Thyroid Disease Data Set Analysis

Ulises Jeremias Cornejo Fandos[1] and Gaston Gustavo Rios[2]

[1] *Licenciatura en Informatica - 13566/7, Facultad de Informatica, UNLP*

[2] *Licenciatura en Informatica - 13591/9, Facultad de Informatica, UNLP*

compiled: July 16, 2018

Resumen

OCIS codes:

http://dx.doi.org/10.1364/XX.99.099999

1. Introducción

1.A. Thyroid Disease Data Set

El data set cuenta con un conjunto de 6 bases de datos. En general, estos conjuntos son muy similares y presentan muchos atributos, *aproximadamente 29 atributos cada conjunto de datos*, siendo la mayoría booleanos o reales. El dominio del problema es el análisis de enfermedades de tiroides y se utiliza para su estudio un conjunto de datos otorgado por Garavan Institute. Cada conjunto de datos cuenta con aproximadamente 2800 ejemplos y una *gran cantidad de datos faltante*.

El conjunto de datos seleccionado permite la clasificación de pacientes entre aquellos que tienen hipotiroidismo y aquellos que no. En su versión presenta un total de 3163 ejemplares de los cuales se conocen 26 atributos.

Información de los atributos

- *age*: Este atributo corresponde a la edad del paciente. En un valor numérico continuo que toma valores de R.
- *sex*: Este atributo corresponde al sexo del paciente. Toma valores del conjunto $\{M, F\}$ y presenta valores faltante.
- *on_thyroxine*: Toma valores del conjunto $\{f, t\}$.
- *query_on_thyroxine*: Toma valores del conjunto $\{f, t\}$.
- *on_antithyroid_medication*: Toma valores del conjunto $\{f, t\}$.
- *thyroid_surgery*: Toma valores del conjunto $\{f, t\}$.
- *query_hyphotiroid*: Toma valores del conjunto $\{f, t\}$.
- *query_hyperthyroid*: Toma valores del conjunto $\{f, t\}$.
- *pregnant*: Toma valores del conjunto $\{f, t\}$.
- *sick*: Toma valores del conjunto $\{f, t\}$.
- *tumor*: Toma valores del conjunto $\{f, t\}$.
- *lithium*: Toma valores del conjunto $\{f, t\}$.
- *goitre*: Toma valores del conjunto $\{f, t\}$.
- *TSH_measured*: Toma valores del conjunto $\{f, t\}$.
- *TSH*: Toma valores de R.
- *T3_measured*: Toma valores del conjunto $\{f, t\}$.
- *T3*: Toma valores de R.
- *TT4_measured*: Toma valores del conjunto $\{f, t\}$.
- *TT4*: Toma valores de R.
- *T4U_measured*: Toma valores del conjunto $\{f, t\}$.
- *T4U*: Toma valores de R.
- *FTI_measured*: Toma valores del conjunto $\{f, t\}$.
- *FTI*: Toma valores de R.
- *TBG_measured*: Toma valores del conjunto $\{f, t\}$.
- *TBG*: Toma valores de R.

1.B. Recolección del conjunto de datos

2. Pre-procesamiento de datos

2.A. Conjunto a analizar

Como se explica en la sección anterior, existen atributos en los cuales se muestra si fue medido un atributo o no. Dado que los valores de estos coinciden con los datos faltantes se opta por la eliminación de dichos atributos. Al mismo tiempo, la cantidad de datos faltantes para la columna correspondiente al atributo TBG es **PORCENTAJE y SOLUCION**. Teniendo en cuenta esto, se opta por la eliminación de dicha

columna dado que se considera que la misma podría llegar a interferir en el proceso del análisis de los datos, quedando finalmente un total de 19 atributos y 3163 ejemplos.

Luego se resuelven los datos faltantes de cada atributo utilizando la media de los mismos para los datos continuos y el valor de mayor frecuencia para los valores nominales. Se mapea el valor correspondiente al atributo sexo para que tome valores del conjunto $\{Male, Female\}$, y aquellos atributos booleanos para que los valores pasen de f a $False$ y de t a $True$.

3. Análisis de datos

Para el análisis de los datos se evalúan distintas gráficas de lo mismos, como histogramas de los datos nominales y gráficas de dispersión para aquellos datos de tipo numérico, además de ciertas métricas que permiten conocer la correlación entre cada uno de ellos. De este modo se permite observar relaciones entre los distintos atributos del conjunto de datos así como también la relación entre estos mismos y la etiqueta, o *label*.

