TESIS DE GRADO
EN INGENIERÍA INDUSTRIAL

APLICACIÓN DE MINERÍA DE DATOS PARA DETERMINAR LA EFICACIA DE LA BRAQUITERAPIA EN EL TRATAMIENTO DE CÁNCER DE PROSTATA

Autor: Diego Reparaz

Directora de Tesis: M. Ing. Paola Britos
Co-Director de Tesis: Dr. Ramón García Martínez

2008

RESUMEN EJECUTIVO

En las últimas décadas, hemos sido testigos de vertiginosos cambios en las tecnologías existentes. En este marco de rápida evolución, no podemos dejar afuera los avances producidos en las ciencias de salud. Frente a dichos cambios, aparecen en la medicina nuevas alternativas para el tratamiento de enfermedades. Al encontrarnos frente a estas nuevas alternativas, comienza a tomar mayor importancia el concepto de calidad de vida, ligada a cada uno de los tratamientos.

Para el tratamiento del cáncer de próstata, existen tratamientos alternativos. Entre ellos se encuentran, la prostatectomía radical (abierta o laparoscópica), radioterapia conformada externa, terapia hormonal y braquiterapia. Para distintos pacientes con estadíos de enfermedad curables, la elección del tratamiento debe contrastarse contra los riesgos de la terapia, la edad del paciente, el comportamiento biológico del cáncer, la calidad de vida y otros factores. Esta situación, pone de manifiesto la necesidad de generar herramientas de toma de decisiones, para maximizar la eficacia a la hora de elegir el tratamiento adecuado para un paciente.

El objetivo de este estudio es encontrar patrones de comportamiento y relaciones entre las variables, de forma tal, de poder predecir de antemano la eficacia de la braquiterapia, para un paciente que padece cáncer de próstata. Para alcanzar este objetivo se utiliza la metodología de minería de datos CRISP-DM. Se realizan distintos experimentos tendientes a la compresión completa del comportamiento del conjunto de datos y caracterización de la información y su posterior clasificación, para ello se utilizan los algoritmos redes bayesianas, mapas auto-organizados e inducción.

El procesamiento de los datos según las distintas estrategias utilizadas permite verificar la consistencia de los resultados. Por otra parte, se logró generar reglas de clasificación para los distintos atributos considerados, estableciendo así limites cuantitativos definidos para sugerir o no la braquiterapia, ante la presencia de un caso testigo.

EXECUTIVE SUMMARY

In the last few decades, we have seen extensive changes in the available technologies in all areas. What is called a *technological revolution* started with the invention of the transistor by J. Bardeen. In particular, the medical sciences field was strongly influenced since the new technologies provided alternative techniques for treating specific pathologies. The concept of *life quality* was born based on the development of such medical techniques.

For the prostate cancer treatment a series of treatments are available. Among them we find: Radical Prostatectomy (Opened or Laparoscopic), External Radiotherapy, Hormonal Therapy, and Brachytherapy. The election of a particular treatment for patients with different pathologies must be based on the risks of the chosen therapy, the age of the patient, the biological behaviour of the cancer, the life quality, etc. The development of new tools for determining a suitable treatment is needed, since the number of variables for determining such a treatment is usually high.

The purpose of this work is to find behaviour patterns and relations within the variables, in order to predict the efficiency of the Brachytherapy treatment for a patient affected with prostate cancer. For this purpose, we use the data mining methodology CRISP-DM. We have performed different experiments in order to fully address the behaviour of the samples, and their subsequent classification using the algorithms of Bayesian networks, self assembled maps and induction.

We were able to verify the consistency of the obtained results by processing the data using different strategies. We have also obtained a set of classification rules for the different considered variables, establishing quantitative limits for suggesting or not the Brachytherapy treatment for any particular case.

AGRADECIMIENTOS

A mi padre, por quien hoy estoy aquí.
A Marcelo, compañero de viaje.
A mi madre y mi hermano, que siguen sosteniéndome y apoyando mis proyectos.
Gracias.

ÍNDICE:

1 INTRODUCCIÓN

En las últimas décadas, hemos sido testigos de vertiginosos cambios en las tecnologías existentes. En este marco de rápida evolución, no podemos dejar afuera los avances producidos en las ciencias de salud. Frente a dichos cambios, aparecen en la medicina nuevas alternativas para el tratamiento de enfermedades. Al encontrarnos frente a estas nuevas alternativas, comienza a tomar mayor importancia el concepto de calidad de vida, ligada a cada uno de los tratamientos.

Para el tratamiento del cáncer de próstata, existen tratamientos alternativos. Entre ellos se encuentran, la prostatectomía radical (abierta o laparoscópica), radioterapia conformada externa, terapia hormonal y braquiterapia. Para distintos pacientes, con estadíos de enfermedad curables, la elección del tratamiento debe contrastarse contra los riesgos de la terapia, la edad del paciente, el comportamiento biológico del cáncer, la calidad de vida y otros factores. Esta situación, pone de manifiesto la necesidad de generar herramientas de toma de decisiones, para maximizar la eficacia a la hora de elegir el tratamiento adecuado para un paciente.

En la actualidad, los especialistas cuentan con algunas herramientas que sirven de ayuda en el momento de tomar una decisión respecto al tratamiento adecuado para cada paciente. La herramienta por excelencia es la experiencia profesional, la cual confiere a los especialistas criterios consensuados a la hora de optar por un tratamiento u otro. Estos criterios se basan en una serie de indicadores, que varían en un rango de valores determinado, y que al interactuar generan un output, que es la decisión del profesional acerca del tratamiento “óptimo”.

Con el afán de brindar un marco teórico a las decisiones, tanto para optimizarlas como para dotar de un valor cuantitativo a la experiencia, es que la ingeniería industrial hace su aporte. Existen dos métodos de ayuda en la toma de decisiones. El primero, más extensamente estudiado y aceptado, es el conocido como Nomogramas (Partin). Los nomogramas son herramientas predictivas basadas en análisis de regresión multivariable. Son representaciones gráficas de modelos estáticos, que utilizan escalas, para calcular el “peso” del valor de cada variable, y luego predecir un determinado punto final (end point). Los puntos finales que se estudian entre otros pueden ser: estadío de la enfermedad, probabilidad de reaparición de la enfermedad [Shariat *et al*., 2005], predicción de retención urinaria aguda o intervención quirúrgica en pacientes con hyperplasia próstatica benigna [Slawin *et al*., 2006]. Las predicciones que se obtienen son resultado de los indicadores individuales de cada paciente. Los nomogramas están formados por una serie de ejes, cada uno de los cuales representa una variable. Las variables varían dentro de una escala, y a cada valor de la variable le corresponde una puntuación dependiendo del impacto que dicha variable tenga en la predicción. El eje final, concentra la puntuación final, que es transformada en la probabilidad de alcanzar el punto final. Este tipo de métodos, debe tener especial

cuidado a la hora de definir como imputar el "peso" al valor de las variables, al descartar variables que puedan resultar importantes, al incorporar variables inadecuadas, entre otras. Los nomogramas son los métodos mas estudiados y por lo tanto existen varios estudios de validación de dichos modelos. Entre otros podemos encontrar: Validación de nomograma para predecir resultados positivos de biopsia en cáncer de próstata [Yanke *et al*., 2005].

El otro método es el de minería de datos [García-Martínez *et al*., 2003]. Éste último, es bastante novedoso para este tipo de aplicaciones, por lo cual existen desarrollos muy puntuales. Existe gran variedad de algoritmos (caracterización, inducción, etc), que tienen la capacidad de aprender de la experiencia. Están formados por nodos de ingreso, nodos ocultos y nodos de salida. Mediante entrenamiento (Supervisado, No Supervisado) el modelo ajusta los pesos de las neuronas ocultas para optimizar la salida. La ventaja de minería frente a los nomogramas es que posee la capacidad de resolver relaciones no lineales complejas entre las variables, sin necesidad de hacer ninguna suposición previa respecto a dichas relaciones. La utilización de este método aún sigue siendo controversial, tanto por lo novedoso para estas aplicaciones como porque no resulta sencillo demostrar que sus resultados sean mejores a los arrojados por los nomogramas [Stephan *et al*., 2005].

En este contexto, el objetivo de este trabajo es caracterizar a los pacientes con cáncer de próstata tratados con braquiterapia, con el fin de optimizar la elección de dicha terapia para futuros pacientes.

2 GENERALIDADES DEL CÁNCER DE PRÓSTATA

2.1 Próstata

La próstata es parte del sistema reproductor del hombre. Está ubicada enfrente del recto y debajo de la vejiga. Una próstata sana es del tamaño de una nuez y tiene la forma de rosca. La uretra (el tubo por el que fluye la orina) pasa a través de la próstata. Si la próstata crece demasiado, comprime la uretra. Esto puede causar problemas urinarios al hacer lento o detener el flujo de la orina desde la vejiga al pene.

La próstata es una glándula que produce parte del fluido seminal. En la eyaculación, el fluido seminal ayuda a transportar los espermatozoides hacia afuera del cuerpo del hombre como parte del semen.

Las hormonas masculinas (andrógenos) hacen que crezca la próstata. Los testículos son la fuente principal de hormonas masculinas, incluyendo la testosterona. Las glándulas suprarrenales también producen testosterona aunque en pequeñas cantidades. En la figura 1 se muestra un detalle de la localización de la próstata en el cuerpo humano y un corte de la próstata.

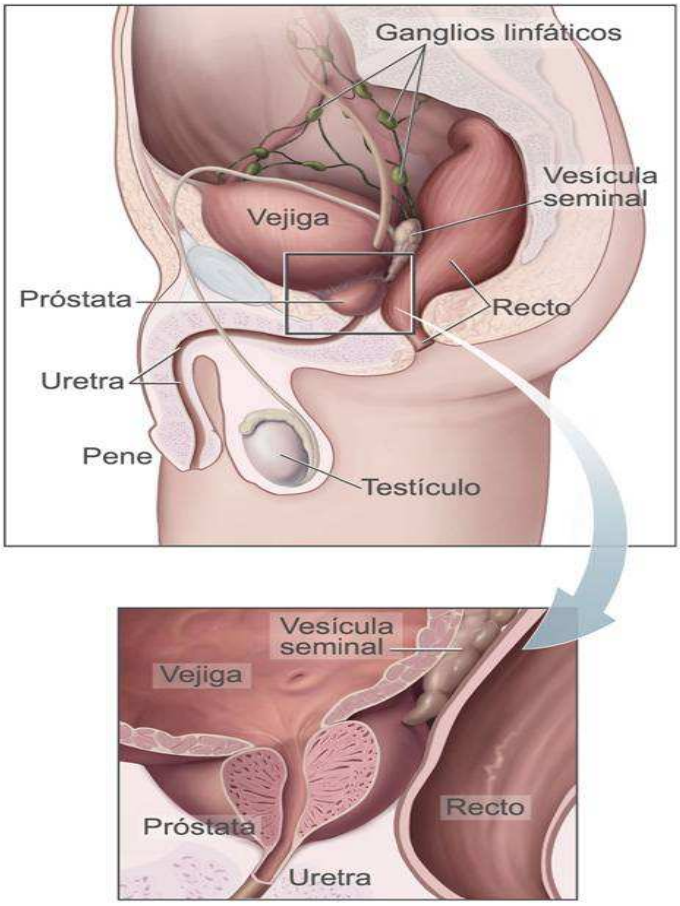


Figura 1. Detalle de localización de la próstata.

Normalmente, las células crecen y se dividen para formar nuevas células cuando el cuerpo las necesita. Cuando las células envejecen, mueren, y células nuevas las reemplazan.

Algunas veces este proceso ordenado se descontrola. Células nuevas se siguen formando cuando el cuerpo no las necesita y las células viejas no mueren cuando deberían morir. Estas células que no son necesarias forman una masa de tejido, que es lo que se llama tumor.

Los tumores pueden ser benignos o malignos:

- Los tumores benignos no son cancerosos.
- Los tumores benignos rara vez son una amenaza para la vida.
- Generalmente, los tumores benignos se pueden operar y pocas veces vuelven a crecer.
- Las células de tumores benignos no invaden tejidos de su derredor.
- Las células de tumores benignos no se diseminan a otras partes del cuerpo.
- Los tumores malignos son cancerosos.
- Los tumores malignos generalmente son más graves que los tumores benignos. Pueden poner la vida en peligro.
- Los tumores malignos pueden extirparse, pero pueden volver a crecer.
- Las células de tumores malignos pueden invadir y dañar tejidos y órganos cercanos.
- Las células de tumores malignos pueden diseminarse a otras partes del cuerpo.
- Las células se diseminan al desprenderse del cáncer original (tumor primario) y entrar en el torrente de la sangre o en el sistema linfático. Ellas invaden otros órganos, forman tumores nuevos y dañan estos órganos.
- Cuando el cáncer se disemina, se llama metástasis

Cuando el cáncer de próstata se disemina (se extiende) fuera de la próstata, las células cancerosas se encuentran con frecuencia en los ganglios linfáticos cercanos. Si el cáncer ha llegado a estos ganglios, es muy probable que las células cancerosas se hayan diseminado a otros órganos.

Cuando el cáncer se disemina (tiene metástasis) desde su sitio original a otra parte del cuerpo, el nuevo tumor tiene el mismo tipo de células anormales y el mismo nombre que el tumor primario.

2.2 Factores de Riesgo

- Edad: La edad es el factor de riesgo más fuerte de cáncer de próstata. Esta enfermedad es rara en hombres menores de 45 años.
- Antecedentes familiares: El riesgo de un hombre de desarrollar cáncer de próstata es mayor si su padre o hermano ha tenido esta enfermedad.
- Raza: El cáncer de próstata es más común en hombres afroamericanos que en hombres blancos, incluyendo hombres blancos hispanos. Es menos común en hombres asiáticos o indígenas americanos.
- Ciertos cambios de la próstata: El tener células anormales, lo que se llama neoplasia intraepitelial prostática, puede aumentar el riesgo de cáncer de próstata. Estas células de la próstata se ven anormales al microscopio.

2.3 Detección del cáncer de próstata

Se realizan dos estudios iniciales para determinar la posible existencia de cáncer.

- Examen rectal digital: Se inserta un dedo en el recto y se siente la próstata a través de la pared del recto para buscar áreas endurecidas o abultadas.
- Análisis de sangre para antígeno prostático específico (PSA): Se examina el nivel de PSA en la muestra de sangre del hombre. Un nivel elevado de PSA es causado con más frecuencia por BPH o prostatitis (inflamación de la próstata). También puede ser el resultado de cáncer de próstata.

El examen rectal digital y el análisis de PSA pueden usarse para detectar un problema de la próstata, pero no pueden mostrar si un problema es cáncer u otro estado menos grave

2.4 Síntomas

- Problemas urinarios
- Inhabilidad para orinar o dificultad para empezar o detener el flujo de orina
- Necesidad de orinar frecuentemente, especialmente durante la noche
- Flujo débil o interrumpido de orina
- Dolor o ardor al orinar
- Dificultad para tener erecciones
- Sangre en la orina o en el semen
- Dolor frecuente en la parte baja de la espalda, las caderas o la parte superior de los muslos

2.5 Diagnostico

Si en los exámenes iniciales se encuentran sospechas de posible cáncer de próstata, se procede a realizar nuevos exámenes:

- Ecografía transrectal: Se inserta una sonda en el recto para buscar áreas anormales. La sonda envía ondas sonoras fuera del alcance del oído humano (ultrasonido). Las ondas sonoras rebotan en la próstata y una computadora usa los ecos para crear una imagen llamada ecografía.
- Cistoscopia: Se mira dentro de la uretra y la vejiga por medio de un tubo delgado y luminoso.
- Biopsia: Una biopsia es la remoción de tejido para buscar células cancerosas. El médico inserta en la próstata una aguja por el recto y retira una pequeña cantidad de tejido (biopsia transrectal). El médico toma muestras de tejido de varias áreas de la próstata. La ecografía puede usarse para guiar la aguja.

El cáncer de próstata se puede clasificar según la etapa o estadío en el que se encuentra. Cada etapa se puede describir usando un número romano (I-IV):

- Etapa I es cáncer que no se puede sentir durante un examen rectal. Se encuentra por casualidad cuando se hace una operación por otra razón, generalmente por hiperplasia prostática benigna. El cáncer está localizado sólo en la próstata.
- Etapa II es cáncer más avanzado, pero no se ha diseminado fuera de la próstata.
- Etapa III es cáncer que se ha diseminado más allá de la capa exterior de la próstata. Se puede encontrar en las vesículas seminales, pero no se ha diseminado a los ganglios linfáticos.
- Etapa IV se refiere a una o varias de las características siguientes:
    - Cáncer que ha invadido la vejiga, el recto u otras estructuras vecinas (con excepción de las vesículas seminales); o
    - Cáncer que se ha diseminado a los ganglios linfáticos; o
    - Cáncer que se ha diseminado a otras partes del cuerpo, tales como los huesos.
    - Cáncer recurrente es cáncer que ha regresado (recurrido) después de tratamiento

2.6 Tratamiento

No existe una forma única que sea la mejor para tratar el cáncer de próstata. El tratamiento de cáncer de próstata depende principalmente de la etapa de la enfermedad, del grado del tumor, de los síntomas del paciente y de su salud en general. Se deben tener en cuenta tanto los beneficios como los efectos secundarios posibles de cada

opción, especialmente los efectos sobre la actividad sexual, sobre los aspectos urinarios y otras preocupaciones acerca de la calidad de vida.

El tratamiento de cáncer de próstata puede incluir: cirugía, radioterapia, terapia hormonal o espera vigilante. Se puede tener una combinación de tratamientos. Si el médico recomienda la espera vigilante, la salud del paciente será observada de cerca y él recibirá tratamiento sólo si se presentan síntomas o si empeoran.

2.6.1 Cirugía

La cirugía es un tratamiento común para cáncer de próstata en etapa inicial. Es un tipo de terapia local. (Afecta las células sólo en el área tratada).

El médico puede extirpar toda la próstata o solo una parte. En algunos casos, el médico puede usar una técnica conocida como cirugía conservadora de nervios. Este tipo de cirugía puede salvar los nervios que controlan la erección. Sin embargo, es posible que los hombres que tienen tumores grandes o tumores que están muy cerca de los nervios no puedan tener esta cirugía.

2.6.2 Radioterapia

La radioterapia usa rayos de alta energía para destruir las células cancerosas. Es un tipo de terapia local. En cáncer de próstata en etapa inicial, la radioterapia puede ser el tratamiento primario (en vez de cirugía). También puede ser usada después de cirugía para destruir cualquier célula cancerosa que quede en el área. En etapas avanzadas, la radioterapia puede ayudar a aliviar el dolor.

