cuantitativa de una variable ordinal, situación típica que se presenta en los procesos por ejemplo de elaboración de tests psicométricos u otras pruebas médicas basadas en criterios cualitativos.
- Elaborar criterios diagnósticos, utilizando en este caso como variables predictoras, posibles síntomas, en lugar de factores de riesgo.
- Someter a validación resultados de ensayos clínicos(R.).

Un análisis de CHAID comienza dividiendo la población en dos o más grupos distintos basado en las categorías del mejor predictor. Divide cada uno de estos grupos en pequeños subgrupos. CHAID visualiza los resultados de la segmentación en forma de un diagrama árbol cuyas ramas (nodos) corresponden a los grupos. Como cada uno de esos grupos se divide además en pequeños subgrupos, el árbol produce nuevos nodos. Entiéndase en este caso que está seleccionando sucesivamente los factores de riesgo más significativamente asociados con la enfermedad y los factores que deben ser fuentes de estratificaciones sucesivas. En cualquier punto en un análisis de CHAID, el árbol muestra el estado actual del análisis.

CHAID divide la población en grupos excluyentes y exhaustivos. Cada individuo de la población queda estar en uno y un solo grupo. A diferencia de otras técnicas de agrupación, como las técnicas de *Clustering*, solo CHAID utiliza una variable dependiente como criterio para la formación de subgrupos (la significación estadística entre la variable dependiente y los predictores es lo que se maneja en el algoritmo de segmentación de CHAID). Esto es, mientras que los segmentos de CHAID se obtienen para predecir una variable dependiente, los clusters no tienen por qué ser predictivos.

2.6 Curvas ROC

Las curvas ROC (del inglés *Receiver Operating Characteristics*) constituyen una manera de examinar el desempeño de un clasificador. Una curva ROC es un gráfico con la

Razón de Falsos Positivos (*FP* = 1-*Sp*) en el eje *X* y la Razón de Verdaderos Positivos (*TP*) en el eje *Y*. Las curvas quedan en el cuadrado [0,1] *x* [0,1]. El vértice superior izquierdo de este cuadrado: (0, 1) representa al clasificador perfecto porque clasifica todos los casos positivos y todos los casos negativos correctamente pues *FP* = 0 y *TP* = 1. El vértice inferior izquierdo (0, 0) representa un clasificador que predice todos los casos como negativos, mientras que el vértice superior derecho (1,1) corresponde a un clasificador que predice todos los casos como positivos. El punto (1, 0) es un clasificador pésimo que resulta incorrecto en todas las clasificaciones.

En muchos casos, un clasificador tiene un parámetro que puede ser ajustado para incrementar *TP* al costo de incrementar *FP* o decrecer *FP* al costo de decrecer *TP*. Cada parámetro puede suministrar un par (*FP*, *TP*) o lo que es lo mismo, un punto sobre este cuadrado y una serie de tales puntos pueden utilizarse para plotear la llamada curva ROC. Un clasificador que no dependa de parámetros, se representa por un punto simple, correspondiente a su par (*FP*, *TP*). El gráfico debajo muestra la curva ROC de dos clasificadores en los que se puede gobernar un parámetro (Clasificadores A y B) y un punto correspondiente a un clasificador simple no paramétrico C.

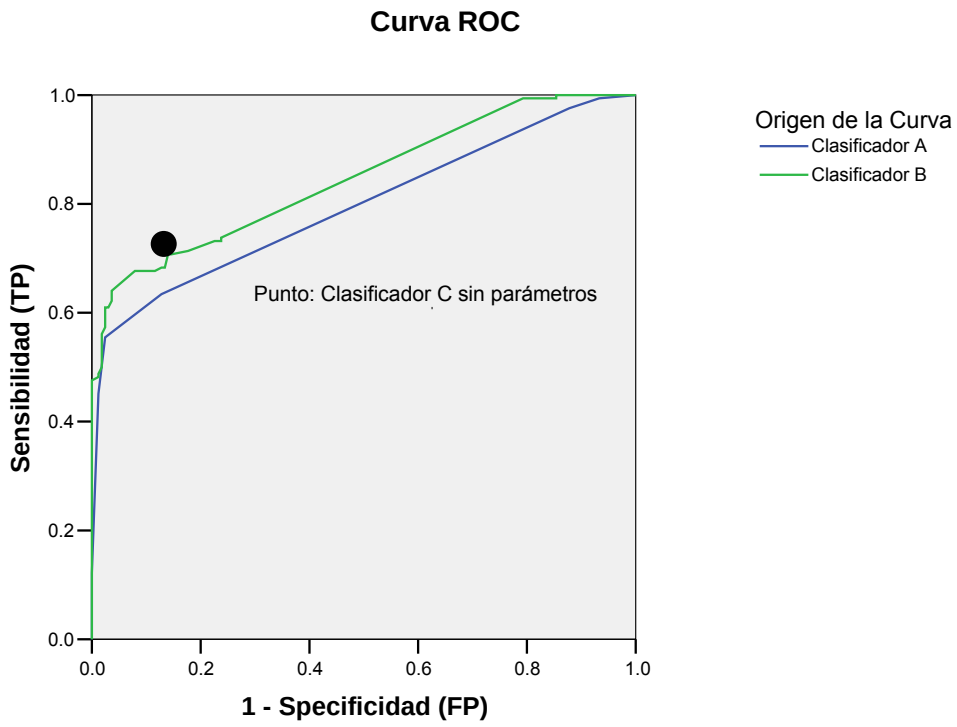
Los hechos fundamentales de las curvas ROC son los siguientes:

- Una curva (o un punto) ROC es independiente de la distribución de las clases o el costo de los errores, esto es, no depende de que en la base de aprendizaje haya más casos negativos que positivos o viceversa
- Una curva ROC resume toda la información contenida en la matriz de confusión ya que *FN* es el complemento de *TP* y *TN* es el complemento de *FP*. Las curvas ROC constituyen una herramienta visual para examinar el equilibrio entre la habilidad de un clasificador para identificar correctamente los casos positivos y el número de casos negativos que están incorrectamente clasificados.

El área bajo la curva ROC puede ser usada como una medida de la exactitud en muchas aplicaciones. Si se comparan dos clasificadores, a través de sendas curvas ROC podemos decidir en general que la de mayor área bajo ella identifica al mejor clasificador, o más precisamente, el clasificador para el cual se pueda obtener un punto más alto en el eje *Y* (mayor ordenada) con un punto más bajo en el eje *X* (menor

abcisa). Para un clasificador no paramétrico, e identificado por un punto ROC, la eficiencia puede medirse en términos de la distancia del punto (*FP, TP*) correspondiente al punto (0, 1). En ambos criterios, pueden introducirse pesos en términos de la importancia relativa de los *FP* o los *TP*. En la figura 2.5, el clasificador B resulta mejor que el A porque tiene un área mayor bajo la curva y el clasificador no paramétrico C, representado por el punto, es mejor que el A, si se atiende bien a reducir los *FN* pero no supera al clasificador B.

Figura 2.5 Curva ROC



Área bajo la curva: Clasificador A=0.796 Clasificador B=0.834

Conclusiones parciales

En este capitulo se presentan varias técnicas estadísticas univariadas y multivariadas. Los métodos univariados (tablas de contingencia con medidas de asociación y pruebas no paramétricas como la U de Mann-Whitney) pueden utilizarse para obtener una caracterización inicial de la población en estudio.

Los métodos multivariados como el análisis discriminante, la regresión logística y los árboles de decisión, son técnicas mas complicadas que pueden utilizarse para obtener los índices de riesgo deseados. En este sentido se necesitará un criterio adicional para seleccionar el mejor de ellos. Las curvas ROC resuelven este problema.

Capítulo 3. Obtención de un índice de alto riesgo cardiovascular en Santa Clara

La Hipertensión arterial (HTA) es una de las enfermedades más comunes que afectan la salud de los individuos adultos en las poblaciones de todas las partes del mundo. Incluso desde hace unos años también se está presentando en la población infantil causando no menos daños que en la adulta.

Debido a su carácter asintomático se le ha denominado la “epidemia silenciosa” pues por lo regular no presenta claras manifestaciones que evidencien su presencia, sin embargo no deja de provocar afectaciones al organismo humano (Rodriguez, 2006).

Al mismo tiempo de representar por sí misma una enfermedad, la hipertensión arterial como ya hemos dicho, constituye un factor de riesgo muy importante para otras enfermedades, fundamentalmente cerebrales, cardiacas y renales, las cuales en caso de que no conlleven al paciente a la muerte, provocan daños irreversibles en órganos tan importantes, ocasionando entonces incapacidad física e intelectual.

Resulta necesario entonces dedicar personal y recursos en investigaciones sobre esta patología par indagar en sus factores predisponentes, descubrir personas que estén en riesgo de padecerla, alertar a los que la padezcan, en fin, llevar a cabo todas las acciones que contribuyan a reducir la prevalencia de esta enfermedad y las graves consecuencias que ella provoca.

En el presente trabajo se pretende obtener un índice de alto riesgo cardiovascular para Santa Clara, para ello se aplican técnicas estadísticas tanto univariadas como multivariadas. Los datos que se utilizan fueron suministrados por el proyecto PROCDEC (UCLV). A continuación se presentan los resultados obtenidos al aplicar dichas técnicas a la muestra utilizada en el estudio.

3.1 Características fundamentales de los datos

La muestra consta de un total de 849 pacientes de los cuales 220 son hipertensos, 219 son hiperreactivos y 410 son normotensos. Se analiza un conjunto de 24 variables aleatorias empleadas en el diagnóstico de alto riesgo cardiovascular. La tabla 3.1 muestra las características fundamentales de las variables aleatorias que son discretas

y la tabla 3.2 muestra las características fundamentales de las variables aleatorias que son continuas.

Tabla 3.1 Variables aleatorias discretas.