Se dispone de las gráficas correspondientes a los atributos en la sección 7.B.1 del apéndice.

Posteriormente, se calcula el índice de correlación lineal entre los atributos, para comenzar así con el análisis de las relaciones entre cada uno de ellos. El cálculo de los mismos se ve reflejado en la siguiente tabla (1).

Atributos	age	TSH	T3	TT4	TU4	FTI
age	1	-0.007	-0.269	-0.091	-0.194	0.015
TSH	-0.007	1	-0.172	-0.310	0.069	-0.244
T3	-0.269	-0.172	1	0.545	0.388	0.294
TT4	-0.091	-0.310	0.545	1	0.323	0.685
T4U	-0.194	0.069	0.388	0.323	1	-0.284
FTI	0.015	-0.244	0.294	0.685	-0.283	1

Table 1. Matriz de correlación lineal

Como se observa en la tabla 1, las tuplas (FTI, TT4) y (T3, TT4) presentan una correlación lineal leve, con un índice de correlación de 0.685 y 0.545 respectivamente.

4. Método Experimental

En esta sección se detalla todo lo referido al estudio y la creación de los distintos modelos de sistemas inteligentes utilizados para el estudio del conjunto de datos elegido.

4.A. Árbol de decisión

Para la creación del modelo de Árbol de decisión, se evalúan las distintas posibilidades permitiendo así la construcción de un modelo con un mayor nivel de cobertura sobre el conjunto de datos empleado para entrenamiento y prueba.

EXPLICAR ELECCIÓN DEL MODELO

Por lo tanto, como se menciona anteriormente, dado que el conjunto de datos presenta atributos numéricos de tipo continuo, no es completamente viable discretizar los mismos en intervalos si es que existe alguna forma de construir un modelo de Árbol de decisión evitando esto.

Se opta finalmente la utilización del algoritmo C4.5 para generar el Árbol de decisión. Se genera el Árbol utilizando el operador W-J48 de rapidminer, utilizando la configuración por defecto de la mayoría de los flags exceptuando el flag C, *confianza*. Luego de probar distintas configuraciones para dicho flag, se observa que dado un conjunto de entrenamiento del 70% del total de los datos, con un 30% de los datos destinado al testing del modelo, el porcentaje de acierto del modelo es 99.59% cuando $C \geq 0.5$, por lo que se configura el flag con C con el valor 0.5.

En la figura (1) se puede observar el modelo obtenido utilizando el algoritmo C4.5.

W-J48

J48 pruned tree

```
FTI <= 63
|   TSH <= 6.2: negative (41.0/2.0)
|   TSH > 6.2
|   |   on_thyroxine = False: hypothyroid (90.0/5.0)
|   |   on_thyroxine = True
|   |   |   TSH <= 23: negative (3.0)
|   |   |   TSH > 23: hypothyroid (12.0/1.0)
FTI > 63
|   TSH <= 6.3: negative (1917.0/1.0)
|   TSH > 6.3
|   |   age <= 59: negative (94.0/2.0)
|   |   age > 59
|   |   |   FTI <= 69
|   |   |   |   T4U <= 0.97: hypothyroid (3.0)
|   |   |   |   T4U > 0.97: negative (4.0)
|   |   |   FTI > 69: negative (50.0/2.0)

Number of Leaves  :     9

Size of the tree :     17
```

Fig. 1. Modelo de Árbol generado utilizando el algoritmo C4.5, con una performance de 99.16%.

En la figura (2) se puede observar la performance del modelo obtenido aplicando el mismo sobre un conjunto

de testing.

accuracy: 99.16%

	true hypothyroid	true negative	class precision
pred. hypothyroid	40	3	93.02%
pred. negative	5	901	99.45%
class recall	88.89%	99.67%	

Fig. 2. Performance del modelo de Árbol generado utilizando el algoritmo C4.5 sobre un conjunto de testeo correspondiente al 30% del conjunto total de datos.

4.B. Reglas de Clasificación

Para definir el algoritmo a utilizar para la creación de reglas se evalúa cada uno de ellos comparando los modelos generados para determinar así cual es más conveniente. Entre los algoritmos evaluados están OneR y PRISM.

