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

Laboratorio 4: Clasificador Bayesiano

Integrantes: Nicole Reyes
Diego Águila

Curso: Análisis de Datos

Profesor: Max Chacón

Ayudante: Francisco Muñoz

24 de Junio de 2019

Tabla de contenidos

Índice de figuras

Índice de cuadros

1. Introducción

Obtener conocimiento de un conjunto de datos es fundamental al momento de realizar un estudio o análisis con respecto a estos, existen diversas técnicas y estrategias en la minería de datos para lograrlo, las obtención de clasificadores bayesianos es una de las técnicas conocidas para lograr esto y permite establecer que la presencia o ausencia de una característica particular, no está relacionada con la presencia o ausencia de otra característica, es decir, un clasificador de bayes considera que cada característica contribuye de manera independiente a la probabilidad de que algo se cumpla.

Esta cuarta experiencia de laboratorio consiste en, mediante el proceso de minería de datos y a través de la aplicación de clasificadores bayesianos obtener conocimiento de la base de datos asignada en la primera experiencia "*Allhypo*" correspondiente al hipotiroidismo.

1.1. Objetivos

1. Profundizar el entendimiento de los clasificadores baysianos como medida de obtención de conocimiento.

2. Corroborar la teoría vista en clases mediante la aplicación del clasificador bayesiano en el lenguaje estadístico R, concretamente en el paquete "$e1071$".

3. Obtener conocimiento de la base de datos "*Allhypo*" a través de esta técnica de minería de datos.

A continuación se define un marco teórico con la descripción del clasificador bayesiano, y la probabilidad apriori y aposteriori. Luego se presentan los resultados de la ejecución del script en R con la obtención de los clasificadores bayesanos para posteriormente realizar un análisis de los resultados y finalizar con las conclusiones respecto a esta experiencia.

2. Marco Teórico

Para contextualizar el ambiente de desarrollo relacionado a esta experiencia se disponen de las definiciones de los siguientes conceptos.

2.1. Clasificador Bayesiano Ingenuo

Este clasificador es de tipo probabilístico el cual es muy utilizado para resolver problemas de clasificación. Su funcionamiento está basado en el Teorema de Bayes, el cual corresponde al teorema que permite encontrar la probabilidad a posteriori de eventos suponiendo que se conoce la probabilidad a priori de este evento, la probabilidad condicional de una o mas condiciones dada el evento y la probabilidad a priori de las condiciones mencionadas anteriormente; como también esta basado en la suposición de que las variables predictoras son independientes, la cual es la razón de que este clasificador se considere ingenuo.

2.2. Probabilidad a priori

Esta corresponda a la probabilidad de ocurrencia de un evento sin condiciones adicionales. Si se ve de forma mas formal: La probabilidad a priori p(c_{ie}) será la probabilidad de que el sujeto se clasifique en la clase c_i, representado de la siguiente manera:

$$p(c_i) = \lim_{n \to +\infty} \frac{n_i}{n} \tag{1}$$

2.3. Probabilidad a posteriori

Corresponde a la probabilidad de ocurrencia de un evento dada ciertas condiciones. En palabras formales, se tiene que es la probabilidad de que un sujeto pertenezca a la clase c_i dado un valor y. Por ejemplo, para dos clases se tiene lo siguiente:

$$\sum_{i=1}^{2} p(c_i/y) = 1 \tag{2}$$

Para obtener la probabilidad $p(c_i/y)$, se requiere conocer las relaciones de probabilidades condicionales.

3. Obtención del Clasificador

3.1. Pre Procesamiento

Al igual que en las experiencias anteriores es fundamental elaborar una limpieza a la base de datos. En esta oportunidad la limpieza está compuesta por las siguientes etapas:

3.1.1. Eliminación de valores vacíos

Dentro de la base de datos, existen celdas con valores nulos (representados con el símbolo "?"), por lo que se opta por eliminar todos los sujetos de la muestra que contengan este símbolo para alguna variable ya que no aportan nada al análisis, para que así no entorpezca el estudio de esta base de datos.

