Universidad de Cienfuegos "Carlos Rafael Rodríguez"

Facultad de Ingeniería

Maestría en Matemática Aplicada

Trabajo de diploma para optar por el título de Máster en Matemática Aplicada

Título: "Modelo estadístico para predecir el padecimiento de pie diabético en pacientes con diabetes mellitus tipo II del municipio de Cienfuegos"

Autor: Ing. Rachel Yanes Seijo

Tutores: DrC. Raúl López Fernández

MSc. Pedro Roberto Suarez Suri

Cienfuegos, Cuba
2015

Agradecimientos

A mis padres por su amor y apoyo incondicional.

A mis tutores Raúl y Roberto por su dedicación, comprensión y enseñanzas durante la realización de este trabajo.

A Mirita por todos por su atenta asesoría y el tiempo dedicado a la hora de redactar el documento.

A los profesores que formaron parte del claustro de la maestría por contribuir a mi crecimiento profesional.

A mis compañeros de trabajo del Centro de Información de la Universidad de Ciencias Médicas de Cienfuegos por toda la ayuda brindada sobre todo a la doctora Milagrito.

A Elvia y a los trabajadores del departamento de Tecnología Educativa de la Universidad de Ciencias Médicas por su apoyo para la impresión y encuadernación de este documento.

Dedicatoria

A todos los que de una forma u otra me han brindado su amor, respeto y cariño.

Resumen

Los modelos estadísticos son de gran importancia en las ciencias de salud, como herramientas útiles para la toma de medidas preventivas que ayudan a sistematizar la mejor evidencia en la búsqueda de variables que expliquen o predigan determinados fenómenos.

La presente investigación ha tenido como propósito la elaboración de un modelo que permita predecir la aparición de pie diabético en pacientes con diabetes mellitus tipo II sobre la base de un conjunto de datos evidenciados en las historias clínicas de los pacientes del municipio de Cienfuegos, atendidos en la Clínica del Diabético en el período de enero del 2010 a diciembre del 2013.

Se utilizaron los factores de riesgos que influyeron de forma significativa con el padecimiento de pie diabético y para la elaboración de modelo se emplearon las técnicas multivariadas de regresión logística binaria, árboles de decisión y análisis de componentes principales para datos categóricos.

Como resultado se obtuvieron cuatro modelos comparables, el primero con regresión logística binaria, el segundo con regresión logística binaria combinada con análisis factorial, el tercero con árbol de decisión a través del algoritmo CHAID (Chi-squared Automatic Interaction Detection) y el cuarto con árbol de decisión a través del algoritmo CRT(Classification and Regression Trees); de los cuales se pudo comprobar que el primer y el tercer modelo fueron superiores a partir de los criterios de comparación considerados con este propósito: porcentaje de clasificación correcta, índice de Kappa, sensibilidad y especificidad.

Por último se estableció la validación a través de la curva ROC (Receiver Operating Characteristic, o Característica Operativa del Receptor) con los modelos seleccionados siendo el modelo con el algoritmo CHAID el más óptimo para la predicción del padecimiento del pie diabético en pacientes con diabetes mellitus tipo II del municipio de Cienfuegos, atendidos en la Clínica del Diabético.

Índice

Introducción

El mundo actual está marcado por un creciente y continuo desarrollo científico tecnológico, que ha transformado el progreso social y económico de los pueblos. Paralelo a esto, los instrumentos de investigación son cada vez más potentes y los investigadores están mejor capacitados, siendo la actividad científica más compleja, productiva y con gran cantidad de recursos.

En los últimos años son pocos los artículos e investigaciones que se realizan sin el apoyo de análisis estadísticos, destacándose las disciplinas que constituyen las ciencias de la salud, en las que muchos investigadores sienten que sus trabajos no tienen suficiente rigor científico si no vienen avalados, al menos, por un proceder estadístico. (Massip, Soler, & Torres, 2011)

La aplicación de herramientas estadísticas es imprescindible en esta rama; no solo como instrumento auxiliar en el enfrentamiento y solución de problemas, sino también como una vía para lograr una mejor comprensión de la literatura científica y una mayor claridad al comunicarse con otros profesionales, a propósito del análisis de los datos.

Con frecuencia, en el ámbito de la salud, es necesario determinar la probabilidad de ocurrencia de un evento, explicar las interrelaciones que existen entre ciertas variables y determinar los factores que afectan a la presencia o ausencia de un episodio adverso determinado. Los modelos estadísticos emergen como un vínculo importante entre la estadística y la práctica médica, definidos como una serie de técnicas matemáticas que combinadas adecuadamente ayudan a sistematizar la mejor evidencia en la búsqueda de variables que expliquen o predigan determinados fenómenos; cuya aplicación es suministrar una explicación matemática simplificada de dicha relación, siendo además válidos en la solución de muchos problemas de la realidad. (González, 2014)

Los modelos estadísticos permiten calcular la probabilidad de que ocurra determinado evento (o episodio adverso), dadas las condiciones o variables consideradas en el modelo, en base a la información de una muestra representativa de la población y extender este resultado a toda la población mientras se mantengan sin cambios sustanciales estas variables en la población, además de ser reproducibles en otras poblaciones de la misma naturaleza en busca de la generalización de los resultados obtenidos.

El desarrollo de modelos estadísticos predictivos en la salud ha crecido significativamente en los últimos años, siendo de gran ayuda en la toma de decisiones y permitiendo la creación de diversos sistemas y herramientas útiles para reducir las incertidumbres, garantizando mejores actuaciones y establecer eficaces medidas de control para la erradicación de las enfermedades. (Berea, et al., 2014)

Sin dudas, los modelos estadísticos pueden ocupar un importante espacio como recurso instrumental aplicable en el mundo de la investigación médico-social y dar respuesta a muchas de las necesidades de la sociedad actual, como es el caso de las enfermedades crónicas no transmisibles (ECNT), que actualmente son consideradas una gran preocupación médica a nivel mundial, debido al constante ascenso de su prevalencia e incidencia en la población adulta, así como por el lugar que ocupan dentro de las principales causas de muerte. (OMS, 2014)

La diabetes mellitus está incluida entre las ECNT como una enfermedad que se distribuye por todo el mundo y conlleva a una gran cantidad de morbilidades. Según estimaciones de la Organización Mundial de la Salud (OMS), la prevalencia de la esta enfermedad en el inicio del siglo XXI la sitúa en el 2,1% de la población mundial y junto a las enfermedades cardiovasculares, el cáncer y las enfermedades respiratorias crónicas, son responsables de la mayoría de las defunciones por ECNT a nivel mundial, con una tasa de mortalidad de aproximadamente el 63%. (OMS, 2014)

La Federación Internacional de Diabetes presentó en el año 2013 evidencias sobre el aumento abrupto de casos a nivel mundial y en particular de diabetes mellitus tipo II, que es la forma más común de dicho padecimiento; por lo general se manifiesta en la vida adulta y representa aproximadamente el 90% de los casos de diabetes. (OMS, 2014) (ALAD, 2013)

El prolongado período preclínico de la diabetes mellitus tipo II, hace que en ocasiones, los pacientes lleguen al diagnóstico con complicaciones crónicas de variada gravedad. Este problema merece una especial consideración, que se magnifica si se estima que entre un 30 y 50% de los pacientes no están diagnosticados, a pesar de tener signos clínicos de larga data que no fueron jerarquizados oportunamente. (OMS, 2014) (Valdés & Camps, 2013)

La expectativa de crecimiento se basa en la alta prevalencia de condiciones que preceden a la diabetes, como la obesidad y la intolerancia a la glucosa. Si los patrones demográficos continúan, se prevé que para el 2035 habrá 471 millones de diabéticos en todo el mundo (8% de la población adulta) de los cuales el 60% estarán en Centro y Suramérica, y sin una acción preventiva concertada, en menos de 25 años serían más de 592 millones. (ALAD, 2013) (Acosta, Hernández, Victorero, & Cruz, 2013)

Estadísticas recientes confirman que en Latinoamérica existen no menos de 15 millones de personas con diabetes mellitus y se prevé que más de 300 millones de personas se enfermarán por esta patología, con mayor impacto en mayores de 65 años. (OMS, 2014)

En Cuba, según el Anuario Estadístico del Ministerio de Salud Pública, al cierre del año 2013, la tasa de prevalencia de diabetes mellitus fue de 53,7 por 1000 habitantes y se registraron 2246 defunciones con este padecimiento, siendo una de las principales causas de muerte en todas las edades, para un tasa de 20,1 por cada 100 000 habitantes. (MINISTERIO DE SALUD PÚBLICA, 2013)

En la provincia de Cienfuegos se informó durante el 2013 una prevalencia de diabetes mellitus de 52 por 1000 habitantes y la tasa bruta de defunciones por esta enfermedad fue de 22,2. (MINISTERIO DE SALUD PÚBLICA, 2013)

Las principales complicaciones de la diabetes mellitus pueden ser metabólicas y además pueden afectar múltiples órganos, entre ellos, los miembros inferiores, siendo el pie el que por diferentes factores presenta mayores afectaciones, con altas probabilidades de presentar ulceraciones que generalmente se tornan complicadas y con infecciones que van desde leves hasta severas; a esta patología se le denomina pie diabético y es una de las complicaciones más frecuentes y significativas de la diabetes mellitus. (Mass, et al., 2014).

En estadísticas mundiales se evidencia que cada año, del 1 al 4% de los pacientes diabéticos padecen úlcera en sus pies; y entre el 10 y el 15%, tienen altas probabilidades de tenerla en algún momento de su vida. (Jiménez, 2014)

Una vez que aparecen las úlceras no sólo se pone en peligro el miembro afectado, sino incluso la vida del paciente puesto que se considera que entre un 15 y el 30 % de los pacientes diabéticos con este padecimiento requiere la amputación del miembro inferior. En Cuba se realizan alrededor de 1000 amputaciones de miembros inferiores cada año. (Bustillo, Feito, Vegoña, Alvarez, & Gurra, 2014)

Según criterios de especialistas, frecuentemente las amputaciones están precedidas por infección y gangrena y en un 40 % de los pacientes amputados se produce una segunda amputación durante los cinco años siguientes a la primera, con una mortalidad entre un 50 y un 60%. (Barnés, et al., 2014)

La necesidad de comprender, prever y estudiar el padecimiento de pie diabético es una cuestión primordial y representa un gran reto médico. Trazar estrategias de intervención para disminuir este padecimiento, puede traducirse en resultados positivos para mejorar la calidad de vida de los pacientes diabéticos, así como desde

la perspectiva socioeconómica, debido a la alta prevalencia de la diabetes en la población laboralmente activa.

El manejo de esta complicación debe ser multidisciplinario, oportuno y eficaz, para reducir potencialmente su progresión y la morbilidad relacionada. Desde este enfoque multidisciplinario, puede entenderse la relevancia de la aplicación de modelos estadísticos, no solo como una herramienta útil para recopilar y describir datos, sino también para comprender e interpretar la magnitud de la influencia los diversos factores de riesgos que determinan de forma significativa en la aparición de este padecimiento y así facilitar evidencias que permitan establecer un mejor pronóstico con diagnóstico oportuno y el tratamiento adecuado según el estado evolutivo del paciente.

El estudio realizado en la Clínica del Diabético del municipio de Cienfuegos, (que es una entidad docente-asistencial a la que son remitidos pacientes de las consultas multidisciplinarias de las diferentes áreas de salud o del Hospital Provincial y se consultan a personas con diabetes complicada que presenten dificultades en su control y manejo), ha recopilado una información estadística considerable sobre el comportamiento de los factores de riesgo del padecimiento de pie diabético, lo cual posibilita profundizar en el estudio del sistema de relaciones entre la aparición del pie diabético en la población de personas con diabetes mellitus y los factores de riesgo.

El estudio bibliográfico realizado evidencia la escasez de estudios estadísticos con el empleo de técnicas multivariadas que permitan pronosticar el padecimiento de pie diabético a partir de los factores de riesgo que influyen con mayor significación y de forma conjunta sobre dicha enfermedad; solo se reflejan artículos de descripción y caracterización del padecimiento, en los que se identifica una alta incidencia de la enfermedad y un alto desconocimiento de esta por parte de los pacientes que presentan dicha complicación. (Garriga, Sánchez, & Vázquez, 2014)

Otros estudios, a pesar de los hechos científicos que aportan, se realizaron en contextos socio-económicos y culturales diferentes, lo que puede hacer variar el

comportamiento de estos factores. En Tijuana México se identifican factores de riesgos que influyen en las manifestaciones tempranas de nefropatía diabética que pueden provocar padecimiento de pie diabético, utilizan regresión logística para la caracterización de estos factores y aplican el cuestionario síntomas de neuropatía diabética (SND) que mide la deformidad del pie y evalúa la sensibilidad con el monofilamento de Semmes-Weinstein para clasificar los pacientes en alto o bajo riesgo de padecer pie diabético.(Márquez, Zonana, Anzaldo, & Muñoz, 2014) En Colombia utilizan el coeficientes de Correlación de Pearson y estadísticos descriptivos para identificar los factores de riesgo de pie diabético a través de la evaluación nutricional y la prevalencia actividades de prevención de este padecimiento por médicos y pacientes. (Pinilla, Barrera, C., & D., 2014)

Teniendo en cuenta la carencia en nuestro contexto de estudios estadísticos con el empleo de técnicas multivariadas, que permitan profundizar en el conocimiento del sistema de relaciones existentes entre la aparición del pie diabético y los factores de riesgo, así como la necesidad de predecir de forma temprana la aparición de esta enfermedad, para apoyar la toma de decisiones sobre políticas terapéuticas y educativas con fines preventivos, con un basamento científico, que aporten elementos para el diseño de estrategias de prevención e intervención a escala individual, familiar y comunitaria

Se plantea como **problema de investigación** la siguiente interrogante ¿Cómo establecer un modelo estadístico que permita predecir la aparición de pie diabético en la población de pacientes con diabetes mellitus de tipo II a partir de factores de riesgo?

Objeto de la investigación: Los modelos estadísticos multivariados de clasificación.

Campo de acción: Los modelos estadísticos multivariados de clasificación con variable dependiente categórica y variables independientes mixtas, que permitan predecir el padecimiento de pie diabético en los pacientes con diabetes mellitus tipo II.

Se plantea como **Objetivo General:** Establecer un modelo estadístico que permita predecir el padecimiento de pie diabético en pacientes con diabetes mellitus tipo II a partir de factores de riesgo.

Para el cumplimiento del Objetivo General se tuvieron en cuenta los siguientes **Objetivos específicos**:

1. Resumir los modelos estadísticos utilizados en las ciencias de la salud que permiten predecir el comportamiento de una variable dependiente categórica a partir de un grupo de variables predictoras mixtas.
2. Analizar la relación entre los factores de riesgo y el padecimiento de pie diabético en los pacientes con diabetes mellitus tipo II del municipio de Cienfuegos atendidos en la Clínica del Diabético.
3. Estimar modelos estadísticos alternativos que permitan predecir el padecimiento de pie diabético en pacientes con diabetes mellitus tipo II a partir de los factores de riesgo.
4. Validar los modelos estadísticos estimados para predecir el padecimiento de pie diabético a partir de los factores de riesgo en pacientes con diabetes mellitus tipo II del municipio de Cienfuegos atendidos en la clínica del diabético.

Como **hipótesis de investigación** se tiene: Si se establece un modelo estadístico predictivo, entonces se conocerá la probabilidad de que un paciente sea afectado por el padecimiento de pie diabético.

- **Variable dependiente:** Probabilidad de padecimiento de pie diabético.
- **Variable independiente:** El modelo estadístico que predice el padecimiento de pie diabético en los pacientes con diabetes mellitus tipo II a partir de los factores de riesgo.

La investigación se **justifica** por la necesidad de prevenir el padecimiento de pie diabético en pacientes con diabetes mellitus tipo II a partir de los factores de riesgo,

lo cual permite disminuir la incidencia, morti-morbilidad y los costos por tratamientos de la enfermedad.

El **aporte práctico** está dado por el establecimiento de un modelo estadístico para predecir el padecimiento de pie diabético en pacientes con diabetes mellitus tipo II a partir de los factores de riesgo, que puede constituir un recurso de apoyo para la toma de decisiones respecto a programas de prevención, educativos o tratamientos tempranos a los pacientes de riesgo.

Metodología aplicada en la investigación

Se trabajó con los siguientes métodos científicos:

Entre los métodos, técnicas y procedimientos del nivel teórico se utilizaron: Histórico-Lógico, Analítico-Sintético, Inductivo-Deductivo.

Teóricos

- Histórico-lógico

  El método Histórico-Lógico permitió analizar la evolución del uso de la estadística en ciencias de la salud, en particular de los modelos estadísticos para predecir el pronóstico de una enfermedad a partir de factores de riesgo y para analizar el comportamiento del pie diabético en los pacientes atendidos en la Clínica del Diabético en el período del año 2010 al 2013.

- Analítico-sintético

  El método Analítico-Sintético se concretó en el estudio de las valoraciones sobre los modelos estadísticos predictivos utilizados en salud y sobre el padecimiento de pie diabético que permitió comprender, resumir, unificar y generalizar los fundamentos teóricos para procesar la información relacionada con los aspectos

abordados en la investigación, así como arribar a conclusiones parciales y generales.

- Inductivo-Deductivo

Estrechamente vinculado al método Analítico Sintético, permitió, teniendo en cuenta los aspectos generales de modelos estadísticos predictivos y el padecimiento de pie diabético, poder dar respuesta a las interrogantes planteadas en el desarrollo del trabajo investigativo, así como a partir de casos particulares inferir proposiciones que posteriormente se deducen, generalizan y particularizan nuevamente en cuanto al padecimiento de pie diabético.

Del nivel empírico se utilizaron: Análisis de documentos, Modelación matemática y Métodos Estadísticos de los niveles descriptivo e inferencial.

Empírico

- Análisis de documentos

El Análisis de documentos facilitó indagar sobre los factores de riesgo del padecimiento de pie diabético y los modelos estadísticos predictivos que con mayor frecuencia son utilizados en las ciencias de la salud.

- Modelación matemática

Con el uso de la modelación matemática se seleccionó, estimó y validó un modelo estadístico predictivo para el pronóstico del padecimiento de pie diabético a partir del conocimiento del comportamiento de los factores de riesgo.

- Métodos Estadísticos:

  – Pruebas de hipótesis Kolmogorov-Smirnov para la verificar la normalidad de las variables; prueba T-Student y U de Mann-Whitney para comparar los grupos de la variable dependiente respecto a las variables

independientes cuantitativas; prueba Chi Cuadrado de independencia y coeficientes de asociación Phi, V Cramer y Spearman, para establecer la existencia de relaciones entre la variable dependiente y las independientes categóricas.

- Se aplicaron el escalamiento óptimo y análisis factorial; modelos de regresión logística binaria y árbol de decisión empleando los algoritmos CHAID y CRT.
- Para la validación del modelo se construye la curva ROC y se interpreta el cálculo de área bajo la curva

El tratamiento de los datos fue a través del paquete de programas estadísticos SPSS v.15

El trabajo quedó estructurado de la siguiente manera: Resumen, Introducción, tres capítulos, Conclusiones, Recomendaciones, Referencias Bibliográficas y Anexos.

En el **Capítulo I** se abordan aspectos teóricos que sustentan la investigación realizada. Se reseña una breve historia de la aplicación de la estadística en las ciencias de la salud. Se exponen de forma resumida algunas de las técnicas utilizadas en esta rama de la ciencia para la elaboración modelos estadísticos predictivos; se profundiza en aquellas que van a ser posteriormente aplicadas para darle cumplimiento a los objetivos propuestos y se presentan, además, reportes de sus aplicaciones vinculados a la salud. Se presentan estudios realizados acerca de la patología del pie diabético.

En el **Capítulo II** se dedica a la descripción del procedimiento para establecer un modelo predictivo del padecimiento de pie diabético en los pacientes con diabetes mellitus de tipo II hospitalizados en la Clínica del Diabético del municipio de Cienfuegos. Se tienen en cuenta en el estudio variables clínicas, sociodemográficas, hábitos de consumo y de antecedentes patológicos y personales que se registran en la clínica. Se exponen los pasos del procedimiento empleado y se describe, además, la muestra de casos utilizada para la obtención del modelo y los criterios para la

selección de las variables predictoras. Se utilizan técnicas multivariadas de regresión logística binaria y árboles de decisión para la elaboración del modelo y, por último, se presenta la variante propuesta al procedimiento para la obtención del modelo predictivo y los criterios a seguir para la comparación de los modelos obtenidos por ambos procedimientos.

En el **Capítulo III** se presentan los resultados relacionados con la aplicación del procedimiento de elaboración del modelo estadístico para predecir el padecimiento de pie diabético en pacientes con diabetes mellitus tipo II a partir de los factores de riesgo en el municipio de Cienfuegos, basado en un modelo que prediga el mismo en la Clínica del Diabético, sobre la base de un conjunto de variables que inciden en la ocurrencia de dicho padecimiento, siguiendo el procedimiento expuesto en el capítulo anterior a partir del uso de la regresión logística y árbol de decisión sobre los componentes obtenidos.