4.B.1. OneR

Para la construcción de este modelo se trabaja con un conjunto normalizado de los datos utilizando normalización Z sobre los datos de tipo numérico. El antecedente de la regla generada por este algoritmo se define a partir del atributo *FTI* como se puede observar en la figura (3).

W-OneR

```
FTI:
        < -1.1047537479703995   -> hypothyroid
        < -0.94914920044515     -> negative
        < -0.9232151091909417   -> hypothyroid
        >= -0.9232151091909417  -> negative
(2169/2214 instances correct)
```

Fig. 3. Modelo generado por el algoritmo OneR.

En la figura (4) se puede observar la performance del modelo obtenido aplicando el mismo sobre un conjunto de testing.

accuracy: 98.10%

	true hypothyroid	true negative	class precision
pred. hypothyroid	39	12	76.47%
pred. negative	6	892	99.33%
class recall	86.67%	98.67%	

Fig. 4. Performance del modelo de reglas generado utilizando el algoritmo OneR sobre un conjunto de testeo correspondiente al 30% del conjunto total de datos.

4.B.2. PRISM

Para la construcción del modelo de reglas utilizando el algoritmo PRISM se discretiza por frecuencia los datos numéricos probando le performance del modelo para distintos intervalos. Finalmente se opta por discretizar en 7 intervalos obteniendo los siguientes resultados.

5. Análisis de Resultados

6. Discusión y Conclusiones

References

[1] https://authors.aps.org/revtex4/.
[2] http://www.opticsinfobase.org/submit/style/jrnls_style.cfm.

7. Apéndice

7.A. Marco Teórico

7.B. Imágenes

7.B.1. Gráficos de los atributos

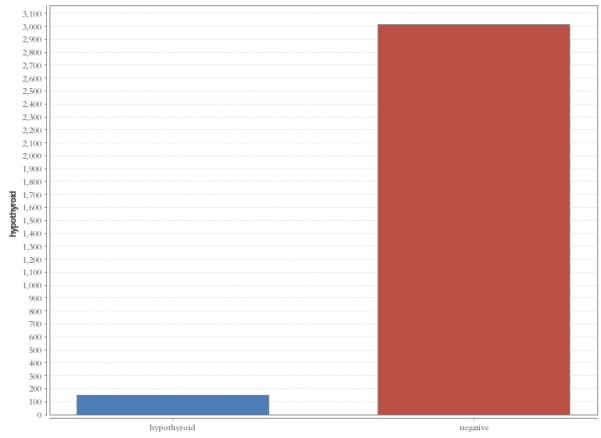

Fig. 5. Gráfico de barras del atributo etiqueta.

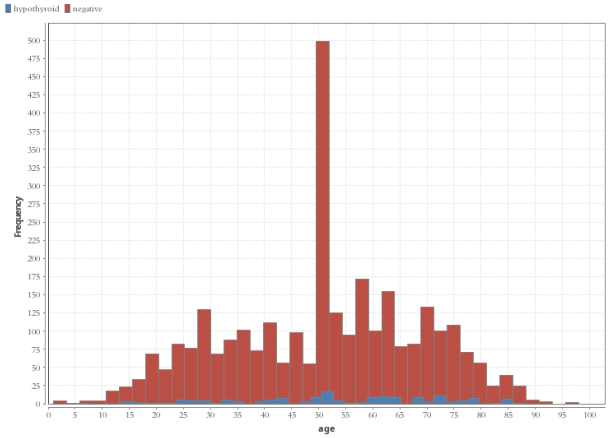

Fig. 6. Histograma del atributo *age*.

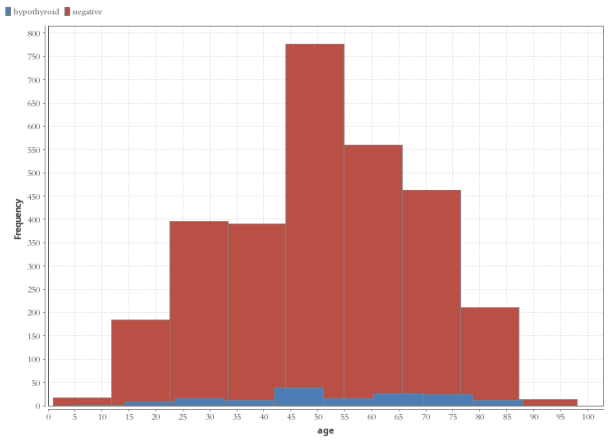

Fig. 7. Histograma del atributo *age*.

