UNIVERSIDAD DE SANTIAGO DE CHILE
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA INFORMÁTICA

INFORME EXPERIENCIA 5 ANÁLISIS DE DATOS

ARBOLES DE DECISIÓN

Integrantes:

Marcela Rivera Castro

Kevin Alvarez Alvial

Profesor:

Max Chacón

Ayudante:

Adolfo Guzmán

Santiago - Chile

18 de abril de 2018

TABLA DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE CUADROS

CAPÍTULO 1. INTRODUCCIÓN

1.1 MOTIVACIÓN

Antiguamente cerca de los años 1880, se tenía completo desconocimiento a cerca de la glándula tiroides y sus funciones. No existía advertencia de lo importante que es para el organismo humano. Entre los conocimientos que se manejaban en aquel momento, se sabía del cretinismo y los casos de mixedema del adulto de Gull, sin embargo no se conocía su origen tiroideo.

En 1883 el cirujano Teodoro Emilio Kocher, realizó una publicación sobre las consecuencias funestas de la tiroidectomía radical. Resolviendo en 1888 que cretinismo, mixedema y "caquexia" postiroidectomía eran síndromes estrechamente relacionados, si no idénticos, y se debían los tres a la pérdida de la función tiroidea.(Aguirre, 2002)

En la actualidad se conocen muchas de las enfermedades relacionadas con la tiroides. En el presente informe se abordará el hipotiroidismo. Hipotiroidismo significa "poca hormona tiroidea". Ocurre cuando la glándula tiroidea esta dañada y no es capaz de producir las hormonas tiroideas suficientes para mantener el metabolismo del cuerpo normal. El exceso de TSH puede causar que la glándula tiroidea aumente de tamaño lo que se llama bocio. Existen otras causas de hipotiroidismo como las tiroiditis autoinmunes o virales que pueden generar el mismo cuadro final pero sin bocio.(de Endocrinología Facultad de Medicina UC, s.f.)

1.2 ORGANIZACIÓN DEL DOCUMENTO

El documento consta de cuatro secciones principales: marco teórico para entender el dominio del problema, desarrollo de la experiencia y finalmente conclusiones, donde se indica el aprendizaje obtenido a partir del desarrollo de la experiencia.

1.3 METODOLOGÍAS Y HERRAMIENTAS UTILIZADAS

- Para el estudio de los datos se utilizará el programa R studio.
- La base de datos a utilizar es: allhypo.data y allhypo.names

CAPÍTULO 2. MARCO TEÓRICO

2.1 ARBOLES DE DECISIÓN

Árboles de decisión es un método de aprendizaje de máquina que proporcionan un conjunto de reglas que se van aplicando sobre los ejemplos nuevos para decidir qué clasificación es la más adecuada a sus atributos. (Caparrini, 2017)

Los árboles están formados por un conjunto de nodos de decisión (interiores) y de nodos-respuesta (hojas), donde cada uno de ellos significan:

- Un nodo de decisión está asociado a uno de los atributos y tiene 2 o más ramas que salen de él, cada una de ellas representando los posibles valores que puede tomar el atributo asociado. De alguna forma, un nodo de decisión es como una pregunta que se le hace al ejemplo analizado, y dependiendo de la respuesta que de, el flujo tomará una de las ramas salientes.

- Un nodo-respuesta está asociado a la clasificación que se quiere proporcionar, y nos devuelve la decisión del árbol con respecto al ejemplo de entrada.

Para comprender mejor que es un árbol de decisión, se recomienda ver la figura 2-1, en ella se puede ver el ejemplo de un árbol de decisión el cual permite saber si una paciente de 30 años, puede sufrir o no de cáncer de mamas, esto haciendo uso de probabilidades, estas probabilidades de los resultados se muestran tras cada nodo de probabilidad (circulo) y la elección en los rectángulos. (Salvador, 2017)

2.2 ALGORITMO ÁRBOL DE DECISIÓN

Los algoritmos más conocidos son el ID3, el C4.5 (Quinlan 1993), C5.0 y CART (Classification And Regression Trees). No obstante, estos algoritmos son variaciones de un algoritmo genérico llamado “Greedy algorithm” que básicamente va desde la raíz hacia abajo (Top-Down) buscando de manera recursiva los atributos que generan el mejor árbol hasta encontrar el óptimo global con una estructura de árbol lo más simple posible. (Berrios, 2014)