Los médicos usan dos tipos de radioterapia para tratar el cáncer de próstata:

Radiación externa: La radiación procede de una máquina. Los pacientes van al hospital o clínica para su tratamiento, generalmente 5 días a la semana durante varias semanas. Algunos hombres con cáncer de próstata reciben radioterapia de conformación tridimensional. Este tipo de radioterapia se apunta más de cerca al cáncer y conserva el tejido normal.

Radiación interna (radiación por implante o braquiterapia): La radiación procede de material radiactivo puesto en semillas, agujas o tubos delgados de plástico colocados directamente en el tejido. El paciente se queda en el hospital. Los implantes permanecen en el sitio generalmente por varios días. Se remueven antes de que el paciente regrese a casa.

Algunos hombres con cáncer de próstata reciben ambas clases de radioterapia.

2.6.3 Terapia hormonal

La terapia hormonal impide que las células cancerosas obtengan las hormonas masculinas que necesitan para crecer. Esto se llama terapia sistémica porque entra en el torrente de la sangre y puede afectar las células cancerosas en todo el cuerpo. La terapia sistémica se usa principalmente para tratar el cáncer que se ha diseminado. Algunas veces este tipo de terapia se usa para tratar de impedir que el cáncer regrese después de la cirugía o de tratamiento con radiación.

El cáncer de próstata que se ha diseminado a otras partes del cuerpo generalmente puede ser controlado con terapia hormonal por un período de tiempo, con frecuencia por varios años. Eventualmente, sin embargo, la mayoría de los cánceres de próstata pueden crecer con muy pocas hormonas masculinas o sin ellas. Cuando esto sucede, la terapia hormonal ya no es efectiva y el médico puede sugerir otras formas de tratamiento que están siendo estudiadas

2.6.4 Espera vigilante

Los pacientes algunas veces escogen la espera vigilante cuando los riesgos y los efectos secundarios posibles de la cirugía, radioterapia o de la terapia hormonal pueden tener más peso que los beneficios posibles.

A los hombres que escogen la espera vigilante se ofrece tratamiento cuando se presentan síntomas o cuando los síntomas empeoran. La espera vigilante se puede aconsejar para hombres con más edad o para hombres que tienen cáncer de próstata y otros problemas médicos graves. También, la espera vigilante se puede sugerir para algunos hombres con cáncer de próstata que se encuentra en un estadio o etapa inicial y que parece estar creciendo lentamente.

3 BRAQUITERAPIA

3.1 Definición y objetivos

La palabra braquiterpia procede del griego brachys que significa "corto". La braquiterapia es el tratamiento radioterápico, que consiste en la colocación de fuentes radiactivas encapsuladas dentro o en la proximidad de un tumor (distancia "corta" entre el volumen a tratar y la fuente radiactiva).

El término braquiterapia, (también llamada curiterapia o radioterapia interna), fue acuñado por Forsell en 1931, para diferenciarla de la radioterapia externa, donde la fuente radiactiva está lejos del volumen a tratar.

Se usa principalmente en tumores ginecológicos, en los que la paciente es hospitalizada y a la que colocan unos dispositivos que contienen sustancias radiactivas en el interior de su cuerpo y se dejan por un determinado tiempo. Las fuentes encapsuladas son isótopos radiactivos en forma de tubos (Cesio 137), alambres (Iridio 192) o semillas (yodo 131, oro 198, paladio 109) que se colocan dentro del tumor o de cavidades de órganos. La mayoría de veces la inserción de estas sustancias se realiza en el quirófano y requiere anestesia local o general para no provocar dolor.

El objetivo de la braquiterapia es administrar dosis altas de radiación al tumor, con dosis escasas a los tejidos normales de alrededor. Tiene el inconveniente de que sólo se puede emplear en el tratamiento de tumores pequeños y que no irradia áreas linfáticas. Tiene la ventaja frente a la radioterapia externa, que los implantes radiactivos ofrecen la posibilidad de administrar una dosis alta al tumor, en un tiempo reducido, y a un volumen bien delimitado alrededor del mismo, con exposición reducida de las estructuras o tejidos adyacentes normales. Aunque la braquiterapia puede ser el único tratamiento radioterápico que reciba el lecho tumoral (braquiterapia exclusiva), como por ejemplo en la braquiterapia de próstata, la mayoría de las veces se combina con la radioterapia externa.

3.2 Tipos de Braquiterapia

Una vez definido que un paciente será intervenido a través de braquiterapia próstatica, resta definir que tipo de braquiterapia es la apropiada para cada paciente. Aquí es donde debe definirse que cantidad de radiación debe utilizarse, que elemento se utilizará, cuanto tiempo se mantendrá el implante, etc. En la siguiente tabla se muestran los distintos tipos de braquiterapia según diferentes criterios de clasificación y su descripción.

Tipo	Descripción
Según la localización de la braquiterapia	• Braquiterapia endocavitaria o endoluminal: En este tipo se introducen unos dispositivos que tienen la forma de la cavidad del órgano a tratar, como son cilindros vaginales, colpostatos, sondas endouterinas, endoesofágicas, endobronquiales, etc.
	• Braquiterapia intersticial: En este tipo se introducen unas agujas huecas a través del área tumoral. Estas agujas pueden hacer de guía para la introducción posterior de tubos huecos de plástico por la que circulará la fuente radiactiva.
	Braquiterapia de contacto superficial: En este tipo los tubos están en contacto, generalmente con la piel, adoptando su forma y sujetos con moldes de cera. Se ha utilizado para epiteliomas de nariz y resto de la cara.
Según el sistema de carga del implante radiactivo	• Braquiterapia de carga inmediata: Utiliza un sistema que se carga al finalizar la colocación de los aplicadores en el tumor, por ejemplo en la braquiterapia de baja tasa de tumores de orofaringe, en los que en el quirófano es necesario sustituir los vectores introducidos en el tumor (lengua, amígdala) bajo anestesia general, por la fuente radiactiva (horquilla o hilos de iridio) es decir, es el implante de la fuente radiactiva en el mismo quirófano, que se realiza cada vez menos.
	• Braquiterapia de carga diferida: Utiliza durante el proceso de implantación intersticial o endocavitaria, vectores o portadores huecos que posteriormente y comprobada por medio de rayos X su adecuada colocación con fuentes ficticias o fantomas, la carga en la misma habitación en donde permanecerá el paciente durante el tratamiento, mediante control remoto. A partir de la década de los 90 su utilización es casi universal y con su empleo se ha reducido drásticamente el riesgo de exposición del personal laboralmente expuesto a las radiaciones
Los equipos de carga diferida automáticos, son sistemas que robóticamente transportan la fuente radiactiva desde un contenedor blindado hasta los aplicadores colocados en el paciente y retornan la fuente automáticamente cuando el tratamiento ha finalizado .Los sistemas de carga diferida de control remoto tienen la ventaja de permitir una mejor dosimetría por emplear una fuente	

Tipo	Descripción
radiactiva móvil, consigue una mejor administración de la dosis ya que se realiza en un corto periodo de tiempo (minutos) y con escasa movilidad de los órganos durante este tiempo.	
Según la tasa de dosis de radiación que se administra en la braquiterapia	• Braquiterapia de Baja Tasa: (Tasa de dosis menores a los 2 Gy/h) En este tipo de braquiterapia, la radiación liberada por unidad de tiempo de la sustancia radiactiva es baja, por lo que el paciente debía permanecer durante varias horas, generalmente dos o tres días aislado en una habitación, para poder recibir una dosis determinada al tumor. Además el personal sanitario se irradiaba al introducir los hilos del material radiactivo dentro de los tubos insertados en el paciente.
	• Braquiterapia de Media Tasa: (Tasa de dosis entre 2 Gy/h y 12 Gy/min)En este tipo de braquiterapia, la radiación liberada por unidad de tiempo de la sustancia radiactiva es intermedia, por lo que el paciente debía permanecer durante varios minutos a varias horas, generalmente entre treinta minutos a dos días aislado en una habitación, para poder recibir una dosis determinada al tumor. Además el personal sanitario se irradiaba al introducir los hilos del material radiactivo dentro de los tubos insertados en el paciente.
	• Braquiterapia de Alta Tasa de dosis: (Tasa de dosis mayores a los 12 Gy/h)En este tipo de braquiterapia se utiliza una sustancia radiactiva que libera mucha radiación en poco tiempo, generalmente Iridio 192 de alta tasa, que tiene muy como volumen (1x4 mm), por lo que se puede introducir por tubos muy finos automáticamente y puede ser controlado desde un ordenador desde otra habitación. Cada sesión de tratamiento dura muy pocos minutos, generalmente menos de 10 minutos, y el personal sanitario no se irradia durante la introducción de los isótopos en los tubos. Las unidades de alta tasa de dosis constan fundamentalmente de una sola fuente muy activa (de 10 curios de actividad). El tratamiento se programa de forma que la fuente radiactiva permanezca tiempos determinados en lugares preestablecidos dentro de los aplicadores, obteniendo al final del tiempo de irradiación, la distribución de dosis deseada.

Tipo	Descripción
Según la temporalidad del implante radiactivo	• Braquiterapia con implante temporal: La fuente radiactiva que se inserta en el tumor se extrae una vez que finaliza el tiempo de radiación que libera la dosis pautada.
	• Braquiterapia con implante permanente: Las fuentes radiactivas encapsuladas permanecen indefinidamente en el cuerpo del paciente y son identificadas en una radiografía simple, por ejemplo en la braquiterapia del cáncer de próstata con I131.

Tabla 1. Tipos de braquiterapia

3.3 Dosis

La dosis prescrita, definida como dosis terapéutica, es referida a una línea de isodosis que aúna los criterios de englobar la totalidad del volumen tumoral prostático, evitando inhomogeneidades inaceptables en la distribución de la dosis dentro del mismo, y considerando la máxima dosis tolerable que pueden y deben recibir las estructuras limitantes radiosensibles de interés (uretra, recto y vejiga). Dicha etapa del procedimiento denominada de planificación y dosimetría, es realizada utilizando un complejo y sofisticado sistema informático en el que la colaboración de un radiofísico especializado además del Radioterapeuta y el Urólogo son claves para la correcta colocación del implante y el éxito definitivo del tratamiento.

Se utiliza el isótopo radiactivo Yodo -125 (125I) en forma de semillas recubiertas por una cápsula de Titanio, cuyas características principales se resumen en la tabla siguiente:

	T1/2 Días	Energía Kev	Actividad Semilla mCi	Tamaño Semilla mm	Tasa de dosis periférica cGy/hora
I-125	59.4	27.4	0.3 - 0.6	4.5	8

Tabla 2. Isótopo I-125

Se recomienda utilizar 125I en tumores con células bien o moderadamente diferenciadas, de crecimiento lento y el 103Pd en tumores mal diferenciados, de crecimiento más rápido. Las características del isótopo Yodo 125 posibilitan que las medidas de radio protección para el paciente y sus acompañantes habituales sean mínimas y de ningún modo incapacitantes. Las excretas, ropas e utensilios no están contaminados, no son radiactivos

3.4 Beneficios de la Braquiterapia

La aceptación en aumento de la Braquiterapia se debe a la experiencia desarrollada y los buenos resultados obtenidos durante casi 15 años, así como al desarrollo de la ecografía transrectal y la incorporación de nuevos y sofisticados sistemas informáticos y los programas de dosimetría que mejoran la técnica de implantación y la distribución de la dosis de irradiación.

Además, únicamente precisa una hospitalización de 24 horas, observando una rápida recuperación del paciente y una mínima morbilidad inmediata y tardía por la implantación de las semillas. No hay que olvidar que el tratamiento más utilizado hasta ahora, la cirugía radical, tiene mayor riesgo de complicaciones (impotencia e incontinencia) y la hospitalización es más prolongada.

La desventaja es la de no obtener información histológica precisa que corrobore la situación exacta de la enfermedad (estudio anatomopatológico) que en ocasiones no coincide con el estadío previamente establecido. Esta información que se consigue con la cirugía (Prostatectomia Radical con Linfadenectomía Regional) permite establecer un pronóstico más exacto.

4 MARCO TEORICO

El Aprendizaje Automático (Machine Learning) se enfrenta con el desafío de construir programas computacionales que automáticamente mejoren con la experiencia. Estos son sistemas capaces de adquirir conocimientos de alto nivel y/o estrategias para la resolución de problemas mediante ejemplos, en forma análoga a la mente humana. A partir de ejemplos provistos por un tutor o instructor y de los conocimientos de base o conocimientos previos, el sistema de aprendizaje crea descripciones generales de conceptos.

La Minería de Datos (Data Mining) busca generar información similar a la que podría generar un experto humano, que además satisfaga el principio de comprensibilidad. El objetivo de éste es descubrir conocimientos interesantes; como patrones, asociaciones, cambios, anomalías y estructuras significativas a partir de grandes cantidades de datos almacenados en bases de datos, data warehouses, o cualquier otro medio de almacenamiento de información. A continuación se describen algunos algoritmos:

- Caracterización: También conocidos como algoritmos de clustering, la caracterización consiste en agrupar un conjunto de datos sin tener clases previamente definidas. Estos algoritmos operan basándose en la similitud de los valores de los atributos de los distintos datos. Este tipo de aprendizaje se realiza en forma no supervisada ya que no se saben de antemano las clases del set de datos de entrenamiento. La caracterización identifica regiones densamente pobladas, de acuerdo a alguna medida de distancia, en un gran conjunto de datos multidimensional. El análisis de clases se basa en maximizar la similitud de las instancias en cada cluster y minimizar la similitud entre clusters. Se utiliza para reconocimiento de patrones, análisis de datos, procesamiento de imágenes entre otras. Como función de la minería de datos, el análisis de clases puede ser utilizado de forma independiente para obtener la distribución del set de datos, para caracterizar cada clase y dividir grupos para su análisis. Alternativamente, puede servir para el preprocesamiento de datos, antes de utilizar otros algoritmos.
- Clasificación: También conocidos como algoritmos de inducción. Algunos utilizan ejemplos como entradas para aplicar sobre ellos un proceso inductivo y así presentar la generalización de los mismos como resultado de salida. Existen dos tipo de ejemplos, los positivos y los negativos. Los primeros fuerzan la generalización, mientras que los segundos previenen que ésta sea excesiva. Se pretende que el conocimiento adquirido cubra todos los ejemplos positivos y ningún ejemplo negativo. Los ejemplos deben ser representativos de los conceptos que se está tratando de enseñar. Además la distribución de las clases en el conjunto de ejemplos de entrenamiento, a partir de los que el sistema aprende, debe ser similar a la distribución existente en los datos sobre los cuales

se aplicará el sistema. Otros utilizan el descubrimiento y observación, para que el sistema forme criterios de clasificación.En este caso se aplica aprendizaje no supervisado, permitiendo que el sistema clasifique la información de entrada para formar conceptos. En estos casos el sistema puede interactuar con el entorno para realizar cambios en el mismo y luego observar los resultados.

- Redes Bayesianas: Una red bayesiana es un grafo acíclico dirigido en el que cada nodo representa una variable y cada arco una dependencia probabilística, en la cual se especifica la probabilidad condicional de cada variable dados sus padres, la variable a la que apunta el arco es dependiente (causa-efecto) de la que está en el origen de éste. La topología o estructura de la red nos da información sobre las dependencias probabilísticas entre las variables pero también sobre las independencias condicionales de una variable (o conjunto de variables) dada otra u otras variables, dichas independencias, simplifican la representación del conocimiento (menos parámetros) y el razonamiento (propagación de las probabilidades). El obtener una red Bayesiana a partir de datos, es un proceso de aprendizaje que se divide en dos etapas: el aprendizaje estructural y el aprendizaje paramétrico [Pearl, 1988; Hernández O.J. et al, 2004]. La primera de ellas, consiste en obtener la estructura de la red bayesiana, es decir, las relaciones de dependencia e independencia entre las variables involucradas. La segunda etapa, tiene como finalidad obtener las probabilidades a priori y condicionales requeridas a partir de una estructura dada.

La Minería de Datos es un caso especial del Aprendizaje Automático, utiliza sus métodos para encontrar patrones, con la diferencia que el escenario observado es una base de datos. En un esquema de Aprendizaje Automático, el mundo real es el entorno sobre el cual se realiza el aprendizaje, estos se traducen en un conjunto finito de observaciones u objetos que son codificados en algún formato legible. El conjunto de ejemplos constituye la información necesaria para el entrenamiento del sistema.

Los sistemas de aprendizaje se clasifican en métodos de caja negra y métodos orientados al conocimiento. Los primeros desarrollan su propia representación de conceptos, realizando cálculos numéricos de coeficientes, distancia, vectores; generalmente no comprensibles por los usuarios. Los segundos tratan de crear estructuras simbólicas de conocimiento que si sean comprensibles para el usuario. Las Redes Neuronales pertenecen al primer grupo y el Aprendizaje Automático al segundo. [García Martínez et al, 2003]

4.1 Fundamentos

4.1.1 Algoritmos de Caracterización

En 1982, Teuvo Kohonen, presentó un modelo de red neuronal, basado en el funcionamiento de neuronas biológicas. La red neuronal diseñada posee la capacidad de formar mapas de características. El objetivo de Kohonen era demostrar que un estímulo externo por sí solo, suponiendo una estructura propia y una descripción funcional del comportamiento de la red, era suficiente para forzar la formación de los mapas.
El modelo tiene dos variantes, LVQ (Learning Vector Quantization) y TPM (Topología Preserving Map) o SOM (Self Organizing Map). Ambas se basan en el principio de formación de mapas topológicos para establecer características comunes entre las informaciones (vectores) de entrada a la red, aunque difieren en las dimensiones de éstos, siendo de una sola dimensión en el caso de LVQ y bidimensional e incluso tridimensional en la red SOM o TPM.
Como se puede ver en la figura 2, el modelo presenta dos capas con N neuronas de entrada y M de salida. Cada una de las N neuronas de entrada se conecta a las M de salida a través de conexiones hacia delante (feedforward). Entre las neuronas de la capa de salida, existen conexiones laterales de inhibición (peso negativo) implícitas, a pesar de no estar conectadas, cada una de estas neuronas va a tener cierta influencia sobre sus vecinas. El valor que se asigne a los pesos de las conexiones feedforward entre las capas de entrada y salida (wij) durante el proceso de aprendizaje de la red va a depender precisamente de esta interacción lateral.

