Aplicación de arboles de decisión sobre una muestra de niños infectados por el Rotavirus utilizando J48.

Objetivo

Poder aplicar los conocimientos obtenidos durante el cursado de la materia, sobre arboles de decisión, a través de la herramienta Weka utilizando del algoritmo J48, sobre una muestra real de casos positivos y negativos de Rotavirus en niños con edad menor a 5 años. Realizar un análisis de los resultados apuntando a encontrar hipótesis que nos ayuden a conseguir alguna predicción sobre la enfermedad o discriminación de grupos que sean mas propensos a enfermarse. También analizar la presencia del fenómeno de sobreajuste al aplicar el algoritmo.

Alcance

Realizar numerosas corridas del algoritmo variando el parámetro de confidence factor, analizando si ocurre el fenómeno de sobreajute y poda. A partir de allí, poder visualizar el crecimiento del árbol y su performance en función del conjunto de entrenamiento y validación.

Discretizar la variable continua edad, que esta representada en la unidad meses, variando la cantidad de intervalos o bins con igual frecuencia o densidad. En función del tipo de desertización analizaremos el crecimiento del árbol y su performance.

Resultados

Sobre la muestra: Rotavirus.

El rotavirus es considerado la principal causa viral de gastroenteritis aguda, en niños, y no depende de países desarrollados o en vías de desarrollo. La experimentación se hizo sobre un conjunto de datos que muestran la información de 410 niños que fueron hospitalizados con cuadros de diarrea. Se tomo una muestra fecal para cada caso. La información obtenida tiene su origen en la capital de Paraguay, Asunción, y proviene tanto de hospitales públicos como privados. Se confirmó la presencia de Rotavirus en 93 casos (casi un 23%) de la muestra.

A continuación una breve descripción de cada variable :

- Día de la semana: Es una variable nominal que se ha rescatado de la fecha original de la muestra. Según la definición de weka:

@attribute dia {Monday,Tuesday,Wednesday,Thursday,Friday,Saturday,Sunday}

- Mes: Tambien es una variable nominal, pero que ha sido discretizada basándose en las estaciones del año. Esto fue hecho, con la idea de poder explicar los casos positivos de Rotavirus en relación con el factor climatico. Según Weka:

@attribute mes {Verano, Otonio, Invierno, Primavera}

- Hospital: Cada instancia de la muestra tuvo su origen en un establecimiento de salud. Esta variable también es nominal y solo cuenta con cuatro categorías:

@attribute hospital {'S. R.','I. P. S.','H. C.','Inst. Priv. Niño'}

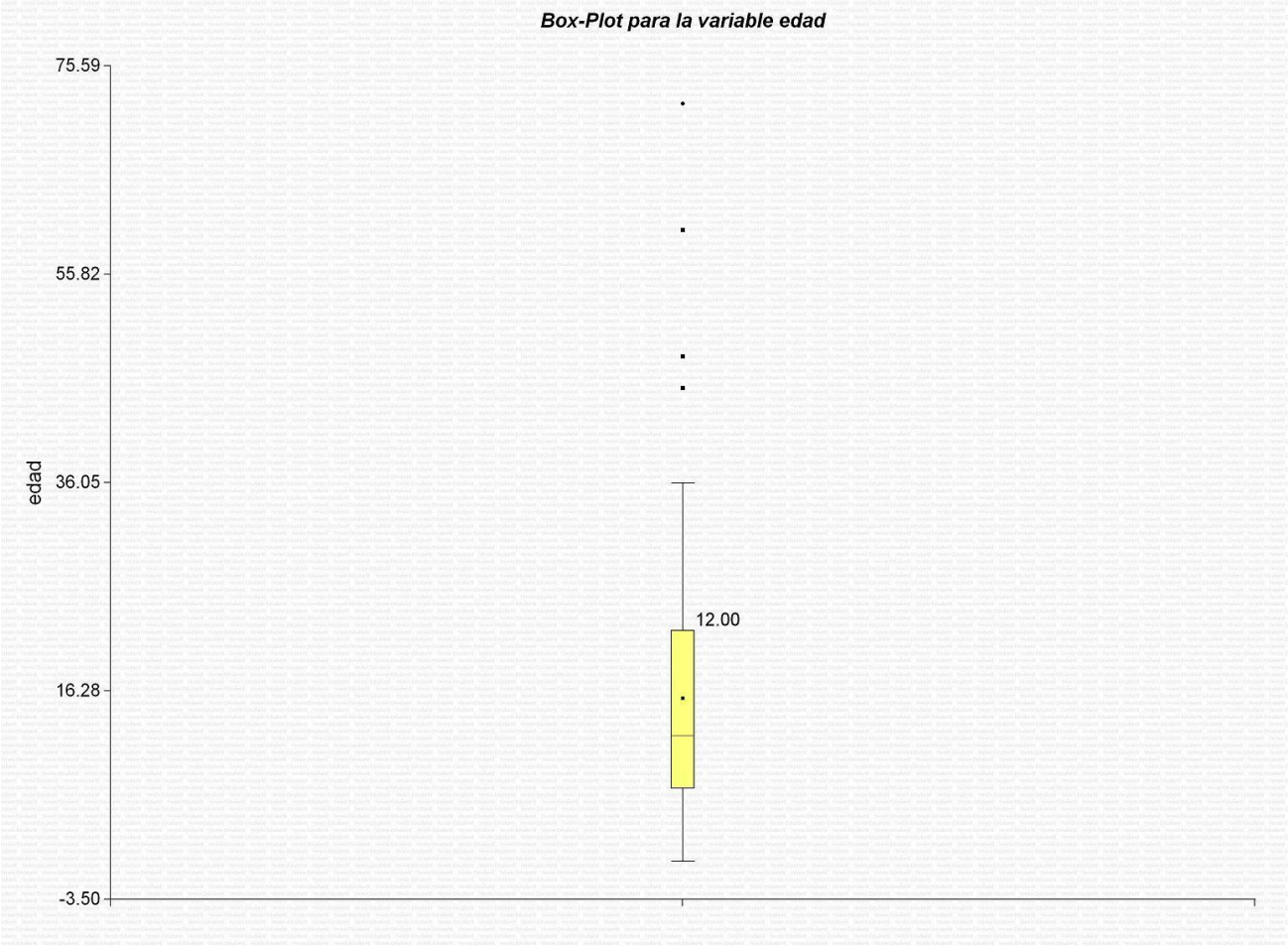
- Sexo: Simplemente la información relacionada al niño.