3.1.2. Eliminación de variables

Dentro del conjunto de datos usados, hay columnas que no formarán parte del estudio, debido a que no aportan información importante para este pudiendo estropear el análisis. Las columnas correspondientes a la variables que son eliminadas son:

- Variable TBG: Corresponde a la proteína fijador de la Tiroxina (T4). Como toda su columna representa valores vacíos, no entrega información relevante debido a que se puede inferir a que no se midió la presencia de esta proteína en ningún paciente de la muestra, por lo tanto se opta por eliminarla de la base de datos.

- Variables de verificación de medición (measured): Indican si el paciente se ha realizado alguna medición, siendo esta información no relevante para el estudio. Lo que si importa es el valor numérico de estas mediciones de las hormonas por litro de sangre en los pacientes. Entonces, se decide dejar fuera las siguientes variables de medición: *TSH measured, T3 measuted, TT4 measured, T4U measured, FTI measured y TBG measured.*

- Variable de fuente: Como también sucede con las variables anteriores, la variable *referral.source* no entrega información que sea pertinente para el análisis ya que solo indica la fuente de la cual se obtuvo los datos de sujeto en cuestión.

3.2. Clase a utilizar

Una vez realizado el pre-procesamiento, se procede a encontrar la Clase a utilizar de acuerdo a lo que se quiere buscar con este análisis utilizando Clasificador Bayeasiano. Para esto, es importante plantear lo que se está buscando en relación a la información que se tiene de la base de datos. Entonces, la pregunta que se obtiene es:

¿Cuál es la probabilidad de que un sujeto padezca Hipotiroidismo a partir de información como la edad, sexo, resultado de sus exámenes hormonales, entre otros datos?

En base a esta pregunta, que una persona padezca de Hipotiroidismo no depende solo de un examen hormonal, por ejemplo, se tiene que esta condición depende de una serie de síntomas, que debido a la investigación que se realizó para la primera experiencia se pudo dar cuenta que los distintos atributos correspondientes a la base de datos *allhypo.data* de alguna manera se relacionan entre sí, como por ejemplo las medidas hormonales como la TSH y la TT4 o T4U. Entonces la conclusión que se obtiene es que con un solo atributo no se puede definir si un sujeto va a padecer o no de Hipotiroidismo. Si una persona tiene niveles altos o bajos de la hormona T3, no quiere decir que padecerá Hipotiroidismo, es necesario tener en cuenta otros factores para tener una certeza de que el diagnóstico será positivo. Dicho lo anterior, como atributo predictor se decide que la variable *State* es la mas indicada, la que muestra el cuadro de Hipotiroidismo para cada sujeto.

3.3. Conjuntos de entrenamiento y de prueba

Para poder llevar a cabo el modelo, se necesita un *Conjunto de entrenamiento* el cual se usa para la construcción y ajuste del modelo de acuerdo a información previa que se tiene. Y también es necesario de un *Conjunto de prueba (o test)* para luego aplicar el modelo. Para el *Conjunto de entrenamiento*, en vez de realizar algún método para obtener este conjunto, se decide ocupar la base de datos *allhypo.data*, lo mismo ocurre para el *Conjunto de prueba*, para evitar implementar un método y simplificar el trabajo, se decide trabajar con la base de datos *allhypo.test* asumiendo que los datos son representativos para los dos conjuntos. Ambas bases de datos utilizadas para estos conjuntos, se debe realizar el pre procesamiento para la limpieza de ellas.

4. Análisis de los resultados

Para observar los resultados del Clasificador, se tiene la Matriz de Confusión, en donde se va comparando la predicción de cada clase con las instancias de las clases reales, que en este caso corresponden las del conjunto de prueba.

A continuación en el Cuadro 1 se presenta la Matriz de confusión del Clasificador Bayesiano ingenuo. Las columnas corresponden a la clase real (conjunto de prueba) y las filas corresponden a las clasificación que se realiza.