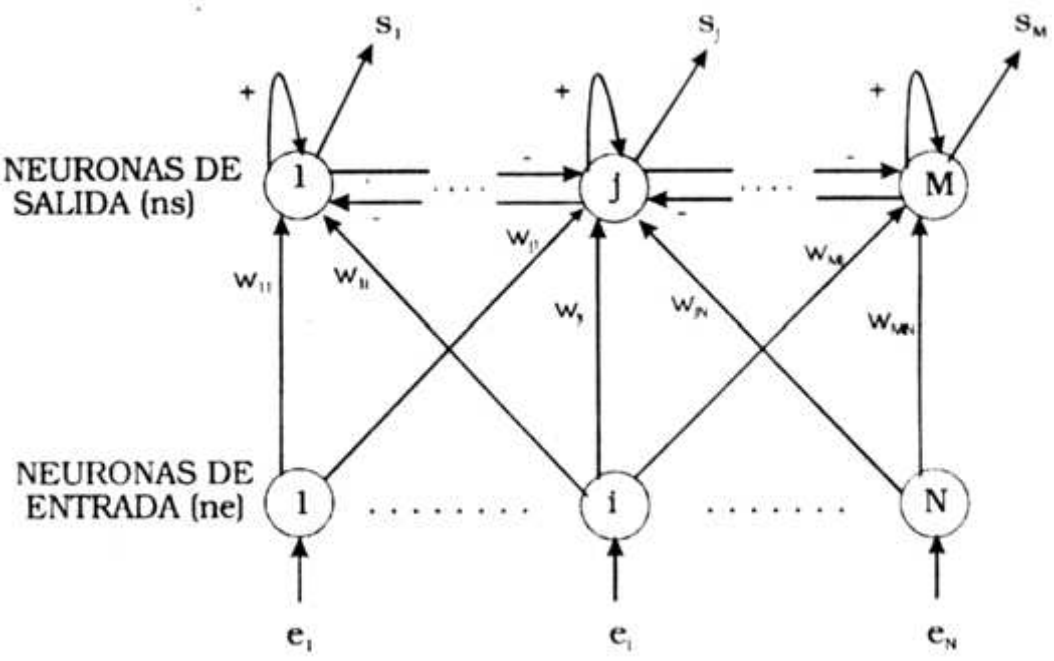


Figura 2. Modelo de Kohonen.

La influencia que cada neurona ejerce sobre las demás es función de la distancia entre ellas, siendo muy pequeñas cuando están muy alejadas. Es frecuente que dicha influencia tenga la forma de un sombrero mejicano, como se puede observar en la figura 3. Se han descubierto conexiones de este tipo entre las neuronas del sistema nervioso central de los animales. [Redes Competitivas, 2000]

Figura 3. Influencia entre neuronas

4.1.2 Algoritmos de Clasificación

El C4.5 forma parte de la familia de los TDIDT (Top Down Induction Trees), junto con su antecesor el ID3. Pertenece a los métodos inductivos del Aprendizaje Automático que aprenden a partir de ejemplos preclasificados. Se utilizan en Minería de datos para modelar las clasificaciones en los datos mediante árboles de decisión. Tanto el ID3 como el C4.5 fueron propuestos por Quinlan, el primero en la década de los ochenta y el segundo en 1993.

El C4.5 es una extensión del ID3, que sólo trabaja con valores discretos en los atributos. El C4.5, en cambio, permite trabajar con valores continuos, separando los posibles resultados en dos ramas: una para aquellos Ai<=N y otra para Ai>N. Se genera un árbol de decisión a partir de los datos mediante particiones realizadas recursivamente, aplicando la estrategia de profundidad-primero (depth-first). El algoritmo considera todas las pruebas posibles que pueden dividir el conjunto de datos y selecciona la prueba que resulta en la mayor ganancia de información. Para cada atributo discreto, se considera una prueba con n resultados, siendo n el número de valores posibles que puede tomar el atributo. Para cada atributo continuo, se realiza una prueba binaria sobre cada uno de los valores que toma el atributo en los datos.

Estos algoritmos han tenido gran impacto en la Minería de Datos. Forman parte del grupo de sistemas de aprendizaje supervisado. Han tenido muy buena performance en aplicaciones de dominio médico, artificiales y el análisis de juegos de ajedrez. Posee un nivel alto de precisión en la clasificación, pero no hace uso del conocimiento del dominio. [García Martínez et al, 2003]

4.1.3 Redes de Bayes

Otra herramienta a disposición pero de naturaleza estocástica son las redes de Bayes. Una red bayesiana es un grafo acíclico dirigido en el que cada nodo representa una variable y cada arco una dependencia probabilística, en la cual se especifica la probabilidad condicional de cada variable dados sus padres, la variable a la que apunta el arco es dependiente (causa-efecto) de la que está en el origen de éste. La topología o estructura de la red nos da información sobre las dependencias probabilísticas entre las variables pero también sobre las independencias condicionales de una variable (o conjunto de variables) dada otra u otras variables, independencias, simplifican la representación del conocimiento (menos parámetros) y el razonamiento (propagación de las probabilidades). Estas redes son utilizadas en diversas áreas aplicación como por ejemplo en medicina [Beinlinch et al., 1989; Hernández O.J. et al, 2004], ciencia [Breese & Blake, 1995; Hernández O.J. et al, 2004], y economía [Hernández O.J. et al, 2004]. Las mismas proveen una forma compacta de representar el conocimiento y métodos flexibles de razonamiento - basados en las teorías probabilísticas - capaces de predecir el valor de variables no observadas y explicar las observadas. Entre las características que poseen las redes bayesianas, se puede destacar que permiten aprender sobre relaciones de dependencia y causalidad, permiten combinar conocimiento con datos [Heckerman et al., 1995; Díaz & Corchado, 1999; Hernández O.J. et al, 2004] y pueden manejar bases de datos incompletas [Heckerman, 1995; Heckerman & Chickering, 1996; Ramoni & Sebastiani, 1996; Hernández O.J. et al, 2004].

4.2 Aplicación

4.2.1 Aplicación de Algoritmos de Caracterización

4.2.1.1 Funcionamiento de una red de Kohonen

Cuando se presenta a la entrada una información cada una de las M neuronas de la capa de salida la recibe a través de las conexiones feedforward con pesos wij.

$$E_k = \left(e_1^{(k)}, e_2^{(k)} .... e_N^{(k)}\right) \qquad (1)$$

También estas neuronas reciben las entradas producto de las interacciones laterales con el resto de las neuronas de salida y cuya influencia dependerá de las distancia a la que se encuentren. Así, la salida generada por una neurona de salida j ante una vector de entrada E_k como el que se puede observar en (1), será:

$$s_j(t+1) = f\left(\sum_{i=1}^{N} w_{ij} e_i^{(k)} + \sum_{p=1}^{M} Int_{pj} s_p(t)\right) \qquad (2)$$

Int_{pj} es una función del tipo sombrero mejicano que representa la influencia lateral de la neurona p sobre la neurona j. La función de activación de las neuronas de salida f , véase ecuación 2 será del tipo continuo, lineal o sigmoidal, ya que esta red trabaja con valores reales.

SOM es una red de tipo competitivo, ya que al presentarse una entrada E_k, la red evoluciona hasta alcanzar un estado estable en el que solo hay una neurona activada, la ganadora. La formulación matemática del funcionamiento de esta red puede simplificarse así:

$$s_j = 1 \rightarrow MIN\left|E_k - W_j\right| = MIN\left(\sqrt{\sum_{i=1}^{N}\left(e_i^{(k)} - w_{ij}\right)^2}\right) \quad \text{y} \quad S_j = 0 \rightarrow RESTO \quad (3)$$

Donde /*E −W/,* obsérvese la ecuación (3) es una medida de la diferencia entre el vector de entrada y el vector de pesos de las conexiones que llegan a la neurona j desde la entrada. Es en estos pesos donde se registran los datos almacenados por la red durante el aprendizaje. Durante el funcionamiento, lo que se pretende es encontrar el dato aprendido más parecido al de entrada para averiguar qué neurona se activará y en que zona del espacio bidimensional de salida se encuentra.

Esta red realiza una tarea de clasificación ya que la neurona de salida activada ante una entrada representa la clase a la que pertenece dicha información; ante otra entrada parecida se activa la misma neurona o una cercana a la anterior, garantizando que las neuronas topológicamente cercanas sean sensibles a entradas físicamente similares. Por esto, la red es muy útil para establecer relaciones antes desconocidas entre conjuntos de datos.

4.2.1.2 Aprendizaje

El aprendizaje es de tipo OFF LINE, por lo que se distingue una etapa de aprendizaje y otra de funcionamiento. En la primera se fijan los pesos de las conexiones feedforward entra las capas de entrada y salida. Emplea un aprendizaje no supervisado de tipo competitivo. Las neuronas de la capa de salida compiten por activarse y sólo una de ellas permanece activa ante una entrada determinada. Los pesos de las conexiones se ajustan en función de la neurona que haya resultado vencedora.

En el entrenamiento, se presenta a la red un conjunto de informaciones de entrada para que ésta establezca las diferentes clases que servirán durante la fase de funcionamiento para realizar la clasificación de los nuevos datos que se presenten. Los valores finales de los pesos de las conexiones feedforward que llegan a cada neurona de salida se corresponderán con los valores de los componentes del vector de aprendizaje que consigue activar la neurona correspondiente. Si existiesen más vectores de entrenamiento que neuronas de salida, más de un vector deberá asociarse a la misma clase. En tal caso, los pesos se obtienen como un promedio de dichos patrones.

Durante el entrenamiento habrá que ingresar varias veces todo el juego de entrenamiento para refinar el mapa topológico de salida consiguiendo que la red pueda realizar una clasificación más selectiva.

El algoritmo de aprendizaje es el siguiente:

1. En primer lugar, se inicializan los pesos (wij) con valores aleatorios pequeños y se fija la zona inicial de vecindad entre las neuronas de salida.
2. A continuación se presenta a la red una información de entrada (la que debe aprender) en forma de vector $E_k = (e_1^{(k)} \dots e_n^{(k)})$ cuyas componentes e_{ik} serán valores continuos.
3. Se determina la neurona vencedora a la salida. Esta será aquella j cuyo vector de pesos Wj sea el más parecido a la información de entrada E_k. Para ello se calculan las distancias entre ambos vectores, una para cada neurona de salida. Suele utilizarse la distancia euclídea o bien la expresión (4), similar pero sin la raíz:

$$d_j = \sum_{i=1}^{N} \left(e_i^{(k)} - w_{ij}\right)^2 \quad 1 \le j \le M \qquad (4)$$

   donde: $e_i^{(k)}$: Componente i-ésimo del vector k-ésimo de entrada. wij: Peso de la conexión entra las neuronas i (de entrada) y j (de salida).
4. Una vez localizada la neurona vencedora j*, se actualizan los pesos de las conexiones feedforward que llegan a dicha neurona y a sus vecinas, como se puede ver en la ecuación (5). Con esto se consigue asociar la información de entrada con cierta zona de la capa de salida.

$$w_{ij}(t+1) = w_{ij}(t) + \alpha(t)\left[e_i^{(k)} - w_{ij}(t)\right] \quad j \in Zona_j(t) \qquad (5)$$

   $Zonaj^*(t)$ es la zona de vecindad de la neurona vencedora j*. El tamaño de esta zona se puede reducir en cada iteración del entrenamiento aunque en la práctica es habitual mantener esa zona fija.

   El término α(t) es el coeficiente de aprendizaje y toma valores entre 0 y 1. Este parámetro decrece con cada iteración. De esta forma, cuando se ha presentado todo el juego de datos un gran número de veces, α tiende a cero y las variaciones de pesos son insignificantes. α(t) suele tener alguna de las expresiones en (6):

$$\alpha(t) = \frac{1}{t} \qquad \alpha(t) = \alpha_1\left(1 - \frac{t}{\alpha_2}\right) \qquad (6)$$

   El proceso se repite un mínimo de 500 veces ($t \ge 500$

4.2.2 Aplicación de Algoritmos de Inducción

4.2.2.1 Construcción de árboles de decisión

Los árboles TDIDT, (ID3 y C4.5), se construyen con el método de Hunt. Se parte de un conjunto *T* de datos de entrenamiento. Dadas las clases *{C1, C2,. . ., Ck}*, existen tres posibilidades:

1. *T* contiene uno o más casos, todos pertenecientes a un única clase *Cj*: El árbol de decisión para *T* es una hoja identificando la clase *Cj* .
2. *T* no contiene ningún caso: El árbol de decisión es una hoja, pero la clase asociada debe ser determinada por información que no pertenece a *T*. Por ejemplo, una hoja puede escogerse de acuerdo a conocimientos de base del dominio, como ser la clase mayoritaria.
3. *T* contiene casos pertenecientes a varias clases: Se refina *T* en subconjuntos de casos que tiendan hacia una colección de casos de una única clase. Se elige una prueba basada en un único atributo, que tiene uno o más resultados, mutuamente excluyentes *{O1, O2,. . ., On}*. *T* se particiona en los subconjuntos *T1, T2,. . ., Tn* donde *Ti* contiene todos los casos de *T* que tienen el resultado *Oi* para la prueba elegida. El árbol de decisión para *T* consiste en un nodo de decisión identificando la prueba, con una rama para cada resultado posible. El mecanismo de construcción del árbol se aplica recursivamente a cada subconjunto de datos de entrenamientos, para que la *i*-ésima rama lleve al árbol de decisión construido por el subconjunto *Ti* de datos de entrenamiento.

A continuación se presenta el algoritmo del método ID3 (antecesor del C4.5) para la construcción de árboles de decisión en función de un conjunto de datos previamente clasificados.

Función ID3: (R: conjunto de atributos no clasificadores, C: atributo clasificador, S: de entrenamiento) devuelve un árbol de decisión;

Comienzo

Si S está vacío,

devolver un único nodo con Valor Falla;

Si todos los registros de S tienen el mismo valor para el atributo clasificador,

Devolver un único nodo con dicho valor;

Si R está vacío, entonces

devolver un único nodo con el valor más frecuente del atributo clasificador en los registros de S [Nota: habrá errores, es decir, registros que no estarán bien clasificados en este caso];

Si R no está vacío, entonces

D ← atributo con mayor Ganancia (D, S) entre los atributos de R;

Sean {dj| j=1,2, .., m} los valores del atributo D;

Sean {Sj| j=1,2, .., m} los subconjuntos de S correspondientes a los valores de dj respectivamente;

Devolver un árbol con la raíz nombrada como D y con los arcos nombrados d1, d2, .., dm que van respectivamente a los árboles

ID3(R-{D}, C, S1), ID3(R-{D}, C, S2), .., ID3(R-{D}, C, Sm);

Fin

4.2.2.2 Cálculo de Ganancia de Información

Cuando los casos en un conjunto *T* contiene ejemplos pertenecientes a distintas clases, se realiza una prueba sobre los distintos atributos y se realiza una partición según el "mejor" atributo. Para encontrar el "mejor" atributo, se utiliza la teoría de la información, que sostiene que la información se maximiza cuando la entropía se minimiza. La entropía determina la azarosidad o desestructuración de un conjunto.

Supongamos que tenemos ejemplos positivos y negativos. En este contexto la entropía de un subconjunto Si, H(Si), puede calcularse según (7) como:

$$H(S_i) = -p_i^+ \log p_i^+ - p_i \log p_i^- \qquad (7)$$

Donde p_i^+ es la probabilidad de que un ejemplo tomado al azar de Si sea positivo.

Esta probabilidad puede calcularse como:

$$p_i^+ = \frac{n_i^+}{n_i^+ + n_i^-} \qquad (8)$$

Si el atributo *at*, veasé la expresión (9), divide el conjunto *S* en los subconjuntos *Si*, *i* = 1,2,….., n, entonces, la entropía total del sistema de subconjuntos, *H(S, at)* se expresa mediante (9):

$$H(S, at) = \sum_{i=1}^{n} P(S_i) H(S\ ) \qquad (9)$$

Donde *H(Si)* es la entropía del subconjunto *Si* y *P*(*Si*) es la probabilidad de que un ejemplo pertenezca a *Si*. Puede calcularse en (10), utilizando los tamaños relativos de los subconjuntos, como:

$$P(S_i) = \frac{|S_i|}{|S|} \qquad (10)$$

La ganancia en información puede calcularse como la disminución en entropía. Es decir:

$$I(S, at) = H(S) - H(S, at) \qquad (11)$$

Donde *H*(*S*), veasé (11), es el valor de la entropía a priori, antes de realizar la subdivisión, y *H*(*S*, *at*) es el valor de la entropía del sistema de subconjuntos generados por la partición según *at.*

El uso de la entropía para evaluar el mejor atributo no es el único método existente o utilizado en *Aprendizaje Automático*. Sin embargo, es el utilizado por Quinlan al desarrollar el ID3 y su sucesor el C4.5.

4.2.2.3 Entropía

Es la cantidad de información que se espera observar cuando un evento ocurre según una distribución de probabilidades. Mide la incertidumbre dada una distribución de probabilidades. Si tomamos un conjunto con elementos positivos y negativos, la entropía variará entre 0 y 1.

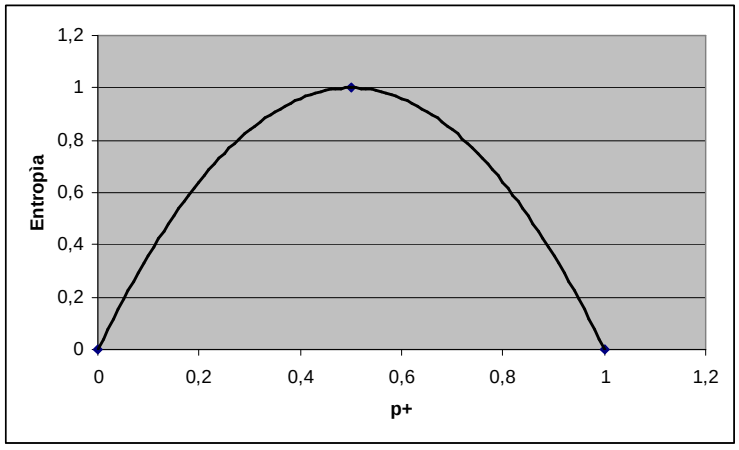


Figura 4. Entropía de un sistema

Como se observa en la figura 4, la entropía es 0 si todos los ejemplos pertenecen a la misma clase, y 1 cuando hay igual número de ejemplos positivos y negativos en el conjunto de datos. Cuanto tenemos *c* clases posibles, el valor máximo de la entropía será *log2c.*

La probabilidad de que un ejemplo tomado al azar pertenezca a la clase *i* y se calcule en base a la frecuencia de los datos de dicha clase en los datos de entrenamiento, se representa en la fórmula (12) según:

$$H(S_i) = \sum_{i=1}^{n} - p_i \log p_i \qquad (12)$$

4.2.2.4 Proporción de ganancia

Favorece a los atributos que tienen muchos valores frente a los que tienen pocos valores. Si se tiene un conjunto de registros con fecha y se particiona según el campo fecha, se obtendrá un árbol perfecto, pero que no servirá para clasificar casos futuros, dato el gran tamaño del mismo. Para resolver esta situación se divide a los datos de

entrenamiento en conjuntos pequeños, con lo cual tendrá una alta ganancia de información.