Capítulo I: La aplicación de modelos estadísticos predictivos en el campo de la salud

En este capítulo se abordan aspectos teóricos que sustentan la investigación realizada. Se reseña una breve historia de la aplicación de la estadística en las ciencias de la salud. Se exponen de forma resumida algunas de las técnicas utilizadas en esta rama de la ciencia para la elaboración modelos estadísticos predictivos; se profundiza en aquellas que van a ser posteriormente aplicadas para darle cumplimiento a los objetivos propuestos y se presentan, además, reportes de sus aplicaciones vinculados a la salud. Se presentan estudios realizados acerca de la patología del pie diabético.

1.1 Breve introducción del uso de la estadística en salud

La aplicación de la estadística se ha extendido a casi todos los campos de la actividad humana al constituirse en una herramienta imprescindible para caracterizar los objetos y fenómenos del mundo y sus interrelaciones. Con el desarrollo de la informática los más complejos procedimientos estadísticos son asequibles a los investigadores, lo que ha generado un auge considerable en el uso de la estadística en las investigaciones. El análisis de los resultados de las investigaciones en el campo de la salud en las últimas décadas muestra un incremento acelerado del uso de los procedimientos estadísticos como consecuencia del desarrollo tecnológico, el incremento de la práctica investigativa y la elevación del nivel científico del personal de la salud.

El inicio de las aplicaciones estadísticas en salud se comienza con el análisis de los registros de nacimiento y de mortalidad sanitarias y coincide con un extraordinario avance de las ciencias naturales que se reflejó por el inglés Thomas Sydenham entre 1650-1676 en las descripciones clínicas de diferentes enfermedades tales como: de la disentería, la malaria, la viruela, la gota, la sífilis y la tuberculosis. Los trabajos de este autor resultaron esenciales para reconocer estas patologías como entidades

distintas y dieron origen al sistema actual de clasificación de enfermedades. (Lilienfeld & Lilienfeld, 1987)

En los siguientes años, el estudio de la enfermedad poblacional, bajo este método, condujo a la elaboración de un sinnúmero de "leyes de la enfermedad", que inicialmente se referían a la probabilidad de enfermar a determinada edad, a la probabilidad de permanecer enfermo durante un número específico de días y a la probabilidad de fallecer por determinadas causas de enfermedad. (Ruiz, 2006)

La búsqueda de "leyes de la enfermedad" fue una actividad permanente hasta el final del siglo XIX que contribuye al desarrollo de la estadística moderna. Durante este proceso, la incursión de la probabilidad en el estudio de la enfermedad fue casi natural, surge así la idea de una "ley de mortalidad" y, poco más tarde, la convicción de que habría leyes para todas las desviaciones sociales: el suicidio, el crimen, la vagancia, la locura y, naturalmente la enfermedad. (López, Corcho, & A.A., 1999)

Se descubren las relaciones entre prevalencia, incidencia y la duración de enfermedades y se fundamenta la necesidad de contar con grupos de casos para lograr inferencias válidas, con lo que se inauguran los conceptos de término medio y normalidad biológica, categorías ampliamente usadas durante la inferencia epidemiológica. (OPS, 1988)

El incremento en la incidencia de enfermedades crónicas ocurrido a mediados del siglo XX también contribuye a ampliar el campo de acción de la estadística en salud, que desde los años cuarenta del siglo XX se ocupa del estudio de la dinámica del cáncer, la hipertensión arterial, las afecciones cardiovasculares, las lesiones y los padecimientos mentales y degenerativos. (Star & Moghadas, 2010)

En esta etapa se demuestra la comprobación de la relación existente entre el consumo de cigarrillos y el cáncer de pulmón; entre radiaciones ionizantes y determinadas formas de cáncer; entre exposición a diversas sustancias químicas y tumores malignos; entre obesidad y diabetes mellitus; entre consumo de estrógenos

y cáncer endometrial; entre uso de fármacos y malformaciones congénitas, y entre sedentarismo e infarto de miocardio. (Star & Moghadas, 2010)

Se definen con mayor precisión los conceptos de exposición, riesgo, asociación, confusión y sesgo y se incorpora el uso de la teoría de la probabilidad, de técnicas avanzadas y modelos estadísticos predictivos como herramientas útiles para la toma de medidas preventivas e identificación de los factores de riesgos. (López, Corcho, & Moreno, 1999)

1.2 Modelos estadísticos aplicados en salud para predecir enfermedades a partir de factores de riesgo.

El desarrollo de modelos de pronóstico para predecir un fenómeno en particular no es una tarea fácil. Obtener modelos de pronóstico que presenten altos niveles de correcta clasificación conlleva a realizar un profundo estudio del problema concreto que se esté tratando y los factores de riesgo asociados. En dependencia de la propia naturaleza de dicho problema, será factible la utilización de unos u otros modelos de pronóstico o predictivos. En los epígrafes siguientes se exponen algunas técnicas estadísticas para la estimación de modelos predictivos.

1.2.1 Factores de riesgo en salud

El **riesgo** es cualquier característica o circunstancia detectable de una persona o grupo de personas que se sabe asociada con un aumento en la probabilidad de padecer, desarrollar o estar especialmente expuesto a un proceso específico. El conjunto de estos riesgos se les denomina **factores de riego**, y cada uno de ellos proviene de diferente naturaleza (biológica, ambiental, de comportamiento, socio-culturales, económica) que sumándose unos a otros desencadenan un aumento del efecto aislado de cada uno de ellos produciendo un fenómeno de interacción. (Echemendía, 2011).

El grado de asociación entre el factor de riesgo y la enfermedad, se cuantifica con determinados parámetros que son:

- **Riesgo individual:** es la relación entre la frecuencia de la enfermedad en los sujetos expuestos al probable factor causal y la frecuencia en los no expuestos.
- **Riesgo relativo:** es una comparación de la frecuencia con que ocurre el daño en los individuos que tienen el atributo o factores de riesgo y la frecuencia con que acontece en aquellos que no la tienen el factor de riesgo
- **Riesgo atribuible:** es parte del riesgo individual que puede ser relacionado exclusivamente con el factor estudiado y no con otros, es una medida útil para mostrar la proporción en que el daño podría ser reducido si los factores de riesgo causales desaparecieran de la población total.
- **Fracción etiológica del riesgo** es la proporción del riesgo total de un grupo, que puede ser relacionada exclusivamente con el factor estudiado y del resto del mundo. (Gonzalez & Agudo, 2003)

Ningún riesgo existe de forma aislada, en todos los lugares las personas están expuestas a lo largo de su vida a una serie prácticamente ilimitada de riesgos para su salud, en forma de enfermedades transmisibles o no transmisibles, traumatismos, productos de consumo, actos violentos o catástrofes naturales.

La prevención y el control de los factores de riesgo constituyen uno de los principios básicos de los profesionales de la salud; siendo necesario e indispensable profundizar en el conocimiento de aquellos que benefician o perjudican el desarrollo y comportamiento del organismo humano. Su correcta interpretación, permite el enfrentamiento adecuado para lograr la conservación de la salud. (González, 2014)

1.2.2 Modelos estadísticos

Una de las formas de representar el conocimiento que poseemos de la realidad es a través de modelos que caractericen de forma simplificada la esencia de los

fenómenos u objetos reales. En particular los modelos estadísticos permiten establecer de forma simbólica el sistema de relaciones existentes entre las variables que describen un fenómeno determinado. (Star & Moghadas, 2010)

Los modelos estadísticos resultan de gran utilidad cuando se trata de estudiar problemas de la vida real y explicar el comportamiento de enfermedades, aun cuando la operación no sea altamente estable. Se expresan por una ecuación matemática de la forma $Y=S+A$ que reproduce los fenómenos que observamos, contienen una parte sistémica (S) y una parte aleatoria (A), y la combinación de ambos componentes debe ser capaz de reproducir el objeto de interés (Y). (Carrasco, 2010)

En las ciencias de la salud los modelos estadísticos con carácter predictivo tienen una amplia aplicación. Estos se obtienen mediante la aplicación de una serie de técnicas estadísticas que se combinan adecuadamente para el análisis de la información registrada de las variables que caracterizan el fenómeno hasta obtener un modelo que evidencie un ajuste adecuado al objeto real. Se utilizan para la creación de modelos epidemiológicos sobre transmisión de enfermedades y los destinados a la evaluación económica, hasta aquellos que ayudan a resolver problemas de toma de decisión o prevención, tales como: la elección de terapéuticas de manera más personalizada, elección de programas de salud más efectivos y la identificación de grupos de factores de riesgo**.** (Echemendía, 2011)

Con frecuencia se aprecian investigaciones con grandes volúmenes de información, en los que para realizar sus análisis es necesario aplicar técnicas estadísticas avanzadas, como las multivariadas que en la actualidad, y gracias a los avances tecnológicos, es posible aplicarlas, ya que en general se encuentran implementadas en la mayoría de los programas estadísticos.

1.2.3 Procedimientos para establecer modelos estadísticos.

Cuando se tiene como propósito estimar un modelo, cuya finalidad sea emplearlo con fines predictivos, se puede sugerir un grupos de acciones válidas independientemente de la naturaleza del problema que se esté abordando. Diferentes autores coinciden que para poder desarrollar, interpretar y validar cualquier modelación estadística, deben tenerse en cuenta los siguientes aspectos:

- **Objetivos del análisis**.

  Se debe plantear y tener claro el problema a enfrentar y los objetivos propuestos, definir los conceptos y las relaciones fundamentales (dependencia o de interdependencia) que se van a investigar; una vez aclarados estos criterios se determinan las variables a observar de tal forma que no exista ambigüedad.

- **Diseño del análisis.**

  Se establece el tamaño de la muestra, las técnicas de estimación y los procedimientos estadísticos a aplicar partiendo de la naturaleza del problema a tratar y el tipo de información que se disponga. En función de estos elementos se opta por uno u otro procedimiento para obtener el modelo de pronóstico deseado

- **Análisis de los supuestos da la técnica a emplear.**

  Se evalúan las hipótesis subyacentes a la técnica multivariante escogida. Dichas hipótesis pueden ser de normalidad, linealidad, independencia, homocedasticidad, etc. También se debe decidir qué hacer con los datos perdidos.

- **Estimación del Modelo.**

  Se realiza el análisis y se evalúa el ajuste a los datos. En este paso pueden aparecer observaciones atípicas cuya influencia se debe analizar. Una vez

estimado el modelo, debe valorarse su significación estadística. Para ello se toman como referencia las herramientas que se adecuen a la técnica en particular con la que se esté tratando y determinar el grado de ajuste de forma tal que se pueda realizar pronósticos confiables a partir del modelo obtenido.

- **Explicación de los resultados.**

Se interpretan los resultados producidos, si estos fuesen decepcionantes o cuestionables, puede ser necesario revisar de nuevo las hipótesis y la estimación; así como alguna posible medida remedial a aplicar en caso de ser posible. Debe considerarse, además, si se cumplen determinados requerimientos basados en la experiencia de investigaciones precedentes.

- **Validación del análisis.**

Para confirmar que los resultados obtenidos son pertinentes; y pueden generalizarse a poblaciones como de las que proceden, es recomendable realizar un procedimiento de validación para determinar su estabilidad. Para esto se puede proceder a seleccionar una muestra adicional si es posible o, por el contrario, dividir la muestra si su tamaño lo permite y comparar si existen diferencias significativas entre uno y otro. Los criterios de comparación dependen de la técnica y particular y de los propósitos de la investigación. (Salvador, 2000) (Hair, Anderson, Tatham, & Black, 2004)

1.2.4 Técnicas multivariadas para elaborar modelos estadísticos de uso más frecuente en salud

Las técnicas estadísticas multivariadas son la base para la elaboración de modelos predictivos; sistematizan la mejor evidencia en la búsqueda de variables o factores de riesgo que expliquen o predigan determinados fenómenos. Se emplean comúnmente en problemas de reducción de dimensionalidad, en donde el conjunto de datos multidimensional es proyectado en un espacio de menor dimensionalidad y

los resultados de estas proyecciones pueden ser empleados para visualización y/o para determinar asociaciones significativas entre variables. (Cuadras, 2014)

Cuando se han observado "p" características numéricas (i=1,..., p) sobre "n" individuos (j=1,..., n) los resultados obtenidos pueden escribirse en una matriz de datos de dimensión p x n

$$x = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \vdots & \vdots \\ x_{p1} & \cdots & x_{pn} \end{bmatrix}$$

Cada columna de esta matriz se refiere a un individuo y constituye un vector X_j

$$x_j = \begin{bmatrix} x_{1j} \\ \vdots \\ x_{pj} \end{bmatrix}$$

Esta matriz de datos se reduce bajo la forma de parámetros: la media, las varianzas y co-varianzas, las desviaciones típicas y los coeficientes de correlación.

- $x_i = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$ Las medias
- $S = \begin{bmatrix} S_{11} & \cdots & S_{1p} \\ \vdots & \vdots & \vdots \\ S_{p1} & \cdots & S_{pp} \end{bmatrix}$ Las varianzas y co-varianzas
- $S_1 = \sqrt{S_{11}}, S_2 = \sqrt{S_{22}}, \ldots, S_P = \sqrt{S_{PP}}$ Las desviaciones típicas
- $R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & & 1 & r_{2p} \\ \vdots & & & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$ Los coeficientes de correlación

A partir de las matrices S y R se realizan casi todos los análisis multivariantes. El elemento esencial de estos análisis es el valor teórico y la combinación lineal de variables con ponderaciones determinadas empíricamente. El investigador especifica las variables, mientras que las ponderaciones son objeto específico de determinación por parte de la técnica multivariante. Un valor teórico de "n" variables ponderadas (X_1, X_2,...., X_n) puede expresarse matemáticamente:

$$valor\ teórico\ = w_1 X_{1+} w_2 X_2 + \cdots + w_n X_n$$

Donde X_n es la variable observada y w_n es la ponderación determinada por la técnica multivariante. El resultado es un valor único que representa una combinación de todo el conjunto de variables que mejor se adaptan al objeto multivariante específico. (Cuadras, 2014)

Algunas de las técnicas más empleadas y conocidas en el análisis multivariante son:

- **Análisis de Regresión.** Se emplea frecuentemente a efectos de clasificación, su objetivo es predecir los cambios en la variable dependiente en repuesta a cambios en varias variables independientes. Existen varios modelos de regresión, pero entre los más usados en el campo de la salud se encuentran la regresión logística binaria y la regresión logística multinomial, que a diferencia de la regresión lineal múltiple exige que la variable dependiente sea categórica.
- **Análisis Discriminante**. Es útil en situaciones donde la muestra total puede dividirse en grupos basándose en una variable dependiente caracterizada por varias clases conocidas. Los objetivos primarios son entender las diferencias de los grupos y predecir la verosimilitud de que una entidad (pacientes) pertenezca a una clase o grupo particular basándose en varias variables independientes. Esta técnica estadística exige que las variables independientes sean cuantitativas, lo que limita considerablemente su

aplicación en problemas del campo de la salud donde predominan las variables con un nivel de medición nominal y ordinal.

- **Análisis Factorial y componentes principales**. Se emplea en la valoración de pautas de variables, es una aproximación estadística que puede usarse para analizar interrelaciones entre un gran número de variables y explicar estas en términos de sus dimensiones subyacentes comunes (factores). El objetivo es encontrar un modo de condensar la información contenida en un número de variables originales en un conjunto más pequeño de variables ficticias con una pérdida mínima de información.
- **Análisis de Clúster.** Clasifica una muestra de entidades (individuos o variables) en un número pequeño de grupos de forma que las observaciones pertenecientes a un grupo sean muy similares entre sí y muy diferentes a las del resto de los grupos. (Hair, Anderson, Tatham, & Black, 2004)

Además, se han desarrollado otras técnicas multivariadas basadas en técnicas de segmentación que han sido poco estudiadas pero se hacen cada vez más populares en su uso, siendo una herramienta valiosa para ajustar objetos con el fin de identificar grupos para una consideración futura, como por ejemplo:

- **Árbol de decisión**. Son particiones secuenciales del conjunto de datos para maximizar las diferencias de los individuos respecto a la variable dependiente, ofrecen una forma concisa de desarrollar grupos que son consistentes en sus atributos pero que varían en términos de la variable dependiente. (Segura, 2012)

Para la correcta utilización de un modelo estadístico este necesita ser evaluado, pues en ocasiones es creado en contextos diferentes donde se ejecutó y es conveniente saber cómo ha sido obtenido, a qué población se aplica, sus posibilidades y también sus deficiencias. La modelación tiene carácter subjetivo durante su formulación, el dominio de la solución de problemas y las variables a

considerar abarca decenas y hasta miles de soluciones; seleccionar unas cuantas es un proceso que lleva la impronta del investigador. (González, 2014)

1.3 Técnicas estadísticas utilizadas en la presente investigación para la elaboración modelos predictivos.

En este epígrafe se van a abordar fundamentalmente la regresión logística binaria y los árboles de decisión por ser una de las herramientas a emplear en la presente investigación, para la elaboración de modelos predictivos; se atenderá también el uso de análisis factorial como técnica auxiliar para garantizar cualidades deseables en los modelos obtenidos. Se expondrán los fundamentos teóricos de cada una de ellas, además de hacer referencia a algunas aplicaciones encontradas en la salud.

1.3.1 Regresión logística binaria

Uno de los elementos que más ha contribuido al avance de la epidemiología en los últimos años ha sido el desarrollo de determinados métodos de análisis como la regresión logística que se encuentra entre las técnicas multivariadas más utilizadas en las ciencias de la salud. (Medina, Palma, Zapata, Pérez, & Zuniga, 2014)

El término regresión fue utilizado por primera vez a finales del siglo pasado por un investigador llamado Francis Galton, el análisis de regresión trata de la dependencia de una variable, la variable dependiente (variable que suscita especial interés) que suele denominarse Y, con respecto a una o más variables explicatorias o independientes que suelen denominarse X_1, X_2, …, X_n, con el objetivo de estimar o predecir la media o valor promedio (poblacional) de la primera con base a valores conocidos o fijados (en muestras repetidas) de las segundas; de ninguna forma se establece una relación formal de causa efecto, solo se trata de una relación matemática empírica. (Cuadras, 2014)

La regresión logística binaria es la técnica más ampliamente utilizada para modelar respuestas discretas, la variable dependiente o respuesta presenta dos categorías,

codificándose con los valores uno (1) y cero (0), respectivamente. Por lo que se refiere a las variables independientes o explicativas, no se establece ninguna restricción, pudiendo ser cuantitativas, tanto continuas como discretas y categóricas, con dos o más modalidades.

Lo que se pretende mediante el modelo de regresión logística es expresar la variable dependiente en términos de probabilidad, utilizando la función logística para estimar la probabilidad de que ocurra el evento en cuestión, dado determinados valores de las variables explicativas ($X_1, X_2, X_3...X_n$) que se presumen relevantes o influyentes, mediante la siguiente formulación:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}$$

Donde $\beta_0, \beta_1, \beta_2, \cdots \beta_n$ son los parámetros del modelo.

Para la estimación de los coeficientes del modelo se recurre al método de máxima verosimilitud y para la realización de inferencias sobre la significación estadística de las variables se emplean básicamente tres métodos: el estadístico de Wald, el estadístico G de razón de verosimilitud y la prueba Score. (Berlanga & Vilà, 2014)

- **El estadístico de Wald**

  Contrasta la hipótesis de que un coeficiente aislado es distinto de 0, y sigue una distribución normal de media 0 y varianza 1. Su valor para un coeficiente concreto viene dado por el cociente entre el valor del coeficiente y su correspondiente error estándar. La obtención de significación indica que dicho coeficiente es diferente de 0 y merece la pena su conservación en el modelo. En modelos con errores estándar grandes, el estadístico de Wald puede proporcionar falsas ausencias de significación.

- **El estadístico G de razón de verosimilitud**

Se trata de ir contrastando cada modelo que surge después de eliminar de forma aislada cada una de las covariables frente al modelo completo. En este caso cada estadístico G sigue una distribución χ^2 con un grado de libertad (no se asume normalidad). La ausencia de significación implica que el modelo sin la covariable no empeora respecto al modelo completo (da igual su presencia o su ausencia), por lo que según la estrategia de obtención del modelo más reducido (principio de parsimonia), dicha covariable debe ser eliminada del modelo ya que no aporta nada al mismo.

Esta prueba no asume ninguna distribución concreta, por lo que es la más recomendada para estudiar la significación de los coeficientes.

- **La prueba Score**

Su cálculo para el caso de una única variable viene dado por:

$$S = \frac{\sum_{i=1}^{n}(y_i - \bar{y})}{\sqrt{\bar{y}(1-\bar{y})}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

En el caso de múltiples covariables hay que utilizar cálculo matricial, si bien no requiere un cálculo iterativo. Este estadístico se incrementa conforme aumenta el número de covariables y también asume una distribución normal con media 0 y varianza 1. Si alcanza significación indica que la covariable debería permanecer en el modelo.