Predicciones	**Compensated Hypothyroid**	**Negative**	**Primary Hypothyroid**
Compensated Hypothyroid	11	1	1
Negative	19	634	3
Primary Hypothyroid	0	2	25
Secondary Hypothyroid	0	0	0

Cuadro 1: Matriz de Confusión del clasificador.

Cabe destacar que en el Cuadro 1 en la columna correspondiente a las clases reales del conjunto de prueba no se aprecia el tipo de Hipotiroidismo denominado *Secondary Hypothyroid*, pero si en la que tiene ver con la clasificación, esto ocurre debido a que la base de datos que corresponde al conjunto de prueba, no se aprecian resultados de este tipo de Hipotiroidismo, solo se encuentran las otras 3 en el conjunto de prueba. Aunque esta situación no se ve a simple vista que no genera problema, se desconoce si realmente lo genera, debido a que está la posibilidad de que igual se realizara la clasificación aunque no se tenga el conjunto para realizarlo lo que provocaría un error en el clasificador. A partir de la Matriz de Confusión mostrada anteriormente, se pueden obtener las precisiones con respecto al rendimiento de este clasificador. Se obtuvieron 4 medidas distintas en donde se mide esta precisión, una correspondiente a una precisión general y las otras 3 en relación a los valores predictivos, la cuales son:

1. **Precisión General**: Ésta tiene una proporción de 0.9626437, correspondiendo a un 96.26 %.

2. **Valor Predictivo "Negative"**: Ésta tiene una proporción de 0.9664634, correspondiendo a un 96.64 %.

3. **Valor Predictivo "Primary Hypothyroid"**: Ésta tiene una proporción de 0.9259259, correspondiendo a un 92.59 %.

4. **Valor Predictivo "Compensated Hypothyroid"**: Ésta tiene una proporción de 0.8461538, correspondiendo a un 84.61 %.

Con respecto a estas medidas, se puede decir que el clasificador tiene una alta precisión, siendo la más alta la del Valor Predictivo *Negative* con un un error del solo del 3,35366 %. Si se realiza una comparación con los otros valores predictivos se tiene que la *Presición General* tiene un error del 3,73563 %, *Primary Hypothyroid* tiene un error del 7,40741 % y *Compensated Hypotyroid* tiene un error del 15,38462 %.

4.1. Comparación con experiencias anteriores

Con respecto al *Método de las K-medias* realmente es complicado realizar una comparación con esta experiencia que tiene que ver con *Clasificador Bayesiano* debido a que lo que se realiza en ambas difieren en su propósito y uso. En donde el objetivo de K-medias esta enfocado en realizar agrupamientos de datos de acuerdo a la distancia que hay entre ellos, en cambio el enfoque del Clasificador Bayesiano está en obtener probabilidades a priori y posteriori (Teorema de Bayes) para luego realizar la clasificación, y se le agrega el concepto de *ingenuo* que tiene que ver con la independencia entre las variables.

Los resultados obtenidos de las K-medias en donde se realizaron dos Clusters, en el primero, se pudo determinar que el factor mas notorio para la agrupación tenía que ver con el sexo del sujeto y para el segundo, se determinó que debido a la variación de las distintas variables correspondiente a las hormonas en las distintas agrupaciones, existe una alta correlación de estas variables. Los resultados correspondientes a esta experiencia, la clasificación se enfocó de acuerdo al cuadro del paciente.

Con respecto a lo realizado con las *Reglas de Asociación*, al igual que lo realizado en esta experiencia, ambas se enfocan en la clase o variable predicha, en donde se trata de

predecir el cuadro de un sujeto de acuerdo variables involucradas (en este caso, las que se encuentran en la base de datos utilizada para las experiencias). Lo que las diferencia, es que ambas utilizan metodologías distintas para realizar lo dicho anteriormente. Para las Reglas de Asociación, lo que se hace es obtener reglas que buscan relacionar distintas variables de tal manera que su asociación en conjunto, tenga como consecuencia la clase, en cambio, con el Clasificador Bayesiano ingenuo lo que se hace es, mediante a probabilidades, se busca la probabilidad a posteriori, que de acuerdo al Teorema de Bayes, se encuentra ésta suponiendo que se conoce la probabilidad a priori del evento.