Variables	Identificador	Valores	Porcentaje
Sexo	Sexo	Masculino	61.6
		Femenino	38.4
Raza	Raza	Blanca	85.5
		Mestiza	14.5
Ingiere bebida alcohólicas	Bebe	Si	50.7
		No	49.3
Hábito de fumar	Fuma	Si	38.7
		No	61.3
Diabetes mellitus	Diabetes	Si	10.3
		No	89.7
Dislipidemia	Dislipid	Si	7.6
		No	92.4
Número de padres con HTA	Nropadre	0	50.9
		1	33.9
		2	15.3
Número de abuelos con HTA	Nroabuel	0	84.5
		1	11.1
		2	3.1
		3	0
		4	1.2
Nuevo diagnóstico	Nuevodiag	Hipertenso	44.1
		Hiperreactivo	25.7
		Normotenso	30.2
Riesgo	Riesgo	Alto	4.8
		No alto	95.2

En nuestro análisis la variable más importante es riesgo, puesto que es la que vamos a contrastar con las demás en las diferentes técnicas estadísticas que apliquemos. En la tabla anterior se observa que las categorías de la variable riego son Alto y No alto, inicialmente no eran éstas sus categorías sino Muy Alto, Alto, Moderado, Bajo y Muy Bajo. Para llevar a cabo el estudio que queremos (obtener un índice de alto riesgo

cardiovascular) fue necesario recodificar la variable riesgo en Alto y No Alto, agrupando en la primera las categorías Muy Alto y Alto, y en la segunda Moderado, Bajo y Muy Bajo. Seguidamente se presenta la tabla de distribución de frecuencias de la variable riesgo recodificada que es para nosotros Riesgo.

Tabla 3.2 Distribución de frecuencias de la variable Riesgo sin pesar los datos.

Riesgo

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Moderado o Bajo	808	95,2	95,2	95,2
	Alto	41	4,8	4,8	100,0
	Total	849	100,0	100,0	

Observe que la cantidad de personas con alto riego es muy pequeña (41) comparada con los 849 casos que tenemos en total, con esto sabemos que los datos están desbalanceados, es decir, hay mucha diferencia de una categoría a otra.

Con el objetivo de equiparar los datos se decidió utilizar una variable peso, que toma los valores: 38, si la clasificación del paciente es de Alto riesgo o 2 si el paciente es de Bajo o Moderado.

Después de la introducción de la variable peso la tabla de distribución de frecuencias de riesgo resulta como sigue:

Tabla 3.3 Distribución de frecuencias de la variables Riesgo pesando los datos.

Riesgo

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Moderado o Bajo	1616	50,9	50,9	50,9
	Alto	1558	49,1	49,1	100,0
	Total	3174	100,0	100,0	

Logramos entonces que exista en los datos un balance entre las categorías de la variable riesgo que era lo que se esperaba obtener.

A continuación presentamos las variables aleatorias continuas de nuestro análisis. En la tabla **3.4**, además de las variables continuas mostramos los valores mínimo y máximo que toma cada una de estas variables, así como su identificador en los datos.

Tabla 3.4 Variables aleatorias continuas.

Variable	Identificador	Mínimo	Máximo
Edad	edad	18	78
Índice de masa corporal	imc	15.54	44.53
TA Sistólica basal	sistbas	80	220
TA Diastólica basal	diastbas	50	130
TA Sistólica basal al 1er minuto	sistmin1	80	230
TA Diastólica basal al 1er minuto	diasmin1	48	140
TA Sistólica basal al 2do minuto	sistmin2	80	230
TA Diastólica basal al 2do minuto	diasmin2	66	140
Presión arterial media	pam	66.67	170
Glicemia	glicemia	2.70	11.10
Triglicéridos	triglice	0.42	7.95
Colesterol total	coltotal	88.94	421.50
Colesterol HDL	coleshdl	13.92	270.69
Colesterol LDL	colesldl	30.55	494.97

3.2 Caracterización inicial de la muestra

Como primer paso en nuestro estudio se realizó un análisis univariado de las variables con las que contamos en la investigación mediante la aplicación de tablas de contingencia y de la prueba no paramétrica U de Mann- Whitney.

3.2.1 Análisis de tablas de contingencia

En este epígrafe se realizó el cálculo del estadístico Chi Cuadrado en tablas de contingencia así como se analizaron los resultados obtenidos. Se hizo la comparación de la variable riesgo con cada una de las variables aleatorias discretas del análisis, ver anexo 1.

Tabla 3.5 Resumen del cálculo de la V de Cramer

	Variable	V de Cramer
Riesgo	Fuma	0.437
	Diabetes mellitus	0.402
	Diagnóstico	0.402
	Sexo	0.369
	Número de abuelos con HTA	0.265
	Dislipidemia	0.237
	Bebe	0.206
	Número de padres con HTA	0.174
	Raza	0.094

Como se puede observar en la tabla anterior se efectuó el cálculo correspondiente a la V de Cramer. Como todas las variables resultaron significativas, se muestran ordenadas según su respectivo valor de la V de Cramer.

Al interpretar estadísticamente el problema valiéndonos del análisis univariado que se acaba de realizar, se puede decir que todas las variables por separado son capaces de diferenciar los grupos, o sea, para todas ellas el test Chi Cuadrado arrojó resultados significativos. Como en todos los casos la significación es igual a 0.00, incluso menor que 0.01, podemos decir que el riesgo depende en gran medida de todas estas variables.

El orden establecido anteriormente para las variables, teniendo en cuenta su respectivo valor de la V de Cramer, responde a que este valor nos ofrece una medida de la fortaleza de la asociación que existe entre la variable riesgo y cada una de ellas. Puede apreciarse que la variable Fuma es la más significativa de todas las variables discretas en el estudio.

3.2.2 Análisis del test U de Mann-Whitney

En este epígrafe se le aplica a los datos que tenemos una prueba no paramétrica: el test U de Mann- Whitney, siguiendo la idea de determinar si existe o no asociación entre la variable riesgo y las demás variables, en este caso como variables predictoras tendremos las continuas.

Tabla 3.6 Resumen del test U de Mann-Whitney

	Variable	Significación asintótica
Riesgo	Edad	0.000
	IMC	0.000
	TA Sistólica	0.000
	TA Diastólica	0.000
	TA Sistólica al	0.000
	TA Diastólica al	0.000
	TA Sistólica al	0.000
	TA Diastólica al	0.000
	PAM	0.000
	Glicemia	0.000
	Triglicéridos	0.000
	Colesterol total	0.000
	Colesterol HDL	0.000
	Colesterol LDL	0.000

La tabla anterior muestra que se calculó la significación asintótica, en este caso como para las variables discretas obtenemos que igualmente todas las variables aleatorias continuas están asociadas a la variable riesgo. El valor de la significación asintótica nos dice que cada variable por separado, está significativamente relacionada con el riesgo.

Para que el estudio que llevamos a cabo arroje resultados verdaderamente confiables no podemos conformarnos con un análisis univariado aunque ello no significa que la información que nos ofrecen las tablas de contingencia, la V de Cramer y el test U de Mann-Whitney sea incorrecta. Este pequeño problema llega a resolverse con un análisis multivariado, mediante el que se comprenderá mejor la relación que existe entre las variables, además de que como resultado se obtendrían modelos con las variables realmente importantes para la predicción.

3.3 Análisis multivariado

3.3.1 Análisis del análisis discriminante

En este epígrafe se aplica la técnica del análisis discriminante, como variable independiente tomamos riesgo y como posibles variables predictoras las demás variables del análisis. Se utilizó un método por pasos con el fin de reducir, de ser posible, el número de variables predictoras.

Se llevó a cabo el análisis discriminante aplicando el método de introducción de las variables paso a paso, este a su vez tiene como opciones varios métodos de este tipo, entre ellos escogimos el de la lambda de Wilks. Como resultado, en el paso número 18 se obtuvieron las variables que finalmente forman parte del modelo, lográndose clasificar correctamente un 94.6% de los casos agrupados originalmente. En la tabla que se muestra a continuación se muestran algunos de los resultados obtenidos.

Tabla 3.7 Resumen de las variables en el modelo discriminante.

Variables in the Analysis

Step		Tolerance	F to Remove	Wilks' Lambda
18	Edad	,886	2049,604	,468
	TA Sistólica basal	,110	96,861	,293
	Fuma	,789	303,772	,311
	Diabetes mellitus	,507	418,209	,322
	Raza	,860	148,343	,297
	Colesterol HDL	,921	102,287	,293
	Indice de masa corporal (IMC)	,700	63,439	,290
	Sexo	,810	50,175	,288
	Colesterol Total	,754	20,798	,286
	Diagnóstico	,301	60,941	,289
	Presión arterial media (PAM)	,018	27,812	,286
	TA Sistólica (al 1er minuto)	,059	34,814	,287
	TA Diastólica basal	,158	20,561	,286
	Triglicéridos	,804	20,351	,286
	Glicemia	,516	12,695	,285
	TA Diastólica (al 1er minuto)	,093	18,305	,286
	TA Diastólica (al 2do minuto)	,026	7,749	,285
	Nro. de padres con HTA	,836	5,455	,284

Por los resultados que observamos en la tabla anterior se puede decir que la variable que más ayuda a separar los grupos de “Alto” y “Moderado o Bajo” riesgo es la variable Edad, pues es la que se introduce en la ecuación en el primer paso. A pesar de que no se presente la tabla completa, en todos los pasos los modelos obtenidos fueron significativos.

Tabla 3.8 Resumen del cálculo de la Lambda de Wilks.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	,284	3982,024	18	,000

La tabla anterior es un resumen del cálculo de la lambda de Wilks. Se muestra el cálculo del test Chi-cuadrado con la hipótesis fundamental de que las medias de los grupos son iguales, el valor es significativo, (.000), por lo que se rechaza esa hipótesis y se concluye que el modelo hallado discrimina los grupos.

A continuación se muestra la tabla de los coeficientes de las funciones discriminantes canónicas estandarizadas, en ella se puede apreciar el orden de importancia de las variables. En este caso la variable más importante es la PAM (Presión arterial media), luego la Edad y en tercer lugar la TA Sistólica basal.

Tabla 3.9 Coeficientes de las funciones discriminantes canónica estandarizada

Standardized Canonical Discriminant Function Coefficients

	Function
	1
Edad	,788
Indice de masa corporal (IMC)	-,198
Sexo	-,164
Raza	-,270
Fuma	,394
Diabetes mellitus	,568
Nro. de padres con HTA	,054
TA Sistólica basal	,614
TA Diastólica basal	-,239
TA Sistólica (al 1er minuto)	-,510
TA Diastólica (al 1er minuto)	,294
TA Diastólica (al 2do minuto)	-,364
Presión arterial media (PAM)	,819
Glicemia	-,104
Triglicéridos	,106
Colesterol Total	,110
Colesterol HDL	-,218
Diagnóstico	,297

La matriz de estructura, que a continuación se presenta, muestra la correlación de cada variable con la función discriminante.