Una alternativa para dividir a los datos es la *ganancia de información*. Esta medida penaliza a los atributos como fecha al incorporar el término de información de la división

$$I\_división(X) = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} \times \log_2\left(\frac{|T_i|}{|T|}\right) \qquad (13)$$

La información de la división, como queda expresada en (13), no es otra cosa que la entropía del conjunto con respecto al atributo *i*. Se define en (14), entonces, a la proporción de ganancia como:

$$proporcion\_de\_ganancia(x) = \frac{I(T,X)}{I\_division(X)} \qquad (14)$$

La información de la división penalizará a aquellos atributos con muchos valores uniformemente distribuidos. Si tenemos *n* datos separados perfectamente por un atributo, la información de la división para ese caso será log2*n*. En cambio, un atributo que divide a los ejemplos en dos mitades, tendrá una información de la división de 1.
Cuando la información de la división es cercana a cero, pueden aplicarse varias heurísticas. Puede utilizarse la ganancia como medida y utilizar la proporción de ganancia sólo para los atributos que estén sobre el promedio.

4.2.2.5 Datos Numéricos

Cuando los árboles de decisión se generan con atributos discretos, la partición del conjunto según el valor de un atributo es simple. Por ejemplo, agrupamos todos los animales que tengan pico, siendo *tiene_pico* un atributo y sus posibles valores *si* y *no.*
Cuando los atributos son continuos, no es tan fácil. Por ejemplo, si queremos partir los días de un mes en función a la cantidad de lluvia caída, es casi imposible que encontremos dos días con exactamente la misma cantidad de precipitaciones caídas. Para ello se aplica *binarización.*
Este método consiste en formar dos rangos de valores de acuerdo al valor de un atributo, que pueden tomarse como simbólicos. Por ejemplo, si en un día hubo 100ml de lluvia, pueden crearse los intervalos [0,100) y [100, +∝) y el cálculo de la entropía se realiza como si los dos intervalos fueran los dos valores simbólicos que puede tomar el atributo.

4.2.2.6 Poda de los árboles generados

Hay varias razones para podar los árboles generados por los métodos de TDIDT, la sobregeneralización, evaluación de atributos poco importantes o significativos; y el gran tamaño del árbol. Ejemplos con ruido, atributos no relevantes, deben podarse ya que sólo agregan niveles en el árbol y no contribuyen a la ganancia de información. Si el árbol es demasiado grande, se dificulta la interpretación, con lo cual hubiera sido lo mismo utilizar un método de caja negra.

Existen dos enfoques para podar árboles: la pre-poda *(preprunning)*, detiene el crecimiento del árbol cuando la ganancia de información producida al dividir un conjunto no supera un umbral determinado y la post-poda *(postprunning)*, se aplica sobre algunas ramas una vez que se ha terminado.

La pre-poda, no pierde tiempo en construir una estructura que luego será simplificada en el árbol final, busca la mejor manera de partir el subconjunto y evaluar la partición desde el punto de vista estadístico mediante la teoría de la ganancia de información, reducción de errores, etc. Si esta evaluación es menor que un límite predeterminado, la división se descarta y el árbol para el subconjunto es simplemente la hoja más apropiada. Tiene la desventaja de que no es fácil detener un particionamiento en el momento adecuado, un límite muy alto puede terminar con la partición antes de que los beneficios de particiones subsiguientes parezcan evidentes, mientras que un límite demasiado bajo resulta en una simplificación demasiado leve.

La post-poda, es utilizada por el ID3 y el C4.5. Una vez construido el árbol se procede a su simplificación según los criterios propios de cada uno de los algoritmos.

4.2.2.7 Principio de Longitud de Descripción Mínima

El fin de los sistemas de aprendizaje es aprender una “teoría” (árboles o reglas de decisión, por ejemplo) del dominio de los ejemplos, predictiva en el sentido de que es capaz de predecir la clase de nuevos instancias.
El Principio de Longitud de Descripción Mínima (MDL) sostiene que la mejor teoría es aquella que minimiza el tamaño y la cantidad de información necesaria para especificar las excepciones. El MDL provee una forma de medir la performance de los algoritmos basándose en los datos de entrenamiento únicamente. Supongamos que un sistema de aprendizaje genera una teoría *T*, basada en un conjunto de entrenamiento *E*, y requiere una cierta cantidad de bits *L[T]* para codificar la teoría. Dada la teoría, el conjunto de entrenamiento puede codificarse en una cantidad *L[E/T]* de bits. *L[E/T]* está dada por la función de ganancia de información sumando todos los miembros del conjunto de entrenamiento. La longitud de descripción total de la teoría es *L[E]+L[E/T]*. El principio MDL recomienda la teoría *T* que minimiza esta suma.

4.2.2.8 Funciones alternativas

La entropía no es la única alternativa para elegir el "mejor" atributo en la partición de datos al momento de construir un árbol de decisión según el método de divide y reinarás. Existen otras medidas alternativas. Una de ellas es la función de pérdida cuadrática: dada una instancia con *k* clases posibles a la que puede pertenecer, el sistema aprendiz devuelve un vector de probabilidades *p1, p2,.....,pk* de las clases de la instancia. Es decir, *pi* indica la probabilidad que tiene la instancia de pertenecer a la clase *i*. Con lo cual, los elementos del vector suman 1.

El resultado verdadero de la clasificación de la instancia será una de las clases posibles, entonces, si lo expresamos en un vector *a1, a2, ...,ak* donde *ai=1* si el elemento es de clase *i* y es 0 en caso contrario.

Entonces, utilizamos la función (15) para evaluar la pérdida de información según cada atributo:

$$\sum_{j=1}^{k}\left(p_j - a_j\right)^2 = 1 + 2p_i + \sum_{j=1}^{k} p_j^2 \quad (15)$$

4.2.2.9 Atributos Desconocidos

El método de Hunt, considera los resultados de todas las pruebas para todos los casos conocidos. Pero cuando los datos están incompletos, podemos tomar dos caminos posibles: descartar una proporción importante de los datos por incompletos y declarar algunos casos como inclasificables, o adaptar los algoritmos para poder trabajar con valores de atributos faltantes. La primera opción es inaceptable. Para la segunda opción, hay tres cuestiones importantes que deben ser tenidas en cuenta:

1. Selección de una prueba en la cual la partición del conjunto de entrenamiento se realiza en base a un criterio heurístico como ser la *ganancia* o la *proporción de ganancia*. Si dos pruebas distintas utilizan atributos con distinta cantidad de valores desconocidos, ¿cómo debe tenerse esto en cuenta al medir su importancia relativa?
2. Una vez que una prueba ha sido seleccionada, los casos de entrenamiento con valores desconocidos para los atributos relevantes no pueden ser asociados con una respuesta particular de la prueba, y, por lo tanto, no pueden asignarse a un subconjunto {*Ti*}.

- ¿Cómo deben tratarse estos casos durante la partición?
- Cuando el árbol de decisión se utiliza para clasificar un caso nuevo, ¿cómo debe proceder el sistema al encontrarse con un valor de atributo desconocido para el nodo de decisión que está tratando de evaluar?

Varios autores han tratado de resolver estos problemas, generalmente rellenando los valores desconocidos con los valores más frecuentes. En un estudio realizado por Quinlan, se comparan las soluciones más comunes a este problema; y se llega a la conclusión de que existen varios enfoques que son notablemente inferiores, pero no existe ningún enfoque que sea claramente superior.

4.2.2.10 Transformación a Reglas de Decisión

Los árboles de decisión demasiado grandes son difíciles de entender porque cada nodo debe ser interpretado dentro del contexto fijado por las ramas anteriores. Cada prueba tiene sentido si se analiza junto con los resultados de las pruebas previas. Cada prueba tiene un contexto único. Puede ser muy difícil comprender un árbol en el cual el contexto cambia demasiado seguido al recorrerlo. Además, la estructura puede hacer que un concepto en particular quede fragmentado, lo cual hace que el árbol sea aún más difícil de entender.

Existen dos maneras de solucionar estos problemas: definir nuevos atributos que estén relacionados con las tareas o cambiar de método de representación, por ejemplo, a reglas de decisión.

En cualquier árbol de decisión, las condiciones que deben satisfacerse cuando un caso se clasifica por una hoja pueden encontrarse analizando los resultados de las pruebas en el camino recorrido desde la raíz. Es más, si el camino fuese transformado directamente en una regla de producción, dicha regla podría ser expresada como una conjunción de todas las condiciones que deben ser satisfechas para llegar a la hoja. Consecuentemente, todos los antecedentes de las reglas generadas de esta manera serían mutuamente excluyentes y exhaustivos. Al hablar de reglas de decisión o de producción nos referimos a una estructura de la forma:

> **Si** *atributo1 = valor X* **y** *atributo2 = valor Y....* **y** *atributo n = valor Z* **Entonces** *clase K*

4.2.3 Aplicación de Redes de Bayes

Vamos y comenzar recordando el teorema de Bayes con una formulación de sucesos, para posteriormente formularlo en términos de variables aleatorias. Viéndose 1a necesidad de ir simplificando las premisas sobre las que se construye en aras de obtener paradigmas que puedan ser de aplicación para la resolución de problemas reales. Teorema (Bayes, 1764) Sean A y B dos sucesos aleatorios cuyas probabilidades se denotan por p(A) y p(B) respectivamente, verificándose que p(B) > 0. Supongamos conocidas las probabilidades a priori de los sucesos A y B, es decir, p(A) y p(B), así como la probabilidad condicionada del suceso B dado el suceso A, es decir p(B|A). La probabilidad a posteriori del suceso A conocido que se verifica el suceso B, es decir p(A|B), puede calcularse a partir de la fórmula 16:

$$\mathrm{P}(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(A)P(B|A)}{P(B)} = \frac{P(A)P(B|A)}{\sum_A P(A')P(B|A')} \qquad (16)$$

La formulación del teorema de Bayes puede efectuarse también para variables aleatorias, tanto unidimensionales como multidimensionales. El teorema de Bayes también puede ser expresado por medio de una notación que usa el número de componentes de cada una de las variables multidimensionales aleatorias X e Y, como expresa la ecuación 17.

$$P(Y = y|X = x) = P\left(Y_1 = y_1,.....Y_m = y_m \middle| X_1 = x_1,....,X_m = x_m\right)$$

$$= \frac{P(Y_1 = y_1,...,Y_m = y_m)P(X_1 = x_1,...,X_m = x_m|Y_1 = y_1,...,Y_m = y_m)}{\sum_{y'1....y'n} P(X_1 = x_1,...,X_m = x_m|Y_1 = y_1',...,Y_m = y_m')P(Y_1 = y_1',...,Y_m = y_m')} \qquad (17)$$

En primer Lugar vamos a considerar que los diagnósticos son excluyentes, es decir, que dos diagnósticos no pueden darse al unísono. Esto trae como consecuencia que en lugar de considerar el diagnóstico como una variable aleatoria m-dimensional, este caso puede verse como una única variable aleatoria unidimensional siguiendo una distribución polinomial con m valores posibles.

Vamos a denotar por $X_1,....X_n$ a las n variables predictoras. Supongamos que todas ellas sean binarias. Denotamos por C la variable objetivo, que suponemos puede tomar m posibles valores. La búsqueda del valor del objetivo mas probable a posteriori, c^*, una vez conocidos los valores de los atributos, $x = (x_1,....,x_n)$, puede plantearse como la búsqueda del estado de la variable C con mayor probabilidad a posteriori. Según expresa la ecuación 18:

$$c^* = \arg\max P(C = c|X_1 = x_1,..., X_m = x_m) \quad (18)$$

El cálculo de $P(C=c| X_1 = x_1,....., X_m = x_m)$ puede llevarse a cabo utilizando el teorema de Bayes, y ya que el objetivo es calcular el estado de C, c^*, con mayor probabilidad a posteriori, no es necesario calcular el denominador del teorema de Bayes. Es decir,

$p(C = c | X_1 = x_1,....., X_m = x_m) \propto p(C=c)\ p(X_1 = x_1,....., X_m = x_m | C = c)$.

Por tanto, en el paradigma en el que los distintos diagnósticos son excluyentes, y considerando que el número de posibles diagnósticos es m, y que cada variable predoctora X_i es dicotómica, tenemos que el número de parámetros a estimar es $(m - 1) + m (2_n - 1)$, de los cuales:

- m - 1 se refiere a las probabilidades a priori de las variable C;
- m (2n - 1) se relacionan con las probabilidades Condicionadas de cada posible combinación de las variables predictoras dado cada posible valor de la variable.

Esta situación hace que el número de parámetros a estimar sea elevado, de modo que hay que imponer suposiciones mas restrictivas.

Vamos finalmente a introducir el paradigma naïve Bayes: diagnósticos excluyentes y hallazgos condicionalmente independientes dado el diagnóstico. El paradigma naïve Bayes se basa en dos premisas establecidas sobre las variables predoctoras (hallazgos, síntomas) y la variaba a predecir (diagnósticos). Dichas premisas son:

1. Los diagnósticos son excluyentes, es decir, la variable C a predecir toma uno de sus m posibles valores: $c_1,.....,c_m$;
2. Los hallazgos son condicionalmente independientes dado el diagnóstico, es decir, que si uno conoce, el valor de 1a variable diagnóstico, el conocimiento del valor de cualquiera de los hallazgos es irrelevante para el resto de los hallazgos. Esta condición se expresa matemáticamente por medio de la fórmula 19:

$$P\left(X_1 = x_1,..., X_m = x_m \middle| C = c\right) = \prod_{i=1}^{n} P\left(X_i = x_i \middle| C = c\right) \qquad (19)$$

Por otra parte teniendo en cuenta la independencia condicional entre las variables predictoras dado el atributo de clase, se tiene según 20:

$$P\left(X_i = x_i \middle| X_{i+1} = x_{i+1},..., X_n = x_n, C = c\right) = P(X_i = x_i \middle| C = c) \qquad (20)$$

Para todo i=1,….,n. De ahí, que se verifique la ecuación.

Por tanto, en el paradigma naïve Bayes, la búsqueda del resultado más probable, c^*, una vez conocidos los valores de las variables ($x_1,.....x_n$) determinado caso, se reduce a 21:

$$c^* = \arg\max_{c} \mathrm{P}\left(\mathrm{C} = \mathrm{c} \middle| \mathrm{X}_1 = x_1,..., X_m = x_m\right) = \arg\max_{c} P\left(C = c)\prod_{i=1}^{n} P\left(X_i = x_i \middle| C = c\right)\right) \quad (21)$$

Suponiendo que todas las variables predictoras son dicotómicas, el número de parámetros necesarios para especificar un modelo naïve Bayes resulta ser (m – 1) + m n, ya que: Se necesitan (m - 1) parámetros para especificar la probabilidad a priori de la variable C. Para cada variable predictora X_i se necesitan m parámetros para determinar las distribuciones de probabilidad condicionadas.

En e1 caso de que las n variables predoctoras $X_1,.....,X_n$ sean continuas, se tiene que el paradigma naïve Bayes se convierte en buscar el valor de la variable C, que denotamos por c^*, que maximiza la probabilidad a posteriori de la variable C, dada la evidencia expresada como una instanciación de las variables $X_1,....,X_n$, esto es, x = ($x_1,....,x_n$). Es decir, el paradigma naïve Bayes con variables continuas trata de encontrar c^* verificando 22:

$$c^* = \arg\max_{c} P\left(C = c\right)\prod_{i=1}^{n} fX_i \middle| C = c\left(x_i \middle| c\right) \qquad (22)$$

donde $f X_i \mid C = c(x_i \mid c)$ denota, para todo i = 1,.....,n, la función de densidad de la variable X_i condicionada a que el valor del diagnóstico sea c.

Suele ser habitual utilizar una variable aleatoria normal (para cada valor de C) para modelar el comportamiento de la variable X_i. En este caso el número de parámetros a estimar es (m – 1) + 2mn:

- m - 1 en relación con las probabilidades a priori p(C = c)
- 2nm en relación con las funciones de densidad condicionadas

Finalmente puede ocurrir que algunos de los hallazgos se recojan en variables discretas mientras que otros hallazgos sean continuos. En tal caso hablaremos del paradigma naïve Bayes con predictoras continuas y discretas.

Supongamos que de las n variables predictoras, n_1 de ellas, $X_1,....,X_n$, sean discretas, mientras que el resto $n - n_1 = n_2$, $Y_1,.....Y_{n2}$, sean continuas. En principio al aplicar directamente la formula del paradigma naïve Bayes correspondiente a esta situación se obtiene de 23:

$$P\left(c|x_1,...,x_{m1},y_1,...,y_{m2}\right) \approx P(c)\prod_{i=1}^{n1} P\left(x_i|c\right)\prod_{j=1}^{n2}\frac{f\left(y_1|c\right)}{\max_y f\left(y_j|c\right)} \qquad (23)$$

5 Desarrollo

La problemática planteada, fue abordada mediante la utilización de la metodología de minería de datos, la cual está basada a su vez, en la metodología CRISP-DM que consta de 5 fases:

I. Comprensión del contexto.
II. Comprensión de los datos.
III. Preparación de los datos.
IV. Modelado.
V. Resultados.

Fase I: Comprensión del contexto

Esta es la primera fase planteada por la metodología, y como tal, requiere la valoración de una serie de factores fundamentales, a partir de los cuales se construye el resto de la metodología. Estos factores suponen la comprensión del contexto, es decir, las circunstancias sobre las cuales se trabajará, el punto de partida. Asimismo, en esta etapa se deberán establecer objetivos claros a ser alcanzados. En esta etapa, también deben ser definidos, los criterios bajo los cuales trabajará el modelo, junto con las expectativas, es decir lo que espera obtenerse del proyecto y bajo que supuestos se plantea el problema y su resolución, es decir que presunciones previas existen. Estos supuestos estarán así, relacionados con las expectativas. Se valorarán que recursos son necesarios para llevar adelante el proyecto y por último se considerarán los riesgos asociados al tratamiento del problema.

Por otra parte, en esta etapa se establecerán los requisitos o subobjetivos para cada uno de los objetivos planteados en el estudio. Para poder determinar los requisitos será necesario estudiar para cada objetivo:

- Objetivos del requisito.
- Supuestos del requisito.
- Expectativas.
- Origen de la información.
- Atributos del requisito.
- Riegos de requisito.