@attribute sexo {M,F}

- Edad en meses: La edad de los niños esta dada en meses. Esta variable era la única para la cual había datos faltantes y la variación total iba entre 3 días y 72 meses.

En el primer ejercicio de Poda, optamos por una desertización a nuestro criterio, que esta basada en la información de un paper[1] al respecto de la enfermedad.
Para el segundo ejercicio, el de desertización, mantuvimos la información original de la variable, representada en meses.

Faltantes: Todos los análisis realizados sobre esta variable estuvieron apoyados en el siguiente gráfico Box-Plot:



Aquí podemos ver que la mediana es 12, y como la distribución no es simétrica, este es el mejor estadístico para la media muestral. Este es el valor utilizado para el rellenado de valores faltantes.

@attribute edad {'0 a 5', '6 a 11', '12 a 23', '24 a 35', '36 a 47', '48 a 72'}
- Rotavirus (positivo o negativo): Esta información de la clase sobre la que se realizaron los experimentos.

@attribute rotavirus {Negativo,Positivo}

Sobreajuste y Poda

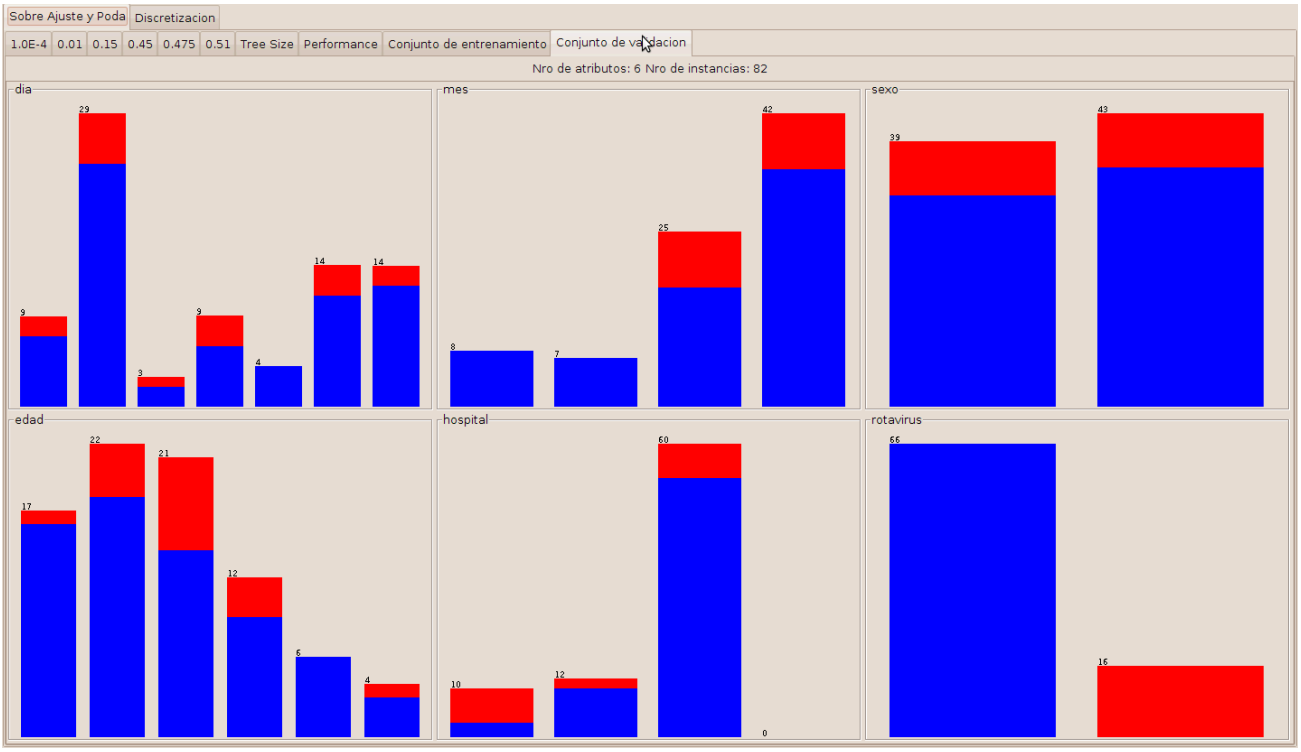
Keywords: Conjunto de entrenamiento y conjunto de validación. Estadísticas. Gráficos.

Conjunto de entrenamiento: Nro de atributos 6, Nro de instancias: 328. Casos positivos de rotavirus: 77.

1 - Gráfico del conjunto de entrenamiento[1]

Conjunto de validación: Nro de atributos 6, Nro de instancias: 82. Casos positivos de rotavirus: 16.

[1]El grafico fue tomado del programa java que acompaña el presente trabajo. Esta apoyado en una clase de weka que detalla estos valores estadisticos



2 - Gráfico del conjunto de entrenamiento[2]

Keywords: Arboles resultantes. Tamaños obtenidos. Variaciones del confidence factor. Gráfico.

La ejecución del algoritmo j48 se ha hecho variando el confidence factor, no en forma equidistante, sino estratégicamente, buscando mas detalle en las zonas donde había mas variabilidad. Esta decisión es producto de la empíria. La corrida inicial del experimento se hizo usando un array de confidence factors que variaba de 0 a 1 en intervalos iguales, aumentando en 0.025 el parámetro. Una vez que se detectaron los puntos de inflexión en el tamaño del árbol, redujimos las distancias entre los confidence factors para esa zona. El resultado esperado era llevar lo mas hacia la izquierda posible el cambio de tamaño en el árbol.

Finalmente, en el programa que adjuntamos al presente trabajo, solo se gráfico aquellos arboles que fueran distintos entre si. Esta es la razón por la cual solo se muestran 6 diferentes arboles.

Confidence Factor	Cantidad de nodos	Cantidad de hojas	Altura del árbol
0.0001	1	1	0
0.01	5	4	1
0.15	11	9	2
0.45	38	30	4

[2]El grafico fue tomado del programa java que acompaña el presente trabajo. Esta apoyado en una clase de weka que detalla estos valores estadisticos