Cuando la covariable es cualitativa con n categorías (siendo n > 2), en el modelo se analizará la significación de cada una de sus n −1 variables ficticias, así como la significación global de la covariable comparando la presencia en bloque frente a la ausencia en bloque de sus n −1 covariables ficticias.

Para obtener una buena ecuación de regresión en el sentido estadístico que no contenga variables de poco peso y con el fin de encontrar el número óptimo de

variables independientes a incluir en el modelo, pueden aplicarse distintos métodos de selección de variables, que son de tipo secuencial y que se puede construir utilizando las siguientes técnicas:

- **Técnica de introducir*(Enter):***Procedimiento para la selección de variables en el que todas las variables de un bloque se introducen en un solo paso obligatoriamente, produce que el proceso de selección de las variables sea manual, partiendo de un modelo inicial, se va evaluando qué variable es la que menos participa en él y se elimina, volviendo a construir un nuevo modelo de regresión aplicando la misma técnica, pero excluyendo la variable seleccionada y aplicando el mismo proceso de selección. Este proceso se repite reiteradamente hasta que se considere que el modelo obtenido es el que mejor se ajusta a las condiciones impuestas y que no se puede eliminar ninguna variable más de las que lo componen.
- **Técnica de selección hacia delante *(Forward)*:** comienza con el modelo que incluye solo el término constante y consiste en ir introduciendo las variables independientes en el modelo según su grado de relación con la variable dependiente y su significación estadística, únicamente, hasta que no se pueda introducir ninguna más que cumpla la condición impuesta.
- **Técnica de pasos hacia atrás (Backward):** el análisis comienza con todas las variables explicativas incluidas en el modelo y se van suprimiendo si cumplen una serie de condiciones definidas a priori hasta que no se pueden eliminar más, es decir que ninguna variable cumpla la condición impuesta.
- **Técnica por pasos *(Stepwise)*:** es una combinación de los dos métodos anteriores. Consiste en ir introduciendo o eliminando variables del modelo si cumplen una serie de condiciones definidas a priori hasta que ninguna variable satisfaga ninguna de las condiciones expuestas de entrada o salida del modelo, además, en cada paso, existe la posibilidad de eliminar una variable ya incluida en la ecuación, si deja de ser significativa en la explicación de la variable respuesta.

Algunas consideraciones que se deben tener en cuenta respecto a las variables independientes que se introducen en el modelo de regresión son las siguientes:

- Si es una variable continua, se introducirá la variable real o bien alguna transformación de ella (logaritmo, cuadrado, etc.), cuando sea necesario.
- Si es una variable categórica con n categorías, n> 2, se debe expresar cada una de estas categorías mediante n-1 variables ficticias o “dummy”. Una variable dummy es una variable construida artificialmente y que únicamente puede tomar los valores 0 o 1. Este recurso permite expresar la misma información contenida en una variable de n categorías mediante la combinación de n-1 variables dummy. Adicionalmente estas variables permiten realizar todas las comparaciones necesarias respecto a las n categorías de la variable original en el modelo de regresión.
- Es conveniente analizar la asociación de la variable respuesta con las variables explicativas así como entre las propias variables explicativas para poder interpretar los coeficientes obtenidos de forma adecuada. Este análisis previo fundamentará la elección de las variables independientes que se añaden en el modelo evitando la multicolinealidad entre los predictores que puede sesgar las interpretaciones y errores típicos inflados. (Aguayo, 2007) (Berlanga & Vilà, 2014)

La valoración de la confusión y/o interacción se realiza mediante el análisis multivariante, es necesario diferenciar los dos fenómenos (confusión e interacción) y comprobarlos por separado.

- La confusión se detecta cuando la Odds Ratio (OR) que evalúa la fuerza de asociación entre la variable independiente y la variable dependiente cambia de forma importante cuando se introduce en la ecuación de regresión una tercera variable.
- La interacción requiere introducir en la ecuación de regresión un término multiplicativo, compuesto por las dos variables independientes que se

presuponen interactúan en su efecto sobre la variable dependiente; y una vez incluido ver si su coeficiente de regresión logística (B) es estadísticamente significativo, en cuyo caso existe un efecto de interacción entre las variables consideradas. (Aguayo, 2007)

La capacidad predictiva del modelo de regresión logística se valora mediante la comparación entre el grupo de pertenencia observado y el pronosticado por el modelo, que clasifica a los individuos en cada grupo definido por la variable dependiente en función de un punto de corte establecido para las probabilidades predichas a partir de los coeficientes estimados y del valor que toman las variables explicativas para cada individuo. (Cuadras, 2014)

Para esto se han desarrollado un conjunto criterios de evaluación que cuantifican el nivel de ajuste del modelo al conjunto de datos. Algunos de estos criterios se exponen a continuación:

La Desvianza

Mide hasta qué punto un modelo se ajusta bien a los datos; cuanto más pequeño sea el valor, mejor será el ajuste. Se define como

$$D = -2\sum_{i=1}^{n}\left(y_i \log\left(\frac{\hat{p}}{y_i}\right) + \left(1-y_i\right)\log\frac{1-\hat{p}}{1-y_i}\right)$$

Si D es mayor que una χ^2 con n-p grados de libertad para un nivel de significación dado, entonces el modelo de regresión logística es confiable

- **La Prueba de Bondad de Ajuste de Hosmer-Lemeshov**

Verifica la hipótesis nula de que el modelo se ajusta a los datos. El estadígrafo de prueba del test se define como:

$$c = \sum_{i=1}^{g} \frac{(O_i - n_i' \overline{p}_i)^2}{n_i' \overline{p}_i \ (1 - \overline{p}_i)}$$

Donde:

g Es el número de grupos;

n_i' Es el número de observaciones en el i–ésimo grupo;

O_i Es la suma de las Y en el i–ésimo grupo;

$\overline{p}_i$ Es el promedio de las p_i en el i–ésimo grupo.

El test consiste en establecer los deciles de riesgo o probabilidad predicha por el modelo de presentar el evento, y en cada una de estas diez categorías se comparan los valores observados y los predichos, tanto para los que tienen el resultado explorado como para los que no lo tienen. Si hay una elevada coincidencia entre observados y esperados existe un buen ajuste. Cuando el test Chi cuadrado de la prueba no es significativo indica que no hay motivos para pensar que los resultados predichos sean diferentes de los observados o que si hay diferencias pueden explicarse razonablemente por el azar o error del muestreo y que el modelo puede considerarse aceptable, o sea, si la probabilidad asociada al estadígrafo de prueba es mayor de 0,05 se considera que el modelo se ajusta a los datos.

A continuación se enuncian algunos ejemplos de aplicaciones de la regresión logística en el campo de la salud

- En el Hospital “Mancha-Centro” de Alcázar de San Juan, se realizó un estudio con el objetivo de determinar las diferencias en el número de cesáreas entre partos espontáneos y partos inducidos. Se calcularon las frecuencias absolutas y relativas de las variables cualitativas y la media y su desviación

estándar para las cuantitativas; se realizó un análisis bivariante entre los factores de confusión con el inicio del parto (espontáneo/ inducido) empleando las pruebas de Chi Cuadrado y T de Student según el tipo de variable; se desarrollaron diversos modelos multivariantes con regresión logística binaria para controlar el sesgo de confusión. Como resultado se determinó que la inducción del parto es un factor de riesgo para una mayor duración de la dilatación con (OR=6,00; IC 95%: 4,02 - 8,95), empleo de analgesia epidural (OR=3,10; IC95%: 2,24 - 4,29) y necesidad de transfusión sanguínea (OR=3,33; IC 95%: 1,70-9,67). (Hernández, Pascual, Baño, Melero, & Molina, 2014)

- En el Hospital General Universitario Ciudad Real, se realizó un estudio para detectar factores predictores de esofagitis eosinofílica en pacientes con impactación esofágica por bolo alimenticio. El análisis estadístico se efectuó mediante los test de la t de Student y de la Chi Cuadrado y se elaboró un modelo de regresión logística. Como resultado se demostró que la edad (especialmente por debajo de 40 años) con OR: 8.67; IC 95%: 2,33-33,01 y la historia de impactaciones previas con OR: 6,77; IC 95%: 1.40-32,67 ayudan a predecir la esofagitis eosinofílica en pacientes con impactación esofágica por bolo alimenticio, facilitando así un diagnostico precoz y evitar complicaciones durante la endoscopia. (Rodríguez, et al., 2013)
- El doctor José Antonio González Pompa llevó a cabo una investigación para evaluar el efecto de factores de riesgo en la ocurrencia del infarto agudo del miocardio en pacientes fumadores. El análisis estadístico se basó en un estudio multivariado con regresión logística para determinar el valor independiente de cada uno de los factores de riesgo. Como conclusión obtuvieron que el mayor grado de dependencia es representado por la hipercolesterolemia (OR 4,20; IC 1,18-14,97), seguido del tiempo de evolución del hábito de fumar (OR 3,60; IC 1, 468,91) y la cantidad de cigarrillos fumados en el día (OR 2,32; IC 1, 02- 4,95).(González & González, 2013)

- En Cuba, la Dra. Consuelo Beatriz Correa Sierra, en su tesis doctoral determinó los factores de riesgo influyentes en la infección por Citomegalovirus (CMV) y otros herpes virus en población con riesgo de desarrollar infección congénita y en receptores pediátricos de trasplante. Para determinar la asociación entre los diferentes factores de riesgos se calculó la razón de prevalencia (RP) y se empleó la prueba estadística exacta de Fisher, considerando significativos los valores de RP>1 y p<0,05, con un nivel de significación para el intervalo de confianza (IC) del 95 %. Además, se realizó un análisis univariado y multivariado mediante el método de regresión logística binaria. Como resultado se obtuvo que la edad menor de 20 años constituye un factor de riesgo para adquirir dicha infección durante el embarazo y la mayoría de los recién nacidos infectados congénitamente por CMV son asintomáticos al nacer. (Correa, 2014)

1.3.2 Árbol de Decisión

El uso de la metodología de árboles de decisión no es muy frecuente aún, pero en los últimos años ha despertado un interés creciente en el campo de la medicina como una herramienta de investigación promisoria y útil para solucionar los problemas que surgen a la hora de obtener información, encontrar patrones y definir tendencias. (Segura, 2012) (Esquerda & Trujillano, 2010)

Se ha visto reflejada en problemas clínicos de diagnóstico o pronóstico generando reglas de decisión, en la selección de las variables en orden de importancia según la información que aportan al modelo definitivo, para seleccionar puntos de corte óptimos en variables cuantitativas y buscar relaciones clínicas entre distintas variables.

Los análisis de clasificación basados en árbol de decisión son técnicas de exploración de datos (data mining) que consisten en estudiar grandes masas de datos con el fin de descubrir patrones no triviales. Los patrones no triviales que se estudiarán habitualmente serán los predictivos y los explicativos.

Un árbol de decisión representa una serie de pautas basadas en ciertas variables explicativas que se muestran según recorremos el árbol, se construyen mediante un algoritmo que va dividiendo los registros de la base de datos (casos u observaciones) en nodos de forma recursiva, de manera que con cada subdivisión las frecuencias relativas de las categorías de la variable dependiente vayan tendiendo a 0 o a 1.

A partir de un nodo raíz, que incluye todos los casos, el árbol se va ramificando en diferentes nodos "hijo" que contienen un subgrupo de casos. El criterio de ramificación (o partición) es seleccionado de manera óptima después de examinar todos los posibles valores de todas las variables predictivas disponibles. En los nodos terminales ("hojas" del árbol) se obtiene una agrupación de los casos de la manera más homogénea posible en cuanto al valor de la variable dependiente.

La creación de un árbol de decisión consiste, básicamente, en ir creando sucesivas subdivisiones del conjunto de datos de acuerdo con un algoritmo determinado, de forma tal que, con cada nueva subdivisión que se realice, mejore la clasificación de la variable dependiente. (Rojo, 2006) (Segura, 2012)

El proceso de construcción de árboles de decisión se esquematiza en 4 fases:

- Construcción del árbol. A partir del nodo raíz, se identifica la variable más adecuada para dividir este nodo en dos nodos hijo, y el punto de corte óptimo. A cada uno de los nodos hijo se le asigna un valor de la variable dependiente, la correspondiente al mayor número de registros de aquel nodo. A su vez, cada uno de los nodos hijo será dividido en otros sucesivos nodos siguiendo la misma metodología.
- Parada del proceso de crecimiento del árbol (se constituye un árbol máximo que sobre ajuste la información contenida en nuestra base de datos). Los nodos hijos ya no pueden subdividirse cuando contienen un único caso, o bien cuando el valor de la variable dependiente es el mismo para todos los casos integrantes del nodo. Cuando cualquiera de estas dos situaciones se cumple en todos los nodos, el desarrollo del árbol se detiene. Por otra parte, es

posible definir criterios adicionales (número máximo de nodos del árbol, número mínimo de casos por nodo) que eviten una excesiva ramificación.

- Podado del árbol haciéndolo más sencillo y dejando sólo los nodos más importantes. Por lo general, un árbol desarrollado según el esquema anterior es excesivamente complejo y ramificado y puede reflejar con demasiada minuciosidad las características de la base de datos utilizada en su construcción. La eliminación de las ramas superfluas proporcionará un árbol más sencillo y, a la vez, con una mayor capacidad de generalización. El proceso de "poda" se verifica según unos criterios predefinidos de coste complejidad: empezando por los últimos nodos, se eliminan aquellos cuya presencia añade más complejidad que efectividad.
- Selección del árbol óptimo con capacidad de generalización. Elegir el árbol óptimo precisa de un sistema de validación (se selecciona el árbol que mejor clasifica en este grupo de validación). La validación puede ser externa (utilizando casos no empleados en el desarrollo del modelo) o interna (validación cruzada). En la validación cruzada se realiza una partición aleatoria (suelen ser 10 partes) del grupo de desarrollo y se utiliza de forma recursiva un subgrupo (9 partes) para generar el árbol y otro (1 parte) para la validación.

Las "medidas" que habitualmente se suelen utilizar para comparar la mejora de cada nueva subdivisión son las siguientes:

- El coeficiente χ^2
- Índice de Gini
- Índice Binario.

El coeficiente χ^2

Este criterio trata de medir la asociación entre dos variables nominales u ordinales; $\mathbf{V_1}$ (con $\boldsymbol{K}$ categorías) y $\mathbf{V_2}$ (con $\boldsymbol{M}$ categorías) y se define como:

$$\chi^2 = \sum_{i=1}^{K} \sum_{j=1}^{M} \frac{(n_{i,j} - n'_{i,j})^2}{n'_{i,j}}$$

Siendo $n_{i,j}$ la frecuencia observada de la celda {***i, j***} y $n'_{i,j}$ es la frecuencia esperada de la celda {***i, j***}.

Valores cercanos a cero de este coeficiente indicaran que no hay asociación entre la variable fila y la variable columna.

Valores grandes de este coeficiente indicaran la existencia de asociación entre las variables fila y columna de la tabla.

Índice de Gini

El Índice de Gini en el nodo ***t*** se define como:

$$g(t) = 1 - \sum_{i=1}^{K} p(i/t)^2$$

Donde ***K*** representa las distintas categorías de la variable dependiente, $p(i/t)$es la proporción de la categoría ***i*** en el nodo ***t.***

Cuando todos los casos del nodo ***t*** pertenecen a la misma categoría, el índice de Gini toma el valor cero, se dice entonces que el nodo se vuelve puro.

Este índice es una medida de impureza en la clasificación de los datos, a medida que vamos clasificando correctamente los datos, el índice de *Gini* va tomando valores cercanos a 0. Para “medir” la mejora de una clasificación debida a la división de los datos en dos grupos, se utiliza el siguiente criterio:

$$\Phi(s,t) = g(t) - p_{iz} * g(t_{iz}) - p_{de} * g(t_{de})$$

Donde **g (t)** es el valor del índice de Gini en el nodo **t, P_{iz}** es la proporción de casos enviados al nodo izquierdo, **P_{de}** es la proporción de los casos enviados al nodo derecho, **g (t_{iz})** es el valor del índice de Gini en el nodo izquierdo y **g (t_{de})** es el valor en el nodo derecho y **s** es la división propuesta.

Valores altos de esta función serán indicios de una buena clasificación y valores bajos indicaran una mala clasificación.

Índice Binario.

La función del criterio binario para la división ***s*** en el nodo ***t*** se define como:

$$\Phi(s,t) = P_{de} * P_{iz}\left[\sum_{i=1}^{K}|P(i/t_{iz}) - P(i/t_{de})|\right]^2$$

Donde ***K*** se refiere a las categorías de la variable dependiente; ***P_{iz}*** es la proporción de casos enviados al nodo izquierdo, ***P_{de}*** es la proporción de los casos enviados al nodo derecho; $P(i/t_{iz})$ es la proporción de la categoría ***i*** en el nodo izquierdo y $P(i/t_{de})$ es la proporción de la categoría ***i*** en el nodo derecho.

El Índice Binario al igual que el Índice de Gini se basa en encontrar la división *s* que maximice este valor, pues valores altos de esta función indicarán buenas particiones.

Los árboles de decisiones disponen de cuatro algoritmos clasificación:

- **CHAID (Chi Square Automatic Interaction Detector)**
- **CHAID Exhaustivo**
- **C&RT o CART**
- **QUEST**

CHAID

Si la variable dependiente es continua, se utiliza la prueba F que permite comparar dos varianzas obteniendo un coeficiente, mayor que uno resultante de dividir la varianza mayor entre la menor. Si la variable dependiente es categórica se utiliza la prueba χ^2**.**

No es un algoritmo binario y por lo tanto tiende a crear un árbol más ancho que con los algoritmos que producen árboles binarios.

Puede trabajar con variables en cualquier nivel de medida. Dada una variable predictora, funde aquellas categorías consideradas estadísticamente homogéneas y deja las categorías heterogéneas inalteradas. A continuación de todas las variables predictoras potenciales elige la que tenga el mayor valor del coeficiente para formar la primera rama del árbol.

CHAID EXHAUSTIVO

Este algoritmo funciona básicamente igual que el anterior, la única diferencia es que realiza un examen más minucioso para realizar la fusión de categorías y, por lo tanto, utiliza más tiempo de cálculo que el anterior.

C&RT o CART

Es un algoritmo que produce árboles binarios. El C&RT divide los datos en dos conjuntos de forma que los datos comprendidos dentro de cada subconjunto sean más homogéneos que en el conjunto anterior.

Las medidas de asociación que utiliza son:

Para variables categóricas: Índice de impureza de Gini e Índice Binario.

Para variables continuas: Test de homogeneidad de varianzas

QUEST

Es un algoritmo que produce árboles binarios; está creado con vistas a la eficiencia en los cálculos, siendo el tiempo de procesamiento más corto que en el C&RT. La variable dependiente debe de tener nivel de medida nominal.

Para variables categóricas: Estadístico basado en χ^2.

Para variables continuas: Prueba F.