Los resultados que muestran las Reglas de Asociación son las reglas que tienen el Hipotiroidismo como consecuente, sin especificar el tipo de Hipotiroidismo (Primary Hypothyroid, Secondary Hypothyroid o Compensated Hypothyroid), en cambio, los resultados del Clasificador son mas precisos de acuerdo a la clase se que eligió para la clasificación.

5. Conclusiones

Para obtener el clasificador bayesiano se realiza en primer lugar una limpieza a la base de datos, en la que se eliminan variables que no entregan información, tanto para el conjunto de datos de entrenamiento, como para el conjunto de prueba. Luego se utiliza el algoritmo naiveBayes para la obtención de las probabilidades dada la variable predictora que sería "State", correspondiente a la clasificación de hipotiroidismo del paciente, obtienen las predicciones con respecto al modelo obtenido con naiveBayes y se representan los resultados a través de una matriz de confusión.

Los objetivos de esta experiencia se cumplen, ya que al analizar los resultados se observa que el clasificador obtenido con respecto a la clase State posee altos valores predictivos y que la precisión general de éste es de 96,26 %.

Para el desarrollo de esta experiencia se presentan algunas dificultades, pero mayormente, la interpretación de los resultados obtenidos por el clasificador bayesiano.

6. Anexos

A continuación se presenta el script en R utilizado para el desarrollo de esta experiencia.

```
library("e1071")
# LIMPIEZA DE LA BASE DE DATOS DEL CONJUNTO DE ENTRENAMIENTO
data.entrenamiento <- read.csv("/home/nicole/allhypo.data",
                    col.names = c("age","sex", "on.thyroxine", "query.thyroxine",
    "on.antithyroid.med",
                                  "sick","pregnant", "thyroid.surgery", "I131", "
    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","state"), header=FALSE
    , sep=",", stringsAsFactors=FALSE)
# Se limpia la variable de estado del paciente eliminando numeros
state <- data.entrenamiento$state
state <- strsplit(state, ".", fixed = TRUE)
state <- lapply(state, `[[`, 1)
state <- unlist(state)
data.entrenamiento$state <- state
# Se eliminan los pacientes que contengan algun NA
# Sex
error = data.entrenamiento$sex == '?'
data.entrenamiento = data.entrenamiento[!error,]
# Age
error = data.entrenamiento$age == '?'
data.entrenamiento = data.entrenamiento[!error,]
# On thyroxine
error = data.entrenamiento$on.thyroxine == '?'
data.entrenamiento = data.entrenamiento[!error,]
# Query on antithyroid thyroxine
```

```
error = data.entrenamiento$query.thyroxine == '?'
data.entrenamiento = data.entrenamiento[!error,]
# On antithyroid medication
error = data.entrenamiento$on.antithyroid.med == '?'
data.entrenamiento = data.entrenamiento[!error,]
# Sick
error = data.entrenamiento$sick == '?'
data.entrenamiento = data.entrenamiento[!error,]
# Pregnant
error = data.entrenamiento$pregnant == '?'
data.entrenamiento = data.entrenamiento[!error,]
# Thyroid surgery
error = data.entrenamiento$thyroid.surgery == '?'
data.entrenamiento = data.entrenamiento[!error,]
# I131 treatment
error = data.entrenamiento$I131 == '?'
data.entrenamiento = data.entrenamiento[!error,]
# Query hypothyroid
error = data.entrenamiento$query.hypothyroid == '?'
data.entrenamiento = data.entrenamiento[!error,]