La identificación de cada uno de estos puntos, para cada uno de los objetivos del proyecto permitirá la comprensión llegar a una correcta interpretación del contexto. En la Tabla 3 se sintetizan las actividades de esta fase según la metodología CRISP-DM:

Tareas Componentes	Actividades Asociadas
Determinar los objetivos	➢ Background ➢ Objetivos del negocio ➢ Criterios de éxito del negocio
Evaluación de la situación	➢ Inventario de Recursos ➢ Requisitos, supuestos y requerimientos ➢ Riesgos y contingencias ➢ Terminología ➢ Costos y beneficios
Determinar los objetivos del proceso de Explotación de Datos	➢ Las metas del proceso ➢ Criterios de éxito de proceso
Realizar el plan de proyecto	➢ Plan de proyecto ➢ Valoración inicial de herramientas

Tabla 3. Tareas y actividades de la Fase I

Fase II: Comprensión de los datos.

Las tareas involucradas para la realización de esta fase de la metodología CRISP-DM hacen básicamente a la comprensión de los datos, para la cual se debe describir, coleccionar, organizar, verificar y limpiar, antes de realizar cualquier análisis sobre ellos. Esta etapa puede consumir mucho tiempo y en este caso, se realiza en forma genérica para la totalidad de los objetivos. Vamos a encontrar cuatro sub fases para esta etapa del estudio. La primera de ellas es la recolección de datos iniciales. Esta tarea básicamente consiste en el acceso a la información propuesta en el plan de recursos. En esta etapa solo se recolecta información, más no se integra en caso de proceder de distintas bases de datos. La segunda sub fase es la descripción de los datos o bien como estos datos son volcados en la base de datos, cuales son los campos que relacionan unas tablas con otras. Este paso es de gran importancia para conservar la consistencia de la información ya que una interpretación equivoca en esta etapa lleva indudablemente a resultados errados. La tercera sub fase es la exploración de los datos. En esta sub etapa se realiza un análisis preliminar de la información. Esta investigación resulta de utilidad para conocer cuales son los datos que se utilizan, su rango de valores, su continuidad o discreción, etc. La cuarta y última subyace es la verificación de la calidad de la información, etapa en la que se observará la consistencia de los datos, la ausencia de ellos, la existencia de outliers, etc. En la Tabla 4 se sintetizan las actividades de esta fase según la metodología CRISP-DM:

Tareas Componentes	Actividades Asociadas
Recolectar los datos iniciales	➢ Reporte de recolección de datos iniciales
Descubrir datos	➢ Reporte de descripción de datos
Exploración de los datos	➢ Reporte de exploración de datos
Verificación de la calidad de datos	➢ Reporte de la calidad de datos

Tabla 4. Tareas y actividades de la Fase II.

Fase III: Preparación de los datos

Esta fase agrupa todas aquellas tareas o actividades relativas a preparación total de los datos para el análisis. Estas tareas básicamente son la selección de los datos, su limpieza, la estructuración de la información, la integración cuando sea necesaria y por último se establecerá el formato final de los datos. A diferencia de la fase I y de la fase II, a partir de aquí, se desarrollan el resto de las fases, es decir III, IV y V para cada uno de los requisitos de cada objetivo del proyecto. En la tabla 5 se sintetizan las tareas y actividades derivadas de la fase III.

Tareas Componentes	Actividades Asociadas
	➢ Dataset ➢ Descripción de dataset
Seleccionar los datos	➢ Inclusión/exclusión de datos
Limpiar los datos	➢ Reporte de calidad de datos
Estructurar los datos	➢ Derivación de atributos ➢ Generación de registros
Integrar los datos	➢ Ubicación de datos
Formato de los datos	➢ Reporte de calidad de datos

Tabla 5. Tareas y actividades de la Fase III.

Fase IV: Modelado

Hasta aquí se tiene la información lista para ser procesada según los criterios y objetivos establecidos en la Fase I. En esta etapa se realizarán los modelos necesarios de procesamiento de información, se pondrán a prueba las hipótesis planteadas, se contrastarán herramientas. Cada modelización irá ligada a un objetivo y según sus requerimientos se construirán los modelos necesarios para procesar la información que ha sido preparada para alcanzar los sub objetivos que permitan llegar al resultado esperado. Esta fase de trabajo puede ser divida en cuatro sub etapas. En primer lugar se selecciona la técnica de modelado, luego se realiza el diseño de pruebas preliminares para posteriormente construir el modelo y finalmente evaluar el modelo. En la tabla 6 se sintetizan las actividades y tareas correspondientes a la Fase IV.

Tareas Componentes	Actividades Asociadas
Seleccionar una técnica de modelado	➢ La técnica modelada ➢ Supuestos de modelo
Generar un plan de pruebas	➢ Plan de puebas
Construir el modelo	➢ Configuración de parámetros ➢ Modelo ➢ Descripción del modelo
Evaluar el modelo	➢ Evaluar el modelo ➢ Revisación de la configuración de parámetros

Tabla 6. Tareas y actividades de la Fase IV.

Fase V: Resultados

En esta última fase de trabajo, se evaluarán los resultados, es decir que se analizará la consistencia de los resultados como así también su interpretación correcta. Posteriormente se realiza un proceso de revisión o corroboración de dichos resultados. Por último, se establecen próximos pasos a seguir.

Tareas Componentes	Actividades Asociadas
Evaluar resultados	➢ Valoración de resultados mineros con respecto al éxito del proyecto
Proceso de revisión	➢ Revisión del proceso
Determinar los próximos pasos	➢ Listar posibles acciones

Tabla 7. Tareas y actividades de la Fase V.

5.1 Metodología de minería de datos

Según lo desarrollado en el punto 5, a continuación se determinan cada una de las fases de la metodología.

5.1.1 Comprensión del contexto

Objetivo del proyecto

En la actualidad, los especialistas médicos no cuentan con herramientas objetivas, que los ayuden a tomar una decisión respecto al tratamiento óptimo para un paciente. Utilizando la información que los especialistas consideran importante, se pretende caracterizar a los pacientes con cáncer de próstata tratados con braquiterapia con el fin de optimizar la elección de dicha terapia para futuros pacientes.

Criterios de éxito del proyecto

Los criterios de éxito del proyecto se identifican con localizar comportamiento en los casos de éxito y fracaso de tal forma que permita establecer indicadores para tratar pacientes con la patología planteada.

Expectativa del proyecto

La expectativa principal de este proyecto es encontrar patrones cuantitativos de comportamiento y a partir de ellos, predecir de forma confiable el éxito o fracaso de la braquiterapia ante un nuevo caso.

Recursos

Desde el punto de vista del dominio de los datos, los recursos humanos con los que se cuenta, tienen que ver con un experto en el área de medicina (urólogo), quien define y establece los criterios médicos a seguir, y expertos en minería de datos para interpretar la información. Por otra parte, se necesita recolectar la información necesaria para llevar adelante la minería y el acceso al software necesario.

Supuestos

Este trabajo se desarrolla bajo de supuestos de distintas clases. En primer lugar, supuestos urológicos y radioterapéuticos, que hacen al tratamiento mismo, por ejemplo se supone que la dosis suministrada a un paciente es siempre adecuada. Asimismo, se consideran supuestos físicos referentes a la calidad y cálculo de las emisiones radiactivas. Se acepta que los mecanismos utilizados en la actualidad son confiables.

Riesgos

Por último, se consideran riesgos de distintas categorías. Por un lado, riesgos derivados de la cantidad de información utilizable con la que se cuenta. Esto representa un riesgo estadístico ya que la poca cantidad de registros puede resultar en una muestra no significativa de pacientes tratados con braquiterapia. Por otro lado se consideran riesgos biológicos, que vienen riesgos vienen dados por el dinamismo del propio cuerpo humano. La inserción de semillas en la próstata del paciente al momento del implante puede garantizar un superficie cacerosa totalmente recubierta por la radiación, sin embargo no se puede determinar con certeza el movimiento posterior de las semillas y en consecuencia su efecto sobre el área irradiada. Por último, se consideran los riesgos físicos. Este tipo de riesgos tienen que ver con los cálculos de radiactividad propiamente dichos. Como la forma de las pastillas modifica y ubicación modifica los valores de cálculo. Para los tipos de riesgos planteados no existe un plan de contingencia a corto plazo, de hecho el experto médico ha tomado las acciones necesarias para que en el futuro se cuente con mayor cantidad de información y con mas caracterización de los pacientes desde el punto de vista del tratamiento como de su contexto.

Objetivos de proceso de explotación de datos (requerimiento)

Encontrar patrones de comportamiento y relaciones entre las variables en pacientes con cáncer de próstata tratados mediante braquiterapia. Se espera con esto determinar el grado de eficiencia del tratamiento en los pacientes. La información para cumplir con este requerimiento ha sido suministrada por el experto médico y cuenta con los datos concernientes a pacientes tratados según la indicación del experto.

5.1.2 Comprensión de los datos

5.1.2.1 Población

La población utilizada para realizar este estudio, está formada en su totalidad por hombres argentinos, a quienes se les detectó cáncer de próstata y fueron tratados según el tratamiento denominado Braquiterapia prostática. Según expresa la metodología CRISP-DM, en su fase II, véase Tabla 2, se analizaron 206 casos de pacientes, que tuvieran seguimiento posterior al tratamiento. Luego de realizar la recolección de información, análisis de datos y limpieza de dichos datos se formó el set definitivo, integrado por 116 personas. Se verifica una gran diferencia entre los datos disponibles y los utilizables para el análisis. Básicamente, muchos de los datos no pudieron verificar la evolución de la aplicación del tratamiento, muchos otros no contenían los valores de las variables características, tema que se verá a continuación en el punto 5.2. En consecuencia, sólo 116 registros cumplen los requisitos necesarios para ser analizables en función de lo que los especialistas consideran importante.
Por otra parte, no puede dejar de observarse que la cantidad de registros utilizados no es lo suficientemente grande para asegurar validez estadística.

5.1.2.2 Variables de análisis

Todo sistema del mundo real, esta caracterizado por una gran cantidad de variables, que interaccionan directa e indirectamente sobre él. Para poder modelar, cualquier sistema del mundo real, es necesario acotarlo, aislarlo y entender su funcionamiento mediante su representación virtual. Como se ha dicho anteriormente, sobre el sistema actúan gran cantidad de factores que influyen sobre él. Para hallar su equivalente en el modelo resulta necesario encontrar cuales son aquellos factores o variables que mayor influencia ejercen sobre el sistema y cuales aquellas que al modificarlas generan mayor apalancamiento, asumiendo que responden al principio de 20-80, es decir que el 20% de la variables generan un efecto del 80% sobre el resultados mientras el que restante 80% solo apalanca un 20% al sistema. Así, la primer tarea fue trabajar junto a especialistas de la salud, para identificar junto a ellos cuales son las variables que mayor influencia tienen sobre el diagnostico, tratamiento y evolución de un paciente que sufre cáncer de próstata.

Luego de esta primera etapa, se tuvieron en cuenta exactamente todas aquellas variables cuantificables o categorizables que resultan indispensables para que los especialistas tomen sus decisiones. Es importante destacar que no se generan juicios previos sobre las variables, sino que los algoritmos son quienes buscarán relaciones subyacentes que no puedan ser observadas a simple vista. Finalmente las variables seleccionadas para el estudio fueron:

- PSA Pre-implante.
- Nivel de Gleason.
- PSA diagnosticado.
- Edad.
- Volumen Ecográfico en gramos.
- Tiempo transcurrido del implante al último seguimiento, Delta T.
- Estadío de la Enfermedad.
- Si el tratamiento fue combinado con alguna otra terapia (SI, NO).

A continuación se explica cada una de las variables citadas:

- PSA: El antígeno prostático específico (frecuentemente abreviado por sus siglas en inglés, PSA), es una sustancia proteica sintetizada por células de la próstata y cuya función es la disolución del coágulo seminal. Es una proteína de síntesis exclusiva en la próstata. Una pequeñísima parte de este PSA pasa a la circulación sanguínea de hombres sanos, y es precisamente este PSA que pasa a la sangre, el que se mide para el diagnóstico, pronóstico y seguimiento del cáncer -tanto localizado como metastásico- y otros trastornos de la próstata, como la prostatitis. Los niveles normales en sangre de PSA en los varones sanos son muy bajos, del orden de millones de veces menos que el semen, y se elevan en la enfermedad prostática. Los valores de referencia para el PSA sérico varían según los distintos laboratorios, aunque normalmente éstos se sitúan en 4 ng/mL. Su producción depende de la presencia de andrógenos y del tamaño de la glándula prostática. Este concepto incluye tanto el PSA preimplante como el diagnosticado.
- Gleason: Sistema de calificación del tejido canceroso de la próstata según su aspecto bajo el microscopio. Los puntajes de Gleason oscilan entre 2 y 10 e indican la probabilidad de diseminación del tumor. Un puntaje de Gleason bajo significa que el tejido canceroso es similar al tejido de próstata normal y el tumor tiene menos probabilidades de diseminarse; un puntaje de Gleason alto significa que el tejido canceroso es muy diferente del normal y lo más probable es que este se disemine.
- Edad: Tiempo biológico de vida desde el nacimiento.
- Volumen ecográfico: Es el volumen de próstata según el calculo realizado por ecografía.
- Delta T: Tiempo definido desde el momento del implante hasta el último control.

- Estadío de la enfermedad: El estadio o etapa de un cáncer es el factor más importante para elegir la opción de tratamiento más adecuada y predecir el pronóstico del paciente. Si se confirma un cáncer, es necesario realizar más pruebas diagnósticas para saber la extensión del cáncer dentro de la próstata.
- Tratamiento combinado: Establece si posteriormente al implante de semillas, se utiliza alguna otra terapia complementaria.

Por último, se tiene la variable Resultado, que puede tomar dos valores: Éxito o Fracaso. El resultado básicamente mide, si al realizarse el último control registrado del paciente, se verifica reaparición del cáncer. En caso de no verificarse reaparición se establece que el tratamiento ha sido exitoso y viceversa. Es una variable binaria, que en el caso de utilización de algoritmos de caracterización, resulta una variable más de análisis, cuyos resultados verifican que las clases obtenidas están muy ligadas al Resultado.
Por otra parte, al utilizar algoritmos de inducción, esta variable actúa como nodo objetivo. Como se ha mencionado previamente, la combinación de algoritmos genera que el Resultado sea utilizado como variable de análisis en la etapa de caracterización y genere clases asociadas a él, pero permita la acción de otras variables. Así en la etapa de inducción, el Resultado se presenta contenido en las distintas clases o clusters. Por otra parte, cuando se analiza según el enfoque de Redes Bayesianas, el Resultado sí actúa como nodo objetivo.

5.1.3 Modelo de datos

Se generaron distintos experimentos para encontrar patrones de comportamiento y de predicción. A continuación se exponen y caracterizan dichos experimentos.
En primer lugar se trabaja en la comprensión completa de los datos del set, mediante la utilización de redes bayesianas. Se utilizó el software Bayeslab y dentro de las opciones que dicho programa ofrece, se utilizó el algoritmo de Naive-Bayes cuya aplicación se puede ver en la sección 4.2.3. Las razones de su elección radican básicamente en la posibilidad de manipular los datos del set, sumado a una interfaz gráfica muy accesible. También presenta una serie de reportes de análisis muy interesantes en cuanto a la información que se pretendía interpretar. Por último, es una herramienta disponible en una versión de prueba.
Luego, se trabaja en la caracterización del dataset, con algoritmos del tipo SOM, es decir aprendizaje automático no supervisado con capacidad de generar mapas bidimensionales de información caracterizada. La herramienta utilizada para realizar este análisis fue el NNclust. La parametrización del software se muestra a continuación en la figura 5.

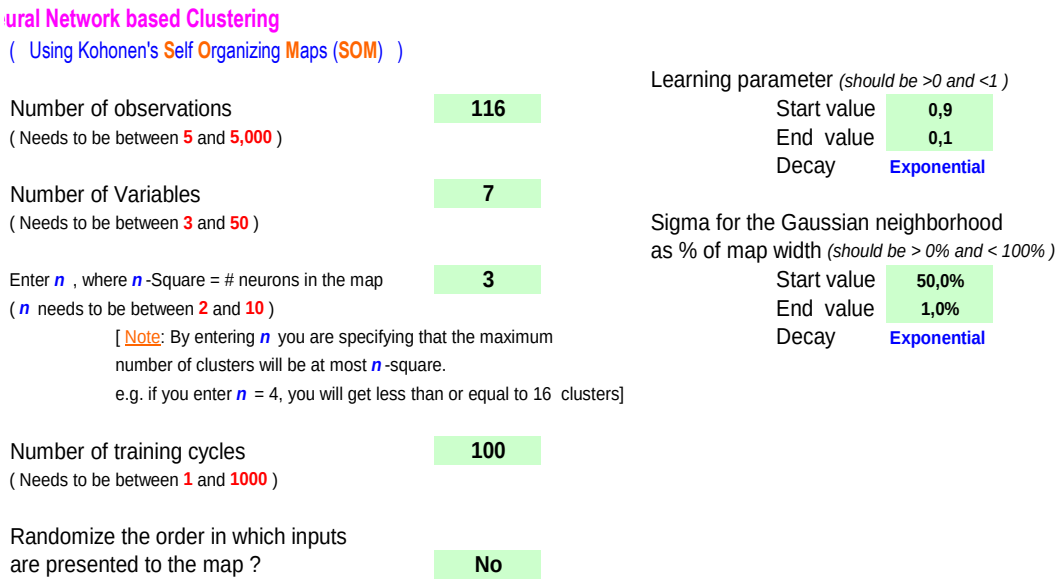


Figura 5. Parametrización del software NNclust.

Por último, se estudia la clasificación o inducción del dataset utilizándose algoritmos de tipo inductivo. El output generado por la clusterización, es decir, las clases generadas, son ahora el nodo objetivo para esta última etapa de análisis. A diferencia de la etapa de clusterización, se categorizan las variables PSA, PSA preimplante y volumen ecográfico, para ser procesadas en el software. En el Anexo se puede ver la categorización mencionada. No se consideran los atributos cualitativos como Estadío de la enfermedad y Tratamiento combinado. Para la realización de este experimento se utiliza el software Sipina Reaserch, que trabaja con una gran variedad de algoritmos. La elección de esta herramienta se debe principalmente a su gran versatilidad. Se testearon diferentes algoritmos para evaluar la respuesta del sistema. Entre ellos:

- ID3-IV
- ChAID
- C4.5

Se trabajaron en sus versiones originales y en sus versiones mejoradas. El algoritmo que mejor ajustó los datos y permitió explicar las predicciones de las variables fue el ID3.

5.1.4 Resultados

5.1.4.1 Resultados obtenidos utilizando Redes Bayesianas

Las redes Bayesianas se representan mediante la utilización de gráficos (arcos y nodos). Los nodos representan a las variables y los arcos que los conectan corresponden a las relaciones probabilísticas entre dichas variables. Las redes de Bayes son estructuras de

aprendizaje automático, que aprenden directamente del set de datos. A continuación se puede ver la representación gráfica de la red en la figura 6.