Ejemplos de aplicación de árboles de decisión

Podemos citar las siguientes investigaciones:

- En los Estados Unidos se utilizó el árbol de decisión para diagnosticar influenza; a través de los signos y síntomas referidos por el paciente, se analizaron datos de 30.668 adultos mayores con el fin determinar segmentos de la población que tienen mayor probabilidad de ser afectados por la influenza. Se realizó a través del algoritmo CART, se desarrollaron tres modelos que maximizan el número de pacientes que no requieren pruebas de diagnóstico previo a las decisiones de tratamiento. El modelo 2 clasifica correctamente el 67% de los pacientes en el grupo de validación en un alto o grupo de bajo riesgo en comparación con sólo el 38% para el modelo 1 y el 54% para el modelo 3. Se selecciona el modelo 2 como el mejor ajuste y clasifica a los pacientes sobre la base de la temperatura y la presencia de escalofríos. (Afonso, et al., 2012)
- En Japón se utilizó en el Departamento de Fisioterapia de la Facultad de Salud y Asistencia Médica de Universidad de Medicina de Saitama para predecir la independencia de la marcha en pacientes con fractura de cuello femoral. Se utilizaron las pruebas T de Student y la prueba de Chi-Cuadrado para probar las diferencias estadísticas entre los grupos. En el análisis multivariado, se utilizaron árboles de clasificación y regresión (CART) para determinar el valor predictivo de las medidas que difieren significativamente entre los grupos. Los análisis del algoritmo CART mostraron que cuando el

poder de extensión de rodilla fue > 0,34 kgf / kg, la puntuación MMSE (the mini-mental state examination) fue >13,5; sin antecedentes de accidente cerebrovascular, el ritmo de la marcha independiente al alta fue del 93,8%. Por el contrario, cuando el poder extensión de la rodilla era ≤0,33 kgf / kg, FRT (functional reach test) fue ≤ 25,5 cm, la puntuación MMSE fue ≤13,5, y la tasa de caminar dependiente al alta fue del 100%. (Arai, Kaneko, & Fujita, 2011)

- En el Hospital General Universitario "Carlos Manuel de Céspedes", en el municipio Bayamo, se crearon reglas que permitieron predecir el riesgo de desarrollar la cardiopatía hipertensiva en individuos hipertensos a partir de factores hemodinámicos y no hemodinámicos; el árbol predijo el riesgo de desarrollar la cardiopatía hipertensiva a 82,598 % de los pacientes. El factor más importante lo constituyó la proteína C reactiva, seguida en orden de importancia por la glucemia, el ácido úrico, el colesterol y la microalbuminuria. (Álvarez, et al., 2014)
- En un estudio realizado en Santa Clara para predecir la evolución hacia la hipertensión arterial en la adultez desde la adolescencia, se realizó la determinación de un modelo predictivo mediante un árbol de decisiones utilizando la técnica de CHAID se obtuvo un árbol de decisiones que fue capaz de clasificar acertadamente al 80 % de los casos. Con dicho modelo se obtuvo un 67,4 % de verdaderos negativos y 86,6 % de verdaderos positivos. Como resultado se plantea que la edad de inicio de la pre hipertensión arterial el peso al nacer, la presencia de obesidad familiar y el ambiente familiar desfavorable fueron las situaciones, que en su interacción desde la adolescencia y hasta la adultez, presentaron mayor fuerza para dicha conversión.(Pérez & Grau, 2014)
- En el Instituto de Medicina Tropical “Pedro Kourí” se realizó una investigación para la clasificación de dengue hemorrágico en la fase temprana de la enfermedad con el objetivo de la búsqueda de signos de alarma de gravedad. Las variables consideradas para la clasificación fueron los signos, síntomas y exámenes de laboratorio al tercer día de evolución de la enfermedad. Se

aplicó el algoritmo de árboles de clasificación y regresión con el algoritmo CART utilizando el índice de Gini; el árbol resultante obtuvo una sensibilidad de 74 % con un error de 0,25. Como resultado se obtuvo que las variables de laboratorio son de mayor importancia para discriminar las formas clínicas de dengue y los resultados del conteo de plaqueta y hemoglobina fueron los que mejor clasificaron esta patología. (Vega, Sánchez, Cortiñas, Castro, & González, 2012)

1.3.3 El análisis factorial

El análisis factorial se emplea en múltiples aplicaciones para reducir la dimensionalidad de un problema y cumplimentar los supuestos de técnicas estadísticas que se aplicarán posteriormente. Su objetivo es el de reducir un conjunto de variables cuantitativas aleatorias (interrelacionadas) en un grupo de factores latentes (independientes), de tal manera que los factores siempre serán, en número, inferiores a las variables iniciales. El éxito de esta técnica queda garantizado en la medida que su resolución cumpla dos requisitos: el principio de parsimonia; y la interpretabilidad de los factores elegidos.

En el modelo factorial, si los factores son inferidos a partir de las variables observadas, cada variable será expresada como una combinación lineal de factores no observables directamente. Se admite que un conjunto de variables aleatorias, X_1, X_2,... X_p, se explicarán por un conjunto de factores comunes, F_1,... F_k (siendo f<v) y p factores únicos, u_1,...u_p de acuerdo con el siguiente modelo factorial lineal. (Cuadras, Arenas, & Fortiana, 1991)

$$\mathbf{X}_1 = a_{11}\mathbf{F}_1 + a_{12}\mathbf{F}_2 + \ldots + a_{1k}\mathbf{F}_k + \mathbf{u}_1$$
$$\mathbf{X}_2 = a_{21}\mathbf{F}_1 + a_{22}\mathbf{F}_2 + \ldots + a_{2k}\mathbf{F}_k + \mathbf{u}_2$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$\mathbf{X}_p = a_{p1}\mathbf{F}_1 + a_{p2}\mathbf{F}_2 + \ldots + a_{pk}\mathbf{F}_k + \mathbf{u}_p$$

Donde $F_1,\ldots, F_k$ (k<p) son los factores comunes y $u_1,\ldots u_p$ los factores únicos o específicos y los coeficientes $\{a_{ij}; i=1,\ldots, p; j=1,..., k\}$ las cargas factoriales.

El análisis factorial busca factores que expliquen la mayor parte de la varianza común que se distingue entre varianza común y varianza única. La varianza común es la parte de la variación de la variable que está compartida con las otras variables y se puede cuantificar con la denominada **comunalidad**. La varianza única es la parte de la variación de la variable que es propia de esa variable denominada la **especificidad**

$$\text{Var}\,(X_i) = \sum_{j=1}^{k} a_{ij}^2 + \psi_i \;= h_i^2 + \psi_i;\; i=1,\ldots,p$$

Donde la comunalidad es

$$h_i^2 \;=\; \text{Var}\left(\sum_{j=1}^{k} a_{ij}F_j\right)$$

Y la especificidad es

$$\psi_i \;=\; \text{Var}(u_i)$$

Además se tiene que

$$\text{Cov}\,(X_i, X_\ell) \;= \text{Cov}\left(\sum_{j=1}^{k} a_{ij}F_j, \sum_{j=1}^{k} a_{\ell j}F_j\right) = \sum_{j=1}^{k} a_{ij}a_{\ell j} \quad \forall i \neq \ell$$

Por lo que son los factores comunes los que explican las relaciones existentes entre las variables del problema.

El elemento identificador de esta técnica es su capacidad en sintetizar información, lo que consigue eliminando del conjunto de variables iniciales aquellas que ofrecen información redundante y aquellas que no se adaptan al modelo de regresión

múltiple en el que se basa esta técnica. En este caso, y a diferencia de la ecuación del modelo de regresión, los factores no son variables simples sino dimensiones que engloban a un conjunto determinado de variables pudiendo ser explicadas linealmente en función de los factores seleccionados. (Cuadras, 2014)

En las aplicaciones de esta técnica estadística frecuentemente se trabaja con un grupo de componentes que expliquen un porcentaje considerable de la variabilidad del conjunto de datos y por estar incorrelacionadas son útiles como predictoras en modelos de regresión lineal múltiple, regresión logística binaria, análisis discriminante, etc.

El análisis factorial analiza las correlaciones lineales entre las variables y si las variables no estuvieran asociadas linealmente, las correlaciones entre ellas serían nulas, no existiendo asociación y, en consecuencia, no tendría sentido seguir aplicando esta técnica.

Para garantizar que los datos se ajustan, estos pueden ser analizados a través diferentes test entre los que se destacan:

- El determinante de la propia matriz de correlaciones (si está por debajo de 0.05 las variables estarán intercorrelacionadas);
- El test de esfericidad de Bartlett, su significación será mejor cuando esté por debajo 0.05. En esta prueba se evalúa la hipótesis nula de que no existe correlación entre las variables; es decir, que la matriz de correlación es la identidad. Al rechazar esta hipótesis, se demuestra que en realidad sí existe algún grado de correlación estadísticamente significativa.
- El índice de Kaiser-Meyer-Olkin (KMO), también conocido como el índice de adecuación de la muestra individual (MSA). indica qué tanta correlación tiene una variable específico con las demás en la matriz. Compara los valores de las correlaciones entre las variables y sus correlaciones parciales Si su valor se aproxima a 1 su significación es elevada. (Cuadras, 2014)

Otras condiciones a tener en cuenta y que garantizan la idoneidad del análisis factorial es determinar cómo están medidas las variable que en general estas deben ser métricas.

En caso de no ser métricas se hace necesario aplicar la técnica de escalamiento óptimo, que consiste en asignar cuantificaciones numéricas a las categorías de cada variable, lo que permite utilizar los procedimientos estándar para obtener una solución con las variables cuantificadas.

Los valores de escala óptimos se asignan a las categorías de cada variable de acuerdo con el criterio de optimización del procedimiento que se esté utilizando. A diferencia de las etiquetas originales de las variables nominales u ordinales del análisis, estos valores de escala tienen propiedades métricas.

En la mayoría de los procedimientos de categorías, la cuantificación óptima de cada variable escalada se obtiene mediante un método iterativo denominado mínimos cuadrados alternantes en el que, después de que se utilicen las cuantificaciones actuales para encontrar una solución, las cuantificaciones se actualizan utilizando dicha solución. A continuación, se utilizan las cuantificaciones actualizadas para buscar una nueva solución, que a su vez se utiliza para actualizar las cuantificaciones y así sucesivamente, hasta que se alcanza algún criterio que indica al proceso que finalice. (Cuadras, 2014)

1.4 Modelos estadísticos empleados en el estudio del padecimiento de pie diabético.

Existen diversos estudios que han analizado los factores de riesgo asociados al pie diabético y a pesar de que se realizaron en contextos diferentes y utilizando variados criterios de inclusión, la mayoría de estas investigaciones manifiestan que hay varios factores de riesgo claves que predisponen al desarrollo del pie diabético tales como: la edad mayor de 50 años, tiempo prolongado de evolución de la enfermedad, antecedentes de úlceras o amputación, presencia de neuropatía, artropatía y/o

vasculopatías, deficiente educación sanitaria en el cuidado de los pies, mal con control metabólico, valore de glicemia elevados, entre otros factores asociados con enfermedades vasculares y con la diabetes. (Márquez, Zonana, Anzaldo, & Muñoz, 2014) (Pinilla, Barrera, Rubio, & Devia, 2014) (Faget, 2014)

A continuación se citan algunos ejemplos de investigaciones que han abordado el tema del padecimiento de pie diabético y los procedimientos estadísticos empleados para la caracterización y detección de los factores de riesgo de dicho padecimiento.

- En Tijuana México se estudiaron los riesgos de padecer pie diabético en pacientes con diabetes Mellitus tipo II en una unidad de medicina de familia; el análisis se realizó utilizando la técnica multivariada de regresión logística y se obtuvo como resultado que el 44% de los pacientes tuvieron alto riesgo para desarrollar pie diabético; los factores de riesgo que presentaron asociación significativa fueron: la evolución de la diabetes mellitus mayor a 10 años, sexo femenino, ingreso mensual familiar < 236 euros y una hemoglobina glucosilada (HbA1c) ≥ 7, 0%. (Márquez, Zonana, Anzaldo, & Muñoz, 2014)
- En Bogotá, Colombia a través del análisis descriptivos y los coeficientes de correlación se realizó un estudio con el objetivo de determinar prevalencia de actividades de prevención por médicos y pacientes del padecimiento de pie diabético e identificar factores de riesgo: metabólicos, alimentarios y estado nutricional; se evidenciaron como factores de riesgo para diabetes Mellitus y pie diabético: hiperlipidemia, hiperglucemia, obesidad y hábitos alimentarios inadecuados. (Pinilla, Barrera, Rubio, & Devia, 2014)
- En Cienfuegos utilizando los estadísticos descriptivos de frecuencia y/o porcentajes se caracterizó el estado de salud de los pies de la población diabética atendida en la consulta de podología del Área de Salud #2 de Cienfuegos, durante los meses de febrero a junio de 2013, de los 243 pacientes seleccionados para el análisis el 95,89 % se identificaron con pie diabético, siendo el grupo de edades más afectado los pacientes mayores de 51 años para un total 138 pacientes.(Garriga, Sánchez, & Vázquez, 2014)

Es criterio común entre los investigadores que han abordado el tema del pie diabético que existe en la población diabética una baja percepción de riesgo y se considera esencial fomentar e intensificar las actividades de educación grupal participativa y trabajo preventivo desde la Atención Primaria, para reforzar los conocimientos que se hayan adquirido de manera individual en relación con los cuidados del pie y sus complicaciones. (Garriga, Sánchez, & Vázquez, 2014) (Pinilla, Barrera, Rubio, & Devia, 2014) (Vicente, et al., 2010)

El pie diabético se ha convertido en un problema de salud que va en aumento no solo en Cuba sino en todo el mundo debido a la prevalencia de diabetes mellitus; resulta de gran interés hacer frente a esta complicación y sus consecuencias, siendo un reto a nivel mundial buscar modelos que permitan predecir este padecimiento y elaborar estrategias para prevenir y tratar el pie diabético de forma oportuna identificando los factores de riesgos que influyen de forma significativa en su incidencia.

1.5 Conclusiones Parciales

- El desarrollo de modelos estadísticos predictivos en la salud ha crecido significativamente en los últimos años, siendo de gran ayuda en la toma de decisiones y permitiendo la creación de diversos sistemas y herramientas útiles para reducir las incertidumbres, garantizando mejores actuaciones y establecer eficaces medidas de control para la erradicación de las enfermedades.
- Las técnicas estadísticas más comúnmente empleadas en el establecimiento de modelos predictivos en el campo de la salud son la regresión logística binaria aunque ese han visto el incremento de las técnicas de segmentación de árboles de decisión con el algoritmo CHAID y CRT.
- Los estudios realizados sobre el padecimiento de pie diabético establecen que los factores de riesgo de esta enfermedad son: la edad mayor de 50 años, tiempo prolongado de evolución de la enfermedad, antecedentes de úlceras o

amputación, presencia de neuropatía, artropatía y/o vasculopatías, deficiente educación sanitaria en el cuidado de los pies, mal control metabólico con elevados valores de glicemia.

Capítulo II Procedimiento para el establecimiento de un modelo predictivo del padecimiento pie diabético a partir de factores de riesgo.

Este capítulo se dedica a la descripción del procedimiento para establecer un modelo predictivo del padecimiento de pie diabético en los pacientes con diabetes mellitus de tipo II hospitalizados en la Clínica del Diabético del municipio de Cienfuegos. Se tienen en cuenta en el estudio variables clínicas, sociodemográficas, hábitos de consumo y de antecedentes patológicos y personales que se registran en la clínica. Se exponen los pasos del procedimiento empleado y se describe, además, la muestra de casos utilizada para la obtención del modelo y los criterios para la selección de las variables predictoras. Se utilizan técnicas multivariadas de regresión logística binaria y árboles de decisión para la elaboración del modelo y, por último, se presenta la variante propuesta al procedimiento para la obtención del modelo predictivo y los criterios a seguir para la comparación de los modelos obtenidos por ambos procedimientos.

2.1 Procedimiento empleado.

Basado en los criterios planteados en el Capítulo I se propone seguir el siguiente procedimiento:

- Definición de la población, la muestra y las variables a emplear.
- Análisis exploratorio de los datos.
- Determinación y obtención de los modelos de pronóstico factibles:
    - Interpretación de los resultados del modelo seleccionado
    - Análisis de la validez de los modelos estimados.
- Comparación y selección del modelo más ventajoso.
- Validación de los modelos en una muestra externa.

A continuación se explica cada uno de estos pasos.

2.1.1 Definición de la población, la muestra y las variables a emplear.

El marco muestral está constituido por los pacientes que han acudido a la Clínica del Diabético del municipio de Cienfuegos, que es un centro de atención y educación a los pacientes diabéticos. Funciona como unidad docente, asistencial y de investigación donde se diseñan acciones de superación para médicos de la familia y se brindan más de 10 mil consultas especializadas cada año. Para el estudio se seleccionan los pacientes con diabetes mellitus de tipo II por ser la de mayor frecuencia e incidencia y los pacientes que son residentes del municipio de Cienfuegos, debido a la estabilidad de estos por la cercanía a la clínica.

Las variables que se utilizan se corresponden con los factores de riesgo establecidos en el Capítulo I y que han sido medidas en la Clínica del Diabético; se caracterizan como variables: sociodemográficas, hábitos de consumo y antecedentes patológicos personales y familiares. Todas estas variables se precisan a continuación:

- **Sociodemográficas**. Se incluyen las variables edad, sexo, nivel de escolaridad y ocupación.
- **Hábitos de consumo**. Se recogen aquellas vinculadas con el consumo de alcohol, café y cigarros.
- **Clínicas**. Aquellas que describen aspectos relacionadas con la patología de la enfermedad, tales como: tiempo de evolución de la enfermedad desde que fue diagnosticada, índice de masa corporal, niveles de glicemia, creatinina, triglicéridos, ácido úrico y colesterol.
- **Antecedentes patológicos personales y familiares**. Incluye la presencia de patologías como: hipertensión arterial, cardiopatía isquémica, hiperlipidemia y claudicación intermitente, además de registrar si en su familia alguien padece o no de diabetes mellitus.

A continuación se precisa la operacionalización de las variables mencionadas y la definición de las escalas utilizadas**.**

VARIABLE	TIPO	ESCALA
Padecimiento de pie Diabético (PD)	Cualitativa Nominal Dicotómica.	0-No padece 1-Si padece
Edad	Cuantitativa	Medida en años cumplidos
Sexo	Cualitativa Nominal Dicotómica.	0-Masculino(M) 1-Femenino(F)
Nivel de escolaridad (nivel_escolar)	Cualitativa Ordinal Politómica	0-Primaria no terminada 1-Primaria 2-Secundaria 3-Obrero calificado 4-Bachiller 5-Técnico Medio 6-Universitario
Ocupación	Cualitativa Nominal Politómica	0-Trabajador 1-Jubilado 2-Ama de casa 3-Desempleado o desocupado
Consumo de bebidas (bebidas)	Cualitativa Nominal Politómica	0- No bebedor 1- Ex bebedor 2- Bebedor ocasional 3- Bebedor
Consumo de cigarros (fuma)	Cualitativa Nominal Dicotómica	0- No fuma 1- Si fuma
Consumo de café (café)	Cualitativa Nominal Dicotómica	0- No 1- Sí
Tiempo de evolución (tiempo)	Cuantitativa	Expresada en años de ser diagnosticado con diabetes mellitus
Índice de masa corporal (IMC)	Cuantitativa	Expresa el índice de masa corporal de los pacientes (Kg/m^2)
Niveles de glicemia (glicemia)	Cualitativa Nominal Dicotómica	0- < 7,00 mmol/l 1- ≥ 7,00 mmol/l

VARIABLE	TIPO	ESCALA
Niveles de colesterol (colesterol)	Cualitativa Nominal Dicotómica	0- ≤ 5,20 mmol/l 1- > 5,20 mmol/l
Niveles de creatinina (creatinina)	Cuantitativa Continua	Expresa los niveles de creatinina medidos en el laboratorio (µmol/l)
Niveles de triglicéridos (triglicéridos)	Cuantitativa Continua	Expresa los niveles de los triglicéridos medidos en laboratorio (mmol/l)
Niveles de ácido úrico (ácido_úrico)	Cuantitativa Continua	Expresa los niveles del ácido úrico medidos en laboratorio (µmol/l)
Antecedentes familiares de diabetes mellitus (APF_DM)	Cualitativa Nominal Dicotómica	0- No 1- Sí
Antecedentes patológicos de hipertensión arterial (APP_HTA)	Cualitativa Nominal Dicotómica	0- No 1- Sí Si
Antecedentes patológicos de cardiopatía isquémica (APP_CI)	Cualitativa Nominal Dicotómica	0- No 1- Sí
Antecedentes patológicos de hiperlipidemia (APP_HIPER)	Cualitativa Nominal Dicotómica	0- No 1- Sí
Antecedentes patológicos de claudicación intermitente (APP_CINTERMITENTE)	Cualitativa Nominal Dicotómica	0- No 1- Sí

2.1.2 Análisis exploratorio de los datos.

En esta etapa se realiza un análisis descriptivo de los datos con el objetivo de conocer el comportamiento de la muestra en cada una de las variables incluidas en el estudio.

La exploración se lleva a cabo empleando distribuciones de frecuencias y porcentajes para las variables cualitativas; para las variables cuantitativas se

describe el promedio con sus respectivas medidas de variabilidad, la desviación típica y los valores mínimo y máximo.

En la selección de las variables participantes en el modelo de predicción se tendrán en cuenta aquellas que presenten relación con la variable dependiente: padecimiento de pie diabético. Para esto se efectúa el análisis bivariado de asociación entre cada una de las covariables explicativas (independientes) y la variable de respuesta (dependiente).

En el caso de las variables independientes de tipo cualitativa o categórica (nominales u ordinales), se emplean las tablas de contingencia para la exploración de posibles asociaciones; se comprueba la significación estadística del contraste asociado al estadístico Chi cuadrado que contrasta la Hipótesis Nula: H_0 = las variables son independientes.

Cuanto mayor sea el valor de Chi-cuadrado, menos verosímil es que la hipótesis sea correcta. De la misma forma, cuanto más se aproxima a cero el valor de Chi-cuadrado, más ajustadas están ambas distribuciones. Si la significación asociada a este estadístico es menor o igual a 0,05 ($p \leq 0,05$) se rechaza la hipótesis nula de independencia y se incluye la variable explicativa en un análisis posterior.

Para las variables independientes cuantitativas, como premisa para la aplicación de técnicas estadísticas paramétricas, se aplica la prueba Kolmogorov-Smirnov para verificar su ajuste a la distribución normal con la variable dependiente padecimiento de pie diabético.

Si la variable tiene distribución normal se utiliza la prueba de comparación de medias en muestras independientes con varianzas desconocidas T-Student; en caso contrario se realiza la prueba no paramétrica U the Mann-Whitney, se comprueba la significación estadística del contraste y, en caso de diferencias significativas ($p \leq 0.05$), se incluye la variable explicativa en un análisis posterior.