Tabla 3.10 Matriz de estructura del análisis discriminante

Structure Matrix

	Function
	1
Edad	,654
TA Sistólica (al 1er minuto)	,380
TA Sistólica basal	,377
TA Sistólica (al 2do minuto)	,351
Fuma	,306
Presión arterial media (PAM)	,301
Diabetes mellitus	,277
Diagnóstico	-,276
TA Diastólica basal	,256
Sexo	-,250
TA Diastólica (al 1er minuto)	,243
TA Diastólica (al 2do minuto)	,225
Triglicéridos	,212
Glicemia	,208
Dislipidemia[a]	,150
Bebe[a]	,143
Nro. de abuelos con HTA	-,121
Colesterol Total	,114
Colesterol HDL	-,109
Indice de masa corporal (IMC)	,094
Colesterol LDL[a]	,064
Raza	-,060
Nro. de padres con HTA	,017

a. This variable not used in the analysis.

Observe que la Edad muestra la correlación más alta (0.654) y variables como Dislipidemia o Bebe, que no fueron seleccionadas, tienen correlaciones muy bajas de 0.150 y 0.143, respectivamente.

Finalmente para mostrar que efectivamente el modelo obtenido es muy bueno, con un 94.6% de los casos bien clasificados, se presenta la tabla de los resultados de la clasificación, se observa que solo clasifica mal un 10.6% en moderado o bajo riesgo.

Tabla 3.11 Porcentajes correctos de la clasificación.

Classification Results[a]

			Predicted Group Membership		
		Riesgo	Alto	Moderado o Bajo	Total
Original	Count	Alto	1444	172	1616
		Moderado o Bajo	0	1558	1558
	%	Alto	89,4	10,6	100,0
		Moderado o Bajo	,0	100,0	100,0

a. 94,6% of original grouped cases correctly classified.

3.3.2 Análisis de la Regresión logística

En este epígrafe se aplica otro modelo de clasificación a los datos. Como que la variable dependiente tiene dos categorías, dentro de las técnicas de regresión conocidas, la más apropiada a aplicar es la regresión logística, pues resulta útil en situaciones en las que deseamos clasificar a los sujetos exactamente en dos categorías según los valores de un conjunto de variables predictoras.

Seguidamente se muestra el modelo obtenido mediante la regresión logística con el método paso a paso y aplicando Forward Condicional.

Tabla 3.12 Modelo obtenido por la regresión logística

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 18	EDAD	.204	.007	799.610	1	.000	1.227
	IMC	-.185	.041	20.026	1	.000	.831
	SEXO	-1.281	.284	20.394	1	.000	.278
	RAZA	-3.348	.422	62.974	1	.000	.035
	FUMA	5.553	.429	167.503	1	.000	257.998
	DIABETES	10.715	1.030	108.172	1	.000	45043.720
	NROPADRE	.810	.213	14.488	1	.000	2.248
	NROABUEL	-1.380	.385	12.827	1	.000	.252
	SISTBAS	.079	.021	14.127	1	.000	1.082
	DIASTBAS	.101	.028	13.117	1	.000	1.106
	SISTMIN1	-.057	.028	4.109	1	.043	.945
	DIASMIN1	.262	.034	60.935	1	.000	1.300
	SISTMIN2	.070	.019	13.417	1	.000	1.072
	GLICEMIA	-1.435	.198	52.635	1	.000	.238
	TRIGLICE	.895	.153	34.087	1	.000	2.447
	COLTOTAL	.007	.003	6.218	1	.013	1.007
	COLESHDL	-.099	.012	66.745	1	.000	.905
	NUEVODIA	3.987	.441	81.630	1	.000	53.872
	Constante	-63.208	5.431	135.471	1	.000	.000

Puede observarse que en el paso 18 es que se obtiene el modelo, todas las variables introducidas son significativas por lo que podemos decir que este es un buen modelo.
La tabla que aparece a continuación muestra los resultados finales de la clasificación. Efectivamente, el modelo obtenido es muy bueno, el porcentaje de casos correctamente clasificados es de un 97.5%, superior al obtenido con el análisis discriminante (94.6%).

Tabla 3.13 Porcentajes de buena clasificación mediante regresión logística

Classification Table[a]

Observed			Predicted		
			Riesgo		
			Alto	Moderado o Bajo	Percentage Correct
Step 1	Riesgo	Alto	1538	78	95,2
		Moderado o Bajo	0	1558	100,0
	Overall Percentage				97,5

a. The cut value is ,500

Solo se clasifican mal 78 pacientes, realmente de alto riesgo como de bajo o moderado.

3.3.3 Análisis del árbol de decisión

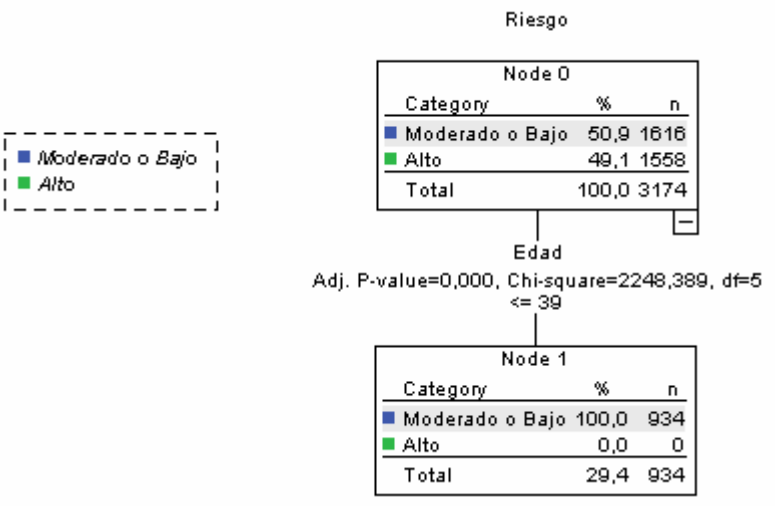
En este epígrafe se aplica la técnica de segmentación CHAID, con la variable riesgo como independiente y las demás variables como posibles predictoras, a pesar de que no todas quedaron incluidas en el modelo presentado por la regresión logística.

En el nodo raíz tenemos el total de casos estudiados: el 50.9% de la muestra representa a las personas que se denominan de alto riesgo cardiovascular y el 49.1% representa a las personas de moderado o bajo riesgo. Como variable que mejor ayuda a diferenciar los grupos tenemos a la Edad.

El árbol consta de 18 nodos terminales, ver anexo 2, que por su complejidad, o mejor dicho, por el número de nodos hijos, lo editamos para mostrarlo por partes. Presentamos a continuación la explicación de cada nodo terminal seguida de la parte del árbol que le corresponde.

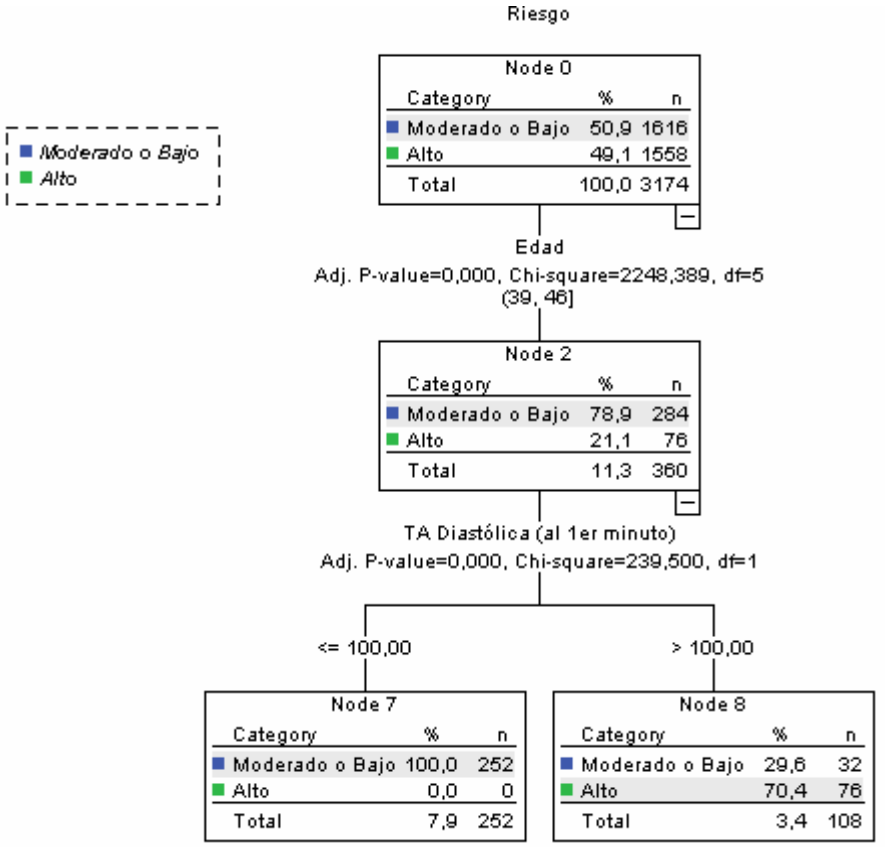
1. Subconjunto formado por 934 de pacientes de hasta 39 años de edad. No hay pacientes de alto riesgo. Se corresponde al nodo 1 del árbol.

Figura 3.1 Nodo 1.