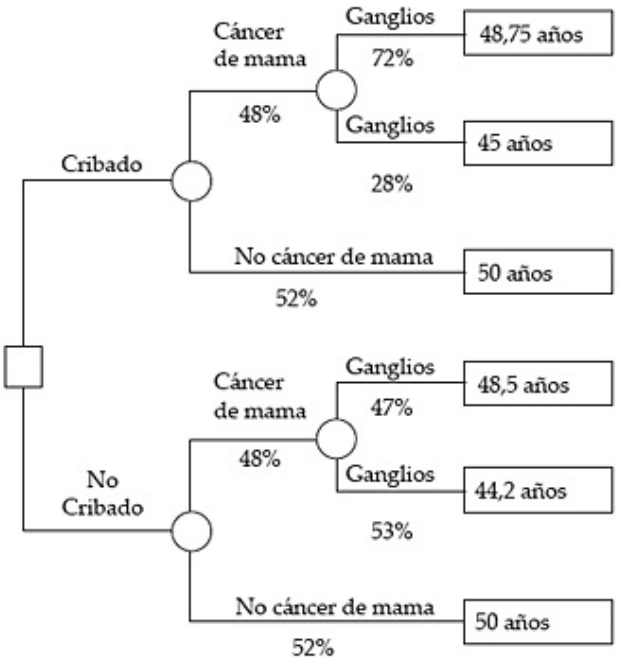

Figura 2-1: Árbol de decisión para la elección de cribado de cáncer de mama familiar.

Cabe destacar que en el presente estudio, se hace uso del algoritmo C5.0 disponible en la librería c50 de R. No obstante, para comprender mejor este algoritmo se procederá por explicar el ID3 y C4.5 ya que el algoritmo utilizado es una mejora de los mencionados recientemente.

2.2.1 Algoritmo ID3

Básicamente ID3 construye un árbol de decisión (DT) desde un set fijo de "ejemplos", el DT generado se usa para clasificar futuros ejemplos. Cada nodo de decisión es una prueba del atributo con otro árbol que comienza a partir de él. El algoritmo ID3 utiliza el criterio de la "ganancia de información" para decidir que atributo va en cada nodo de decisión. (Berrios, 2014)

2.2.2 Algoritmo C4.5

Básicamente es una versión avanzada del algoritmo ID3 en el que se incluyen las siguientes capacidades o ventajas:

1. Manejo de valores continuos y discretos: se crea un umbral para después dividir el atributo entre los que están sobre y bajo el umbral.

2. Tiene la capacidad de manejar valores de atributos faltantes.

3. Una vez construído el árbol, se finaliza con una "poda" de las ramas con el fin de simplicar el árbol generando resultados más fáciles de entender y haciéndolo menos dependiente de la data de prueba.

2.2.3 Algoritmo C5.0

Este algoritmo se basa en los anteriores siendo una mejora del C4.5, éste algoritmo construye los árboles en base a un conjunto de datos de entrenamiento optimizado bajo el criterio de ganancia de información.

Las mayores ventajas se encuentran en la eficiencia del tiempo de construcción de árbol, el uso de memoria y la obtención de árboles considerablemente más pequeños que en el C4.5 con la misma capacidad predictiva. (Berrios, 2014)

Cabe destacar que se ha escogido el C5.0 por la gran ventaja de obtener un árbol simple y con una alta capacidad predictiva.

2.3 GANANCIA DE INFORMACIÓN

Esta medida estadística mide que tan bien un atributo divide los ejemplos de entrenamiento en cada una de las clases seleccionando aquella con más información (información útil para separar) (Berrios, 2014). Para saber quién entrega más información se hace uso de la Entropía de Shannon, que de alguna forma mide el grado de incertidumbre de una muestra (Caparrini, 2017). Para medir la incertidumbre, se utiliza la ecuación 2.1.

$$E(S) = \sum_{i=1}^{C} -p_i log_2(p_i) \tag{2.1}$$

Donde S es el conjunto de muestras (el sistema analizado), C es el número de diferentes clasificaciones que usamos, y cada p_i es la proporción de ejemplos que hay de la clasificación i en la muestra.