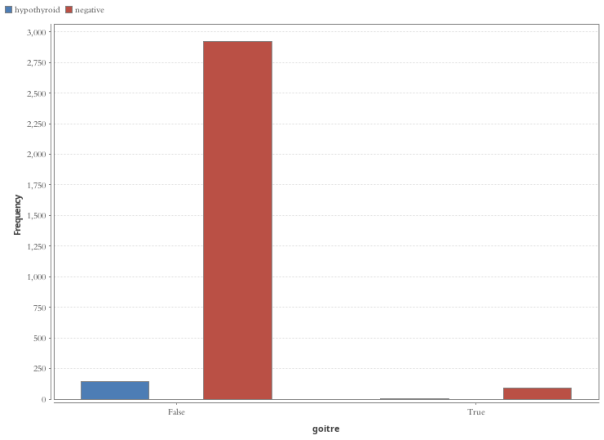

Fig. 8. Histograma del atributo *goitre*.

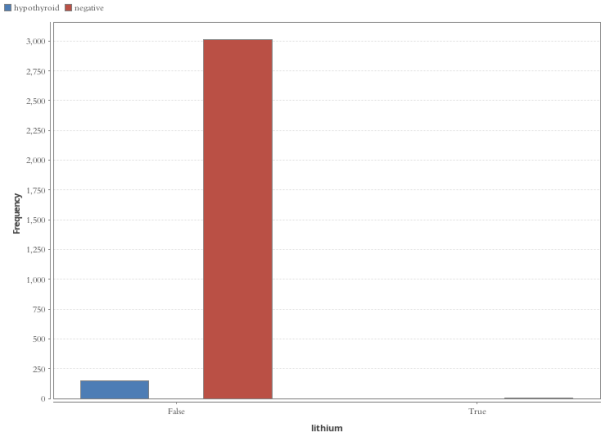

Fig. 9. Histograma del atributo *lithium*.

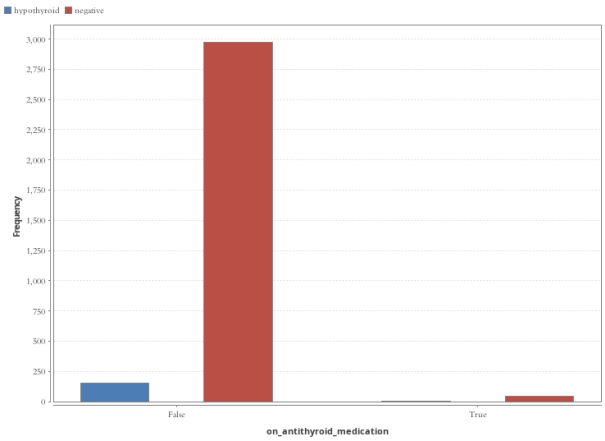

Fig. 10. Histograma del atributo *on_antithyroid_medication*.

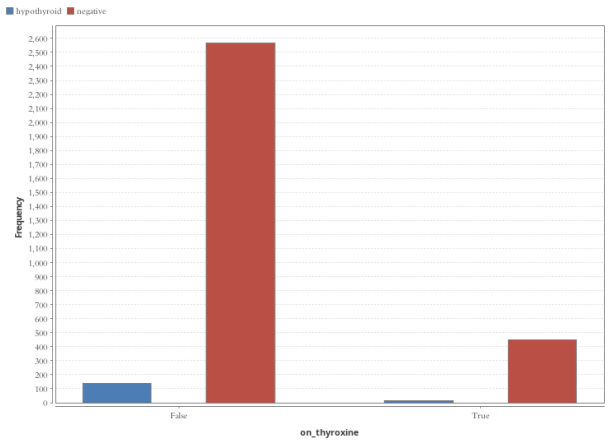

Fig. 11. Histograma del atributo *on_thyroxine*.