Este estudio exploratorio permite establecer cuáles de las variables independientes muestran relaciones estadísticamente significativas con la variable dependiente padecimiento de pie diabético, lo cual permite utilizarlas como variables candidatas para los modelos predictivos.

Con el objetivo de evitar el problema de la multicolinealidad en los modelos que se estimen, cuando se necesita separar la influencia de las variables independientes sobre la dependiente, es útil el análisis de las relaciones que existen entre las variables independientes; es decir, determinar si los cambios en una de las variables influyen en los cambios de la otra. Cuando el grado de asociación sea superior a 0.8 se evidencia la existencia de una correlación fuerte y con ello el problema de la multicolinealidad; en este caso se debe analizar y valorar la posibilidad y beneficios de excluir una de las variables en el análisis de regresión. (Aguayo, 2007)

Para la evaluación de la correlación se analizan los siguientes coeficientes de asociación según el tipo de variable:

- Entre variables categóricas dicotómicas: se utiliza el estadístico Phi y entre variables categóricas politómicas se utiliza V de Cramer y en presencia de variables ordinales Rho de Spearman.
- Entre variables cuantitativas: se utiliza el estadístico Pearson (r) o Spearman según criterio de normalidad.

2.1.3 Determinación y obtención de los modelos de pronóstico factibles

En este epígrafe se proponen los modelos que se analizarán y se explica el proceso para su establecimiento.

La naturaleza de la matriz de datos que resulta particionada por columnas y por casos, donde las variables (columnas) se particionan según una variable dependiente categórica y variables independientes y donde los casos se particionan según las categorías de la variable dependiente (padecimiento de pie diabético, no

padecimiento de pie diabético), unido al hecho de ser las variables independientes nominales, ordinales y cuantitativas, y en correspondencia con los objetivos de la investigación donde se propone establecer un modelo predictivo, corresponde aplicar preferentemente una de las técnicas estadísticas siguientes: regresión logística Binaria, técnicas de segmentación arboles de decisión utilizando el método o algoritmo CHAID (Chi-square AID) o con el CRT(Classification and Regression Trees).

En este paso se establecen y analizan los modelos tentativos seleccionados y se lleva a cabo un estudio comparativo de los resultados obtenidos hasta obtener el modelo que satisfaga en mayor grado los objetivos propuestos en cuanto a los indicadores considerados.

2.1.3.1. Regresión logística binaria

Para la obtención de este modelo con regresión logística binaria en una primera variante se introducen todas las variables independientes seleccionadas en el análisis exploratorio de los datos.

En correspondencia con el objetivo del estudio de encontrar un modelo con la máxima capacidad predictiva, se ha considerado la **técnica de selección hacia delante** que contrasta la entrada basándose en la significación del estadístico de puntuación y contrasta la eliminación basándose en la probabilidad del estadístico de Wald.El estadístico de Wald permite una prueba χ^2para contrastar la hipótesis nula de que el coeficientede cada variable independiente es cero; se identifican aquellas variables cuyos coeficientes resulten significativamente diferentes de 0 ($p \leq 0.05$). (Berlanga & Vilà, 2014).

Se comprueba la bondad del ajuste del modelo de regresión respecto a los datos a través del estadígrafo Chi Cuadrado de Hosmer y Lemeshow, y se calcula la R^2 de Nagelkerke para evaluar la interrelación entre los cambios de la variable dependiente por la unidad de cambios de cada una de las variables independientes.

2.1.3.2 Regresión logística con análisis factorial categórico.

En esta segunda variante de regresión logística binaria para la selección de las variables que se incluirán en el modelo, primeramente se aplica un análisis factorial de componentes principales categórico sobre todas las variables independientes que fueron seleccionadas en el análisis exploratorio, se utiliza el método de normalización principal por variables y el criterio para la selección de componentes de los valores propios mayores que 1 buscando la mayor cantidad posible de varianza explicada. Para garantizar que los datos se ajustan al modelo de análisis factorial se tiene en cuenta los criterios planteados en el capítulo anterior respeto al test de Barlett, el indicador KMO y la diagonal de la correlación anti imagen.

Sobre las variables independientes conformadas por los factores obtenidos se realiza un análisis de regresión logística binaria siguiendo el mismo procedimiento explicado anteriormente.

2.1.3.3 Árbol de decisión CHAID

En la regresión logística el orden de importancia sucesiva de las variables puede sugerir algunas interacciones a considerar, pero esto no es totalmente claro. Por otra parte la inclusión o no de una variable nominal, depende en gran medida de la forma en que esta es codificada y los resultados pueden verse afectados por esto. Las técnicas segmentación de árboles de decisión resuelven estas dificultades.

En el caso concreto de este estudio se realiza el árbol de decisión utilizando los algoritmos **CHAID** y **CRT**, por ser de los más extendidos en ciencias de la salud. (Segura, 2012)

Para la obtención del modelo mediante el algoritmo CHAID se incluyen todas las variables independientes seleccionadas en el análisis exploratorio que tuvieron relación con la variable dependiente padecimiento de pie diabético. El análisis de sensibilidad se basa en los criterios de razón de verosimilitud y los otros parámetros

utilizados y requeridos por el programa fueron: control del tamaño del árbol (hoja con registros superiores a 30 y profundidad del árbol menor a 8).

2.1.3.4 Árbol de decisión CRT

Para la elaboración del modelo utilizando el algoritmo CRT se incluyen todas las variables independientes seleccionadas en el análisis exploratorio que tuvieron relación con la variable dependiente padecimiento de pie diabético, el análisis de sensibilidad se realiza a través del índice de Gini y los otros parámetros utilizados y requeridos por el programa fueron: control del tamaño del árbol (hoja con registros superiores a 30 y profundidad del árbol menor a 8), no selección del criterio de partición inicial, asignación de probabilidades previas iguales para todas las categorías y control del podado por reducción de error estándar.

2.1.4 Selección del o los modelos predictivos con mejor ajuste

Una vez elaborados los modelos estadísticos utilizando las técnicas multivariantes explicadas en los epígrafes anteriores se realiza una tabla comparativa entre cada uno de los modelos elaborados y selecciona el de mejores resultados predictivos y el más adecuado para cumplimentar el objetivo propuesto.

Un modelo que se estime con fines de predicción debe caracterizarse por la estabilidad de sus parámetros en la población donde será empleado y en consecuencia mostrar una capacidad predictiva que justifique su uso en toda la población. Como recurso adicional para validar la calidad del ajuste de un modelo pueden emplearse además de los propios desarrollados para este modelo, indicadores para la evaluación de escalas de clasificación, que permiten valorar su calidad desde diferentes puntos de vista y pueden ser empleados en la evaluación de la capacidad predictiva de modelos estadísticos.

Los criterios a emplear en este análisis van a ser, fundamentalmente los siguientes:

- por ciento de clasificación correcta.
- sensibilidad
- especificidad
- índice Kappa

A continuación se discuten cada uno de estos recursos

Indicadores para la evaluación de escalas de clasificación

Para el cálculo de estos coeficientes se parte de la tabla de clasificación que aportan cada uno de los modelos elaborados. En una tabla de 2X2 se muestra la distribución de los individuos según la categoría de pertenencia real según la información de la muestra (gold estándar), *y=0* (ausencia de la característica) y *y=1* (presenta de la característica), conjuntamente con la categoría en la cual es clasificado por el modelo o escala que se utiliza.

Tabla 1.1 Tabla de clasificación de los modelos predictivos

Grupo actual	Grupo estimado		Total marginal
	y=0	y=1	
y=0	N_{11}	N_{12}	**$N_{11}+N_{22}$**
y=1	N_{21}	N_{22}	**$N_{21}+N_{22}$**
Total	**$N_{11}+N_{21}$**	**$N_{12}+N_{22}$**	**N (Total Pacientes Estudiados)**

Fuente: Libro Informática Médica Tomo II

Esta tabla permite comprobar cómo clasifica el test (el modelo obtenido) a los individuos de la muestra en comparación con el gold estándar (la realidad, lo observado).

A partir de esta tabla de clasificación se definen los conceptos siguientes:

- La sensibilidad (Sb) del modelo está representada por aquel porcentaje de individuos o pacientes que, habiendo presentado el evento, hayan sido clasificados por el modelo correctamente. Indica lo bueno que es el modelo para identificar a los pacientes que poseen esta característica. La expresión aritmética para el cálculo del coeficiente de sensibilidad es:

$$Sb = \frac{N_{22}}{N_{12} + N_{22}}$$

- La Especificidad (Es) del modelo sería aquella proporción de individuos o pacientes que, no habiendo presentado la característica, son clasificados por el modelo como que no presenta la característica. Indica, hasta qué punto el modelo es bueno para identificar a los individuos que no van a sufrir el evento. Matemáticamente se expresa

$$Es = \frac{N_{11}}{N_{11} + N_{21}}$$

- El porciento del 'valor predictivo de un resultado positivo' (VPP), es el que viene determinado por aquellos pacientes que, habiendo sido clasificados con la característica, realmente la hubieran presentado. Es decir:

$$VPP = \frac{N_{22}}{N_{21} + N_{21}} * 100\%$$

- El porciento de 'valor predictivo de un resultado negativo' (VPN), es el que viene determinado por aquellos pacientes que, habiendo sido clasificados como 'no enfermedad' realmente no la hubieran presentado. Es decir

$$VPN = \frac{N_{11}}{N_{11} + N_{12}} * 100\%$$

- El porciento de pacientes bien clasificados (PT) que corresponde a aquellos pacientes que mediante la probabilidad estimada pertenecen a su respectiva categoría

$$PT = \frac{N_{11} + N_{22}}{N} * 100\%$$

Teniendo en cuenta los resultados de estos coeficientes los modelos se comparan según los valores del PT, VPP y Sb. Se selecciona el modelo con mejor por ciento de PT Y VPP, y se valora el que presente mayor Sb, ya que el objetivo del trabajo es la predicción y según plantean diferentes autores para establecer hipótesis diagnósticas hacen falta test de alta sensibilidad evitando que escapen positivos (enfermos). (Pita & Pértegas, 2003)

Índice kappa

Este indicador a considerar para elegir el modelo de mejor resultado es el índice Kappa (k) que se usa para evaluar la concordancia o reproducibilidad de las escalas de clasificación. Relaciona el acuerdo que exhiben los modelos respecto a los valores observadores, más allá del obtenido al azar, con un acuerdo potencial más allá del azar.

En esencia, el proceso de elaboración del índice Kappa (K) consiste en calcular la diferencia entre la proporción de acuerdo observado (P_o) y la proporción de acuerdo esperado por azar (P_e).

$$K = \frac{P_o - P_e}{1 - P_e}$$

Siendo:

$$P_o = \frac{N_{11} + N_{22}}{N} \; P_e = \sum_{i=1}^{n} (P_{i1} * P_{i1})$$

Donde:

n = número de categorías

i = número de la categoría (de 1 hasta n)

P_{i1} = proporción de ocurrencia de la categoría i en el grupo actual

P_{i2} = proporción de ocurrencia de la categoría i en el grupo estimado.

En la interpretación del índice Kappa (k) se debe tener en cuenta que el índice depende del acuerdo observado, pero también de la prevalencia del carácter estudiado y de la simetría de los totales marginales. Si este es igual a cero, entonces el grado de acuerdo que se ha observado puede atribuirse enteramente al azar, si la diferencia es positiva, ello indica que el grado de acuerdo es mayor que el que cabría esperar si solo estuviera operando el azar y viceversa y en el caso (ciertamente improbable) en que la diferencia fuera negativa, entonces los datos estarían exhibiendo menos acuerdo que el que se espera solo por concepto de azar. (Abraira, 2001)

Un criterio para valorar el valor del coeficiente Kappa es el siguiente

Kappa (k) Grado de acuerdo	Kappa (k) Grado de acuerdo
< 0,00	Sin Acuerdo
0,00-0,20	Insignificante
0,21-0,40	Mediano
0,41-0,60	Moderado
0,61-0,80	Sustancial
0,81-1,00	Casi Perfecto

2.1.5 Validación del modelo

La evaluación de la capacidad predictiva del modelo seleccionado se realiza con una muestra de prueba con 185 pacientes de la misma población que no fueron incluidos en la elaboración del modelo predictivo, decisión que se sustenta en el hecho de que en la etapa no ha cambiado la metodología para la medición de las variables independientes en esta población, al tiempo que permite verificar que en el breve tiempo transcurrido no se producen cambios sustanciales en la población por causas socioeconómicas y medioambientales que desaconsejen la extrapolación de los modelos estimados en los años próximos.

Se construye la curva ROC que genera estadísticos que resumen el rendimiento o la efectividad del modelo con el objetivo de validar la utilidad de las variables pronósticas que ante un par de individuos, uno enfermo y otro sano, los clasifique correctamente.

Curva ROC (Receiver Operating Characteristic Curve), considerada útil para definir utilidad de variables pronósticas. Desde el punto de vista estadístico se relacionan una variable continua que actúa como predictora, con una dicotómica.

La evaluación de la curva se realiza a través del examen visual en tanto más alejada del eje de las abscisas más eficiente resultará la función para la predicción se calcula el área y se evalúan la capacidad predictiva del modelo teniendo en cuenta la siguiente escala que corresponde a los valores obtenidos del área bajo la curva (AB). (Cerda & Cifuentes, 2011) (López & Píta, 1998)

- AB= 1 test perfecto
- 0,9 < AB <1 test excelente
- 0,8 < AB ≤ 0,9 test bueno
- 0,7 < AB ≤ 0,8 test regular
- 0,5 < AB ≤ 0,7 mal test
- AB= 0,5 no relación

- AB < 0,5 relación inversa

El cálculo del área bajo consiste en la aplicación de la siguiente formula

$$AB = 1 - \frac{U}{n_1 n_0}$$

Donde:

U corresponde al cálculo de la U de Mann-Whitney respecto a los valores esperados.

n_1 y n_0 son respectivamente el número esperado de “1” y “0”.

2.2 Conclusiones Parciales

- Las variantes introducidas al procedimiento para establecer modelos predictivos constituye un intento en la búsqueda de mejores modelos de pronóstico del padecimiento de pie diabético para los pacientes con diabetes mellitus de tipo II en cuanto a su capacidad de pronóstico y la parsimonia de los mismos, aplicables en otros tipos de enfermedades.
- La aplicación de la prueba de estabilidad con el empleo indicadores para la evaluación de escalas de clasificación, permiten valorar la calidad predictiva de los modelos estadísticos desde diferentes puntos de vista lo cual constituye un elemento que valida la extrapolación de los modelos obtenidos a otros grupos de pacientes.

Capítulo III: Resultados

En el Capítulo III se presentan los resultados relacionados con la aplicación del procedimiento de elaboración del modelo estadístico para predecir el padecimiento de pie diabético en pacientes con diabetes mellitus tipo II a partir de los factores de riesgo en el municipio de Cienfuegos, basado en un modelo que prediga el mismo en la Clínica del Diabético, sobre la base de un conjunto de variables que inciden en la ocurrencia de dicho padecimiento, siguiendo el procedimiento expuesto en el capítulo anterior a partir del uso de la regresión logística y árbol de decisión sobre los componentes obtenidos.

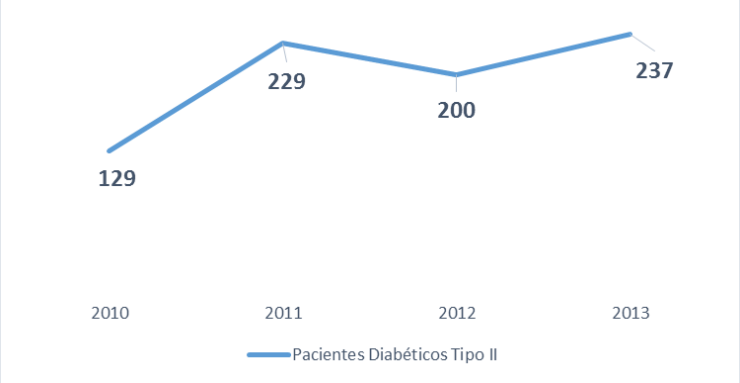
3.1 Análisis descriptivo

Para establecer un modelo estadístico que permita predecir el padecimiento de pie diabético a partir los factores de riesgo se siguen los aspectos abordados en el capítulo anterior. En la investigación se trabaja con toda la información generada en los años comprendidos entre el 2010 y el 2013, considerando representativo el tipo de diabetes (tipo II) y el municipio de residencia de los pacientes (Cienfuegos).

Los datos con los que se trabaja se corresponden a los registros médicos de los pacientes atendidos en la Clínica del Diabético almacenados en una base de datos en formato.xls que contiene los registros de los pacientes según las variables s operacionalizadas en el capítulo anterior.

Se estudiaron un total de 795 pacientes diabéticos tipo II del municipio de Cienfuegos y a continuación se expone el comportamiento en el periodo de tiempo establecido y se describe la distribución de los mismos por sexo y edad

Figura 1: Cantidad de pacientes diabéticos de tipo II del municipio Cienfuegos atendidos en la Clínica del Diabético distribuido por años correspondiente al periodo de estudio.



Fuente: Registros médicos de los pacientes

Como se refleja en la **Figura 1**, desde el año 2011 hubo un incremento de los pacientes diabéticos tipo II, siendo el año 2013 el de mayor frecuencia con un total de 237 (29,8%).

En el **Anexo 1** se especifica en un resumen el comportamiento de cada una de las variables que se utilizan en el estudio; se muestran los estadísticos descriptivos (media, desviación típica, mínimo y máximo) y/o los índices de frecuencia y porcentajes según el tipo de variable.

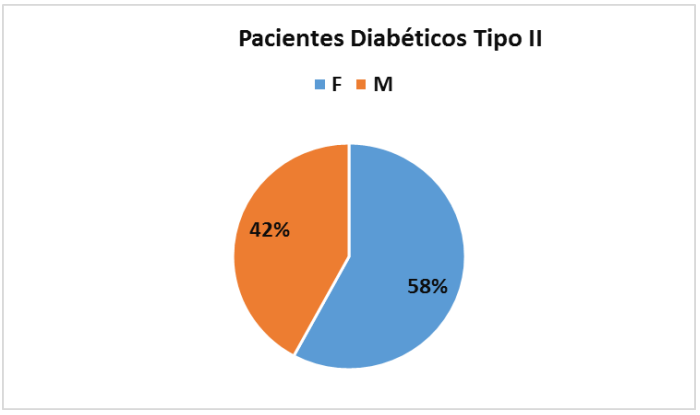
Figura 2: Representación de los pacientes diabéticos de tipo II del municipio Cienfuegos con padecimiento de pie diabético atendidos en la Clínica del diabético

Fuente: Registros médicos de los pacientes

En la **Figura 2** se puede apreciar que 630 pacientes (79,2%) manifestaron el padecimiento de pie diabético y 165 (20,8%) no lo presentaron. La alta frecuencia de esta complicación de la diabetes ha sido reconocida a nivel mundial por su impacto sobre los sistemas de salud. (Márquez, Zonana, Anzaldo, & Muñoz, 2014) (Faget, 2014)

Comportamiento de las variables socio-demográficas (Anexo 1)

Figura 3: Distribución por sexo



Fuente: Registros médicos de los pacientes

En la **Figura 3**, en la distribución por sexo, es válido decir que existe un predominio de los pacientes de sexo femenino 460 (58%), con respecto a los del sexo masculino siendo este 335 (42%); estos resultados están en relación con lo encontrado en la mayoría de los países, refiriéndose a que la prevalencia de diabetes es más elevada en las mujeres que en los hombres. (De la Paz, Proenza, Gallardo, Fernández, & Mompié, 2012)

Tamayo (2014) en su estudio "Influencia de un tratamiento integral de pie diabético en la disminución del índice de amputaciones de los pacientes atendidos en la unidad de pie diabético del hospital provincial general docente Riobamba, durante el período enero – septiembre de 2013", se reporta que 52,90% de los pacientes afectados con pie diabético son femeninos.

Peralta (2014) en su investigación “Estratificación y prevalencia de riesgo de pie diabético en pacientes hospitalizados en el área de medicina interna en el hospital Abel Gilbert Pontón en la ciudad de Guayaquil durante el periodo agosto del 2013 a febrero del 2014”, también refleja el predominio del sexo femenino (58%) sobre el masculino en este padecimiento.

Respecto a las edades de los pacientes se puede resaltar que el promedio fue de 53,84 años siendo la distribución estándar de 11,49 con extremos de 20 a 79 años (**Anexo 1**); varios autores coinciden que el pie diabético es una de las complicaciones principales en la edad reproductiva con una edad promedio a los 51 años, y es común encontrar en estos pacientes dicho padecimiento dada la larga evolución de la diabetes mellitus. (Pinilla, Barrera, Rubio, & Devia, 2014) (Rosales, et al., 2012) (ALAD, 2013)

Los niveles de escolaridad que predominan son bachiller con un total de 271 (34,1%) y secundaria 235 (29,6%), solo 137 (17,2%) tiene nivel superior. Según criterio se ha manifestado en otras investigaciones el grado de escolaridad puede tener relación con el inadecuado control metabólico y deficiencia en el cuidado de los pies, pese a recibir educación diabetológica. Diferentes autores plantean que el grado de instrucción influye en la comprensión de los conocimientos recibidos por los especialistas. (Tamayo, 2014) (Herrera, 2014)

Respecto a la ocupación la mayoría de los pacientes son trabajadores 435 (54,7%) y el 44, 3% corresponde con personas jubiladas 174(21,9%) y amas de casa (22,4%).