En el caso particular de una clasificación binaria (ejemplos positivos / negativos), la fórmula anterior queda como:

$$E(S) = -Plog_2(P) - Nlog_2(N) \tag{2.2}$$

Donde P y N son, respectivamente, la proporción de ejemplos positivos y negativos. (Caparrini, 2017)

Cabe destacar que cuando E(S) = 0, entonces el set S está perfectamente clasificado.

Finalmente de la definición anterior se desprende que la "ganancia de información" corresponde a la diferencia en entropía entre antes de haber separado la data S por un atributo versus después de hacerlo, es decir, en cuanto se redujo la incertidumbre en el set S después de dividirla en el atributo "A":

$$IG(S, A) = E(S) \sum_{i=1}^{C} \frac{|S_v|}{|S|} E(x) \tag{2.3}$$

Donde S_v el set de S para el cual el atributo A tiene el valor v. Los elementos $|S_v|$ y $|S|$ corresponden al número de observaciones en S_v y S respectivamente. (Berrios, 2014)

CAPÍTULO 3. OBTENCIÓN DEL ÁRBOL

3.1 PRE-PROCESAMIENTO

Dado a que en el dataset utilizado para el estudio existen variedad de datos tomados de forma aleatoria, es necesario realizar un análisis previo para verificar que los datos sean consistentes y no hayan anomalías que afecten al resultado del estudio.

3.1.1 Eliminación de registros NA

Dentro de los datos del dataset puede ocurrir que algunos de éstos datos sean nulos para ciertas variables, es decir, que no fueron tomados o que posteriormente fueron eliminados, éstos datos son representados en el dataset como "?" y es a lo que se llama registro NA o desconocido. En este caso, si una observación tiene un registro NA, ésta se elimina.

3.1.2 Eliminación de variables

La eliminación de variables puede darse debido a razones variadas, pero específicamente, para este caso el criterio utilizado, es que la variable no contenga ningún dato, que contenga solo registros desconocidos (NA) o que para el objetivo del estudio ésta no entregue mucha información.

3.1.2.1 TBG

Esta variable dentro del dataset tiene mediciones que no fueron realizadas, ya que todos los datos son registros desconocidos, dado esto, la variable no entrega información y por ende se puede eliminar.

3.1.2.2 Variables de medición

En el dataset, existen variables booleanas para señalar si un examen de hormona fue realizado o no, esta variable es usada para filtrar los registros NA, dado a que si es falsa, significa que el dato es desconocido, éstas variables son:

TSH measured, **T3 measured**, **TT4 measured**, **T4U measured**, **FTI measured**.

Estas variables por si mismas y luego del filtro, no entregan información útil para el objetivo del estudio, por lo que se tornan innecesarias y pueden ser eliminadas.

3.1.2.3 Fuente de referencia

Esta variable indica la fuente de la cual se obtuvo los datos de un sujeto (u observación) en particular, por lo que esta variable no entrega información útil, por lo que se puede

eliminar.

3.2 TRANSFORMACIÓN DE LOS DATOS

Para obtener el árbol, es necesario transformar los datos a variables nominales, para este caso, se hace una transformación de las variables continuas haciendo cortes, por lo que la nueva variable indica el nivel de acuerdo al valor que tenía la variable continua. Las transformaciones fueron realizadas de acuerdo a lo realizado en el laboratorio de reglas de asociación para poder realizar una comparación posteriormente entre los resultados.

3.2.1 Edad

Para el caso de la variable de edad, dentro de ésta se pueden encontrar tres grupos: niños, adultos y ancianos, los cuales se dividen de acuerdo a los rangos de edad, los cuales son:

Tabla 3.1: Rangos de edad.

Grupo	**Mínimo**	**Máximo**
Niño	0	17
Adulto	18	59
Anciano	60	-

Para representar estos datos, se agregan tres variables binarias nuevas a la base de datos, las cuales son: age.child, age.adult, age.oldman, las cuales representan si el sujeto es niño, adulto o anciano. Cabe destacar que solo una de estas puede ser verdadera por observación.