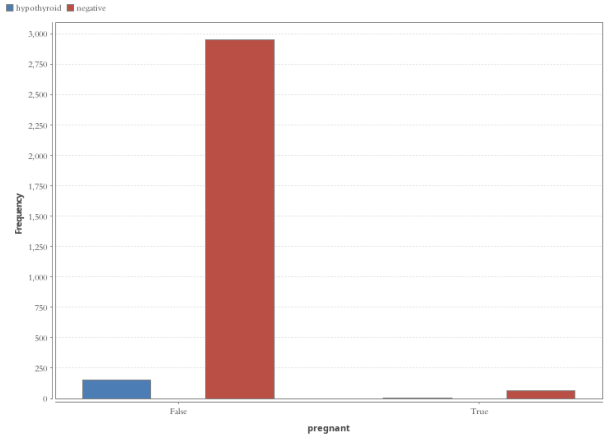

Fig. 12. Histograma del atributo *pregnant*.

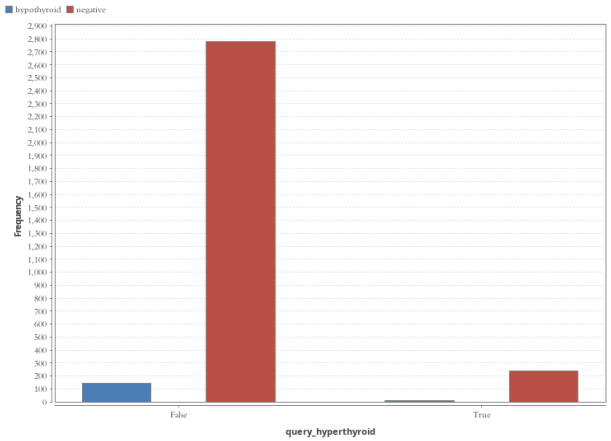

Fig. 13. Histograma del atributo *query_hyperthyroid*.

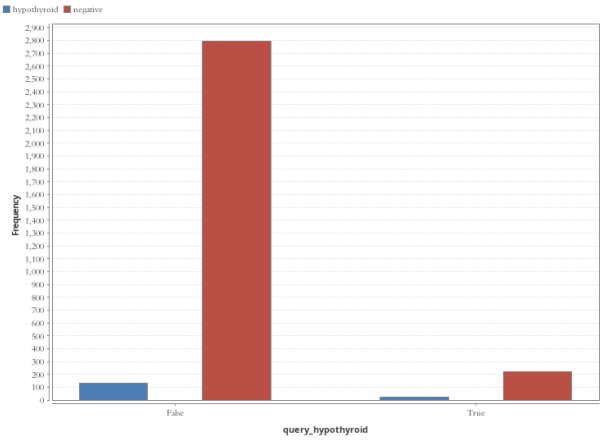

Fig. 14. Histograma del atributo *query_hypothyroid*.

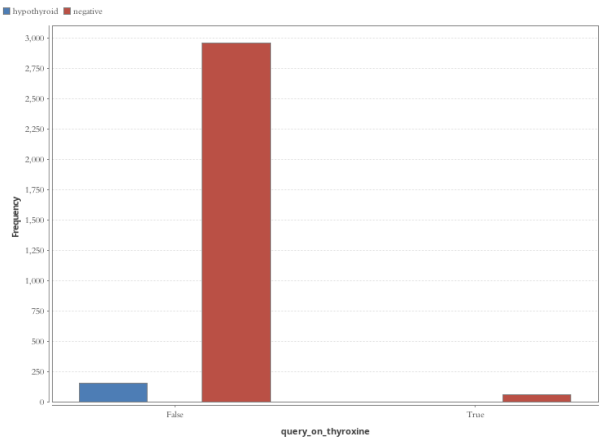

Fig. 15. Histograma del atributo *query_on_thyroxine*.

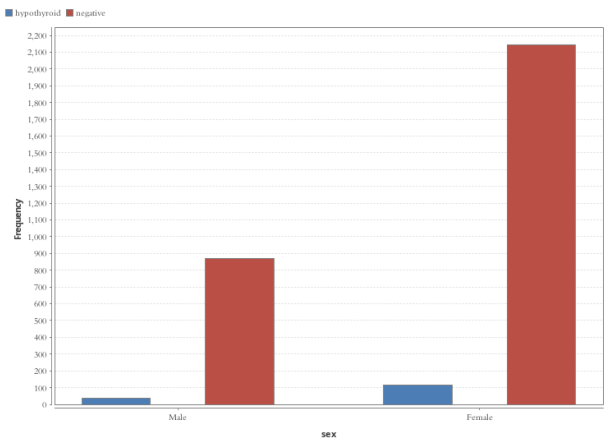

Fig. 16. Histograma del atributo *sex*.