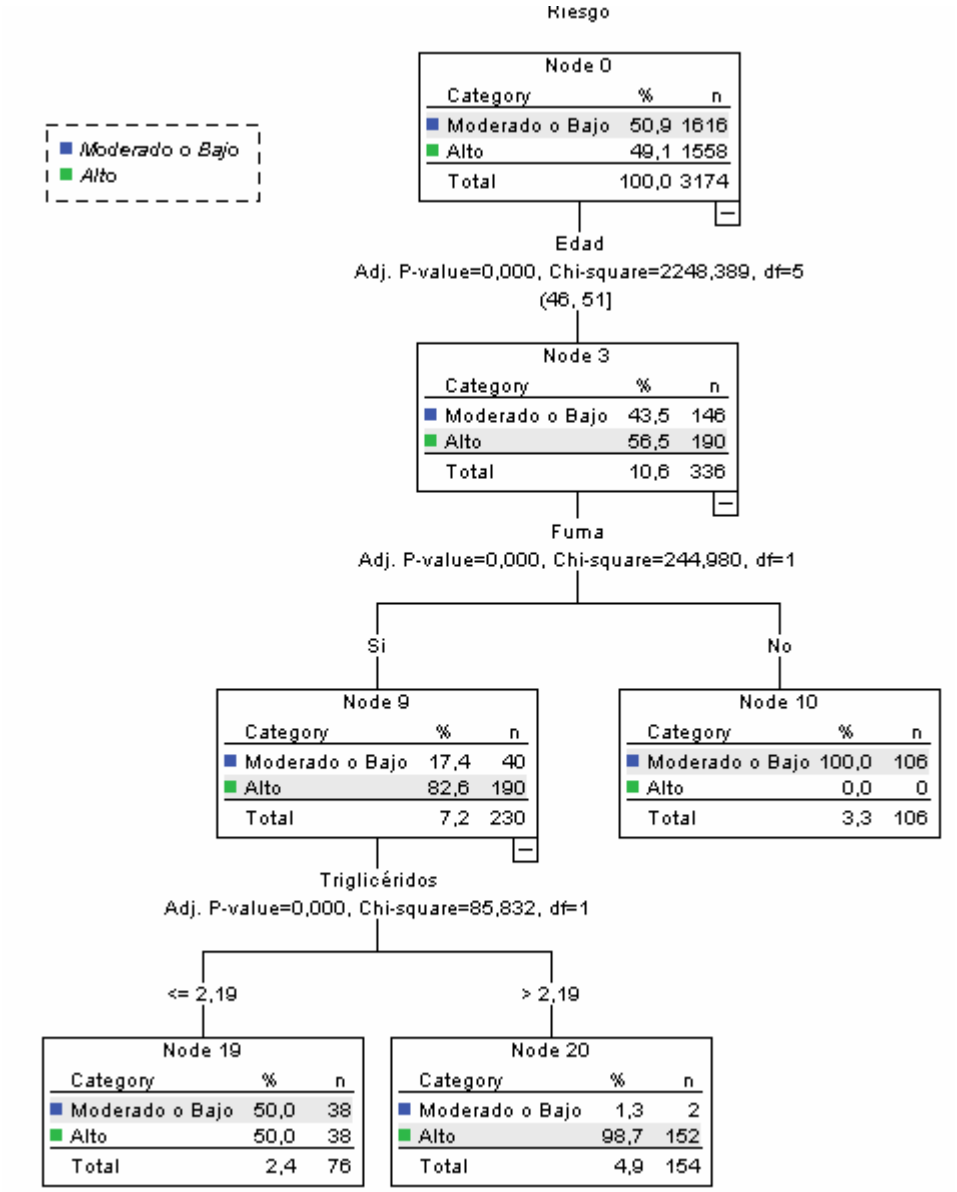
2. Subconjunto formado por 252 de pacientes entre 39 y 46 años de edad, tienen valores de TA Diastólica al primer minuto hasta 100 y no hay pacientes de alto riesgo. Se corresponde al nodo 7 del árbol.
3. Subconjunto formado por 108 de pacientes entre 39 y 46 años de edad, tienen valores de TA Diastólica al primer minuto superiores a 100 y los pacientes de alto riesgo representan un 70.4%. Se corresponde al nodo 8 del árbol.

Figura 3.2 Nodo 7 y 8.



4. Subconjunto formado por 76 de pacientes entre 46 y 51 años de edad que son fumadores, con los triglicéridos hasta de 2.19 y hay un 50% de pacientes de alto riesgo. Se corresponde al nodo 19 del árbol.
5. Subconjunto formado por 154 de pacientes entre 46 y 51 años de edad que son fumadores, con los triglicéridos por encima de 2.19 y hay 98.7% de pacientes de alto riesgo. Se corresponde al nodo 20 del árbol.
6. Subconjunto formado por 106 pacientes entre 46 y 51 años de edad que no son fumadores y no hay pacientes de alto riesgo. Se corresponde al nodo 10 del árbol.

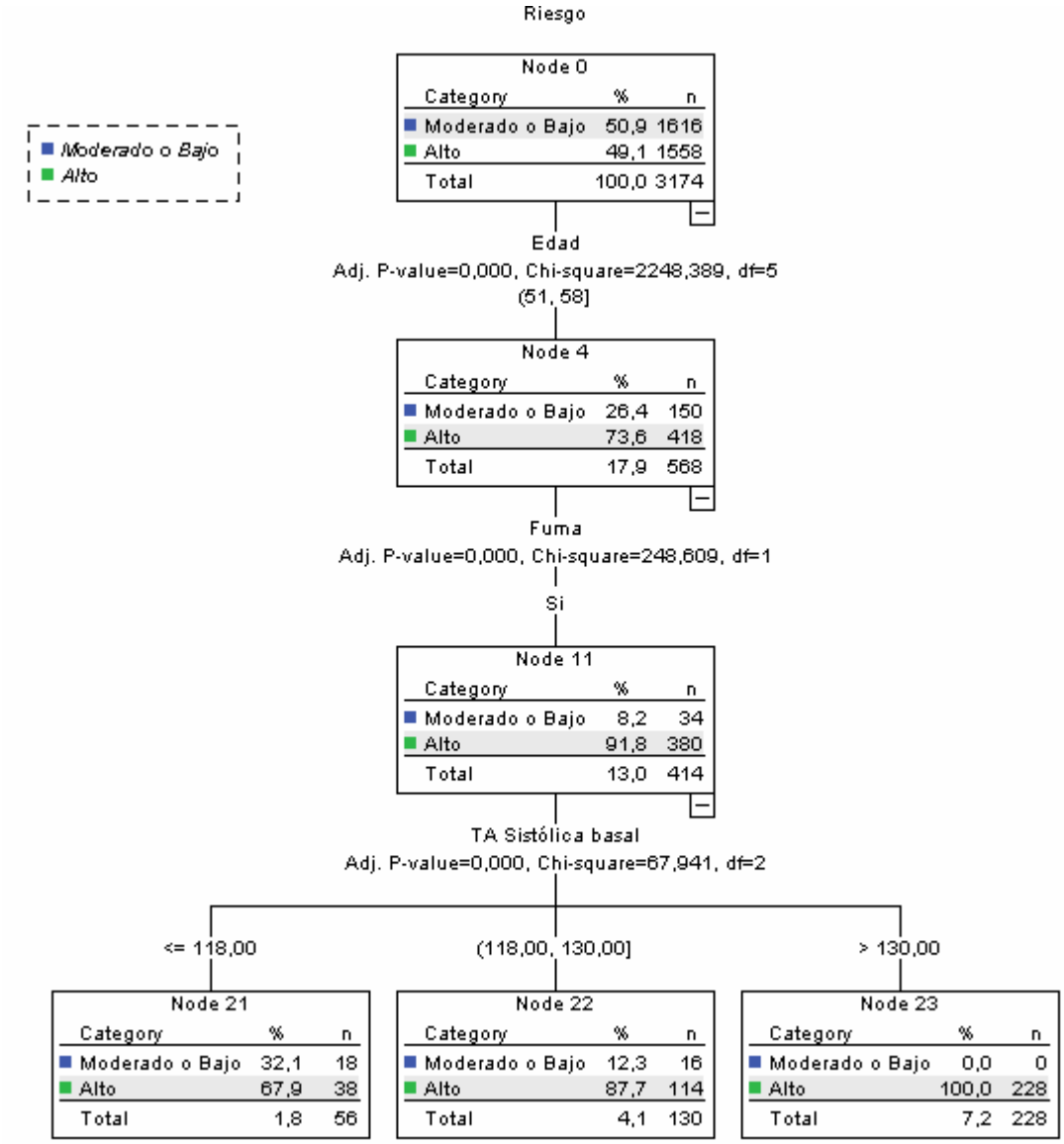
Figura 3.3 Nodo 19, 20 y 10.



7. Subconjunto formado por 56 de pacientes entre 51 y 58 años de edad que son fumadores, con valores de la TA Sistólica basal hasta 118 y hay un 67.9% de pacientes de alto riesgo. Se corresponde al nodo 21 del árbol.
8. Subconjunto formado por 130 de pacientes entre 51 y 58 años de edad que son fumadores, con valores de la TA Sistólica basal entre 118 y 130, además hay un 87.7% de pacientes de alto riesgo. Se corresponde al nodo 22 del árbol.

9. Subconjunto formado por 228 de pacientes entre 51 y 58 años de edad que son fumadores, con valores de la TA Sistólica basal superiores a 130 y todos son de alto riesgo. Se corresponde al nodo 23 del árbol.