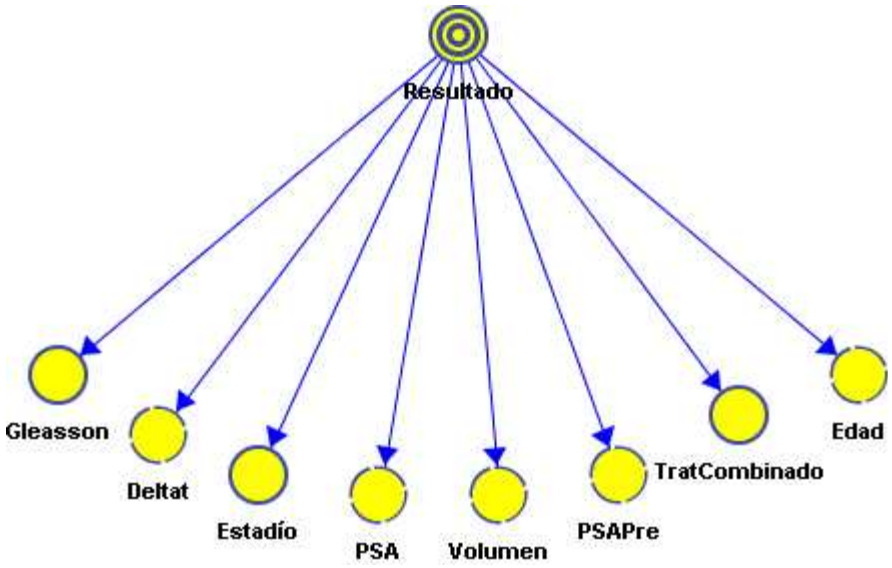


Figura 6. Diagrama de atributos y clase.

La herramienta utilizada para la realización de este análisis fue BayesiaLab. Las razones de su elección radican básicamente en la posibilidad de manipular los datos del set, sumado a una interfaz gráfica muy accesible. También presenta una serie de reportes de análisis muy interesantes en cuanto a la información que se pretendía interpretar. Por último, es una herramienta disponible en una versión de prueba.

Las variables o atributos utilizados para generar el análisis fueron:

- Gleason
- PSA Inicial
- PSA Preimplante
- Edad
- Volumen Ecográfico de la Próstata (Grs)
- Estadío de la enfermedad
- Tiempo transcurrido hasta último control: Delta T.
- Tratamiento combinado: Si se complemento la braquiterapia con otro tratamiento.

Por su parte, el nodo objetivo como puede observarse en la figura 2 es el Resultado. Las variables cuantitativas se procesaron en forma continua, sin requerimientos de categorización, mientras que otras como: Tratamiento combinado, estadío de enfermedad, se analizaron categorizadas.

El algoritmo que se emplea para el análisis es el de Nive-Bayes, que utiliza el criterio de minimización de MDL (minimum description lenght). Este valor permite medir la calidad de la red respecto de la base de datos. Esta formado por dos partes: La primera

evalúa la estructura de la red, mientras que la segunda cuantifica cuan bien la red se ajusta a los datos. El beneficio de la primera parte del MDL radica en la implementación del criterio de Occam. Este es, por ejemplo, el que permite elegir la hipótesis más simple cuando todos los otros factores son idénticos. Así pues, a partir de los datos de la base de pacientes tratados con braquiterapia, se generó la red bayesiana correspondiente. A continuación se pueden ver, en la figura 7, los cuadros resultantes del panel de control de la red. En ellos se puede ver la distribución de probabilidad de las variables monitoreadas, en el proceso de inferencia estadística según los criterios mencionados.

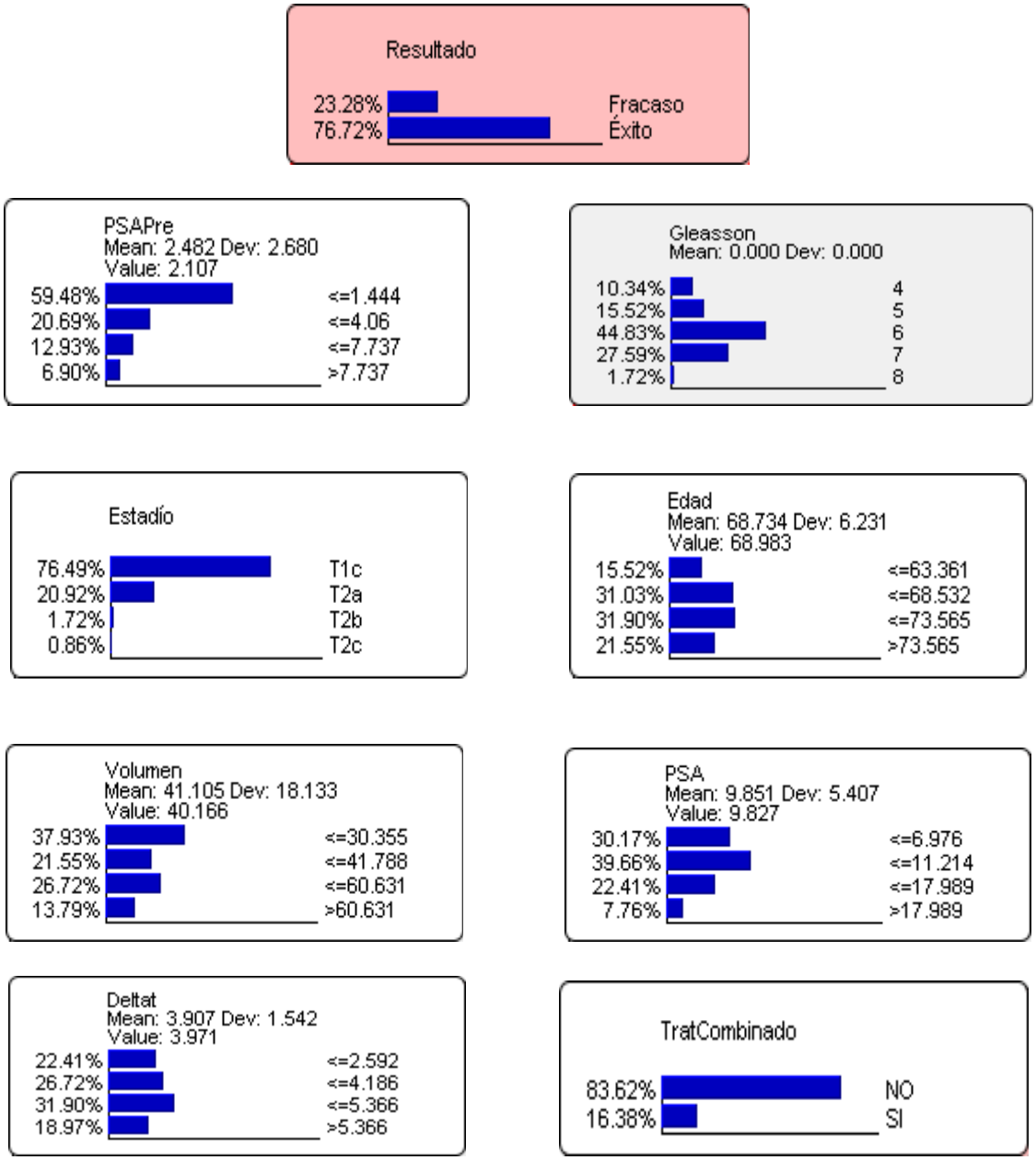


Figura 7. Tablas de resultados del panel de control.

El primer análisis que se realiza permite entender la distribución a priori de los datos y cual es su factor de ocurrencia en la totalidad de la información. De aquí, se establecen las condiciones iníciales de trabajo, a partir de la cuales se analizarán las variaciones para alcanzar una meta preestablecida, ya sea Éxito o Fracaso.

Se puede establecer que el 76.72% de los pacientes reaccionaron exitosamente al tratamiento. El 23.28% de los casos restantes fracasaron. De dicho 76. 72% de casos exitosos, el 59.48% presentaba PSA preimplante menor a 1.44, mientras que en su mayoría (76.49%) los pacientes pertenecen al estadío de enfermedad T1c. Por su parte la edad presenta su ocurrencia bastante distribuida, sin un grupo mayoritario, en tanto que el Gleason encuentra casi la mitad de los casos (44.83%) en el valor 6.

Como se observa en la figura 3, cada cuadro se corresponde con una variable. Se pueden ver las frecuencias de ocurrencia de los valores del set de datos. Así pues, para el caso de PSA Preimplante el 59,48% de los datos presentan un valor inferior a 1,44, mientras que el 20% se encuentra entre dicho valor y 4,06. Para el Gleason el 44,83% de los datos se ubican en el valor 6, mientras que el 7 tiene un frecuencia de 27,59%. Así pues, para cada atributo, se visualiza la distribución probabilística del set de datos, o dicho de otra forma los valores que tomas los arcos que unes los nodos.

Hasta aquí, se ha visto la frecuencia de ocurrencia de cada intervalo de valores o categoría asociada a cada uno de los atributos de análisis. Para entender con mayor profundidad la calidad de la información presentada, resulta de interés analizar la contribución global de los atributos. Cada variable ejerce un impacto diferente sobre la respuesta final del algoritmo, que se denomina contribución global. Así pues, el Gleason contribuye a la respuesta final del algoritmo en un 29,37%. En la tabla 8, pueden observarse los pesos relativos de cada una los atributos y sus contribuciones globales.

Child	Relative Weight	Global Contribution
Gleason	1.00	29.37%
PSA Pre implante	0.88	25.95%
Estadío	0.51	15.12%
Edad	0.40	11.81%
Volumen Ec.	0.34	10.20%
PSA	0.13	4.07%
Delta T	0.10	3.14%
Trat. Combinado	0.01	0.36%

Tabla 8. Contribuciones globales de las variables.

Las variables de mayor influencia, como se puede ver en la tabla 8 son: el Gleason, el PSA Preimplante, el estadío de la enfermedad y la edad, siguiendo en forma descendente en importancia con: El volumen, el PSA, el Delta T y el Tratamiento combinado.

Sin embargo, en una primera experimentación no puede determinarse cuales son las relaciones entre las variables, si existe alguna dependencia entre ellas. Se analiza esta situación en la sección 5.1.4.1.2.

Hasta aquí se ha presentado la distribución a piori de los datos y cual es la significación de cada una de las variables en el resultado final. No debe perderse de vista que el objetivo final es poder generar una herramienta de predicción y análisis. Desde este punto de vista, es necesario seguir adelante para forzar al sistema a encontrar un respuesta exitosa, es decir que la probabilidad de tener Éxito no sea el 76.72%, sino que represente el 100% y analizar como se modifican los valores de las variables mas significantes.

De esta manera en la se analiza la distribución de probabilidad necesaria para poder predecir la clase Éxito y como se modifican los valores de la distribución. En la figura 8 se verifica la condición de Éxito.

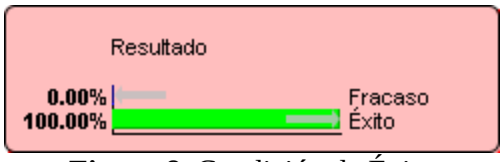
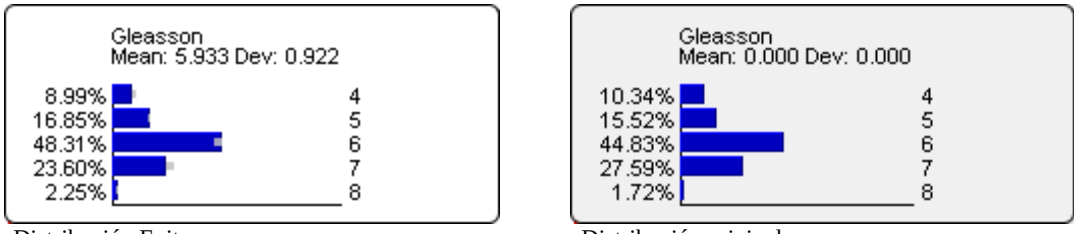


Figura 8. Condición de Éxito

Gleason:

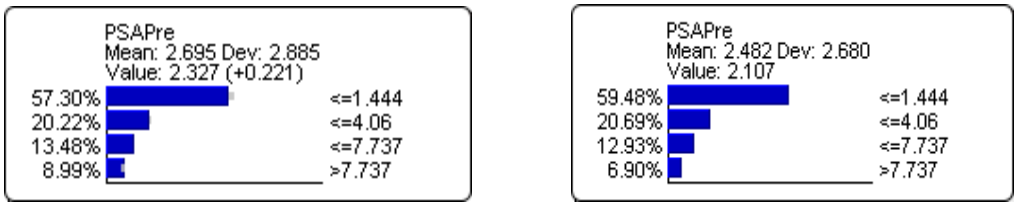


Distribución Exitosa Distribución original

Figura 9. Distribución exitosa y original para el Gleason.

En la figura 9 se observa que la variación de la variable Gleason en su condición de Éxito respecto de la distribución original de probabilidad el valor 6 se incremente desde 44,83% a 48,31%, es decir un 3,48%. El valor correspondiente a Gleason 7 disminuye un 4%, siendo esta disminución la mayor variación porcentual de la variable.

PSA Preimplante:

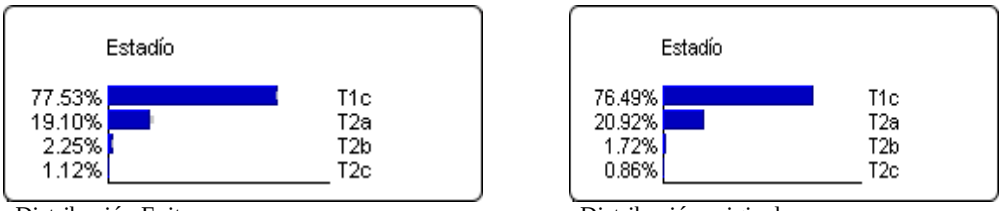


Distribución Exitosa Distribución original

Figura 10. Distribución exitosa y original para el PSA preimplante

En la figura 10 puede observarse que las variaciones relativas a esta variable también son pequeñas, no superando en ningún caso el 2,18%. Es notoria la acumulación de información en valores de PSA preimplante inferiores al 4,06, totalizando el 77,52% de los datos en dicho rango.

Estadío de la enfermedad:



Distribución Exitosa Distribución original

Figura 11. Distribución exitosa y original para el Estadío de la enfermedad.

En la figura 11 se observan las variaciones entre la distribución original y la condición de Éxito. Se verifican variaciones menores al 2% en la distribución para el caso de tratamiento exitoso

Edad:

Distribución Exitosa Distribución original

Figura 12. Distribución exitosa y original para la edad.

Se verifica un incremento de aproximadamente el 3% para los valores menores a 73,56 años y mayores que 68.5 años, como se puede observar en la figura 12, mientras que para valores inferiores a 68.5 compensa el incremento anterior. En esta caso también se observan variaciones pequeñas.

El resto de las variables presentan menor importancia respecto de su contribución global al resultado. Cabe mencionar la poca variación general que se observa en los valores de las variables cuando se fuerza la condición de éxito. Esto pone de manifiesto la no existencia de rangos de valores claramente estratificados para cada estado final, sino que la combinación aleatoria de valores de diferentes atributos, llevan a uno otro resultado. Dada esta situación, se analiza en la sección 5.1.4.1.1. cuales son las variables de mayor peso relativo, para identificar otros patrones de comportamiento.

5.1.4.1.1 Análisis de las variables significativas

En función de los resultados obtenidos para el conjunto total de datos, se estudia el comportamiento de las variables cuando se consideran las de mayor significación, según se menciona en la sección anterior. El criterio de corte es arbitrario se establece para aquellos atributos que presentan un peso relativo mayor a 0,5. Se toma este valor, como límite tal, capaz de explicar más de la mitad de la respuesta obtenida. Asimismo, las tres variables mas importantes presentan un salto en cuanto a su contribución global, como puede observarse en la tabla 3, siendo mayormente manifiesta en el Gleason y en el Estadío de la enfermedad. Siguiendo los lineamientos de estos criterios, las variables que resultan consideradas para el análisis son: Gleason, PSA preimplante, Estadío de la enfermedad.

Se procesan los datos mediante el software Bayesilab, bajo el algoritmo de Naive-Bayes. El resultado que se observa en la figura 13 es consistente con el obtenido según la figura 6 ya que se utiliza el mismo set de datos.

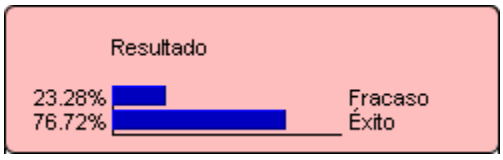


Figura 13. Distribución de datos.

La distribución para cada una de los atributos considerados se ve a continuación en la figura 14:

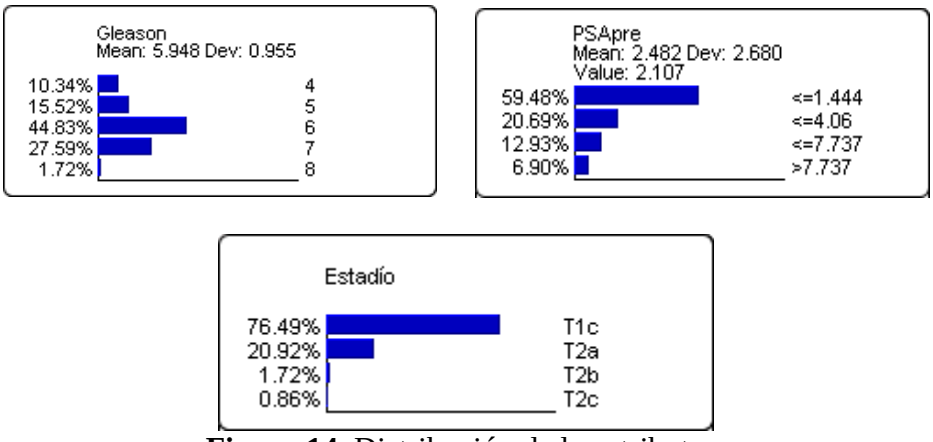


Figura 14. Distribución de los atributos.

Del mismo modo que sucede con el Resultado, la frecuencia de ocurrencia se repita al considerarse solo las tres variables más significativas. Resulta relevante, verificar como se modifican los valores de contribución para encontrar relaciones subyacentes que no hayan quedado en evidencia en el análisis de todas las variables. En la tabla 9 se presentan los valores de contribución global y peso relativo de cada variable.

Child	Relative Weight	Global Contribution
Gleason	10.000	43.43%
PSApre	0.8836	38.37%
Estadío	0.4192	18.20%

Tabla 9. Contribución Global y peso relativo de las variables.