Comportamiento de las variables de hábitos de consumo (Anexo 1)

En los hábitos de consumo la ingestión del café resulta la de mayor frecuencia con 474 pacientes que representan el 59,6%; en cuanto al consumo de cigarros 603 (75,8%) refieren no fumar, al igual que con las bebidas alcohólicas 519 pacientes (82,4%) refieren no tener el vicio. Según estos resultados la población tiene un buen comportamiento respecto al hábito de consumo, ya que estos factores de riesgo son predisponentes para desarrollar pie diabético, principalmente el tabaquismo, como

cita la Asociación Latinoamericana de diabetes (2010), dado que influye en la fisiopatología de la úlcera isquémica por disminuir en aporte de oxígeno a las células. (Tamayo, 2014)

Comportamiento variables antecedentes patológicos personales y familiares (Anexo 1)

Respecto a los antecedentes patológicos el más frecuente en este estudio fue la cardiopatía isquémica con 686 pacientes (86,3%), seguido de la Hiperlipidemia que se refleja en 684 (86,0%), la hipertensión arterial en 496 (62,4%), y la Claudicación Intermitente en 12 (1,9%).

La diabetes mellitus constituye un factor de riesgo elevado para la aparición de otras enfermedades como la hipertensión arterial y la cardiopatía isquémica, que, en orden de frecuencia, son los antecedentes que se recogen con más asiduidad en la literatura. (Sun, et al., 2012)

Bustillo y Col, (2014) en el artículo “Resultados del tratamiento con Heberprot-P a pacientes con diagnóstico de pie diabético en el Municipio de Ranchuelo”, expone que los antecedentes patológicos personales que se presentaron con mayor frecuencia fueron la hipertensión arterial para un 46,5%, la cardiopatía isquémica 37,7% y la hiperlipidemia en un 24,4%.

Los antecedentes familiares de diabetes Mellitus se manifestaron en la mayoría de los pacientes para un total de 792 que representa el 99,6%.

Comportamiento de las variables clínicas

En cuanto a los niveles de la glicemia se refleja que la mayoría de los pacientes presentan niveles altos con 434 casos (54,6%) con valores mayores que 7,00 mmol/l y el comportamiento del colesterol 498 (62,6%) presentan niveles menores de 5,20 mmol/l. El promedio del índice de masa corporal es de 29,135 Kg/m^2 con una desviación estándar de 6,0 Kg/m^2 y el tiempo de evolución de padecer diabetes mellitus 4,7 años con una desviación estándar de 6,99 años.

Los valores de creatinina como promedio fueron de 65,12 µmol/l con una desviación estándar de 27,72 µmol/l, los triglicéridos de 1,9 mmol/l con desviación estándar de 1,4 mmol/l y el ácido úrico 283,45 µmol/l con desviación estándar de 103,65 µmol/l.

Según el comportamiento de las variables clínicas y planteamientos de expertos en salud existen algunas irregularidades en el control metabólico en la mayoría de estos pacientes, se clasifican con inadecuados y de riesgo los valores de glicemia, IMC y triglicéridos. Las otras variables como el colesterol y ácido úrico se comportaron de forma adecuada. (Faget, 2014)

3.2 Relación entre las variables independientes y la variable dependiente

Para la selección de las variables que se incluyen los modelos se realiza el análisis bivariado entre la variable dependiente (padecimiento pie diabético) con cada una de la variables independientes que corresponden con las variables socio demográficas, clínicas, hábitos de consumo y antecedentes patológicos y familiares.

Análisis de las variables cualitativas

En el **Anexo 2** se comprueba la significación del contraste asociada al estadístico Chi cuadrado entre las variables independientes cualitativas y el padecimiento de pie diabético. Se puede apreciar que para un nivel de confianza del 95% ($p \leq 0.05$) presentan asociación significativa las variables fuma (p=0,038), café (p=0,027), bebidas (p=0,043), APP_CI (p=0,00), APP_HIPER (p=0,00), glicemia (p=0,022) y colesterol (p=0,029)

Análisis de las variables cuantitativas.

En relación a las variables cuantitativas se analizan los supuestos de normalidad como premisa para aplicar pruebas paramétricas de comparación de medias; como se muestra en el **Anexo 3**, ninguna variable tiene distribución normal, por lo que se aplica la prueba no paramétrica U the Mann-Whitney para comprobar la significación

del contraste entre las variables independientes cuantitativas y el padecimiento de pie diabético.

Como se puede apreciar en el **Anexo 4** para un nivel de confianza del 95% ($p \leq 0.05$) presentaron asociación significativa las variables edad (p=0,00) y tiempo (p=0,001).

Análisis entre todas las variables que presentan relación con el padecimiento de pie diabético

En el capítulo II se plantea que para evitar una correlación excesiva entre las variables independientes se lleva a cabo una correlación entre las variables independientes con asociación significativa con el padecimiento de pie diabético, con el objetivo de verificar aquellas que se encontraran relacionadas entre sí para manejar la multicolinealidad entre las mismas.

Como se aprecia en el **Anexo 5** en algunas de estas variables se presentan correlación, que no resulta una correlación fuerte según el valor prefijado en el capítulo anterior (mayor que 0,8), y así se descarta en un primer momento que hubiera este tipo de inconveniente en el momento de ingresar las variables en el modelo multivariante final.

Por todo lo anterior dicho se decide incluir en la elaboración de los modelos las siguientes variables: fuma, café, bebidas, APP_CI, APP_HIPER, colesterol, glicemia, edad y tiempo.

3.3 Modelos multivariados para la predicción del padecimiento de pie diabético

3.3.1 Modelo de regresión logística binaria (RL1)

La regresión logística binaria se aplica con el objetivo de cuantificar cómo influye en la probabilidad de aparición del padecimiento de pie diabético, la presencia o no de los factores de riesgo identificados en las etapas anteriores, mediante los coeficientes estimados de las variables incluidas en el modelo.

Para el análisis de regresión incluimos las variables descritas en los apartados anteriores que tuvieron relación con la variable dependiente padecimiento de pie diabético. Se crean variables dummy según las categorías de las variables independientes de tipo cualitativas y se utiliza el método hacia adelante Wald. Los resultados de modelo de regresión se muestran en el **Anexo 6**

Tabla 3.1. Resultados de la regresión logística binaria (Regresión 1)

	B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95,0% para EXP(B)	
							Inferior	Superior
Edad	,020	,008	6,481	1	,011	1,021	1,005	1,037
Tiempo	,050	,018	7,660	1	,006	1,051	1,015	1,089
APP_CI	2,078	,598	12,081	1	,001	7,991	2,475	25,796
APP_HIPER	2,123	,597	12,640	1	,000	8,356	2,592	26,934
Constante	-,210	,420	,250	1	,617	,811		

En la **Tabla 3.1** se muestran las variables estadísticamente significativas con (coeficientes p<=0,05) que en este caso corresponden a las variables edad (p=0, 011), tiempo (p=0,006), APP_CI (p=0,001) y APP_HIPER (p=0,000) por tanto la ecuación de regresión para predecir el padecimiento de pie diabético queda estructurada de la siguiente forma

$$P = \frac{1}{1 + e^{(0{,}020 * \text{edad} + 0{,}050 * \text{tiempo} + 2{,}078 * \text{APP\_CI} + 2{,}123 * \text{APP\_HIPER} - 0{,}210)}}$$

Se comprueba la bondad del ajuste del modelo respecto a los datos a través del estadígrafo Chi Cuadrado de Hosmer y Lemeshow **(Anexo 6).** La probabilidad asociada al estadígrafo es de 0,442 siendo mayor que 0,05 por lo que no hubo diferencias entre los valores esperados y observados con independencia de la prevalencia del suceso y se considera que el modelo se ajusta a los datos.

Para evaluar la interrelación entre los cambios de la variable dependiente por la unidad de cambios de cada una de las variables independientes, se calcula la R^2 de Cox y Snell (0,095) y la R^2 de Nagelkerke (0,148) con un resultado significativo, el modelo explica entre el 0,095 y el 0,148 de la variable dependiente;(ver tabla Resumen de los modelos del **Anexo 6)** .Estos análisis previos indican que algunas de las variables independientes funcionan en el modelo de regresión logística como variables predictoras.

Según los resultados de la tabla de clasificación del **Anexo 6** las variables incluidas en el modelo clasifican correctamente la variable dependiente padecimiento pie diabético con un porcentaje global de 79,2 %, con lo cual el modelo se puede considerar bueno.

La probabilidad de padecer pie diabético en la población estudiada está significativamente relacionada con la edad, tiempo de evolución de padecer diabetes, antecedentes patológicos de cardiopatía isquémica e hiperlipidemia. Los resultados coinciden con los criterios de diferentes autores que han estudiado este padecimiento e incluyen estas variables como factores de riesgo unidos a otros factores que no se registran en la Clínica del Diabético de Cienfuegos. (Garriga, Sánchez, & Vázquez, 2014) (Pinilla, Barrera, Rubio, & Devia, 2014)

3.3.2 Modelo de regresión logística binaria con análisis factorial (RL2)

En muchas situaciones, un número pequeño de factores común pueden representar un gran porcentaje de la variabilidad de las variables originales. La habilidad para expresar la covarianza entre las variables en términos de un número pequeño de factores significativos es de gran ayuda para profundizar en los datos que son analizados; con el fin de profundizar en los datos que son analizados y buscar posible interdependencia entre las variables de estudio se procede a realizar una nueva regresión logística incluyendo como covariables los componentes creados con análisis factorial.

Por ser la mayoría de las variables objeto de estudio de tipo categórico, se hace necesario realizar un escalamiento óptimo, para posteriormente utilizar las técnicas tradicionales de reducción de datos, lo cual se realiza con la ayuda del paquete estadístico SPSS v.15 a través de la opción reducción de datos – escalamiento óptimo, quedando asignadas cuantificaciones numéricas a las categorías de cada variable.

A partir de estos resultados se procede al análisis factorial de componentes principales teniendo en cuenta los criterios planteados en los capítulos anteriores

El coeficiente de adecuación KMO=0,602 cae en el rango de aceptación (superior a 0,50). El test de esfericidad de Bartlett verifica la matriz de correlaciones no es identidad (ver **Anexo 7**). La matriz anti-imagen muestra valores muy bajos y los coeficientes KMO bastante altos en su diagonal, por lo que con este análisis se puede concluir que el procedimiento factorial que sigue a continuación puede proporcionarnos conclusiones satisfactorias.

Utilizando el método de los componentes principales se obtienen cuatro componentes con valores propios mayores que la unidad, que explican el 60,13 % de la varianza total (ver **Anexo 7**), lo cual se considera aceptable, estando en correspondencia con el criterio que plantea que los factores que se extraen deben representar por lo menos un 60% de la varianza. (Hair, Anderson, Tatham, & Black, 2004)

La matriz rotada de los pesos factoriales se obtienen según el procedimiento ortogonal VARIMAX, logrando minimizar el número de variables con saturaciones elevadas en varios factores según se muestra en la tabla Matriz de componentes rotados(a) del **Anexo 7**.

En este estudio se ha elegido la pauta de rechazar las variables que poseen pesos factoriales inferiores a 0,50, pues supondrían que estas aportan poco a la definición del factor. A partir de los criterios antes mencionados, se interpretan las correlaciones entre las variables y los factores según la solución rotada por el

método VARIMAX, que minimiza el número de variables con carga elevada en cada componente, por lo tanto facilita la definición de las mismas. El comentario se realiza sobre la solución rotada que permite que la interpretación de los factores sea más fácil.

La matriz de pesos factoriales rotadas muestra que todas las variables saturan en algún factor (según VARIMAX), quedando en los cuatro factores cargas factoriales superiores al valor de 0,50 preestablecido. A continuación se exponen las componentes resultantes:

- Factor 1: Se recogen las variables relacionadas con antecedente patológico cardiopatía isquémica, hiperlipidemia y hábito de consumo de café
- Factor 2: Se recogen las variables relacionadas con los valores de la glicemia y el tiempo de padecer diabetes mellitus
- Factor 3: Se recogen las variables relacionadas con consumo de bebidas alcohólicas y la edad
- Factor 4: Solo la variable relacionada con los valores del colesterol

Con estos componentes se realiza una regresión logística siguiendo los criterios planteados en el epígrafe anterior.

Tabla 3.2. Resultados de la regresión logística binaria (Regresión 1)

	B	**E.T.**	**Wald**	**gl**	**Sig.**	**Exp(B)**	**I.C. 95, 0% para EXP(B)**	
							Inferior	Superior
Factor 1	-,661	0,114	33,470	1	0,000	0,516	0,413	0,646
Factor 2	0,367	0,099	13,757	1	0,000	1,443	1,189	1,752
Factor 4	0,270	0,098	7,635	1	0,006	1,310	1,082	1,586
Constante	1,511	0,101	221,776	1	0,000	4,529		

En la **Tabla 3.2** se muestran las variables estadísticamente significativas con (coeficientes p<=0,05) que en este caso corresponden al Factor 1 (p=0,000), Factor 2 (p=0,000) y Factor 4 (p=0,006), por tanto la ecuación de regresión para predecir el padecimiento de pie diabético queda estructurada de la siguiente forma

$$P = \frac{1}{1 + e^{(1{,}511 - 0{,}661 * factor1 + 0{,}367 * factor2 + 0{,}270 * factor4)}}$$

Se comprueba la bondad del ajuste del modelo respecto a los datos a través del estadígrafo Chi Cuadrado de Hosmer y Lemeshow **(Anexo 8).** La probabilidad asociada al estadígrafo es de 0,062 siendo mayor que 0,05 por lo que no hubo diferencias entre los valores esperados y observados con independencia de la prevalencia del suceso y se considera que el modelo se ajusta a los datos.

Para evaluar la interrelación entre los cambios de la variable dependiente por la unidad de cambios de cada una de las variables independientes, se calcula la R^2 de Cox y Snell (0,073) y la R^2 de Nagelkerke (0,114) con un resultado significativo, el modelo explica entre el 0,073 y el 0,114 de la variable dependiente;(ver tabla Resumen de los modelos del **Anexo 8)** .Estos análisis previos indican que algunas de las variables independientes funcionan en el modelo de regresión logística como variables predictoras.

Según los resultados de la tabla de clasificación del **Anexo 8** las variables incluidas en el modelo clasifican correctamente la variable dependiente padecimiento pie diabético con un porcentaje global de 79,5 %, con lo cual el modelo se puede considerar bueno.

3.3.3 Modelo con algoritmo CHAID

El Análisis de sensibilidad se basó en los criterios de razón de verosimilitud, los otros parámetros utilizados (requeridos por el programa) fueron: control del tamaño del árbol (hoja con registros superiores a 30 y profundidad del árbol menor a 8).

Los resultados para el algoritmo CHAID se exponen en el **Anexo 9** donde se obtuvo un modelo compuesto por 8 segmentos para una clasificación global de 80 %.

A continuación se presenta la descripción de cada segmento.

Segmento 1 (97,3%)	Pacientes con padecimiento de hiperlipidemia
Segmento 2 (97,3%)	Pacientes con cardiopatía isquémica
Segmento 3 (50%)	Pacientes con edad ≤ 50 años y con tiempo de padecer diabetes mellitus ≤ 0,74
Segmento 4 (71,1%)	Pacientes con edad ≤ 50 años y con tiempo de padecer diabetes mellitus ≥0,74
Segmento 5 (72,2%)	Pacientes con edad > 50 años y con glicemia < 7,0
Segmento 6 (83,6%)	Pacientes con edad > 50 años y con glicemia >7,0 no fumador y con tiempo de evolución con diabetes mellitus ≤ 2, 49 años
Segmento 7 (98,4%)	Pacientes con edad > 50 años y con glicemia >7,0 no fumador y con tiempo de evolución con diabetes mellitus >2, 49 años
Segmento 8 (76,8%)	Pacientes con edad > 50 años y con glicemia >7,0 fumador

3.3.4 Modelo con algoritmo CRT

Para la elaboración del modelo utilizando el algoritmo CRT se utiliza el índice de Gini, los otros parámetros utilizados (requeridos por el programa) fueron: control del tamaño del árbol (hoja con registros superiores a 30 y profundidad del árbol menor a 8), no selección del criterio de partición inicial y asignación de probabilidad previa igual para todas las categorías y control del podado por reducción de error estándar. Los resultados para el algoritmo CRT se exponen en el **Anexo 10** donde se obtuvo un modelo compuesto por 6 segmentos para una clasificación global de un 56,9% (no es buena clasificación).

A continuación se presenta la descripción de cada segmento.

Segmento 1 (97,3%)	Pacientes con padecimiento de hiperlipidemia
Segmento 2 (97,3%)	Pacientes con cardiopatía isquémica
Segmento 3 (57,4%)	Pacientes ≤ 0, 54 años con diabetes Mellitus
Segmento 4 (73,5%)	Pacientes con más de 0, 54 años con diabetes Mellitus, ácido

	y edad menor que 50 años
Segmento 5 (77,0%)	Pacientes con más de 0, 54 años con diabetes Mellitus, edad mayor que 50 años y la glicemia < 7
Segmento 6 (96,3%)	Pacientes con más de 0,54 años con diabetes Mellitus, edad mayor que 50 años y la glicemia mayor que 7

3.4 Selección del modelo más ventajoso

Para la selección del modelo más ventajoso se utiliza la tabla de clasificación de cada uno los modelos elaborados, (**Anexo 6, 8, 9 y 10**); como se propone en el capítulo II para la comparación se tendrán en cuenta los valores del por ciento de pacientes bien clasificados (PT), el por ciento del valor predictivo de un resultado positivo (VPP), y se calculan los valores de sensibilidad (Sb) y el índice de kappa (K)

A continuación en la **Tabla 3.3** se expresan los modelos elaborados con los criterios mencionados en una tabla comparativa

Tabla 3.3 Tabla comparativa de los modelos estadísticos elaborados

Modelo	PT	VPP	Sb	K
RL1	0,792	0,838	0,893	0,420
RL2	0,795	1,00	0,794	0,019
Árbol de decisión CHAID	0,800	0,850	0,891	0,429
Árbol de decisión CRT	0,564	0,470	0,958	0,228

Teniendo en cuenta los valores de la Tabla 3.3 se puede apreciar, respecto al porcentaje de clasificación, que los modelos elaborados con RL1 y el algoritmo CHAID se comportan de forma similar, aunque este último es ligeramente superior con un 80 % de clasificación correcta. El modelo de árbol de decisión con algoritmo CRT se decidió excluirlo de la comparación, ya que solo clasifica correctamente al 56,4% y se considera como un porcentaje muy bajo respecto a los otros modelos.

De los valores de VPP el de mejor resultados es el modelo de RL2 con un 100%, este modelo pronostica a los pacientes con padecimientos de pie diabético

correctamente, aunque los modelos con RL1 y CHAID presentan valores adecuados con un 85% y 83,8% respectivamente.

Respecto a los resultados de la sensibilidad el modelo RL1 y el algoritmo CHAID fueron los que clasificaron correctamente a los pacientes que tenían padecimiento de pie diabético con valores de 89,3% y 89,1% respectivamente;

De acuerdo al coeficiente Kappa, se aprecian valores muy próximos entre el modelo RL1 y el algoritmo CHAID siendo de 0,420 y 0,429 respectivamente, además de ser estos modelos los que mejores valores del índice Kappa presenta.

A partir del análisis de la tabla comparativa con los modelos elaborados, se observa que el modelo RL1 y el modelo con árbol de decisión CHAID se comportan de forma similar y son los que presentan resultados más estables para aplicarlos en la Clínica del Diabético para predecir el padecimiento de pie diabético en pacientes con diabetes mellitus de tipo II residentes en el municipio de Cienfuegos.

3.5 Validación del modelo seleccionado

Para la validación de los resultados se utiliza una muestra prueba para la evaluación de la capacidad predictiva de los modelos elegidos, los pacientes seleccionados cumplen con los requisitos del estudio y no pertenecen a la muestra utilizada en la elaboración del modelo.

Se realiza un análisis del comportamiento de la curva ROC construida con los puntos de corte definidos por los modelos elaborados. Se efectúa el análisis a través el examen visual de la curva (**Anexo 11**) y los valores del área bajo la curva.