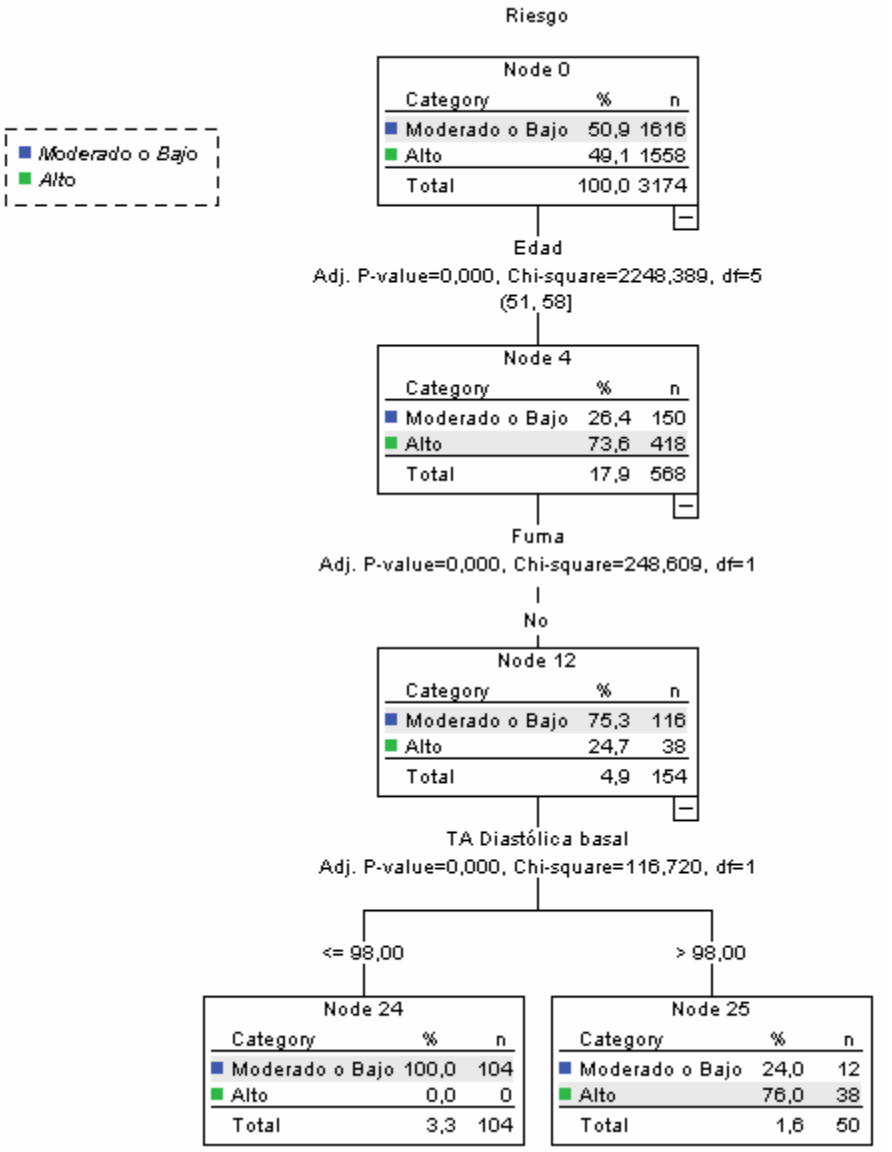
Figura 3.4 Nodo 21,22 y 23.



10. Subconjunto formado por 104 de pacientes entre 51 y 58 años de edad que no son fumadores, con valores de la TA Diastólica basal hasta 98 y no hay pacientes de alto riesgo. Se corresponde al nodo 24 del árbol.

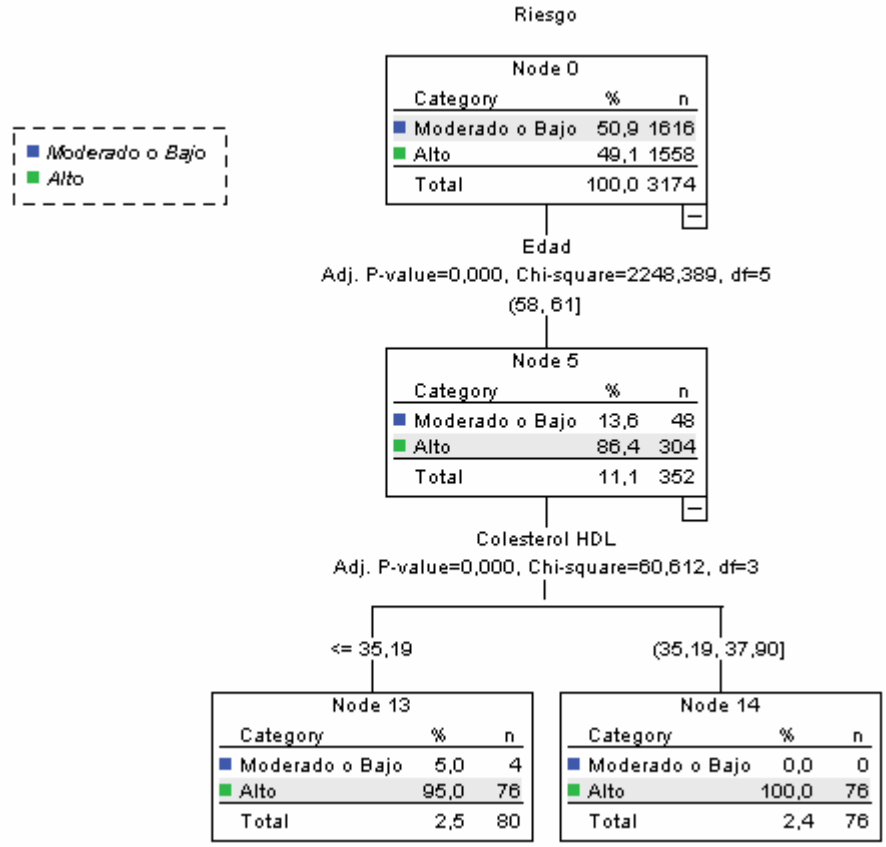
11. Subconjunto formado por 50 de pacientes entre 51 y 58 años de edad que no son fumadores, con valores de la TA Diastólica basal superiores a 98 y hay un 76% de pacientes de alto riesgo. Se corresponde al nodo 25 del árbol.

Figura 3.5 Nodo 24 y 25.



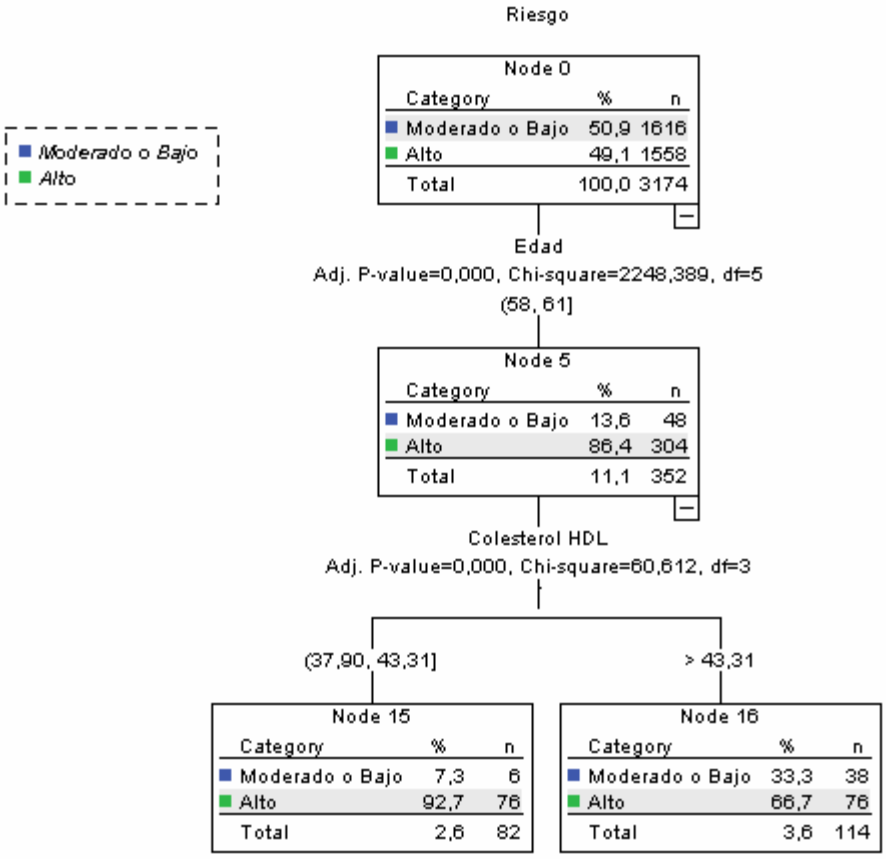
12. Subconjunto formado por 80 de pacientes entre 58 y 61 años de edad, con valores del Colesterol HDL hasta 35.19 y hay un 95% de pacientes de alto riesgo. Se corresponde al nodo 13 del árbol.
13. Subconjunto formado por 76 de pacientes entre 58 y 61 años de edad, con valores del Colesterol HDL entre 35.19 y 37.90, todos son de alto riesgo. Se corresponde al nodo 14 del árbol.

Figura 3.6 Nodo 13 y 14.



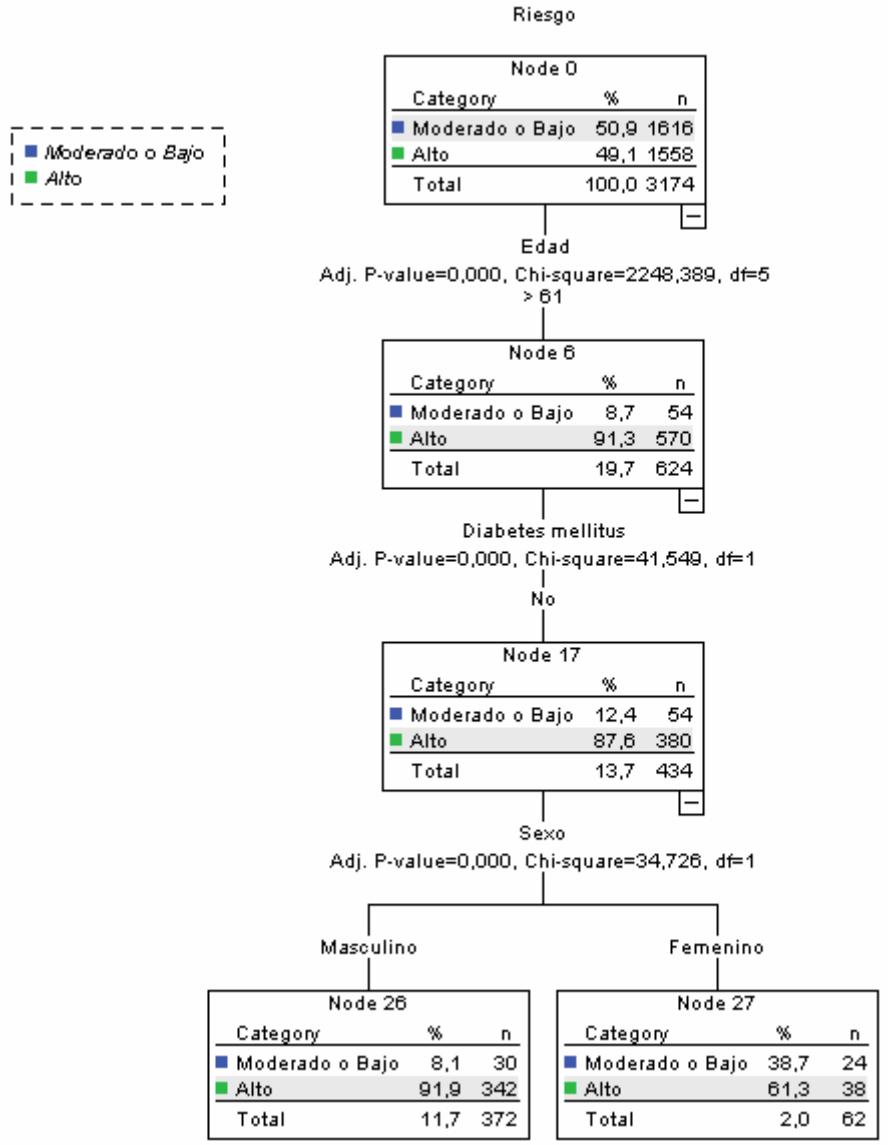
14. Subconjunto formado por 82 de pacientes entre 58 y 61 años de edad, con valores del Colesterol HDL entre 37.90 y 43.31, hay un 92.7% de pacientes de alto riesgo. Se corresponde al nodo 15 del árbol.
15. Subconjunto formado por 114 de pacientes entre 58 y 61 años de edad, con valores del Colesterol HDL superiores a 43.31, hay un 66.7% de pacientes de alto riesgo. Se corresponde al nodo 16 del árbol.