3.2.2 Hormonas

Para el caso de las hormonas se analiza si ésta esta en el rango aceptable, para esto se puede verificar si la medición está por debajo del rango normal, o por encima, para ello se crean dos variables binarias nuevas en la base de datos para cada hormona con el sufijo "$under$" para indicar si el valor está por debajo del normal, u "$over$" para indicar si está por encima de lo normal. Cabe destacar que si un sujeto tiene niveles hormonales normales ambas variables van a ser falsas.

Los rango para la transformación son:

Tabla 3.2: Rangos de valores para hormonas.

Hormona	Rango Mínimo (under)	Rango Máximo (over)
TSH	0.4	4.0
T3	1.07	3.37
TT4	64	164
T4U	0.7	1.8
FTI	33.108	135.191

3.3 ÁRBOL

Para obtener el árbol, se hace uso de la biblioteca C50. Ésta retorna un objeto de la clase C5.0 el cual puede devolver el árbol generado, la matriz de confusión, reglas, entre otros.

Para un correcto uso de esta biblioteca, es necesario definir bien la raíz del árbol, es decir, el atributo experto que permitirá obtener resultados robustos y más cercanos a la realidad, en este caso se ha escogido la variable clasification (ahora llamada hypothyroid), ya que ésta permite separar de mejor forma un paciente sano de uno enfermo. Para hacer el llamado al procedimiento que permite obtener el árbol, basta con hacer entrega de la base de datos y el atributo experto, para que la biblioteca haga la implementación del árbol. Finalmente el árbol obtenido es el de la figura 3-1.

Dentro de árbol se puede observar la importancia de cada variable en el modelo, para una mayor precisión, dado a que el modelo está basado en árboles de decision de Quinlan, se puede obtener la entropía de las variables mas importantes, que fueron usadas para la construcción de éste, y ver cual descrimina mejor la variable de clase, en la tabla 3.3 se muestran los valores.

Tabla 3.3: Entropía de variables.

Variable	Entropía
TSH.over	98.15 %
on.thyroxine	21.78 %
TT4.over	21.16 %

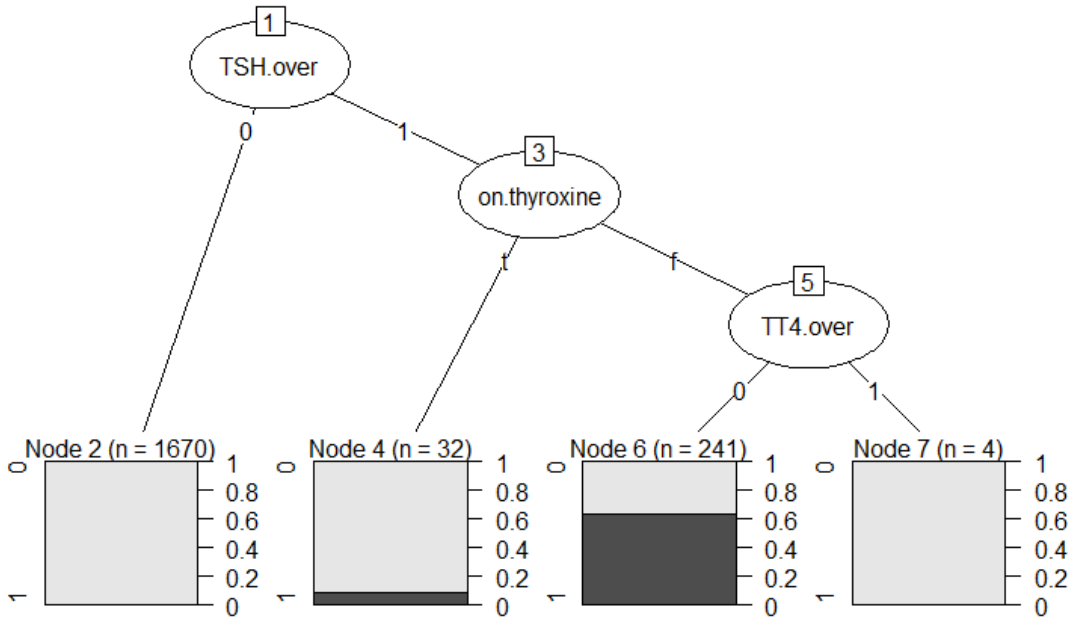

Figura 3-1: Árbol obtenido.