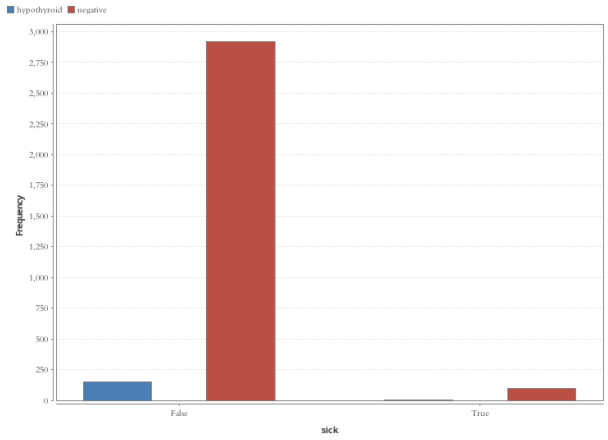

Fig. 17. Histograma del atributo *sick*.

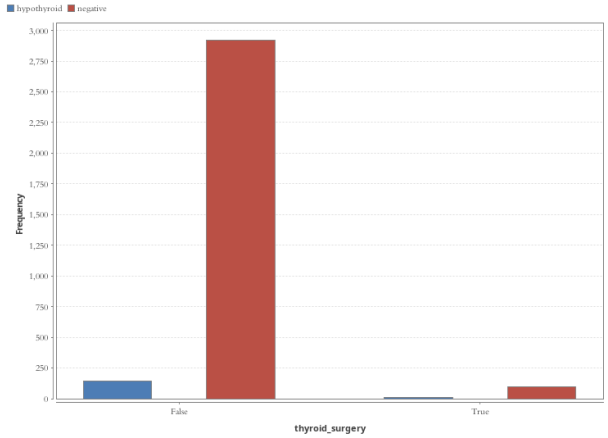

Fig. 18. Histograma del atributo *thyroid_surgery*.

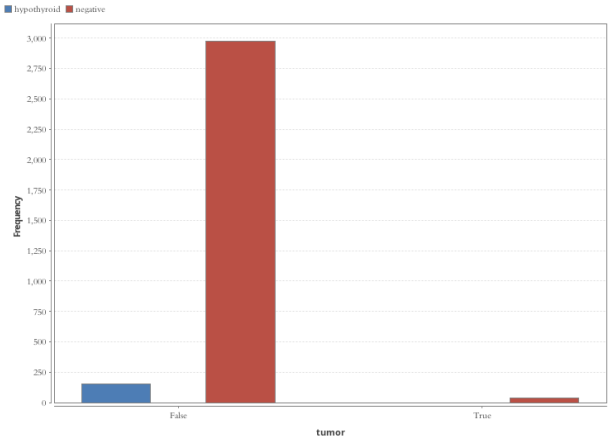

Fig. 19. Histograma del atributo *tumor*.

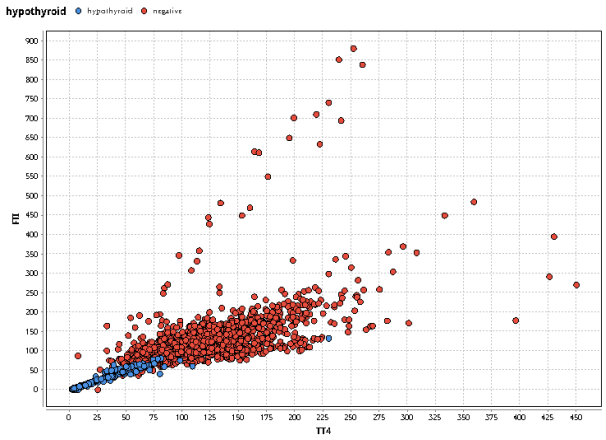

Fig. 20. Diagrama de dispersión de los atributos *TT4, FTI*.