Figura 3.7 Nodo 15 y 16.



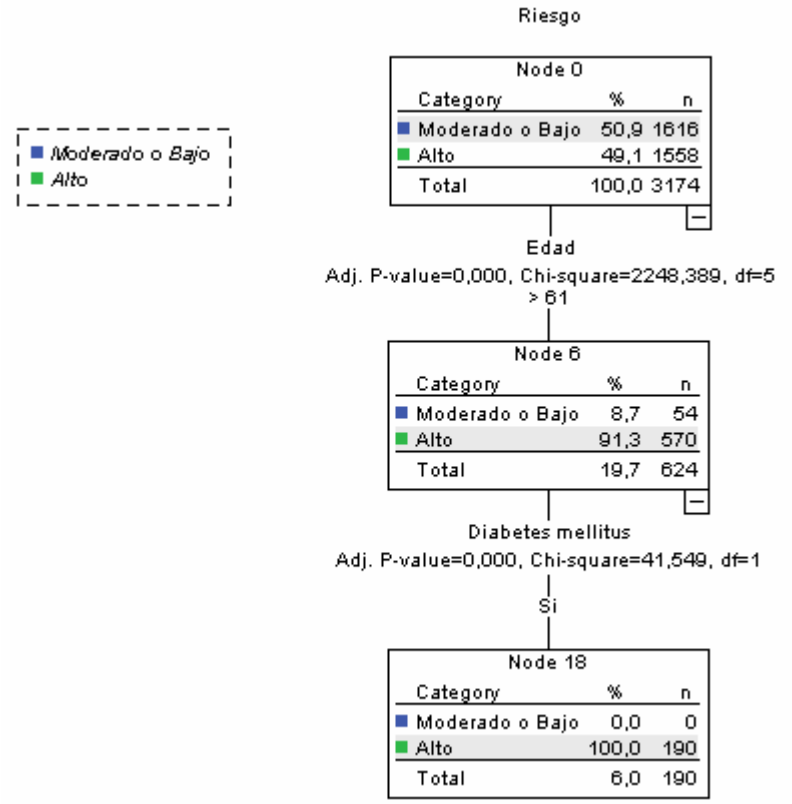
16. Subconjunto formado por 372 hombres mayores de 61 años de edad, que no padecen de Diabetes mellitus, hay un 91.9% de ellos que son de alto riesgo. Se corresponde al nodo 26 del árbol.
17. Subconjunto formado por 62 mujeres mayores de 61 años de edad, que no padecen de Diabetes mellitus, de ellas son de alto riesgo un 61.3%. Se corresponde al nodo 27 del árbol.

Figura 3.8 Nodo 26 y 27.



18. Subconjunto formado por 190 pacientes mayores de 61 años de edad, que padecen de Diabetes mellitus y todos son de alto riesgo. Se corresponde al nodo 18 del árbol.

Figura 3.9 Nodo 18.



Con este árbol de decisión se han obtenido un conjunto de reglas de clasificación dadas por la explicación de cada nodo del árbol. Por ejemplo, del nodo 18, se tiene que dentro de las personas mayores de 61 años, las que padecen de diabetes mellitus tienen un alto riesgo de padecer enfermedades cardiovasculares.

Seguidamente se muestra un resumen de la caracterización de los pacientes de alto riesgo obtenida mediante el árbol. Existe alto riesgo con probabilidad 1 o al menos igual a 0.5 cuando:

i. El paciente se incluye en el rango de edad de 39 a 46 años y tiene la TA Diastólica(al primer minuto) con valor mayor que 100 (alto riesgo 70.4%).

ii. El paciente tiene de 46 a 51 años de edad, es fumador y tiene los triglicéridos hasta en 2.19 (alto riesgo 50%) y superiores a 2.19 (alto riesgo 98.7%).

iii. El paciente tiene de 51 a 58 años de edad, es fumador y tiene la TA Sistólica basal con valor hasta 118 (alto riesgo 67.9%), entre 118 y 130 (alto riesgo 87.7%) y superior a 130 (alto riesgo 100%).

iv. El paciente se incluye en el rango de edad entre 51 y 58 años, no fuma pero tiene la TA Diastólica basal con valor superior a 98 (alto riesgo 76%).

v. El paciente se incluye en el grupo de edad entre 58 y 61 años, tiene valor de Colesterol HDL hasta 35.19 (alto riesgo 95%), de 35.19 a 37.90 (alto riesgo 100%), de 37.90 a 43.31 (alto riesgo 92.7%) y superior a 43.31 (alto riesgo 66.7%).

vi. El hombre mayor de 61 años de edad que no padece de diabetes mellitus (alto riesgo 91.9%).

vii. La mujer mayor de 61 años de edad que no padece de diabetes mellitus (alto riesgo 61.3%).

viii. El paciente mayor de 61 años de edad que padece de diabetes mellitus (alto riesgo 100%).

A continuación se muestran los porcentajes de clasificación obtenidos aplicando la técnica CHAID:

Tabla 3.14 Porcentajes de buena clasificación para la técnica CHAID

Classification

Observed	Predicted		
	Alto	Moderado o Bajo	Percent Correct
Alto	1434	182	88,7%
Moderado o Bajo	38	1520	97,6%
Overall Percentage	46,4%	53,6%	93,1%

Growing Method: CHAID
Dependent Variable: Riesgo

Como se puede observar se obtuvo un porcentaje correcto de clasificación de un 93.1%, además el modelo obtenido mediante la aplicación de la técnica del árbol de decisión clasifica incorrectamente solo a 38 pacientes de riesgo moderado o bajo como de alto y a 182 pacientes de alto como de moderado o bajo riesgo.

3.3.4 Análisis de curva ROC

En este epígrafe se hizo la comparación de la efectividad de los modelos mediante la curva ROC de cada uno de ellos. Para la regresión logística inicialmente con el punto de corte en 0.5 se observa en la tabla 3.15 que cuando la sensibilidad disminuye de 1 a 0.976 la especificidad se mantiene en 0.40. Esto indica que con el valor de corte 0.60 se puede lograr un balance entre la sensibilidad y la especificidad.

Tabla 3.15 Coordenadas de la curva ROC para RL

Coordinates of the Curve Test Result Variable(s): Predicted probability		
Positive if Greater Than or Equal To[a]	Sensitivity	1 - Specificity
,0000000	1,000	1,000
,5782770	1,000	,042
,5962891	1,000	,041
0,6119968	1,000	0,40
,6273006	,976	,040
,6316687	,976	,038
,6358961	,976	,037
,6490822	,976	,036
,6669040	,976	,035

Luego de identificar esta posible ventaja se repitió la regresión logística pero esta vez cambiando el punto de corte de 0.5 a 0.60. Se obtuvo lo siguiente:

Tabla 3.16 Porcentaje de buena clasificación para RL con punto de corte 0.60

Classification Table(a)

Observed			Predicted		
			Riesgo Dicotómico		Percentage Correct
			Moderado o Bajo	Alto	
Step 18	Riesgo Dicotómico	Moderado o Bajo	1552	64	96,0
		Alto	0	1558	100,0
	Overall Percentage				98,0

a The cut value is ,600

Como se puede ver en la tabla anterior el porcentaje de buena clasificación se eleva de un 97.5% que se había obtenido anteriormente a un 98%, al igual que se disminuyen los mal clasificados de alto riesgo (FP) a 64. Evidentemente es un hecho positivo que se

logre reducir los FP de 78 a 64, manteniendo a los FN en cero y la clasificación perfecta de los individuos de alto riesgo. A fines del diagnóstico, los resultados que se obtuvieron tienen un gran valor y desde el punto de vista de la regresión logística se obtuvieron ligeramente mejores resultados.

El gráfico de la figura 3.10 muestra las curvas ROC de los 3 clasificadores vistos anteriormente. Evidentemente, la regresión logística muestra mejores resultados que los otros clasificadores, no obstante el análisis discriminante y el árbol de decisión muestran buenos resultados. Compárese además el área bajo las curvas, tabla 3.17, y véase que el área bajo la curva de la regresión logística es superior a las del árbol de decisión y el análisis discriminante.