El modelo tambien puede ser expresado como un modelo basado en reglas, por lo que se pueden extraer éstas desde el árbol. Las reglas obtenidas de acuerdo al modelo fueron las siguientes:

Tabla 3.4: Reglas obtenidas del árbol.

Antecendentes	**Consecuente**	**Lift**
TSH.over = 0	class = 0	1.08076
TT4.over = 1	class = 0	1.08142
on.thyroxine = t	class = 0	1.06419
TSH.over = 1, on.thyroxine = f, TT4.over = 0	class = 1	7.91028

El consecuente "$class$" hace referencia a la variable de clase, en este caso "$hypothyroid$", que señala si un sujeto en cuestión tiene hipotiroidismo o no.

CAPÍTULO 4. ANÁLISIS DE RESULTADOS

4.1 ÁRBOL OBTENIDO

De acuerdo a lo obtenido en el modelo basado en arboles de decisión, se puede deducir que la variable más importante es la que hace referencia a valores altos de hormona TSH, siendo ésta una razón suficiente, pero no totalmente precisa, para discriminar si el paciente tiene hipotiroidismo, según el modelo para mejorar la precisión de la predicción es necesario incluir las variables TT4.over y on.thyroxine, ya que, por entropía, también describen parte del fenómeno en el modelo.

En consecuencia de lo anterior, si se analizan las reglas obtenidas del árbol (tabla 3.4) se puede observar que la que tiene mayor lift, por lo que tiene mayor calidad, es la última regla, la cual tiene como antecedentes las tres variables anteriormente mencionadas, y describen que, si se cumple la regla, el sujeto en cuestión tiene hipotiroidismo.

4.2 COMPARACIÓN CON REGLAS DE ASOCIACIÓN

Se puede hacer una comparación entre el método actual de clasificación y el método de reglas de asociación estudiado anteriormente, esto es posible, ya que de la representación del árbol se pueden extraer reglas de acuerdo a las decisiones y los valores de las variables para llegar a una conclusión de la variable de clase.

En el modelo actual se obtuvieron 4 reglas, de las cuales solo una se puede considerar causal para la clasificación, mientras que en la experiencia anterior con reglas de asociación, se obtuvieron un set de reglas, las cuales fueron filtradas por redundancia y calidad, las reglas obtenidas se pueden observar a continuación.

1. $\{I131.treatment=f,\ TSH.over=1,\ FTI.under=1\} \Rightarrow \{hypothyroid=1\}$
2. $\{thyroid.surgery=f,\ TSH.over=1,\ FTI.under=1\} \Rightarrow \{hypothyroid=1\}$
3. $\{I131.treatment=f,\ TSH.over=1,\ TT4.under=1,\ FTI.under=1\} \Rightarrow \{hypothyroid=1\}$
4. $\{thyroid.surgery=f,\ TSH.over=1,\ TT4.under=1,\ FTI.under=1\} \Rightarrow \{hypothyroid=1\}$

Se puede observar que la variable mas importante es TSH.over, la cual participa en todo el set de reglas y concuerda con lo presentado en el método actual de árboles de decisión,

otra variable que coincide, no explícitamente, pero si de forma semántica, es la que hace referencia a TT4, en árboles de decisión ésta se expresa como TT4.over=0 y en reglas de asociación como TT4.under=1, lo cual es semánticamente lo mismo, haciendo referencia a un nivel bajo de hormona tiroxina (T4). Existe discordancia en los modelos, para las otras variables, como es on.thyroxine, thyroid.surgery y FTI, dado que éstas aparecen solo en uno de los dos modelos, es decir, mientras son consideradas relevantes en uno, para el otro no describen el cuadro, no siendo buenas para discriminar la condición de un sujeto.

De acuerdo a la comparación realizada entre los métodos, es posible afirmar que la clasificación del cuadro de hipotiroidismo está principalmente definido por el nivel de hormona TSH, la cual presenta niveles altos, y también en parte por los niveles de hormona T4, que se presentan como bajos.