# Query hyperthyroid
error = data.entrenamiento$query.hyperthyroid == '?'
data.entrenamiento = data.entrenamiento[!error,]
# Lithium
error = data.entrenamiento$lithium == '?'
data.entrenamiento = data.entrenamiento[!error,]
# Goitre
error = data.entrenamiento$goitre == '?'
data.entrenamiento = data.entrenamiento[!error,]
# Tumor
error = data.entrenamiento$tumor == '?'
data.entrenamiento = data.entrenamiento[!error,]
# Hypopituitary
error = data.entrenamiento$hypopituitary == '?'
data.entrenamiento = data.entrenamiento[!error,]
# Psych
```

```
error = data.entrenamiento$psych == '?'
data.entrenamiento = data.entrenamiento[!error,]
# TSH
error = data.entrenamiento$TSH == '?'
data.entrenamiento = data.entrenamiento[!error,]
# T3
error = data.entrenamiento$T3 == '?'
data.entrenamiento = data.entrenamiento[!error,]
# TT4
error = data.entrenamiento$TT4 == '?'
data.entrenamiento = data.entrenamiento[!error,]
# T4U
error = data.entrenamiento$T4U == '?'
data.entrenamiento = data.entrenamiento[!error,]
# FTI
error = data.entrenamiento$FTI == '?'
data.entrenamiento = data.entrenamiento[!error,]
# Se eliminan las variables que indican si se realizo la medicion de la
    hormona
data.entrenamiento$T3.measured <- NULL
data.entrenamiento$TSH.measured <- NULL
data.entrenamiento$TT4.measured <- NULL
data.entrenamiento$T4U.measured <- NULL
data.entrenamiento$FTI.measured <- NULL
data.entrenamiento$TBG.measured <- NULL
#Se elimina la variable TBG que no fue medida para ningun paciente
data.entrenamiento$TBG <- NULL
#Se elimina la variable referral.source la cual no es relevante
data.entrenamiento$referral.source <- NULL
#
#
# LIMPIEZA DE LA BASE DE DATOS DEL CONJUNTO DE PRUEBA
data.prueba <- read.csv("/home/nicole/allhypo.test",
                        col.names = c("age","sex", "on.thyroxine", "query.
    thyroxine","on.antithyroid.med",
```

```
96                                      "sick","pregnant", "thyroid.surgery", "
    I131", "query.hypothyroid",
97                                      "query.hyperthyroid", "lithium", "goitre"
    , "tumor", "hypopituitary",
98                                      "psych", "TSH.measured", "TSH", "T3.
    measured", "T3", "TT4.measured",
99                                      "TT4","T4U.measured", "T4U", "FTI.
    measured","FTI", "TBG.measured",
100                                     "TBG", "referral.source","state"), header
    =FALSE, sep=",", stringsAsFactors=FALSE)
101 # Se limpia la variable de estado del paciente eliminando numeros
102 state <- data.prueba$state
103 state <- strsplit(state, ".", fixed = TRUE)
104 state <- lapply(state, `[[`, 1)
105 state <- unlist(state)
106 data.prueba$state <- state
107 # Se eliminan los pacientes que contengan algun NA
108 # Sex
109 error = data.prueba$sex == '?'
110 data.prueba = data.prueba[!error,]
111 # Age
112 error = data.prueba$age == '?'
113 data.prueba = data.prueba[!error,]
114 # On thyroxine
115 error = data.prueba$on.thyroxine == '?'
116 data.prueba = data.prueba[!error,]
117 # Query on antithyroid thyroxine
118 error = data.prueba$query.thyroxine == '?'
119 data.prueba = data.prueba[!error,]
120 # On antithyroid medication
121 error = data.prueba$on.antithyroid.med == '?'
122 data.prueba = data.prueba[!error,]
123 # Sick
124 error = data.prueba$sick == '?'
125 data.prueba = data.prueba[!error,]
126 # Pregnant