Figura 3.10 Comparación de las curvas ROC de los tres métodos

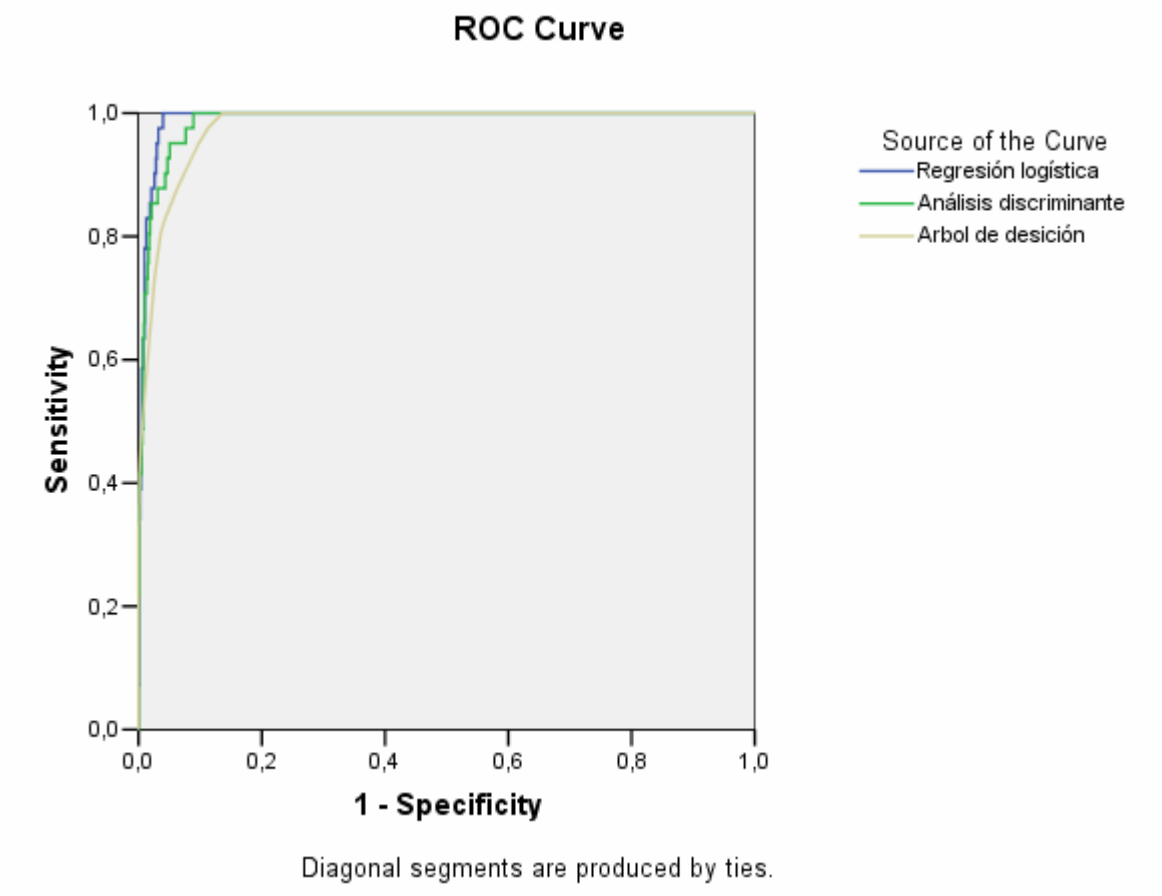


Tabla 3.17 Área bajo la curva de RL con punto de corte 0.60

Area Under the Curve

Test Result Variable(s)	Area
Regresión logística	,991
Análisis discriminante	,987
Arbol de desición	,979

The test result variable(s): Arbol de desición has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

3.4 Análisis de los datos originales

En el presente trabajo hemos obtenido modelos de clasificación para los datos ponderados y con el objetivo de validar los mismos se ha realizado un análisis con los datos originales, es decir, sin ponderar casos.

3.4.1. Análisis discriminante

En el análisis discriminante se había obtenido un buen modelo que clasificaba correctamente un 94,6% de la muestra. Se quiere verificar si el modelo obtenido clasifica correctamente la muestra sin utilizar la ponderación.

En el proceso de obtener el modelo en el SPSS previamente se marcó la opción ***Predicted group membership***, lo que nos permitirá ahora utilizar el modelo que se obtuvo ponderando casos para la validación con los datos originales. Para esto se hace una tabla de contingencia con la variable riesgo original contra la variable que se crea al marcar la opción ***Predicted group membership*** del SPSS. Luego se calculan los por cientos de casos bien clasificados, tanto de alto riesgo como de no alto.

A continuación se muestra la tabla de contingencia obtenida.

Tabla 3.18 Tabla de contingencia Riesgo vs Predicted group(AD)

Riesgo * Predicted Group for Analysis 1 Crosstabulation

		Predicted Group for Analysis 1		
		1	2	Total
Riesgo	1	722	86	808
	2	0	41	41
Total		722	127	849

Al calcular los por cientos de clasificación se obtuvo para los casos bien clasificados un 89.9% y para los mal clasificados un 10,1%. Observe además que los clasificados como

de “alto” riesgo, que son 41, están correctamente clasificados, no siendo así con los de “no alto” puesto que se clasifican 86 como de alto riesgo.

3.4.2. Regresión logística

Para este caso se repite el proceso llevado a cabo en el análisis discriminante, con la diferencia de que para obtener el modelo de clasificación aplicando regresión logística en el SPSS la opción que se marca es ***Group membership***, que será la que nos permitirá utilizar el modelo obtenido para la validación.

La tabla de contingencia para este caso es la siguiente:

Tabla 3.19 Tabla de contingencia Riesgo vs Predicted group(RL)

Riesgo * Predicted group Crosstabulation

		Predicted group		
		1	2	Total
Riesgo	1	769	39	808
	2	0	41	41
Total		769	80	849

El por ciento de casos bien clasificados es de un 95.4% y el de casos mal clasificados es de 4.6%. Estas cifras observe que se diferencian grandemente con respecto a las obtenidas para el análisis discriminante: aumenta el por ciento de bien clasificados y disminuye el de mal clasificados, además se observa la clasificación perfecta de los pacientes de “alto” riesgo cardiovascular.

3.4.3 Árboles de decisión

En la obtención del árbol de decisión en el SPSS, la opción que se marca para utilizar el modelo obtenido ponderando casos para la validación es ***Predicted Value***, el resto del procedimiento es el mismo que se ha seguido en los subepígrafes anteriores.

La tabla de contingencia para este caso es la siguiente:

Tabla 3.20 Tabla de contingencia Riesgo vs Predicted Value

Riesgo * Predicted Value Crosstabulation

		Predicted Value		
		1	2	Total
Riesgo	1	717	91	808
	2	1	40	41
Total		718	131	849

El porcentaje de casos bien clasificados obtenido para esta tabla es de 89.2% y no están correctamente clasificados un 10.8%. Se observa que se ha clasificado incorrectamente a una persona de alto riesgo como de no alto, esto constituye un problema desde el punto de vista clínico para el paciente, no estar correctamente clasificado lo excluye de todo tipo de tratamiento especial a los pacientes de alto riesgo cardiovascular.

Conclusiones parciales

En este capítulo se han aplicado varias técnicas estadísticas univariadas y multivariadas. Los métodos univariados se utilizan para obtener una caracterización inicial de la población en estudio.

Los métodos multivariados como el análisis discriminante, la regresión logística y los árboles de decisión se utilizan para obtener los índices de riesgo cardiovascular deseados. De ellos se puede decir que el modelo obtenido mediante regresión logística es el mejor desde el punto de vista de los por cientos de clasificación puesto que se obtiene con él un 98% de casos correctamente clasificados (cambiando el punto de corte a 0.60). Podemos comprobar esto utilizando el criterio de las curvas ROC. Efectivamente, de la comparación de las curvas ROC de cada uno de los modelos obtenidos podemos concluir que el mejor clasificador es el de la regresión logística.

Conclusiones

Como resultado fundamental del trabajo se obtuvieron tres índices de alto riesgo cardiovascular, aplicando tres técnicas de estadística multivariada diferentes: el análisis discriminante, la regresión logística y los árboles de decisión.

Los modelos obtenidos clasifican bien los datos aunque uno de ellos en mayor medida que los otros. A partir de aquí se pudo determinar el mejor índice, mediante las curvas ROC y resultó ser el modelo obtenido al aplicar regresión logística.

Analizando el mejor modelo obtenido, (regresión logística con el punto de corte igual a 0.60), se concluye que los factores de riesgo más importantes son: la edad, el índice de masa corporal (IMC), el sexo, la raza, hábito de fumar, padecimiento de diabetes, número de padres y abuelos con HTA, la TA Sistólica y Diastólica basal, la TA Sistólica y Diastólica basal al 1er minuto, la TA Sistólica basal al 2do minuto, glicemia, triglicéridos, el colesterol total y HDL, así como diagnóstico.

Recomendaciones

Para continuar desarrollando esta línea de investigación, se recomienda:

1. Aplicar métodos de validación cruzada a los modelos hallados.
2. Hacer mediciones en el tiempo para poder aplicar modelos de supervivencia y de esa forma poder ajustar el modelo de Framingham original a la población de Santa Clara.
3. Adquirir, por criterios de expertos, una clasificación del riesgo en: bajo, moderado y alto, para poder obtener índices de riesgo más generales.

Referencias Bibliográficas

SPSS 10 para Windows. Manual de usuarios. Cap. 12.

(1994) CHAID para SPSS sobre Windows. Técnicas de segmentación basadas en razones de verosimilitud Chi-cuadrado. .

3000, M.

ALBERTO MORALES SALINAS, J. M., YAQUELÍN LUNA, YURI MADRAZO, NORMA GONZÁLEZ, RAIMUNDO CARMONA, YISEL VILLANUEVA, EMILIO GONZÁLEZ RODRÍGUEZ, CARLOS MARTÍNEZ ESPINOSA. Riesgo coronario en trabajadores sin antecedentes de cardiopatía isquémica.

ALVAREZ A, D. L., LOPEZ V, PRIETO DIAZ MA Y SUAREZ S. (2005) Comparación de los modelos SCORE y Framingham en el calculo de alto riesgo cardiovascular para muestra de varones de 45 y 65 años de Asturias. .

ANDERSON KM, W. P., ODELL PM AND KANNEL WB (1991) Un update coronary risk profile. A statement for health professionals. .

BAENA-DÍEZ JM, G.-L. M., DE LA POZA-ABAD M, HERNANDEZ-IBANEZ R, MUNOZ-RUBIO A AND GARCÍA-REY Z. (2006) Estimation of overall cardiovascular risk from coronary risk. A cohort study.

BOWMAN TS, S. H. A. G. J. (2006) Effect of age on blood pressure parameters and risk of cardiovascular death in men. .