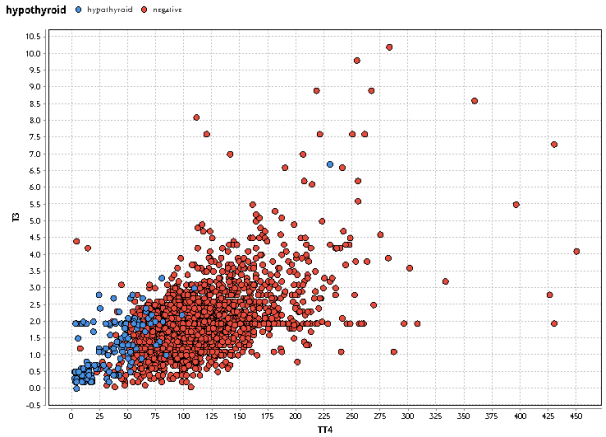

Fig. 21. Diagrama de dispersión de los atributos *TT4, T3*.

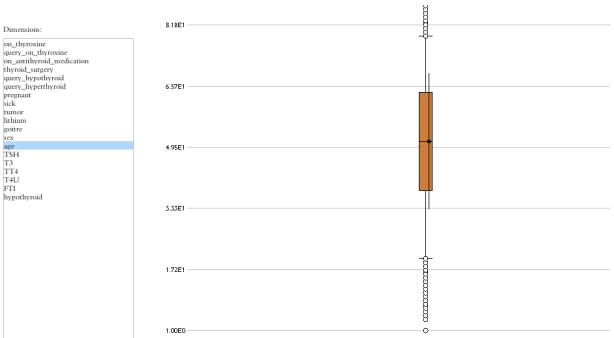

Fig. 22. Diagrama de cajas del atributo *age*.

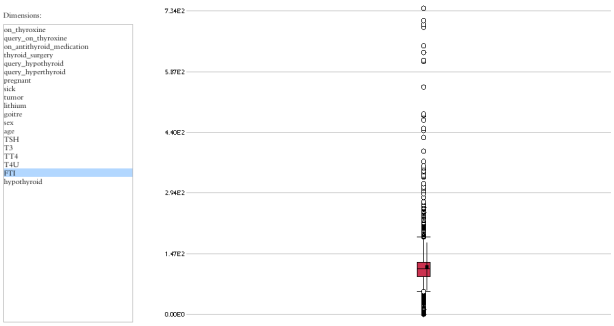

Fig. 23. Diagrama de cajas del atributo *FTI*.

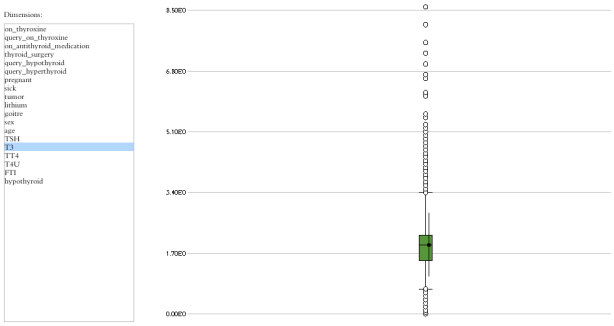

Fig. 24. Diagrama de cajas del atributo *T3*.

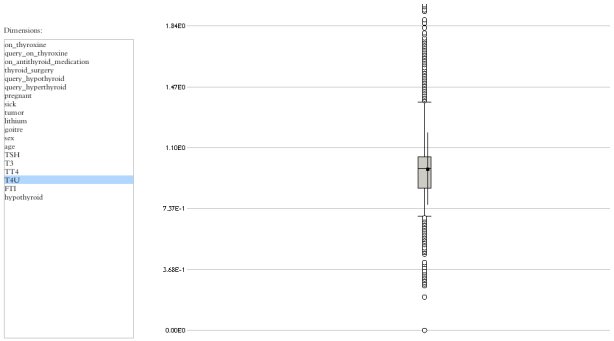

Fig. 25. Diagrama de cajas del atributo *T4U*.

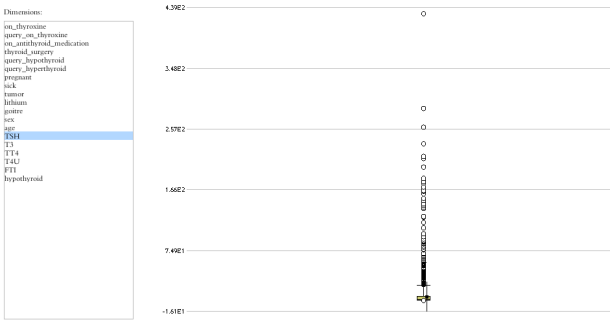

Fig. 26. Diagrama de cajas del atributo *TSH*.