CAPÍTULO 5. CONCLUSIONES

Dentro de los aprendizajes de esta experiencia, es importante destacar la efectividad de los árboles de decisión, ya que éste a partir del nivel de información que aportan las variables, decide como clasificar y generar un árbol sencillo de analizar y con una alta precisión.

Otra gran ventaja es la simpleza que hay para implementar el algoritmo C5.0, esto porque R posee librerías que implementan de manera eficiente el algoritmo, permitiendo al programador y analizador abstraerse de todo el procedimiento para generar el árbol. Por otro lado, como se mencionó en el marco teórico, C5.0 permite trabajar con variables reales y/o continuas (implementado desde el algoritmo C4.5) permitiendo que se adapte mejor a la base de datos que se utiliza en este estudio, ya que en experiencias anteriores esta fue una dificultad a causa de las reglas redundantes que se obtuvieron con reglas de asociación.

Por otro lado, al comparar la implementación de ambos métodos, es posible notar una gran diferencia en la cantidad de reglas obtenidas, esto porque con reglas de asociación se obtuvieron de 8367 reglas, las cuales pueden ser filtradas y determinar cuales son más relevantes gracias al uso de lift. No obstante, con árboles de decisión se obtuvieron 4 reglas, donde aparece como la más importante TSH lo cual coincide en ambas implementaciones.

Otra ventaja del uso de árboles es la capacidad que tiene para manejar bases de datos con valores nulos, no obstante esto no fue explotado ya que como parte del preprocesamiento se incluyó eliminar de ante mano estos datos. Esto se hizo con la intención de asegurar que no haya información errónea en la base de datos, que pueda condicionar algunos casos.

Finalmente, al analizar los resultados obtenidos fueron concluyentes en cuanto a lo esperado, dado a que se observa una clara relación entre los niveles hormonales y padecer el cuadro de hipotiroidismo, ya que se puede observar que las variables mas involucradas en las reglas con mejores medidas de calidad son las que hacen referencia a los niveles de hormonas TSH y TT4, aún así en ninguna de las implementaciones se entregó información sobre los niveles hormonales de hormona T3, por lo que se puede decir que no tiene una

relación determinista con el cuadro.

Como aprendizaje, se obtuvo un mejor acercamiento con los algoritmos de clasificación ya que se han visto diversos métodos como reglas de asociación, clasificador Bayesiano y árboles de decisión, todos con el objetivo de clasificar sujetos según características propias del problema en estudio.

CAPÍTULO 6. BIBLIOGRAFÍA

Aguirre, C. P. (2002). Emil Theodor Kocher (1841-1917). Recuperado desde http://www.historiadelamedicina.org/kocher.html

Berrios, C. D. (2014). "APLICACIÓN DE ÁRBOLES DE DECISIÓN PARA LA ESTIMACIÓN DEL ESCENARIO ECONÓMICO Y LA ESTIMACIÓN DE MOVIMIENTO LA TASA DE INTERÉS EN CHILE". Recuperado desde http://repositorio.uchile.cl/bitstream/handle/2250/117556/Dupouy%5C%20Berrios%5C%20Carlos.pdf;sequence=1

Caparrini, F. S. (2017). Aprendizaje Inductivo: Árboles de Decisión. Recuperado desde http://www.cs.us.es/~fsancho/?e=104

de Endocrinología Facultad de Medicina UC, D. (s.f.). HIPOTIROIDISMO. Recuperado desde http://redsalud.uc.cl/ucchristus/VidaSaludable/Glosario/H/hipotiroidismo.act

Matsudo, N. L. (2001a). ÁRBOLES DE DECISIÓN, UNA TÉCNICA DE DATA MINING DESDE UNA PERSPECTIVA INFORMÁTICA y ESTADÍSTICA. Recuperado desde https://www.dc.uba.ar/academica/tesis-de-licenciatura/2001/matsudo.pdf

Matsudo, N. L. (2001b). ÁRBOLES DE DECISIÓN, UNA TÉCNICA DE DATA MINING DESDE UNA PERSPECTIVA INFORMÁTICA y ESTADÍSTICA. Recuperado desde https://www.dc.uba.ar/academica/tesis-de-licenciatura/2001/matsudo.pdf

Salvador, J. C. (2017). Medida de los resultados. Recuperado desde http://www.gestion-sanitaria.com/3-medida-resultados.html