BREIMAN, F., OLSHEN AND STONE (1984) Classification and Regression Trees.

BUITRAGO F, C.-B. L., DÍAZ-HERRERA N, CRUCES-MURO E, ESCOBAR-FERNÁNDEZ M Y SERRANO-ARIAS JM. (2007) Comparación de las tablas REGICOR y SCORE para la clasificación del riesgo cardiovascular y la identificación de pacientes candidatos a tratamiento hipolipemiante o antihipertensivo. . *Revista Española de Salud Pública.*

COCA A, D. A., ESMATJES E, LLISTERRI JL, ORDOÑEZ J, GOMIS R, GONZALEZ-JUANATEY JR, MARTIN-ZURRO A; GRUPO PREVENCAT. (2006) Treatment and control of cardiovascular risk in primary care in Spain.

CONROY RM, P. K., FITZGERALD AP, SANS S, MENOTTI A AND DE BACKER G, ON BEHALF OF THE SCORE PROJECT GROUP. (2003) Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE Project. .

COX, D. R. (1972) Regression models and life tables. *Journal of de Royal Statistical Society B***,** 187-220.

CRISTÓBAL J, L. F., FUENTE J, GONZÁLEZ-JUANATEY JR, VÁZQUEZ-BELLÉS P Y VILA M. (2005) Ecuación de Framingham de Wilson y ecuación de REGICOR. Estudio comparativo. . *Revista Española de Salud Pública.*

DE BACKER G, A. E., BORCH-JOHNSEN K, BROTONS C, CIFKOVA R AND DALLONGEVILLE J (2003) Third Joint Task Force of European and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of eight societies and by invited experts). European guidelines on cardiovascular disease prevention in clinical practice. .

FISHER (1924) Obtención del Chi-Cuadrado de razón de verosimilitud.

GARCÍA, L. G. V. R. (2007) Muestreo para correlaciones por contingencia y de Pearson. Santa Clara, Cuba., Universidad Central Marta Abreu de Las Villas.

GRUNDY SM, P. R., GREENLAND P, SMITH S JR, FUSTER V. (1999) Assessment of cardiovascular risk by use of multiple-risk-factor assessment equations: a statement for

healthcare professionals from the American Heart Association and the American College of Cardiology. .

HALL, C. (2004) Logistic Regression: Who survived the sinking of the Titanic?

MAHALANOBIS (1936) Métodos para hallar funciones discriminantes.

MAIQUES A, A. F., TAIX MF, ALBERT X, MARTÍ EA AND COLLADO A. (2004) Riesgo cardiovascular del SCORE comparado con el de Framingham. Consecuencias del cambio propuesto por las Sociedades Europeas.

MANCIA G, D. B. G., DOMINICZAK A, CIFKOVA R, FAGARD R AND GERMANO G (2007) Guidelines for the Management of Arterial Hypertension: The Task Force for the Management of Arterial Hypertension of the European Society of Hypertensión (ESH) and of the European Society of Cardiology (ESC). .

MANTEL, N., AND HEANZEL, W. (1950) *Journal of National Cancer Institute,* 22(4)**,** 719-747.

MARRUGAT J, S. P., D'AGOSTINO R, SULLIVAN L, ORDOVAS J AND CORDÓN F (2003) Estimación del riesgo coronario en España mediante la ecuación de Framingham calibrada. . *Revista Española de Salud Pública.*

MORALES, A. R. (2003) Determinación de los factores pronósticos que influyen en la mortalidad infantil. Santa Clara, Cuba., Universidad Central Marta Abreu de Las Villas.

PEARSON., N. Y. (1928) Obtención del Chi- Cuadrado de razón de verosimilitud.

PÉREZ G, P. A., SALA J, ROSET PN, MASIÁ R, MARRUGAT J, AND THE REGICOR INVESTIGATORS. (1998) Acute myocardial infarction case fatality, incidence and mortality rates in a population registry in Gerona, Spain, 1990-1992.

PRESS, W. (2005) NUMERICAL RECIPES in C++. The Art of Scientific Computing. Second Edition.

QUINLAN, R.

R., G. Independencia de variables y medidas de asociación. Capítulo 3. Primera parte

R., G. Independencia de variables y medidas de asociación. Capítulo 3. Segunda parte.

RAO (1952) Obtención del estadístico V de Rao.

RODRIGUEZ, S. E. C. (2006) Tensoft II v2.0: Sistema informativo para el diagnóstico de la hipertensión arterial sobre bases estadísticas. Santa Clara, Cuba., Universidad Central Marta Abreu de Las Villas.

SIEGEL, S. (1970) *Diseño experimental no paramétrico,* La Habana.

STEVENS RJ, K. V., ADLER AI, STRATTON IM, HOLMAN RR ON BEHALF OF THE UNITED KINGDOM PROSPECTIVE DIABETES STUDY (UKPDS) GROUP. (2001) The UPKDS risk engine: a model for the risk of coronary heart disease in type II diabetes (UKPDS 56). .

UCLV, U. D. O. Y. "Proyección del Centro de Desarrollo Electrónico hacia la Comunidad" (PROCDEC)

VICÉNS OTERO, J. A. E. M. M. (2005) Análisis de datos cualitativos. .

VILLAR-ÁLVAREZ F, M.-G. A., BROTONS-CUIXART C, TORCAL-LAGUNA J AND BANEGAS-BANEGAS JJ (2005) Recomendaciones preventivas cardiovasculares en atención primaria.

WANG W, L. E., FABSITZ RR, DEVEREUX R, BEST L, WELTY TK AND HOWARD BV. (2006) A longitudinal study of hypertension risk factors and their relation to cardiovascular disease: the Strong Heart Study. .

WEISS, N. A. (2002) *Elementary Statistics,* Arizona.

WILSON P; D'AGOSTINO R, L. D., BELANGER A, SILBERSHATZ H AND KANNEL W. (1998) Prediction of Coronary Heart Disease Using Risk Factor Categories. .

WILSON PWF, D. A. R., LEVY D, BELANGER AM, SILBERSHATZ H, KANNEL WB. (1998) Prediction of coronary heart disease using risk factor categories.

ZAMORA RODRÍGUEZ, L. (1997) Una técnica de segmentación aplicada a la epidemiología. Santa Clara, Cuba. , Universidad Central Marta Abreu de Las Villas.

Anexos

Anexo 1: Tablas de contingencia de las variables discretas.

Sexo

Tabla de contingencia

		Sexo		Total
		Masculino	Femenino	
Riesgo	1	722	894	1616
	2	1254	304	1558
Total		1976	1198	3174

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	432.883	1	.000		
Corrección por continuidad	431.360	1	.000		
Razón de verosimilitud	447.581	1	.000		
Estadístico exacto de Fisher				.000	.000
Asociación lineal por lineal	432.747	1	.000		
N de casos válidos	3174				

Raza

Tabla de contingencia

		Raza		Total
		Blanca	Mestiza	
Riesgo	1	1356	260	1616
	2	1406	152	1558
Total		2762	412	3174

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	28.165	1	.000		
Corrección por continuidad	27.607	1	.000		
Razón de verosimilitud	28.489	1	.000		
Estadístico exacto de Fisher				.000	.000
Asociación lineal por lineal	28.156	1	.000		
N de casos válidos	3174				

Bebe

Tabla de contingencia

		Bebe		
		0	1	Total
Riesgo	1	964	652	1616
	2	608	950	1558
Total		1572	1602	3174

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	135.039	1	.000		
Corrección por continuidad	134.215	1	.000		
Razón de verosimilitud	136.022	1	.000		
Estadístico exacto de Fisher				.000	.000
Asociación lineal por lineal	134.997	1	.000		
N de casos válidos	3174				

Fuma

Tabla de contingencia

		Fuma		
		0	1	Total
Riesgo	1	1320	296	1616
	2	608	950	1558
Total		1928	1246	3174

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	605.351	1	.000		
Corrección por continuidad	603.564	1	.000		
Razón de verosimilitud	629.287	1	.000		
Estadístico exacto de Fisher				.000	.000
Asociación lineal por lineal	605.161	1	.000		
N de casos válidos	3174				

Diabetes Mellitus

Tabla de contingencia

		Diabetes mellitus		
		0	1	Total
Riesgo	1	1604	12	1616
	2	1102	456	1558
Total		2706	468	3174

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	513.470	1	.000		
Corrección por continuidad	511.204	1	.000		
Razón de verosimilitud	529.780	1	.000		
Estadístico exacto de Fisher				.000	.000
Asociación lineal por lineal	513.308	1	.000		
N de casos válidos	3174				

Dislipidemia

Tabla de contingencia

		Dislipidemia		
		0	1	Total
Riesgo	1	1588	28	1616
	2	1330	228	1558
Total		2918	256	3174

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	178.061	1	.000		
Corrección por continuidad	176.325	1	.000		
Razón de verosimilitud	199.927	1	.000		
Estadístico exacto de Fisher				.000	.000
Asociación lineal por lineal	178.005	1	.000		
N de casos válidos	3174				

Nro. de padres con HTA

Tabla de contingencia

		Nro. de padres con HTA			
		0	1	2	Total
Riesgo	1	784	664	168	1616
	2	836	418	304	1558
Total		1620	1082	472	3174

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	95.757	2	.000
Razón de verosimilitud	96.779	2	.000
Asociación lineal por lineal	2.369	1	.124
N de casos válidos	3174		

Nro. de abuelos con HTA

Tabla de contingencia

		Nro. de abuelos con HTA					
		0	1	2	3	4	Total
Riesgo	1	1180	326	86	2	22	1616
	2	1444	76	38	0	0	1558
Total		2624	402	124	2	22	3174

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	223.629	4	.000
Razón de verosimilitud	245.363	4	.000
Asociación lineal por lineal	166.262	1	.000
N de casos válidos	3174		

Diagnóstico

Tabla de contingencia

		Diagnóstico			
		Hipertenso	Hiperreactivo vascular	Normotenso	Total
Riesgo	1	390	420	806	1616
	2	950	342	266	1558
Total		1340	762	1072	3174

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	513.141	2	.000
Razón de verosimilitud	533.182	2	.000
Asociación lineal por lineal	510.938	1	.000
N de casos válidos	3174		

Anexo 2: Árbol de decisión