El comportamiento de las variables no sufre modificación significativa en cuanto a su contribución global al eliminar los atributos menos significativos. En la Tabla 10 se muestra una comparativa entre los valores originales con todas las variables y la proyección de los valores del análisis de los atributos significativos.

	A	B		Desvío
Gleason	43.43%	29,37%	→ 41,70%	1,73%
PSApre	38.37%	25,95%	→ 36,84%	1,53%
Estadío	18.20%	15,12%	→ 21,47%	-3,27%
		70,44%	100%	

Tabla 10. Comparativa variables significativas.

La columna A muestra los valores del análisis realizado a las tres variables significativas. La columna B, por su parte, presenta el resultado de la contribución global del análisis de todas las variables. Se proyectan dichos resultados sobre un total de 100%. También se puede observar el desvió entre lo proyectado y lo calculado. Las diferencias son poco significativas, así pues, se concluye que no existen modificaciones importantes respecto del análisis original con todas las variables. De todas maneras se menciona el origen de las diferencias encontradas según la proyección. El estadío de Enfermedad pierde importancia cuando se analizan los atributos significativos, mientras que el Gleason y el PSA preimplante la incrementan.

Por otra parte, si cabe destacar la gran variación que presenta el peso relativo de cada variable, según expresa la Tabla 4. Se observa que el orden de importancia de las variables se mantiene comparando la tabla 7 con la tabla 6, y viene dado por:

1. Gleason
2. PSA Preimplante
3. Estadío de la enfermedad.

Sin embargo, al estudiar las tres variables significativas se produce una modificación en el peso relativo de las variables asignando un valor de 10 para el Gleason, 0,88 para el PSA Preimplante y 0,44 para el Estadío. En este nuevo escenario el Estadío de la enfermedad ya no explica mas de 0,5 como se establecía en el criterio de división de variables significativas. Es decir, que en este análisis más exhaustivo de variables significativas se modifican estos pesos relativos, dejando al Gleason como variable muy superior en relación con el PSA Preimplante y el Estadío.

5.1.4.1.2 Análisis de dependencia

Como se mencionaba anteriormente en la sección 5.1.4.1, existen variables que podrían estar relacionadas con otras. El efecto resultante de dicha relación es que la información que trae una variable se encuentre contenida o relacionada con la información de otra de variable. Esto podría generar que la contribución global de una variable parezca a priori poco significativa. En un primer análisis cualitativo de las variables elegidas y según con expuesto por los expertos, podrían existir relación entre el PSA original y el PSA preimplante. La diferencia conceptual entre ambos, radica en que el PSA original es el valor de antígeno prostático específico al momento de la detección del cáncer, mientras que el PSA preimplante cuantifica el valor del antígeno previo a la braquiterapia. El proceso que diferencia a estos valores es la medicación que reciben los pacientes para disminuir el valor del antígeno y la respuesta biológica del organismo ante dicha medicación. Así pues, se analiza el set de datos con las tres variables significativas y el PSA original. En la Tabla 11 se puede observar el valor de la contribución global luego de análisis, mientras que la Tabla 12, muestra los valores correspondientes al análisis cuando no se considera el PSA preimplante. El contraste de estas dos experiencias busca verificar si realmente existe correlación significativa en el PSA original y el PSA preimplante.

Child	Relative Weight	Global Contribution
Gleason	10.000	40.96%
PSApre	0.8836	36.20%
Estadío	0.4192	17.17%
PSA	0.1384	5.67%

Tabla 11. Valores de peso relativo y contribución global.

Child	Relative Weight	Global Contribution
Gleason	10.000	64.20%
Estadío	0.4192	26.91%
PSA	0.1384	8.89%

Tabla 12. Valores de peso relativo y contribución global.

La Tabla 11 verifica que los valores de los atributos PSA y PSA preimplante, mantienen su proporción relativa respecto del análisis con la totalidad de las variables. Por otra parte, en la segunda experiencia, es decir, cuando no se tomo en cuenta el atributo PSA preimplante, según expresa la Tabla 12, la participación relativa del PSA tampoco toma valores significativos. Así pues, resulta que la información aportada por el PSA preimplante no se superpone con la información aportada por el PSA. Dicho de otra manera, según los resultados expuestos en la tabla 9 y en la tabla 10, no existe correlación entre ambas variables.

Por otra parte, se realizó un estudio de correlación sobre el set de datos, para corroborar los resultados obtenidos en el análisis anterior. En la tabla 13 se muestran los valores de correlación entre las variables.

	PSApre	*PSA*
PSApre	1	
PSA	$ 0,1435	1

Tabla 13. Coeficientes de correlación entre PSA y PSA preimplante.

El coeficiente de correlación entre el PSA y el PSA preimplantes es de 0,14. Recordemos que el valor del coeficiente de correlación es siempre mayor o igual a 0 y menor o igual a 1, siendo 1 el valor de máxima correlación y disminuyendo hasta llegar a 0. Se concluye entonces que el nivel de correlación entre las variables no es significativo.

5.1.4.1.3 Análisis de variables poco significativas

Ya se han considerado la totalidad de las variables, las variables significativas, se han analizado posibles relaciones de dependencia. Por último, y para tener un visión completa del problema y de las variables involucradas, se estudia a continuación el comportamiento de aquellas variables que en análisis general realizado en la sección 5.1.4.1. resultan poco significativas. El objetivo de este análisis es observar la evolución del sistema si se quitan las variables significativas y en consecuencia el peso que éstas ejercen sobre el resultado del análisis general, y analizar fenómenos que puedan generarse y no verse atenuados por los atributos de mayor peso. Los resultados de los pesos relativos de las variables y sus contribuciones se muestran en la tabla 14.

Child	Relative Weight	Global Contribution
Edad	10.000	39.94%
VolPros	0.8637	34.49%
PSA	0.3443	13.75%
DeltaT	0.2656	10.61%
TratCom	0.0302	1.20%

Tabla 14. Valores de peso relativo y contribución global.

Como puede verse en la tabla 12 se pueden observar tres grupos. El primero de ellos formado por la Edad y el volumen ecográfico prostático con un contribución global de mayor importancia con valores de 39.94% y de 34.39% respectivamente. En un segundo grupo, podemos agrupar el PSA y el Delta T, con valores de 13.75% y 10. 61% respectivamente. Por ultimo, se presenta el atributo Tratamiento Combinado quedando relegado muy por debajo del resto de las variables. Comparativamente la división antes mencionada se corresponde con la división generada en el análisis de la totalidad de las variables, véase sección 5.1.4.1. Lógicamente sus valores por la ausencia de las variables significativas, pero su comportamiento desde la perspectiva de la contribución global es el mismo.

Sin embargo, cabe destacar el valor del peso relativo que presenta la variable Edad y en menor escala pero también en buena medida el volumen prostático. Son variables que contribuyen poco al resultado final, pero que sin duda cuando influyen lo hacen de

forma significativa. No parecen ser variables que definan el resultado de un tratamiento, pero sin duda al estar presentes harán notar su presencia. Este análisis se verá posteriormente en la sección 5.1.4.2.3.

5.1.4.2 Resultados obtenidos utilizando Algoritmos de Caracterización y Clasificación

5.1.4.2.1 Resultados de la caracterización de Datos

En primer lugar se trabajo con algoritmos de caracterización. La utilización de esta herramienta permitió generar mapas autoorganizados capaces de generar grupos de datos que luego fueron utilizaron como inputs para aplicar algoritmos de inducción, como se verá en la sección 5.4.2.3. La clusterización se realizó utilizando la herramienta NNclust, capaz de trabajar con la cantidad y calidad de datos del set, presentando interfases adecuadas para los requerimientos del problema y disponible para su utilización. Básicamente, el objetivo de esta etapa de experimentación fue no generar clases o clusters previos según preconceptos, sino permitir la aparición de relaciones subyacentes que no sean a priori visibles por el analista.
Así pues, el algoritmo elegido procesó los datos según una red SOM de tipo competitivo. Se alcanzó un estado estable para cada entrada, generando un total de 3 clases. En la Tabla 15 se pueden ver los valores medios de cada clase, correspondientes a cada variable, mientras que en la Tabla 16, se observan las varianzas propias de cada valor medio.

	Cluster Means		
	Cluster 1	**Cluster 2**	**Cluster 3**
Edad	65,0	64,0	70,0
Delta T	1,7	2,4	4,4
PSA diag.	9,9	8,6	9,6
Gleason	6,0	6,2	5,9
PSA preimp.	0,1	0,5	2,4
Volumen ecografico	33,4	37,0	41,5
Resultado	Fracaso	Fracaso	Éxito

Tabla 15. Centroides obtenidos mediante la Caracterización.

	Cluster Variances		
	Cluster 1	Cluster 2	Cluster 3
Edad	24,7	14,3	26,5
Delta T	0,7	3,7	1,5
PSA diag.	3,2	30,4	21,5
Gleason	0,0	0,6	0,9
PSA preimp.	0,0	1,7	9,0
Volumen ecografico	112,2	176,1	294,6

Tabla 16. Varianzas obtenidas mediante la Caracterización

5.1.4.2.2 Primera Interpretación de los Clusters

A continuación se presenta la evaluación de los clusters obtenido:

Cluster 3 (83%): Es el que posee mayor cantidad de registros, agrupando el 83% de ellos. Está directamente asociado a los casos de tratamiento exitoso, a tal punto que todos los casos de Éxito están incluidos en esta clase, mientras que no se verifica presencia de fracaso alguno. Como se puede observar el Tabla 13, la clase 3 se caracteriza por un tiempo promedio Delta T de 4,4 años, lo que lo que posiciona en un estadío estable de cura. Dicho de otro modo, todos aquellos casos exitosos presentan un tiempo promedio sin reaparición de cáncer de cuatro años y medio. La varianza de este valor es de 1,5 años, generando un mínimo de 3 años y un máximo de 6. Se destaca este resultado, porque si no hubo reaparición hasta esta etapa, es poco probable que reaparezca. El Gleason, por su parte, se encuentra en todos los casos por debajo de 7. El PSA preimplante es muy mayor de las otras clases, pero presenta una varianza muy grande.
Cluster 2 (13%): Es el que posee la mayor cantidad de fracasos, agrupando el 13% de los datos y el 78,94% de los fracasos. En cifras absolutas, agrupa 15 registros correspondientes al valor Fracaso de la variable Resultado. La característica de los valores de sus atributos es que se encuentran distribuidos en un amplio rango de valores, es decir que sus variables no se encuentran sesgadas, presentando valores muy grandes para las varianzas, según se puede observar en la Tabla 14, por lo que resulta apresurado sacar conclusiones respecto de su distribución.
Cluster 1 (4%): Esta clase posee solo 4 registros, que están en su totalidad asociados al valor Fracaso de la variable Resultado. A priori caracterizan el ruido del sistema, ya que son Fracaso, pero su patrón de comportamiento no es similar a aquellos registros agrupados en la clase 2. A diferencia de la clase 2, los valores de sus atributos se encuentran muy centrados en todas sus variables, menos en la edad y en el volumen ecográfico. Puede ser esta la causa que divide a los Fracasos en dos grupos diferenciados.

5.1.4.2.3 Resultados de la clasificación de Datos

Una vez realizado el proceso de clusterización, definidas las clases y sus características generales, se realiza un proceso inductivo sobre el set de datos. El output generado por la caracterización, es decir las clases generadas, vease Tabla 13, serán ahora el nodo objetivo o variable a predecir en esta última etapa de análisis. Para la realización de este experimento se utilizó el software Sipina Reaserch, que trabaja con una gran variedad de algoritmos. La elección de esta herramienta se debió principalmente a su gran versatilidad, es decir que otorga la posibilidad de ensayar distintos experimentos, y la comparación posterior de sus respuestas dan al software gran flexibilidad, posibilidad de contraste y verificación de los resultados obtenidos. El algoritmo que mejor ajustó los datos y permitió explicar las predicciones de las variables fue el ID3, desarrollado

anteriormente en la sección 4.2.2. Así pues, en esta etapa, el objetivo principal es procesar los datos utilizando ID3, para encontrar reglas de decisión que permitan predecir el Éxito o Fracaso de la braquiterapia ante la presencia de un nuevo caso de un paciente con cáncer de próstata diagnosticado. El algoritmo generó reglas de decisión. Dichas reglas pueden ser divididas en dos grupos diferenciados para su análisis. Un primer grupo que aglomera las reglas de Éxito y un segundo grupo que expresa las reglas de Fracaso. A continuación se definen ambos grupos:

Reglas de Éxito:

1. ***Si PSApre < 1.50 y Edad < 69.50***
   Entonces Cluster es [3] con confianza 55,88%.
2. ***Si PSApre < 1.50 y Edad >= 69.50 y PSA < 18 y Gleason =< 7***
   Entonces Cluster es [3] con confianza 100%.
3. ***Si PSApre >= 1.50 y Vol. Ec. >= 45***
   Entonces Cluster es [3] con confianza 100%.
4. ***Si PSApre >=1.50 y Vol. Ec. < y Edad >= 61.50 y PSA < 18***
   Entonces Cluster es [3] con confianza 100%.

Las reglas generadas presentan continuidad en muchos de los valores de sus variables.

Así pues, si el PSA preimplante es menor que 1,5 y la edad es menor que 69,5 entonces el tratamiento resulta exitoso. Sin embargo cuando la edad supera los 69,5 años, toma importancia que valores toman tanto el PSA diagnosticado como el Gleason. Para establecer una condición de Éxito el PSA debe ser menor que 18 y el Gleason menor o igual que 7.

De forma análoga, en el caso que el PSA preimplante sea mayor que 1,5 y el tratamiento resulte exitoso adquieren significación el valor de variables como Volumen ecográfico, Edad, PSA diagnosticado. Esto, se había vislumbrado en forma intuitiva en la sección 5.1.4.1.3. y aquí se justifica cuantitativamente. Es decir que como se dijo, atributos como Edad y Volumen prostático aparecen estableciendo límites en las condiciones pero no son definitorios como resulta el caso del Gleason que se verá más adelante. Entonces, para PSA preimplante menor a 1,5 y volumen ecográfico mayor a 45 el tratamiento resulta exitoso. Pero si el volumen ecográfico es menor a 45, la edad debe ser mayor a 61,5 y el PSA diagnosticado menor a 18, para determinar un caso exitoso.

Reglas de Fracaso:

1. ***Si PSApre < 1.50 y Edad >= 69.50 y PSAdiag < 18 y Gleason >= 8***
   Entonces Cluster es [2] con confianza 100%.
2. ***Si PSApre >= 1.50 y Vol. Ec. < 45 y Edad < 61.50***
   Entonces Cluster es [1] con confianza 100%.

3. ***Si PSApre >= 1.50 y Vol. Ec < 45 y Edad >= 61.50 yand PSA > = 18***
   Entonces Cluster es [1] con confianza 100%.
4. ***Si PSApre < 1.50 y Edad >= 69.50 y PSA >= 18***
   Entonces Cluster es [1] con confianza 100%.

La regla 1 de Fracaso debe ser mirada en contraste a la regla 2 de Éxito. Resultan análogas pero opuestas. La diferencia radica básicamente en el valor de la variable Gleason. Cuando dicho valor es inferior o igual a 7 se verifican casos de Éxito, sin embargo, cuando el valor del atributo supera tal valor se convierte en condición de fracaso de Fracaso. Aquí se ve claramente como el Gleason depende sólo de si mismo para determinar condiciones de Éxito o de Fracaso, mientras que las variables de menor contribución global, pero explicativas en cuanto a su peso relativo, generan límites pero no son capaces de generar por si solas tratamientos exitosos o fracasos. El resto de las variables se mantienen en el mismo rango de valores que en el caso de Reglas de Éxito. PSA preimplante menor a 1,5, edad mayor a 69,5 y PSA diagnosticado menor a 18. Para el valor del atributo PSA preimplante mayor a 1,5 y volumen ecográfico menor a 45 aparecen dos casos. Si la edad es menor a 61,5 años el tratamiento fracasa, pero si la edad supera 61,5 para que fracase el PSA diagnosticado debe ser mayor a 18.
Por ultimo encontramos una última regla ligada también al fracaso. En el caso que el PSA preimplante sea menor que 1,5 y la edad supere los 69,5 años para que fracase el PSA diagnosticado debe ser mayor que 18. Si esto sucede no influye que valor tenga el Gleason.

6 CONCLUSIONES

6.1 Conclusiones del problema

Resulta interesante observar que los resultados obtenidos a través de estas herramientas son consistentes con los resultados surgidos de las Redes Bayesianas:

- Tanto en el análisis realizado mediante redes bayesianas como en el análisis de caracterización-inducción, el PSA preimplante resulta una variable de alto nivel de significación.
- En ambos análisis el valor del atributo Gleason resulta definitorio, es decir que el resultado final de un caso quedará determinado por el valor de esta variable independientemente de otros factores.
- Variables como el Volumen ecográfico y la edad tienen un segundo nivel de importancia, haciéndose representativas cuando los valores de Gleason y PSA preimplante no definen un estado.

6.2 Conclusiones del aprendizaje

- La minería de datos basada en la metodología CRISP-DM se aplica satisfactoriamente a este proyecto.
- El aporte de nueva información hará que el sistema sea más confiable y representativo. El aprendizaje es acumulativo y continuo.
- Se verifican relaciones entre las variables y agrupación de datos que no habrían podido ser definidas sin el uso de minería de datos.

6.3 Futuras líneas de investigación

- En cuanto a la minería de datos, se deben profundizar los estudios, estableciendo criterios más sensibles de decisión. Para esto es fundamental contar con bases datos de gran cantidad y calidad de información.
- Respecto de la técnica médica, se debe trabajar en la forma de poder establecer o fijar la ubicación de los isótopos en la próstata.

7 BIBLIOGRAFÍA

[García Martínez et al, 2003] García Martínez, R.; Servente, M.; Pasquín, D.; 2003. ***Sistemas Inteligentes,*** Capítulo 1: "Aprendizaje Automático", Capítulo 2 "Redes Neuronales Artificiales"; Nueva Librería, Buenos Aires, Argentina.

[NNclass, 1998] Angshuman Saha; Testis Doctoral ***Application of Ridge Regression for Improved Estimation of Parameters in Compartmental Models***; Departamento de Estadística; Universidad de Washingtong; Agosto 1998 http://www.geocities.com/adotsaha/NN/SOMinExcel.html. Agosto 1998.

[Nnclust, 1998] Angshuman Saha; Testis Doctoral ***Application of Ridge Regression for Improved Estimation of Parameters in Compartmental Models***; Departamento de Estadística; Universidad de Washintong; Agosto 1998 http://www.geocities.com/adotsaha/NN/SOMinExcel.html. Agosto 1998.