CAPÍTULO 7. ANEXO: CÓDIGO EN R

```
#   rboles   de decisi n
library(C50)

preprocessing <- function(rawdata){
  #delete id from clasification|id column
  d <- c()
  for(i in rawdata$`clasification|id`){
    d <- c(d, strsplit(i, ".", fixed = TRUE)[[1]][1])
  }
  colnames(rawdata)[30] <- "clasification"
  rawdata$clasification <- as.factor(d)
  #data pre-processing
  #delete NA values
  data <- rawdata[(rawdata$age!="?" & rawdata$sex!="?" & rawdata$on.
      thyroxine!="?" & rawdata$query.on.thyroxine!="?" & rawdata$on.
      antithyroid.medication!="?" & rawdata$sick!="?" & rawdata$pregnant
      !="?" & rawdata$thyroid.surgery!="?" & rawdata$I131.treatment!="?"
       & rawdata$query.hypothyroid!="?" & rawdata$query.hyperthyroid!="?
      " & rawdata$lithium!="?" & rawdata$goitre!="?" & rawdata$tumor!="?
      " & rawdata$hypopituitary!="?" & rawdata$psych!="?" & rawdata$TSH.
      measured!="f" & rawdata$T3.measured!="f" & rawdata$TT4.measured!="
      f" & rawdata$T4U.measured!="f" & rawdata$FTI.measured!="f"),]
  #delete variable TBG
  data$TBG.measured <- NULL
  data$TBG <- NULL
  #delete measuring variables
  data$TSH.measured <- NULL
  data$T3.measured <- NULL
  data$TT4.measured <- NULL
  data$T4U.measured <- NULL
  data$FTI.measured <- NULL
  #delete referral.source variable
```

```
data$referral.source <- NULL

#data format transform
#nominal variables
data$sex <- as.factor(data$sex)
data$on.thyroxine <- as.factor(data$on.thyroxine)
data$query.on.thyroxine <- as.factor(data$query.on.thyroxine)
data$on.antithyroid.medication <- as.factor(data$on.antithyroid.
    medication)
data$sick <- as.factor(data$sick)
data$pregnant <- as.factor(data$pregnant)
data$thyroid.surgery <- as.factor(data$thyroid.surgery)
data$I131.treatment <- as.factor(data$thyroid.surgery)
data$query.hypothyroid <- as.factor(data$query.hypothyroid)
data$query.hyperthyroid <- as.factor(data$query.hyperthyroid)
data$lithium <- as.factor(data$lithium)
data$goitre <- as.factor(data$goitre)
data$tumor <- as.factor(data$tumor)
data$hypopituitary <- as.factor(data$hypopituitary)
data$psych <- as.factor(data$psych)

#continuous variables
data$age <-as.numeric(data$age)
data$TSH <- as.numeric(data$TSH)
data$T3 <- as.numeric(data$T3)
data$TT4 <- as.numeric(data$TT4)
data$T4U <- as.numeric(data$T4U)
data$FTI <- as.numeric(data$FTI)

#continuous variables and clasification to binary
#age values
child.adult_border <- 18
adult.oldman_border <- 60
#min values
```

```
TSH.min <- 0.4
T3.min <- 1.07
TT4.min <- 64.0
T4U.min <- 0.7
FTI.min <- 33.108
#max values
TSH.max <- 4.0
T3.max <- 3.37
TT4.max <- 154.0
T4U.max <- 1.8
FTI.max <- 135.191

#vectors(zeros)
child <- integer(length(data[[1]]))
adult <- integer(length(data[[1]]))
oldman <- integer(length(data[[1]]))
TSH.under <- integer(length(data[[1]]))
T3.under <- integer(length(data[[1]]))
TT4.under <- integer(length(data[[1]]))
T4U.under <- integer(length(data[[1]]))
FTI.under <- integer(length(data[[1]]))
TSH.over <- integer(length(data[[1]]))
T3.over <- integer(length(data[[1]]))
TT4.over <- integer(length(data[[1]]))
T4U.over <- integer(length(data[[1]]))