```

```
error = data.prueba$pregnant == '?'
data.prueba = data.prueba[!error,]
# Thyroid surgery
error = data.prueba$thyroid.surgery == '?'
data.prueba = data.prueba[!error,]
# I131 treatment
error = data.prueba$I131 == '?'
data.prueba = data.prueba[!error,]
# Query hypothyroid
error = data.prueba$query.hypothyroid == '?'
data.prueba = data.prueba[!error,]
# Query hyperthyroid
error = data.prueba$query.hyperthyroid == '?'
data.prueba = data.prueba[!error,]
# Lithium
error = data.prueba$lithium == '?'
data.prueba = data.prueba[!error,]
# Goitre
error = data.prueba$goitre == '?'
data.prueba = data.prueba[!error,]
# Tumor
error = data.prueba$tumor == '?'
data.prueba = data.prueba[!error,]
# Hypopituitary
error = data.prueba$hypopituitary == '?'
data.prueba = data.prueba[!error,]
# Psych
error = data.prueba$psych == '?'
data.prueba = data.prueba[!error,]
# TSH
error = data.prueba$TSH == '?'
data.prueba = data.prueba[!error,]
# T3
error = data.prueba$T3 == '?'
data.prueba = data.prueba[!error,]
# TT4
```

```
error = data.prueba$TT4 == '?'
data.prueba = data.prueba[!error,]
# T4U
error = data.prueba$T4U == '?'
data.prueba = data.prueba[!error,]
# FTI
error = data.prueba$FTI == '?'
data.prueba = data.prueba[!error,]
# Se eliminan las variables que indican si se realizo la medicion de la
    hormona
data.prueba$T3.measured <- NULL
data.prueba$TSH.measured <- NULL
data.prueba$TT4.measured <- NULL
data.prueba$T4U.measured <- NULL
data.prueba$FTI.measured <- NULL
data.prueba$TBG.measured <- NULL
#Se elimina la variable TBG que no fue medida para ningun paciente
data.prueba$TBG <- NULL
#Se elimina la variable referral.source la cual no es relevante
data.prueba$referral.source <- NULL
#
#
# Una vez limpiada la base de datos, se transforman los datos correspondientes

# CONJUNTO DE ENTRENAMIENTO
# Variables continuas
data.entrenamiento$T3 <- as.numeric(data.entrenamiento$T3)
data.entrenamiento$age <- as.numeric(data.entrenamiento$age)
data.entrenamiento$TSH <- as.numeric(data.entrenamiento$TSH)
data.entrenamiento$TT4 <- as.numeric(data.entrenamiento$TT4)
data.entrenamiento$T4U <- as.numeric(data.entrenamiento$T4U)
data.entrenamiento$FTI <- as.numeric(data.entrenamiento$FTI)
# Variables binarias
data.entrenamiento$sex <- as.factor(data.entrenamiento$sex)
data.entrenamiento$sick <- as.factor(data.entrenamiento$sick)
data.entrenamiento$I131 <- as.factor(data.entrenamiento$I131)
```

```r
data.entrenamiento$state <- as.factor(data.entrenamiento$state)
data.entrenamiento$tumor <- as.factor(data.entrenamiento$tumor)
data.entrenamiento$psych <- as.factor(data.entrenamiento$psych)
data.entrenamiento$goitre <- as.factor(data.entrenamiento$goitre)
data.entrenamiento$lithium <- as.factor(data.entrenamiento$lithium)
data.entrenamiento$pregnant <- as.factor(data.entrenamiento$pregnant)
data.entrenamiento$on.thyroxine <- as.factor(data.entrenamiento$on.thyroxine)
data.entrenamiento$hypopituitary <- as.factor(data.entrenamiento$hypopituitary
    )
data.entrenamiento$query.thyroxine <- as.factor(data.entrenamiento$query.
    thyroxine)
data.entrenamiento$thyroid.surgery <- as.factor(data.entrenamiento$thyroid.
    surgery)
data.entrenamiento$query.hypothyroid <- as.factor(data.entrenamiento$query.
    hypothyroid)
data.entrenamiento$query.hyperthyroid <- as.factor(data.entrenamiento$query.