[CTree, 1998] Angshuman Saha; Testis Doctoral ***Application of Ridge Regression for Improved Estimation of Parameters in Compartmental Models;*** Departamento de estadística; Universidad de Washintong; Agosto 1998 http://www.geocities.com/adotsaha/CTree/CtreeinExcel.htm

[Contraceptive Method Choice; 1987] National Indonesia Contraceptive Prevalence Contraceptive Method Choice; 1987. http://www.ics.uci.edu/~mlean/MLSummary.html

[Solar Flare Databases, 1989] Gary Bradshaw; Solar Flare Databases; 1989. http://www.ics.uci.edu/~mlean/MLSummary.html

[Redes Competitivas, 2000] Universidad Tecnológica de Pereira, Colombia, 2000 http://ohm.utp.edu.co/neuronales

[Servente et al, 2002] Servente, M; Dr. García Martínez, R; Tesis Doctoral ***Algoritmos TDIDT aplicado a la minería de datos inteligentes*** , Universidad de Buenos Aires, 2002

[Pearl, J., 1988] Pearl, J.; 1988; ***Probabilistic reasoning in intelligent systems***. Morgan Kaufmann, San Mateo,CA.

[Hernández O.J. et al, 2004] Hernández Orallo, J.; Ferri Ramírez, C.; Ramírez Quintana J.;2004;***Introducción a la minería de datos***; Capítulo 10: "Métodos Bayesianos"; PEARSON EDUCACION

[Beinlich, I. et al, 1989] Beinlich, I.; Suermondt, H.; Chavez, R.; Cooper, G.; 1989; ***The Alarm Monitoring System: A case study with two probabilistic inference techniques for belief networks***. In proceedings of the 2º European Conference on Artificial Intelligence in Medicine.

[Blurock, E., 1996] Blurock, E.; 1996; ***The ID3 Algorithm, Research Institute Institute for Symbolic Computation;*** www.risc.unilinz.ac.at/people/blurock/analysis/manual/document ; Australia.

[Breese & Blake, 1995] Breese, J.; Blake, R.; 1995; ***Automating computer bottleneck detection with belief nets***. Proceeding of the conference on Uncertainty in artificial Intelligence. Morga Kaufmann, San Francisco, CA.

[Fiszelew, A. y García *M*artínez, R. 2002] *Generación Automática de Redes Neuronales con Ajustede Parámetros Basado en Algoritmos Genéticos.* Revista del Instituto Tecnológico de Buenos Aires, 26: 76-101.

[Shariat, S; Zippe, C; Ludecke, G;Boman, H; Sanchez, M; Casella, R; Mian, C;Friedrich, M; Eissa, S; Akaza, H; Sawczuk, I; Serreta, V; Huland, H; Hedelin, H; Rupesh, R; Miyanaga, N; Sagalowsky, A; Wians, F; Roerhborn, C; Lotan, Y; Perrote, P; Benayoun, S; Marberger, M; Karakiewicz, P. 2005] *Nomograms including Nuclear matrix Protein 22 for prediction of disease recurrence and progression in patients with Ta, T1 or CIS transitional cell Carcinoma of bladder.* The Journal of Urology, Volume 173, Number 5, May 2005. The Journal of Urology, 173 (5): 1518-1525.

[Slawin, K; Kattan, M; Roehrborn, C; Wilson, T. 2006] *Development of nomogram to predict acute Urinary Retention or Surgical Intervention with or without dutasteride therapy, in men with benign prostatic hyperplasia.* Urology, 67(1): 84-89

[Stephan, C; Cammann, H; Jung, K. 2005] *Artificial Neural networks: Has the time come for their use in prostate cancer patients.* Nature, 2: 262-263.

[Yanke, B; Gonen, M; Scardino, P; Kattan, M. 2005] *Validation of nomogram predicting positive repeat biopsy for prostate cancer.* The Journal of Urology, 173(2): 421-424.

8. ANEXO

8.1 Set de datos para caracterización

A continuación se presenta el conjunto de datos utilizado en el proceso de análisis:

Nro	Edad	Delta T	PSA	Gleason	Estadío clínico	PSA preimp.	Vol. Ec.	Trat. Com.	Resultado
1	61	6	6,8	5	T1c	1,1	25	NO	Fracaso
2	63	5	25,5	7	T2a	1,1	44	NO	Fracaso
3	62	6	6,9	6	T1c	0,1	29,6	NO	Éxito
4	65	6	7,1	6	T1c	5,1	77	NO	Éxito
5	61	6	12,5	4	T1c	7,3	58	NO	Éxito
6	71	6	9,5	5	T1c	1	26,7	NO	Éxito
7	58	6	22,4	6	T1c	5	70	NO	Éxito
8	68	6	16	6	T1c	5,5	67,8	NO	Éxito
9	65	6	10,4	5	T1c	2,1	65	NO	Éxito
10	65	6	4	6	T1c	0,7	68	SI	Éxito
11	68	5	6,1	4	T1c	0,6	46,8	SI	Éxito
12	70	6	5,4	5	T2a	2,4	29,3	NO	Fracaso
13	76	5	6,7	4	T1c	0,3	35,3	NO	Éxito
14	78	5	11,3	4	T1c	1,6	76	NO	Éxito
15	71	4	13	6	T1c	0,3	48,9	NO	Éxito
16	71	4	14,4	5	T1c	8,3	48	NO	Éxito
17	68	3	17,3	4	T1c	5,4	32,5	NO	Fracaso
18	63	5	8,3	6	T1c	5,9	31	NO	Éxito
19	67	4	6	7	T1c	0,1	29	NO	Fracaso
20	62	3	5,9	7	T1c	1,8	22,3	NO	Éxito
21	67	5	10,8	4	T2a	0,11	42,8	NO	Fracaso
22	71	3	10,2	5	T1c	2,6	28	NO	Éxito
23	73	5	16,4	5	T1c	0,1	50,1	NO	Éxito
24	62	5	11,4	6	T1c	0,2	40,6	SI	Éxito
25	77	5	28	6	T1c	4,6	31,5	SI	Éxito
26	71	2	24,2	7	T1c	0,5	19,5	NO	Fracaso
27	67	4	22,76	8	T1c	3,45	21,9	SI	Éxito
28	66	5	8,8	7	T1c	2,5	19,8	NO	Éxito
29	65	5	7,8	4	T1c	1,7	26	NO	Fracaso
30	63	4	9,2	6	T2a	6,2	33	NO	Fracaso
31	75	5	8	6	T1c	0,2	17,8	NO	Fracaso
32	69	5	2,9	5	T2a	0,1	38,3	NO	Éxito
33	76	4	9,5	7	T1c	0,1	32	NO	Fracaso
34	73	4	15	4	T1c	0,8	20,3	NO	Éxito
35	77	3	7,06	5	T1c	0,1	21	NO	Éxito
36	69	4	16,9	7	T1c	0,79	77	NO	Éxito
37	65	5	1	7	T2a	0,1	20	NO	Éxito
38	80	5	10,9	4	T1c	10	42,7	NO	Éxito
39	68	5	5,5	6	T1c	3,6	70,9	NO	Éxito
40	66	5	13	6	T1c	0,1	17	NO	Éxito
41	70	5	4,6	7	T2a	0,1	30,2	SI	Fracaso
42	67	5	7,3	6	T1c	11,2	49	NO	Éxito

Nro	Edad	Delta T	PSA	Gleason	Estadío clínico	PSA preimp.	Vol. Ec.	Trat. Com.	Resultado
43	75	5	12	7	T2c	0,5	94	NO	Éxito
44	71	4	9,5	6	T1c	3,8	27	NO	Fracaso
45	52	4	3,2	6	T2a	0,1	32	NO	Éxito
46	76	4	3,3	7	T2a	3,3	17,5	NO	Éxito
47	69	0	7,6	6	T1c	0,1	43,5	NO	Éxito
48	70	5	13,8	7	T1c	0,7	42,5	NO	Fracaso
49	75	4	8,1	7	T2b	12,3	80,6	NO	Éxito
50	75	4	6,5	7	T2a	1,51	74	SI	Fracaso
51	79	4	12,3	5	T2a	2	28	NO	Éxito
52	71	4	11	7	T1c	0,4	31	NO	Éxito
53	68	2	10	6	T1c	0,1	30	NO	Fracaso
54	66	4	5,4	6	T1c	0,4	30	NO	Éxito
55	71	4	8,4	6	T1c	0,1	40	NO	Éxito
56	61	4	9,5	7	T1c	0,7	37	SI	Éxito
57	66	4	6,1	6	T1c	6,1	43,8	NO	Éxito
58	67	4	7,1	5	T1c	2,6	42	NO	Éxito
59	71	4	9	6	T1c	0,1	27	NO	Éxito
60	68	4	10,3	6	T1c	1,8	42	NO	Éxito
61	70	4	5,7	6	T1c	0,4	29	NO	Éxito
62	69	4	23	6	T1c	0,2	39	SI	Éxito
63	76	4	19	7	T1c	3,9	61	SI	Fracaso
64	65	2	12	6	T1c	0,1	40	SI	Éxito
65	71	4	7,36	6	T2a	2,7	65	NO	Éxito
66	74	4	6,7	5	T1c	0,1	26	NO	Fracaso
67	66	4	6,2	7	T1c	6,2	30	NO	Fracaso
68	67	3	12,2	4	T1c	0,1	74	NO	Fracaso
69	69	0	4,2	6	T1c	0,1	27,8	NO	Éxito
70	64	1	9,2	6	T2a	0,2	25	NO	Fracaso
71	71	3	6,4	6	T1c	0,1	49	NO	Éxito
72	78	4	10	5	T1c	2,5	53	NO	Éxito
73	64	3	15,6	7	T2a	0,1	30	SI	Éxito
74	69	2	8,7	7	T1c	3,6	35,4	NO	Éxito
75	71	4	8,2	7	T1c	5,1	49	NO	Éxito
76	67	3	7,7	5	T2a	3,8	82,3	NO	Éxito
77	71	4	9,7	4	T1c	9,7	53,8	NO	Éxito
78	70	3	9,3	6	T1c	9,3	50	NO	Éxito
79	71	4	9,5	7	T1c	0,1	26,3	NO	Éxito
80	70	3	8,6	6	T1c	1,8	24,8	SI	Éxito
81	71	3	12	7	T1c	4,8	31	NO	Éxito
82	70	3	5,8	6	T1c	5,2	46,5	NO	Éxito
83	67	3	7,6	7	T1c	0,1	34,8	NO	Fracaso
84	72	3	12,9	6	T1c	0,1	67,8	NO	Éxito
85	66	3	19	7	T1c	10,6	57,4	SI	Éxito
86	67	3	6,2	6	T1c	6,2	25,4	NO	Éxito
87	66	1	8,1	7	T1c	0,1	38	NO	Éxito
88	78	0	15,4	6	T2a	0,1	41,9	NO	Éxito
89	75	3	10,2	6	T1c	0,1	52,4	NO	Éxito
90	78	3	3,2	6	T2a	0,1	22	NO	Éxito
91	61	0	18,2	7	T2a	4,1	21	SI	Éxito
92	74	3	4,9	6	T2a	0,1	30	NO	Éxito
93	62	3	9,2	7	T2a	0,1	29	NO	Éxito

Nro	Edad	Delta T	PSA	Gleason	Estadío clínico	PSA preimp.	Vol. Ec.	Trat. Com.	Resultado
94	75	3	10,2	6	T2b	1,15	50	NO	Éxito
95	81	1	12	6	T1c	0,1	35	NO	Fracaso
96	72	3	6	6	T1c	0,1	24,9	NO	Éxito
97	70	3	7,5	6	T1c	3,5	45	NO	Fracaso
98	73	3	15,3	7	T1c	0,1	33	SI	Éxito
99	68	2	5,4	7	T1c	0,7	55	NO	Éxito
100	74	0	8,29	4	T1c	8,29	35	NO	Éxito
101	71	2	11,3	6	T1c	2	49	SI	Éxito
102	78	2	0,9	5	T2a	0,6	60	NO	Éxito
103	61	2	6,85	5	T1c	0,1	29	NO	Éxito
104	67	2	9,2	6	T1c	2	35	NO	Éxito
105	76	2	4,1	7	T2a	0,7	46	NO	Éxito
106	73	2	12,4	5	T1c	0,8	20	NO	Éxito
107	56	1	9,8	6	T1c	0,1	37,7	NO	Fracaso
108	65	1	3,2	5	T1c	0,1	27,7	NO	Éxito
109	71	1	5,1	8	T2a	0,1	43	NO	Éxito
110	59	1	6,24	6		0,1	38	NO	Fracaso
111	62	1	11,3	7	T1c	0,1	46,4	SI	Fracaso
112	74	1	7,5	6	T1c	0,1	28	NO	Éxito
113	58	1	10	6	T1c	0,1	20	NO	Éxito
114	66	1	8,2	6	T1c	0,1	27	NO	Éxito
115	67	1	3,4	6	T2a	0,1	25	NO	Éxito
116	64	1	14	6	T1c	0,1	40,5	SI	Éxito

8.2 Criterio de categorización de variables

A continuación se presentan las categorizaciones utilizadas, para la etapa de clasificación, en las variables PSA, PSA preimplante y Volumen ecográfico prostático:

Categorías	PSA
> 0 y <= 2	1
> 2 y <= 4	2
> 4 y <= 6	3
> 6 y <= 8	4
> 8 y <= 10	5
> 10 y <= 12	6
> 12 y <=16	7
> 16 y <= 20	8
> a 20	9

Categorias	PSA Preimpl.
> 0 y <= 1	1
> 1 y <= 2	2
> 2 y <= 4	3
> 4 y <= 6	4
> 6 y <= 8	5
> a 8	6

Categorias	Vol. Prost.
> 17 y <= 30	1
> 30 y <= 60	2
> 60 y <= 80	3
> a 80	4

8.3 Set de datos para clasificación

A continuación se presenta el conjunto de datos utilizado en el proceso de análisis:

Nro	Cluster	Edad	Delta T Años	PSA diag.	Gleason	PSA preimp.	Vol. Ec.
1	1	71	3	9	7	1	1
2	1	67	4	9	8	3	1
3	1	61	6	4	5	2	1

Nro	Cluster	Edad	Delta T Años	PSA diag.	Gleason	PSA preimp.	Vol. Ec.
4	1	61	1	8	7	4	1
5	1	58	1	5	6	1	1
6	2	69	1	4	6	1	2
7	2	68	2	5	6	1	1
8	2	65	2	6	6	1	2
9	2	69	1	3	6	1	1
10	2	64	1	5	6	1	1
11	2	66	2	5	7	1	2
12	2	61	2	4	5	1	1
13	2	56	2	5	6	1	2
14	2	65	2	2	5	1	1
15	2	71	2	3	8	1	2
16	2	59	2	4	6	1	2
17	2	62	1	6	7	1	2
18	2	66	1	5	6	1	1
19	2	67	1	2	6	1	1
20	2	64	1	7	6	1	2
21	3	63	6	9	7	2	2
22	3	62	6	4	6	1	1
23	3	65	6	4	6	4	3
24	3	61	6	7	4	5	2
25	3	71	6	5	5	1	1
26	3	58	6	9	6	4	3
27	3	68	6	7	6	4	3
28	3	65	6	6	5	3	3
29	3	65	6	2	6	1	3
30	3	68	6	4	4	1	2
31	3	70	6	3	5	3	1
32	3	76	6	4	4	1	2
33	3	78	6	6	4	2	3
34	3	71	5	7	6	1	2
35	3	71	5	7	5	6	2
36	3	68	4	8	4	4	2
37	3	63	6	5	6	4	2
38	3	67	4	3	7	1	1
39	3	62	4	3	7	2	1
40	3	67	6	6	4	1	2
41	3	71	4	6	5	3	1
42	3	73	6	8	5	1	2
43	3	62	6	6	6	1	2
44	3	77	6	9	6	4	2
45	3	66	6	5	7	3	1
46	3	65	6	4	4	2	1
47	3	63	5	5	6	5	2
48	3	75	5	4	6	1	1
49	3	69	5	2	5	1	2
50	3	76	4	5	7	1	2
51	3	73	5	7	4	1	1
52	3	77	4	4	5	1	1
53	3	69	5	8	7	1	3
54	3	65	5	1	7	1	1

Nro	Cluster	Edad	Delta T Años	PSA diag.	Gleason	PSA preimp.	Vol. Ec.
55	3	80	5	6	4	6	2
56	3	68	5	3	6	3	3
57	3	66	5	7	6	1	1
58	3	70	5	3	7	1	2
59	3	67	5	4	6	6	2
60	3	75	5	6	7	1	4
61	3	71	5	5	6	3	1
62	3	52	5	2	6	1	2
63	3	76	5	2	7	3	1
64	3	70	5	7	7	1	2
65	3	75	5	5	7	6	4
66	3	75	5	4	7	2	3
67	3	79	4	7	5	2	1
68	3	71	5	6	7	1	2
69	3	66	5	3	6	1	1
70	3	71	5	5	6	1	2
71	3	61	5	5	7	1	2
72	3	66	5	4	6	5	2
73	3	67	5	4	5	3	2
74	3	71	5	5	6	1	1
75	3	68	5	6	6	2	2
76	3	70	4	3	6	1	1
77	3	69	5	9	6	1	2
78	3	76	5	8	7	3	3
79	3	71	4	4	6	3	3
80	3	74	4	4	5	1	1
81	3	66	5	4	7	5	1
82	3	67	4	7	4	1	3
83	3	71	4	4	6	1	2
84	3	78	4	5	5	3	2
85	3	64	4	7	7	1	1
86	3	69	3	5	7	3	2
87	3	71	4	5	7	4	2
88	3	67	4	4	5	3	4
89	3	71	4	5	4	6	2
90	3	70	4	5	6	6	2
91	3	71	4	5	7	1	1
92	3	70	4	5	6	2	1
93	3	71	4	6	7	4	2
94	3	70	4	3	6	4	2
95	3	67	4	4	7	1	2
96	3	72	4	7	6	1	3
97	3	66	4	8	7	6	2
98	3	67	3	4	6	5	1
99	3	78	1	7	6	1	2
100	3	75	3	6	6	1	2
101	3	78	3	2	6	1	1
102	3	74	4	3	6	1	1
103	3	62	4	5	7	1	1
104	3	75	3	6	6	2	2
105	3	81	2	6	6	1	2

Nro	Cluster	Edad	Delta T Años	PSA diag.	Gleason	PSA preimp.	Vol. Ec.
106	3	72	3	3	6	1	1
107	3	70	3	4	6	3	2
108	3	73	3	7	7	1	2
109	3	68	3	3	7	1	2
110	3	74	1	5	4	6	2
111	3	71	3	6	6	2	2
112	3	78	3	1	5	1	2
113	3	67	2	5	6	2	2
114	3	76	2	3	7	1	2
115	3	73	2	7	5	1	1
116	3	74	1	4	6	1	1