Tabla 4.Área bajo la curva

Variables resultado de contraste	Área	Error típ.(a)	Sig. asintótica(b)	Intervalo de confianza asintótico al 95%	
	Límite inferior	Límite superior	Límite inferior	Límite superior	Límite inferior
RL1 1	,697	,021	,000	,656	,738
CHAID	,732	,020	,000	,694	,770

a Bajo el supuesto no paramétrico
b Hipótesis nula: área verdadera = 0,5

Como se aprecia en la tabla 4 el área el área bajo la curva para el modelo con el algoritmo CHAID es superior al modelo RL1 con un valor de 0,732 considerado como la capacidad predictiva del modelo aceptable a pesar que según la escala definida en el capítulo anterior es regular. A partir del análisis de los resultados de la comparación de ambos modelos, se selecciona el Modelo elaborado con algoritmo CHAID para aplicarlo en la Clínica del Diabético para predecir el padecimiento de pie diabético en pacientes con diabetes mellitus de tipo II.

3.6 Conclusiones Parciales

- Se elaboraron modelos predictivos para el padecimiento de pie diabético en pacientes con diabetes mellitus tipo II en el municipio de Cienfuegos a partir de los factores de riesgos que inciden en la ocurrencia de dicho padecimiento.
- El análisis comparativo entre los RL1 y CHAID arroja resultados similares sobre la base de los indicadores para la evaluación de escalas de clasificación, aunque en la validación con la curva ROC el modelo CHAID muestra resultados levemente superiores, siendo este el seleccionado para aplicar en la Clínica del Diabético
- El modelo propuesto es adecuado y posibilita realizar pronósticos con un aceptable porcentaje de aciertos.

Conclusiones

- Como resultado del proceso de búsqueda de un modelo estadístico que permita predecir la aparición de pie diabético en la población de pacientes con diabetes mellitus de tipo II a partir de factores de riesgo y haciendo uso de las técnicas multivariantes se puede establecer que el uso de los árboles de decisión a través del algoritmo CHAID constituye una variante que garantiza una capacidad predictiva adecuada.
- Con el análisis de documento se pudo apreciar que los modelos estadísticos son de gran importancia en las ciencias de salud, como herramientas útiles, siendo el análisis de regresión logística binaria y árboles de decisión con los algoritmos CHAID y CRT como las técnicas que más se utilizan en salud para elaborar modelos predictivos.
- Los factores de riesgos de mayor influencia en los pacientes con diabetes mellitus de tipo II atendidos en la clínica del diabético son: los antecedentes patológicos de hiperlipidemia y la cardiopatía isquémica, la edad avanzada de los pacientes, el tiempo de evolución de padecer diabetes mellitus, los niveles en sangre del colesterol y la glicemia y los hábitos de consumo de café, bebidas alcohólicas y el fumar.
- .La aplicación de la prueba de estabilidad con el empleo indicadores para la evaluación de escalas de clasificación, permiten valorar la calidad predictiva de los modelos estadísticos a través de: porcentaje de clasificación correcta, índice de Kappa, sensibilidad y especificidad
- La validación de los resultados través de la curva ROC genera estadísticos que resumen la efectividad del modelo y la utilidad de las variables pronósticas que ante un par de individuos, uno enfermo y otro sano, los clasifique correctamente.

Recomendaciones

- Continuar con el estudio de los modelos predictivos en nuevas investigaciones de esta naturaleza con vista a establecer posibles regularidades en los resultados.
- Utilizar la metodología de árboles de decisión con el algoritmo CHAID en el campo de la medicina como una herramienta de investigación para encontrar patrones y definir tendencias.

Referencias

Abraira, V. (2001). El índice kappa. *SEMERGEN-Medicina de Familia*, *27*(5), 247-49.

Acosta, L., Hernández, A., Victorero, G., & Cruz, L. (2013). DIABETES MELLITUS, PIE DIABETICO, HEBERPROT-P; INTERACCION EN LOS SERVICIOS DE PODOLOGIA. *Revista Cubana de Tecnología de la Salud*

Afonso, A., M.H., E., Gonzales, R., Stein, J., Genton, B., & Senn, N. (2012). The use of classification and regression trees to predict the likelihood of seasonal influenza. *Family practice*, cms020.

Aguayo, M. (2007). *Cómo hacer una Regresión Logística con SPSS© "paso a paso". (I).* Sevilla: FABIS. Cómo hacer una Regresión Logística con SPSS" paso a paso"(I). DocuWeb-fabis. Huelva: Fundación Andaluza Beturia para la Investigación en Salud. Sevilla, España.

ALAD. (2013). *Asociación Latinoamericana de Diabetes. Guías ALAD sobre el diagnóstico, control y tratamiento de la diabetes mellitus tipo II con medicina basada en evidencia*. *Revista de la ALAD*, 1-142. Obtenido de http://issuu.com/alad-diabetes/docs/guias_alad_2013?e=3438350/5608514

Álvarez, A., González, J., Maceo, L., Frómeta, A., Bárzaga, S., & Cervantes, A. (2014). Árbol para predecir el desarrollo de la cardiopatía hipertensiva. *Revista Cubana de Medicina*, *53*(3), 266-281.

Arai, T., Kaneko, S., & Fujita, H. (2011). Decision trees on gait independence in patients with femoral neck fracture. *Nihon Ronen Igakkai zasshi. Japanese journal of geriatrics*, *48*(5), 539-544.

Barnés, J., García, Y., Valdés, C., Reynaldo, D., Savigne, W., & LLanes, A. (2014). *Folleto Complementario para el Manejo Integral del Pie Diabético en la Comunidad.* In *Morfovirtual 2014*.

Bembibre, R., Cabrera, J., Suárez, R., & Concepción, E. (2007). Caracterización y factores pronósticos de la enfermedad cerebrovascular en la Provincia de Cienfuegos. *Medisur*, *2*(2), 15-24.

Berea, R., Rivas, R., Pérez, M., Palacios, L., Moreno, J., & Talavera, J. (2014). Del juicio clínico a la regresión logística múltiple. Rev Med Inst Mex Seguro Soc, 52(2), 192-7.

Berlanga, V., & Vilà, R. (2014). Cómo obtener un Modelo de Regresión Logística Binaria con SPSS. *REIRE(Revista d'Innovació i Recerca en Educació)*, 7(2), 105-118

Bustillo, M. J., Feito, T., Vegoña, F., Alvarez, Y., & Gurra, B. (2014). Resultados del tratamiento con heberprot-P a pacientes con diagnóstico de pie diabético en el Municipio de Ranchuelo. *Acta Médica del Centro*, 8(2).

Carrasco, S. (2010). El análisis estadistico de la información (modelización). En S. Carrasco, *Análisis de datos en Economía.*

Cerda, J., & Cifuentes, L. (2011). Uso de curvas ROC en investigación clínica: Aspectos teórico-prácticos. *Revista chilena de infectología, 29*(2), 138-141.

Correa, C. (2014). *Detección temprana de la infección por Citomegalovirus y otros herpes virus en población con riesgo de desarrollar infección congénita y en receptores pediátricos de trasplante cubanos.* La Habana: Ciencias Médicas.

Cuadras, C. (2014). *Nuevos Métodos de Análisis Multivariante.* CMC Editions. Barcelona, España.

Cuadras, C. M., Arenas, C., & Fortiana, J. (1991). *Multicua. Paquete no standard de Análisis Multivarante.* Barcelona: Departamento de Estadística. Universidad de Barcelona.

De la Paz, K., Proenza, L., Gallardo, Y., Fernández, S., & Mompié, A. (2012). Factores de riesgo en adultos mayores con diabetes mellitus. *MEDISAN*, 16(4), 489-497.

Echemendía, B. (2011). Definiciones acerca del riesgo y sus implicaciones. *Revista Cubana de Higiene y Epidemiologia*, 49(3), 470-481.

Esquerda, A., & Trujillano, J. (2010). *Árboles de clasificación para la generación de reglas de decisión clínica.* Jano: Medicina y humanidades, (1758), 75.

Faget, O. (2014). *Prevención del Pie Diabético.* La Habana: La Clínica del Pie Diabético.

Garriga, L., Sánchez, M., & Vázquez, M. (2014). Caracterización del estado de salud de la población diabética del Área 2 en Cienfuegos. *Revista Finlay*, 4(2), 79-84.

Gonzalez, C., & Agudo, A. (2003). Factores de riesgo: Aspectos generales. En A. Martin, & P. Cano, *Atención primaria. Conceptos, organización y práctica clínica.* (págs. 752-63). España: Elsevier.

González, J., & González, J. (2013). Factores de riesgo para la ocurrencia de infarto agudo del miocardio en pacientes fumadores. *Revista Cubana de Salud Pública*, 39(4), 679-688.

González, V. (2014). Aspectos críticos del empleo en salud de modelos estadísticos de clasificación. *Revista Médica Electrónica*, 36, 742-751.

Hair, J., Anderson, R., Tatham, R., & Black, W. (2004). *Análisis Multivariante.* Madrid: PEARSON. Hair, J.; Anderson, R.; Tatham, R.;Black, W. (2004). Análisis Multivariante (Vol. 5a).Madrid: PEARSON, Pretice Hall.

Ham, N., Whiting, D., Guariguata, L., Aschner, P., Forouhi, N., Hamblenton, I., & Li, R. (2013). *IFD Diabetes Atlas.Sixth Edition.* Bruselas: International Diabetes Federation.

Hernández, A., Pascual, A., Baño, A., Melero, M. R., & Molina, M. (2014). Diferencias en el número de cesáreas en los partos que comienzan espontáneamente y en los inducidos. *Revista Española de Salud Pública*, 88(3), 383-393.

Herrera, M. (2014). *Prevalencia de amputacion en pacientes con pie diabetico infectado en el año 2012 del Hospital de especialidades de Guayaquil "Dr. Abel Gilbert Ponton" .* (Doctoral dissertation).Guayaquil : Universidad Católica de santiago de Guayaquil .

Jiménez, G. M. (2014). Nivel de conocimientos del paciente diabético sobre la prevención del pie diabético. *Revista Medisur*, 54-59 *5*(2), 40-43.

Lilienfeld, A., & Lilienfeld, D. (1987). Fundamentos de epidemiología. *Addison Wesley Iberoamericana*, 138.

López, I., & Píta, S. (1998). Curvas ROC. *Atención Primaria en la Red, 5*(4), 229-235.

López, M., Corcho, B., & Moreno, A. (1999). Notas históricas sobre el desarrollo de la epidemiología y sus definiciones. *Rev Mex Pediatr*, 66(3), 110-114.

Márquez, S., Zonana, A., Anzaldo, M., & Muñoz, J. (2014). Riesgo de pie diabético en pacientes con diabetes mellitus tipo 2 en una unidad de medicina de familia. *Medicina de Familia SEMERGEN*, 40(4), 183-188 .

Mass, G., Cabrera, T., Torres, F., Vidal, G., Moya, A., & Alonso, J. (2014). Efectividad del Heberprot P en la úlcera de pie diabético en un área de salud. *Revista Finlay*, 4(2), 85-89.

Massip, J., Soler, S., & Torres, R. (2011). Uso de la estadística en la Revista Cubana de Higiene y Epidemiología, 1996-2009. *Revista Cubana de Higiene y Epidemiología*, 49(2), 276-291.

Medina, M., Palma, M., Zapata, R., Pérez, N., & Zuniga, M. (2014). Cincuenta años de uso de la estadística en la investigación en salud pública. Una proyección al 2029. *Ciencia y Humanismo en la Salud*, *1*(3), 89-103.

MINISTERIO DE SALUD PÚBLICA. (2013). *Anuario estadístico de Salud.* La Habana.

OMS. (Marzo de 2014). *Diabetes*. Obtenido de http://www.who.int/topics/diabetes_mellitus/es/index.html

OMS. (Marzo de 2014). *Enfermedades no transmisibles*. Obtenido de http://www.who.int/chp/chronic_disease_report/part1/es/

OPS. (1988). El desafío de la Epidemiología. Desafío de la epidemiología; problemas y lecturas seleccionadas *Publicación Científica*, (Vol. 505), 3-17.

Peralta, M. (2014). *Estratificación y prevalencia de riesgo de pie diabético en pacientes hospitalizados en el área de medicina interna en el hospital Abel Gilbert Pontón en la ciudad de Guayaquil durante el periodo agosto del 2013 a febrero del 2014.* (Doctoral dissertation) Guayaquil: Universidad Católica Santiago de Guayaquil.

Pérez, G., & Grau, R. (2014). Predicción de la evolución hacia la hipertensión arterial en la adultez desde la adolescencia. *RCIM Revista Cubana de Informática Médica*, 4(1), 43-53.

Pinilla, A., Barrera, M. P., Rubio, C., & Devia, D. (2014). Actividades de prevención y factores de riesgo en diabetes mellitus y pie diabético. *Acta Médica Colombiana*, 39(3), 250-257.

Pita Fernández, S., & Pértegas Díaz, S. (2003). Pruebas diagnósticas: Sensibilidad y especificidad. *Cad Aten Primaria*, *10*, 120-4.

Rodríguez, J., López, B., Verdejo, F., Martín, C., Hernández, A., Lorente, R., Olmedo, J. (2013). Factores predictores de esofagitis eosinofílica en impactación esofágica por bolo alimenticio. *Revista de Gastroenterología Mexicana*, 78(1),5.

Rojo, J. (2006). *Árboles de Clasificación y Regresión.* Madrid: Instituto de Economía y Geografía.

Rosales, M., Bonilla, J., Gómez, A., Gómez, C., Pardo, J., & Villanueva, L. (2012). Factores asociados al pie diabético en pacientes ambulatorios. Centro de Diabetes Cardiovascular del Caribe. Barranquilla (Colombia). Revista *Salud Uninorte*, 28(1), 65-74.

Ruiz, J. (2006). Historia de las Estadisticas de Salud. *Gaceta Médica Boliviana* 29(2), 72-77.

Salvador, M. (2000). *Leccion de estadística .* Obtenido de Introducción al Análisis Multivariante: http://www.5campus.com/leccion/anamul

Segura, A. (2012). Aplicación de árboles de decisión en la salud pública. *Revista CES Salud Pública*, 3(1), 94-103.

Star, L., & Moghadas, S. (2010). The role of mathematical modelling in public health planning and decision making. *Purple Paper*, 6.

Sun, J. H., Tsai, J. S., Huang, C. H., Lin, C. H., Yang, H. M., Chan, Y. S., & Huang, Y. Y. (2012). Risk factors for lower extremity amputation in diabetic foot disease categorized by Wagner classification. *Diabetes research and clinical practice*, 95(3), 358-363

Tamayo, M. (2014). *Influencia De Un Tratamiento Integral De Pie Diabético En La Disminución Del Índice De Amputaciones De Los Pacientes Atendidos En La Unidad De Pie Diabético Del Hospital Provincial General Docente Riobamba, Durante El Período Enero – Septiembre De 2013.* Ambato: Universidad Técnica de Ambato.

Valdés, E., & Camps, M. l. (2013). Características clínicas y frecuencia de complicaciones crónicas en personas con diabetes mellitus tipo 2 de diagnóstico reciente. *Revista Cubana de Medicina General Integral*, 29(2), 121-131.

Vega, B., Sánchez, L., Cortiñas, J., Castro, O., & González, M. (2012). Clasificación de dengue hemorrágico utilizando árboles de decisión en la fase temprana de la enfermedad. *Revista Cubana Medicina Tropical*, 64(1), 35-42..

Vicente, B., Zerquera, G., Rivas, E., Muñoz, J., Gutiérrez, Y., & Castañedo, E. (2010). Nivel de conocimientos sobre diabetes mellitus en pacientes con diabetes tipo 2 . *Medisur*, 8(6), 412-418.

Anexos

ANEXO 1

Padecimiento pie diabético(PD)	Frecuencia	Porcentaje
No	165	20,8
Si	630	79,2
Total	795	100,0

VARIABLES SOCIO DEMOGRÁFICAS

	Valor Mínimo	Valor Máximo	Media Promedio	Desviación. Típica.
Edad	20	79	53,84	11,49

		Frecuencia	Porcentaje	Porcentaje acumulado
Sexo	F	460	57,9	57,9
	M	335	42,1	57,9
Escolaridad	Primaria no terminada	36	4,5	4,5
	Primaria	64	64	12,6
	Secundaria	235	29,6	42,1
	Obrero calificado	1	0,1	42,3
	Bachiller	271	34,1	76,4
	Técnico Medio	51	6,4	82,8
	Universitario	137	17,2	100,0
Ocupación	Trabajador	435	54,7	54,7
	Jubilado	174	21,9	76,6
	Ama de casa	178	22,4	99,0
	Desempleado o desocupado	8	1,0	100,0

VARIABLES ANTECEDENTES PATOLÓGICOS PERSONALES Y FAMILIARES

		Frecuencia	Porcentaje	Porcentaje acumulado
APF_DM	No	3	0,4	99,6
	Si	792	99,6	100,0
APP_HTA	No	299	37,6	37,6
	Si	496	62,4	100,0
APP_CI	No	109	13,7	13,7
	Si	686	86,3	100,0
APP_HIPER	No	111	14,0	14,0
	Si	684	86,0	100,0
APP_CINTERMITENTE	No	12	1,5	1,5
	Si	783	98,5	100,0

VARIABLES HÁBITOS DE CONSUMO

		Frecuencia	Porcentaje	Porcentaje acumulado
Fuma	No	**603**	**75,8**	**75,8**
	Si	**192**	**24,2**	**100,0**
Café	No	**321**	**40,4**	**40,4**
	Si	**474**	**59,6**	**100,0**
Bebidas	No	**658**	**82,8**	**82,8**
	Ex bebedor	**3**	**0,4**	**83,1**
	Ocasional	**33**	**4,2**	**87,3**
	Si	**101**	**12,7**	**100,0**

VARIABLES CLÍNICAS

		Frecuencia	Porcentaje	Porcentaje acumulado
Glicemia	<7,00	**361**	**45,4**	**45,4**
	7,00+	**434**	**54,6**	**100,0**
Colesterol	<= 5,20	**498**	**62,6**	**62,6**
	5,21+	**297**	**37,4**	**100,0**

	Valor Mínimo	Valor Máximo	Media Promedio	Desviación. Típica.
IMC	**15,62**	**69,57**	**29,1354**	**6,02491**
Tiempo	**0,48**	**55,00**	**4,7024**	**6,98895**
Creatinina	**13,00**	**350,00**	**65,1231**	**27,72668**
Triglicérido	**0,14**	**14,80**	**1,9209**	**1,42199**
Acido_urico	**30,40**	**940,00**	**283,4586**	**103,65012**

ANEXO 2

Análisis de las variables cualitativas

Prueba Chi Cuadrado

	Valor	gl	Sig. asintótica (bilateral)
SEXO	2,81	1,00	0,093
ESOLARIDAD	7,805	7,00	0,350
OCUPACION	2,75	3,00	0,432
FUMA	4,30	1,00	0,038
CAFÉ	4,87	1,00	0,027
BEBIDAS	8,125	3,00	0,043
APP_DM	0,79	1,00	0,37
APP_HTA	0,30	1,00	0,58
APP_CI	24,89	1,00	0,00
APP_HIPER	25,56	1,00	0,00
APP_CINTERMITENTE	3,19	1,00	0,07
GLICEMIA	5,275	1,00	0,022
COLESTEROL	4,759	1,00	0,035

ANEXO 3

Análisis de las variables cuantitativas

Pruebas de normalidad

	PD	Kolmogorov-Smirnov(a)			Shapiro-Wilk		
		Estadístico	**gl**	**Sig.**	**Estadístico**	**gl**	**Sig.**
edad	No	0,073	165	0,030	0,988	165	0,174
	Si	0,054	630	0,000	0,991	630	0,000
IMC	No	0,081	165	0,010	0,972	165	0,002
	Si	0,086	630	0,000	0,925	630	0,000
tiempo	No	0,313	165	0,000	0,610	165	0,000
	Si	0,265	630	0,000	0,664	630	0,000
creatinina	No	0,085	165	0,006	0,941	165	0,000
	Si	0,159	630	0,000	0,698	630	0,000
triglicérido	No	0,213	165	0,000	0,714	165	0,000
	Si	0,168	630	0,000	0,696	630	0,000
Acido_urico	No	0,080	165	0,013	0,926	165	0,000
	Si	0,050	630	0,001	0,977	630	0,000

a Corrección de la significación de Lilliefors

ANEXO 4

Análisis de las variables cuantitativas

Estadísticos de contraste(a)

	edad	IMC	tiempo	creatinina	triglicérido	acido_urico
U de Mann-Whitney	42022,500	47011,000	41033,500	51317,500	50361,000	46834,500
W de Wilcoxon	55717,500	245776,000	54728,500	250082,500	249126,000	245599,500
Z	-3,792	-1,890	-4,184	-0,250	-0,615	-1,958
Sig. asintót. (bilateral)	0,000	0,059	0,000	0,802	0,539	0,051

a Variable de agrupación: PD

ANEXO 5

Determinación de colinealidad entre variables independientes,coeficientes de correlación

	X1	X2	X3	X4	X5	X6	X7	X8	X9
X1		0,105	0,140	-0,054	-0,007	0,001	0,010	-0,140	-0,290
X2	0,105		0,074	0,157	0,154	0,022	0,46	0,168	0,064
X3	0,140	0,074		0,071	0,154	0,072	0,053	-0,124	-0,056
X4	-0,054	0,157	0,071		0,219	0,092	0,029	-0,158	-0,078
X5	-0,007	0,154	0,154	0,219		0,061	0,145	-0,107	-0,060
X6	0,001	0,022	0,072	0,092	0,061		0,046	0,046	0,191
X7	0,010	0,460	0,053	0,029	0,145	0,046		0,097	0,057
X8	-0,140	0,168	-0,124	-0,158	-0,107	0,046	0,097		0,138
X9	-0,290	0,064	-0,056	-0,078	-0,060	0,191	0,057	0,138	

X1 Fuma
X2 Café
X3 Bebidas
X4 APP_CI
X5 APP_HIPER
X6 Glicemia
X7 Colesterol
X8 Edad
X9 tiempo

ANEXO 6

Regresión logística 1

Pruebas omnibus sobre los coeficientes del modelo

		Chi-cuadrado	gl	Sig.
Paso 1	Paso	35,551	1	,000
	Bloque	35,551	1	,000
	Modelo	35,551	1	,000
Paso 2	Paso	25,314	1	,000
	Bloque	60,865	2	,000
	Modelo	60,865	2	,000
Paso 3	Paso	11,788	1	,001
	Bloque	72,653	3	,000
	Modelo	72,653	3	,000
Paso 4	Paso	6,535	1	,011
	Bloque	79,188	4	,000
	Modelo	79,188	4	,000

Prueba de Hosmer y Lemeshow

Paso	Chi-cuadrado	gl	Sig.
1	,000	0	.
2	,263	1	,608
3	15,399	8	,052
4	7,910	8	,442

Resumen de los modelos

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	776,444	,044	,068
2	751,129	,074	,115
3	739,342	,087	,136
4	732,807	,095	,148

a La estimación ha finalizado en el número de iteración 6 porque las estimaciones de los parámetros han cambiado en menos de,001.