FTI.over <- integer(length(data[[1]]))

#change data to binary
for(i in 1:length(data[[1]])){
  #age
  if(data$age[i] < child.adult_border){
    child[i] <- 1
  }else if(data$age[i] >= child.adult_border & data$age[i] < adult.
      oldman_border){
```

```
    adult[i] <- 1
  }else if(data$age[i] >= adult.oldman_border){
    oldman[i] <- 1
  }
  #hormones
  if(data$TSH[i] >= TSH.max){
    TSH.over[i] <- 1
  }else if(data$TSH[i] <= TSH.min){
    TSH.under[i] <- 1
  }
  if(data$T3[i] >= T3.max){
    T3.over[i] <- 1
  }else if(data$T3[i] <= T3.min){
    T3.under[i] <- 1
  }
  if(data$TT4[i] >= TT4.max){
    TT4.over[i] <- 1
  }else if(data$TT4[i] <= TT4.min){
    TT4.under[i] <- 1
  }
  if(data$T4U[i] >= T4U.max){
    T4U.over[i] <- 1
  }else if(data$T4U[i] <= T4U.min){
    T4U.under[i] <- 1
  }
  if(data$FTI[i] >= FTI.max){
    FTI.over[i] <- 1
  }else if(data$FTI[i] <= FTI.min){
    FTI.under[i] <- 1
  }
}

#Discretizaci n de las variables
#data$age <- cut(data$age, breaks =  seq(0,100,20))
```

```
#data$TSH <- cut(data$TSH, breaks = c(0, 0.5, 4.7, Inf), labels = c
    ("Inf", "Nor", "Sup"))
#data$T3 <- cut(data$T3, breaks = c(0, 0.9, 2.8, Inf), labels = c("
    Inf", "Nor", "Sup"))
#data$TT4 <- cut(data$TT4, breaks = c(0, 58, 161, Inf), labels = c("
    Inf", "Nor", "Sup"))
#data$T4U <- cut(data$T4U, breaks = c(0, 0.8, 1.3, Inf), labels = c
    ("Inf", "Nor", "Sup"))

data$clasification <- ifelse(data$clasification %in% c("primary
    hypothyroid", "secondary hypothyroid", "compensated hypothyroid"),
     1, 0)

#replace vectors on data frame
data$age <- NULL
data$TSH <- NULL
data$T3 <- NULL
data$TT4 <- NULL
data$T4U <- NULL
data$FTI <- NULL
data$age.child <- as.factor(child)
data$age.adult <- as.factor(adult)
data$age.oldman <- as.factor(oldman)
data$TSH.over <- as.factor(TSH.over)
data$T3.over <- as.factor(T3.over)
data$TT4.over <- as.factor(TT4.over)
data$T4U.over <- as.factor(T4U.over)
data$FTI.over <- as.factor(FTI.over)
data$TSH.under <- as.factor(TSH.under)
data$T3.under <- as.factor(T3.under)
data$TT4.under <- as.factor(TT4.under)
data$T4U.under <- as.factor(T4U.under)
data$FTI.under <- as.factor(FTI.under)
data$clasification <- as.factor(data$clasification)
```

```
  names(data)[names(data) == "clasification"] <- "hypothyroid"

  return(data)
}

#Data read
rawdata <- read.csv("allhypo.data", header = FALSE, sep = ",",
    stringsAsFactors = FALSE)
colnames(rawdata) <- c("age", "sex", "on.thyroxine", "query.on.
    thyroxine", "on.antithyroid.medication", "sick", "pregnant", "
    thyroid.surgery", "I131.treatment", "query.hypothyroid", "query.
    hyperthyroid", "lithium", "goitre", "tumor", "hypopituitary", "psych
    ", "TSH.measured", "TSH", "T3.measured", "T3", "TT4.measured", "TT4"
    , "T4U.measured", "T4U", "FTI.measured", "FTI", "TBG.measured", "TBG
    ", "referral.source", "clasification|id")

#Preprocessing for data
data <- preprocessing(rawdata)
data_tree <- subset(data, select = -hypothyroid)



#Model
model <- C5.0(data_tree, data$hypothyroid)
model_rules <- C5.0(data_tree, data$hypothyroid, rules = T)

#Summary
summ_model <- summary(model_rules)

#Tree plot
plot(model)
```