    hyperthyroid)
data.entrenamiento$on.antithyroid.med <- as.factor(data.entrenamiento$on.
    antithyroid.med)

#CONJUNTO DE PRUEBA
# Variables continuas
data.prueba$T3 <- as.numeric(data.prueba$T3)
data.prueba$age <- as.numeric(data.prueba$age)
data.prueba$TSH <- as.numeric(data.prueba$TSH)
data.prueba$TT4 <- as.numeric(data.prueba$TT4)
data.prueba$T4U <- as.numeric(data.prueba$T4U)
data.prueba$FTI <- as.numeric(data.prueba$FTI)
# Variables binarias
data.prueba$sex <- as.factor(data.prueba$sex)
data.prueba$sick <- as.factor(data.prueba$sick)
data.prueba$I131 <- as.factor(data.prueba$I131)
data.prueba$state <- as.factor(data.prueba$state)
data.prueba$tumor <- as.factor(data.prueba$tumor)
data.prueba$psych <- as.factor(data.prueba$psych)
data.prueba$goitre <- as.factor(data.prueba$goitre)
```

```
data.prueba$lithium <- as.factor(data.prueba$lithium)
data.prueba$pregnant <- as.factor(data.prueba$pregnant)
data.prueba$on.thyroxine <- as.factor(data.prueba$on.thyroxine)
data.prueba$hypopituitary <- as.factor(data.prueba$hypopituitary)
data.prueba$query.thyroxine <- as.factor(data.prueba$query.thyroxine)
data.prueba$thyroid.surgery <- as.factor(data.prueba$thyroid.surgery)
data.prueba$query.hypothyroid <- as.factor(data.prueba$query.hypothyroid)
data.prueba$query.hyperthyroid <- as.factor(data.prueba$query.hyperthyroid)
data.prueba$on.antithyroid.med <- as.factor(data.prueba$on.antithyroid.med)
#
#
# SE PROCEDE A REALIZAR LO NECESARIO PARA EL CLASIFICADOR BAYESIANO INGENUO

# Se realizan las probabilidades dada la variable preditora que es la State (
    clasificaci n del Hipotiroidismo)
modelo <- naiveBayes(state ~., data = data.entrenamiento)
# Se realizan las predicciones con respecto al modelo anterior del conjunto de
     Prueba.
predicciones <- predict(object = modelo, newdata=data.prueba, type = "class")
# Se obtiene la matriz de confusion con los resultados anteriores y la
    variable predictora
matriz.de.confusion <- table(predicciones, data.prueba$state)
# A partir de la matriz de confusion se calculan las distintas precisiones con
     respecto a la clasificacion del Hipotiroidismo
# (correspondiente a la variable "State")
precision.en.general <- sum(diag(matriz.de.confusion)) / sum(matriz.de.
    confusion)
# Precision para Primary Hypothyroid (Hipotiroidismo Primario)
precision.primary <- sum(matriz.de.confusion[3,3]) / sum(matriz.de.confusion
    [3,])
# Precision para Negative
precision.negative <- sum(matriz.de.confusion[2,2]) / sum(matriz.de.confusion
    [2,])
# Presicion para Compensated Hypothyroid (Hipotiroidismo compensado o
    subclinico)
```

```
255 precision.compensated <- sum(matriz.de.confusion[1,1]) / sum(matriz.de.
    confusion[1,])
```

Bibliografía

Chacón, M. (2017). Clasificación bayesiana. [Online] http://www.udesantiagovirtual.cl/moodle2/mod/resource/view.php?id=156965.

Richter-Walsh, S. (2017). Clasificacion naive bayes en r (parte 2). [Online] https://www.r-bloggers.com/naive-bayes-classification-in-r-part-2.

Sucar, L. E. (2012). Sesión 6: Clasificadores bayesianos. [Online] https://ccc.inaoep.mx/~esucar/Clases-mgp/pgm06-clasif-2012.pdf.

Vega, J. B. M. (2018). Naïve bayes con r para clasificación de texto. [Online] https://rstudio-pubs-static.s3.amazonaws.com/378731_524d5e2c353c4dbebc8a23e4235a383a.html#ajustando-naive-bayes.