Tabla de clasificación(a)

Observado			Pronosticado		
			PD		Porcentaje correcto
			No	Si	
Paso 4	PD	No	102	63	61,81
		Si	102	528	83,80
					79,2

a El valor de corte es,500

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95,0% para EXP(B)	
								Inferior	Superior
Paso 1(a)	APP_HIPER	2,413	,592	16,610	1	,000	11,172	3,500	35,663
	Constante	1,170	,090	169,260	1	,000	3,222		
Paso 2(b)	APP_CI	2,164	,596	13,200	1	,000	8,708	2,709	27,988
	APP_HIPER	2,192	,595	13,559	1	,000	8,956	2,788	28,769
	Constante	1,043	,092	128,297	1	,000	2,839		
Paso 3(c)	tiempo	,055	,018	9,348	1	,002	1,057	1,020	1,095
	APP_CI	2,164	,597	13,156	1	,000	8,702	2,703	28,014
	APP_HIPER	2,159	,596	13,108	1	,000	8,665	2,692	27,890
	Constante	,831	,110	56,934	1	,000	2,297		
Paso 4(d)	edad	,020	,008	6,481	1	,011	1,021	1,005	1,037
	tiempo	,050	,018	7,660	1	,006	1,051	1,015	1,089
	APP_CI	2,078	,598	12,081	1	,001	7,991	2,475	25,796
	APP_HIPER	2,123	,597	12,640	1	,000	8,356	2,592	26,934
	Constante	-,210	,420	,250	1	,617	,811		

a Variable(s) introducida(s) en el paso 1: APP_HIPER.
b Variable(s) introducida(s) en el paso 2: APP_CI.
c Variable(s) introducida(s) en el paso 3: tiempo.
d Variable(s) introducida(s) en el paso 4: edad.

ANEXO 7

KMO y prueba de Bartlett

Medida de adecuación muestral de Kaiser-Meyer-Olkin.		,602
Prueba de esfericidad de Bartlett	**Chi-cuadrado aproximado**	208,900
	gl	28
	Sig.	,000

Varianza total explicada

Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción			Suma de las saturaciones al cuadrado de la rotación		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	1,639	20,491	20,491	1,639	20,491	20,491	1,389	17,458	17,458
2	1,120	13,997	34,488	1,120	13,997	34,488	1,212	15,245	32,703
3	1,026	12,829	47,317	1,026	12,829	47,317	1,133	14,168	46,871
4	1,007	12,813	60,13	1,007	12,813	60,13	1,059	13,259	**60,13**
5	,926	11,474	71,604						
6	,822	10,175	81,779						
7	,762	9,501	91,28						
8	,698	8,72	100						

Método de extracción: Análisis de Componentes principales.

Matriz de componentes rotados(a)

	Componente			
	1	2	3	4
APP_CI(cuantificación)	**,708**	-,066	,057	,095
café(cuantificación)	**-,661**	,002	-,024	-,044
App_Hiper(cuantificación)	**,590**	-,019	,047	-,334
Glicemia(cuantificación)	-,073	**,765**	,238	,026
tiempo(cuantificación)	,001	**,748**	-,283	,002
bebidas(cuantificación)	-,045	,104	**,849**	,057
edad(cuantificación)	-,308	,224	**-,518**	,160
colesterol(cuantificación)	,019	,020	-,021	**,952**

Método de extracción: Análisis de componentes principales.
Método de rotación: Normalización Varimax con Kaiser.
a La rotación ha convergido en 4 iteraciones.

ANEXO 8

Regresión logística caso 2

Pruebas omnibus sobre los coeficientes del modelo

		Chi-cuadrado	gl	Sig.
Paso 1	Paso	37,742	1	,000
	Bloque	37,742	1	,000
	Modelo	37,742	1	,000
Paso 2	Paso	14,721	1	,000
	Bloque	52,463	2	,000
	Modelo	52,463	2	,000
Paso 3	Paso	7,963	1	,005
	Bloque	60,426	3	,000
	Modelo	60,426	3	,000

Resumen de los modelos

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	774,253	,046	,072
2	759,532	,064	,100
3	751,569	,073	,114

a La estimación ha finalizado en el número de iteración 6 porque las estimaciones de los parámetros han cambiado en menos de,001.

Prueba de Hosmer y Lemeshow

Paso	Chi-Cuadrado	gl	Sig.
1	19,197	8	,014
2	20,179	8	,010
3	14,838	8	,062

Tabla de clasificación(a)

Observado			Pronosticado		
			PD		Porcentaje correcto
			No	Si	
Paso 3	PD	No	2	163	1,2
		Si	0	630	100,0
					79,5

a El valor de corte es,500

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95,0% para EXP(B)	
								Inferior	Superior
Paso 1(a)	FAC1_2	-,617	,110	31,722	1	,000	,539	,435	,669
	Constante	1,444	,096	225,621	1	,000	4,240		
Paso 2(b)	FAC1_2	-,638	,112	32,540	1	,000	,528	,424	,658
	FAC2_2	,364	,099	13,654	1	,000	1,439	1,186	1,745
	Constante	1,485	,099	223,485	1	,000	4,416		
Paso 3(c)	FAC1_2	-,661	,114	33,470	1	,000	,516	,413	,646
	FAC2_2	,367	,099	13,757	1	,000	1,443	1,189	1,752
	FAC4_2	,270	,098	7,635	1	,006	1,310	1,082	1,586
	Constante	1,511	,101	221,776	1	,000	4,529		

a Variable(s) introducida(s) en el paso 1: FAC1_2.
b Variable(s) introducida(s) en el paso 2: FAC2_2.
c Variable(s) introducida(s) en el paso 3: FAC4_2.

ANEXO 9

Riesgo

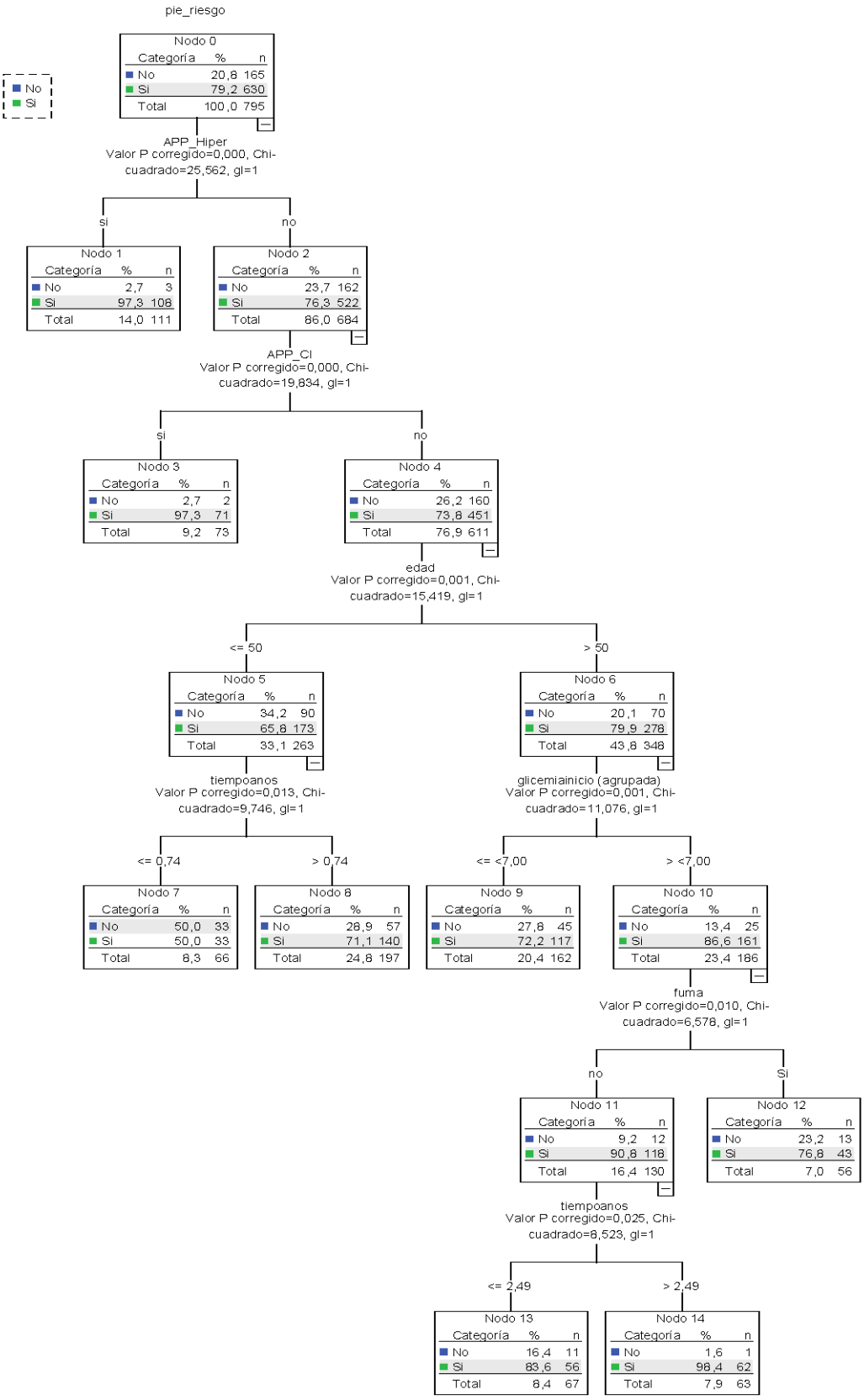
Estimación	Error típico
,208	,014

Método de crecimiento: CHAID
Variable dependiente: PD

Clasificación

Observado	Pronosticado		
	No	Si	Porcentaje correcto
No	100	65	60,60%
Si	94	536	85,08%
			80%

Método de crecimiento: CHAID
Variable dependiente: PD


pie_riesgo
No
Si
Nodo 0
Categoría % n
No 20,8 165
Si 79,2 630
Total 100,0 795
APP_Hiper
Valor P corregido=0,000, Chi-cuadrado=25,562, gl=1
si
no
Nodo 1
Categoría % n
No 2,7 3
Si 97,3 108
Total 14,0 111
Nodo 2
Categoría % n
No 23,7 162
Si 76,3 522
Total 86,0 684
APP_CI
Valor P corregido=0,000, Chi-cuadrado=19,834, gl=1
si
no
Nodo 3
Categoría % n
No 2,7 2
Si 97,3 71
Total 9,2 73
Nodo 4
Categoría % n
No 26,2 160
Si 73,8 451
Total 76,9 611
edad
Valor P corregido=0,001, Chi-cuadrado=15,419, gl=1
<= 50
> 50
Nodo 5
Categoría % n
No 34,2 90
Si 65,8 173
Total 33,1 263
Nodo 6
Categoría % n
No 20,1 70
Si 79,9 278
Total 43,8 348
tiempoanos
Valor P corregido=0,013, Chi-cuadrado=9,746, gl=1
glicemiainicio (agrupada)
Valor P corregido=0,001, Chi-cuadrado=11,076, gl=1
<= 0,74
> 0,74
<= <7,00
> <7,00
Nodo 7
Categoría % n
No 50,0 33
Si 50,0 33
Total 8,3 66
Nodo 8
Categoría % n
No 28,9 57
Si 71,1 140
Total 24,8 197
Nodo 9
Categoría % n
No 27,8 45
Si 72,2 117
Total 20,4 162
Nodo 10
Categoría % n
No 13,4 25
Si 86,6 161
Total 23,4 186
fuma
Valor P corregido=0,010, Chi-cuadrado=6,578, gl=1
no
Si
Nodo 11
Categoría % n
No 9,2 12
Si 90,8 118
Total 16,4 130
Nodo 12
Categoría % n
No 23,2 13
Si 76,8 43
Total 7,0 56
tiempoanos
Valor P corregido=0,025, Chi-cuadrado=8,523, gl=1
<= 2,49
> 2,49
Nodo 13
Categoría % n
No 16,4 11
Si 83,6 56
Total 8,4 67
Nodo 14
Categoría % n
No 1,6 1
Si 98,4 62
Total 7,9 63

ANEXO 10

Riesgo

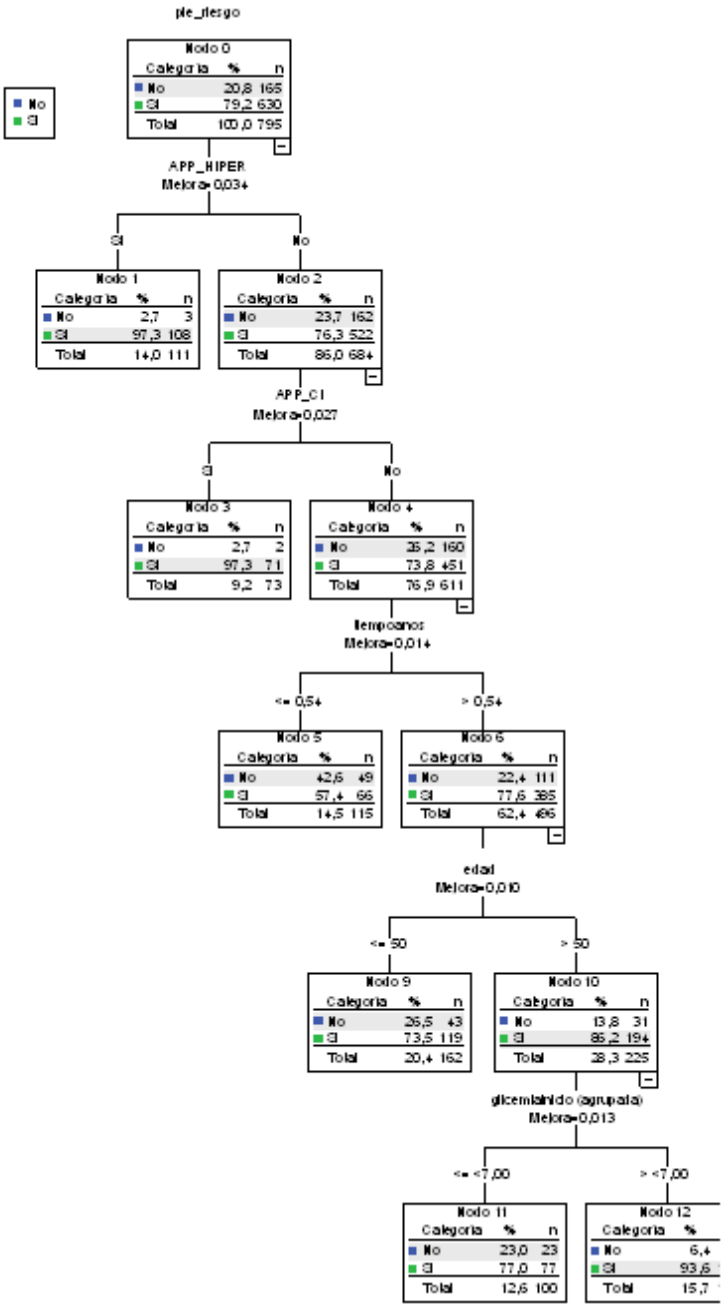
Estimación	Error típico
,304	,014

Método de crecimiento: CRT
Variable dependiente: PD

Clasificación

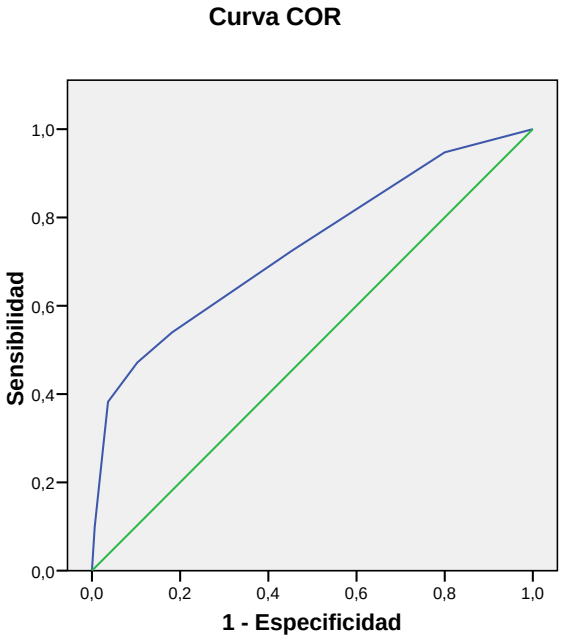
Observado	Pronosticado		
	No	Si	Porcentaje correcto
No	152	13	92,1%
Si	334	296	47,0%
			56,4%

Método de crecimiento: CRT
Variable dependiente: PD


pie_riesgo
No
Sí
Nodo 0
Categoría % n
No 20,8 165
Sí 79,2 630
Total 100,0 795
APP_HIPER
Mejora=0,034
Sí
No
Nodo 1
Categoría % n
No 2,7 3
Sí 97,3 108
Total 14,0 111
Nodo 2
Categoría % n
No 23,7 162
Sí 76,3 522
Total 86,0 684
APP_CI
Mejora=0,027
Sí
No
Nodo 3
Categoría % n
No 2,7 2
Sí 97,3 71
Total 9,2 73
Nodo 4
Categoría % n
No 26,2 160
Sí 73,8 451
Total 76,9 611
tiempoanos
Mejora=0,014
<= 0,54
> 0,54
Nodo 5
Categoría % n
No 42,6 49
Sí 57,4 66
Total 14,5 115
Nodo 6
Categoría % n
No 22,4 111
Sí 77,6 385
Total 62,4 496
edad
Mejora=0,010
<= 50
> 50
Nodo 9
Categoría % n
No 26,5 43
Sí 73,5 119
Total 20,4 162
Nodo 10
Categoría % n
No 13,8 31
Sí 86,2 194
Total 28,3 225
glicemiainicio (agrupada)
Mejora=0,013
<= 7,00
> 7,00
Nodo 11
Categoría % n
No 23,0 23
Sí 77,0 77
Total 12,6 100
Nodo 12
Categoría %
No 6,4
Sí 93,6
Total 15,7

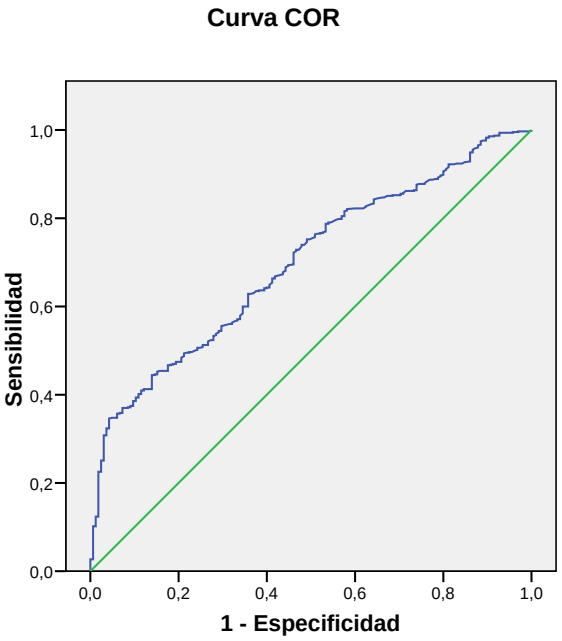
ANEXO 11

Curva ROC Modelo con algoritmo CHAID



Los segmentos diagonales son producidos por los empates.

Curva ROC Modelo con Regresión Logística 1



Los segmentos diagonales son producidos por los empates.