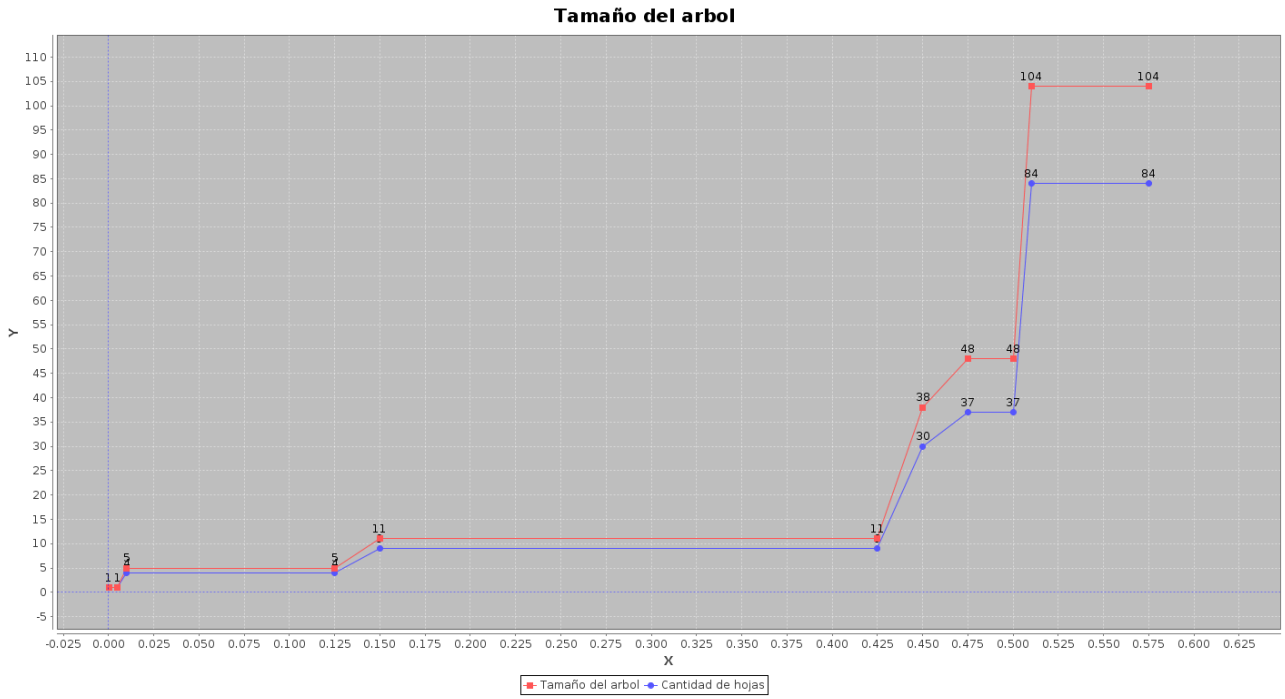
0.475	48	37	4
0.051	104	84	5

Valores de confidence factor utilizados:

```
            float[]  confidenceFactorsValids  =  {  0.0001f,  0.0002f,  0.0003f,
0.0004f,
                        0.0005f,  0.001f,  0.0013f,  0.0015f,  0.0017f,  0.005f,
0.01f,
                        0.025f,  0.05f,  0.075f,  0.1f,  0.125f,  0.15f,  0.16f,
0.17f,
                        0.175f,  0.2f,  0.225f,  0.25f,  0.275f,  0.3f,  0.325f,
0.35f,
                        0.375f,  0.4f,  0.425f,  0.45f,  0.475f,  0.5f,  0.51f,
0.525f,
                        0.53f, 0.54f, 0.55f, 0.575f };
```

Hay que destacar que para todos los valores de confianza se observa que la mayoría de los nodos son hojas, lo cual nos da una idea de que la forma del árbol, más ancho que profundo. Y esto puede pasar porque estamos trabajando con pocas variables, ya que es imposible que el árbol pueda llegar a tener mas de 5 niveles.

Un punto a destacar aquí es la falta de variaciones en el tamaño del árbol para los valores superiores a 0.55 de confidence factor. Probando el experimento con otros dataset se concluyo en desestimar valores superiores a 0.55 de confidence factor, por que siempre daban el mismo resultado.



3 - Gráfico del tamaño del árbol en función del confidence factor[3]

Keywords: Performance del entrenamiento. Performance del test. Fenómeno de sobreajuste. Particularidades. Gráfico.

Aquí es importante detallar el efecto de sobreajuste u overfitting. En las pruebas iniciales, donde el archivo tenia algún ordenamiento arbitrario, el efecto esperado de sobreajuste no siempre estaba presente. Esto derivo en varios intentos de modificaciones al experimento. Cambiar los tamaños de los conjuntos de entrenamiento y validación, ordenar cronologicamente el dataset, separar un 80/20 el archivo pero que llevaran partes proporcionales en relación a la cantidad de casos de Rotavirus positivo, etc... .

El mejor caso donde se pudo observar el fenómeno de sobreajuste u overfitting, fue en la separación aleatoria . Este fenómeno se pudo observar para un confidence factor aproximado de 0.422869 (*Este es un valor aproximado, debido a que sale como una deducción del gráfico*).

4 - Performance del conjunto de entrenamiento y del conjunto de validación en función del confidence factor.[4]

Particularidades a destacar:

- Con un confidence factor de **0.01** podemos ver que el 99% de los casos positivos que fueron clasificados correctamente tienen origen en el hospital S.R.(Hospital San Roque). Si miramos las

[3]El grafico fue tomado del programa java que acompaña el presente trabajo. Para poder confeccionarlo se uso la biblioteca JFreeChart.

[4]El grafico fue tomado del programa java que acompaña el presente trabajo. Para poder confeccionarlo se uso la biblioteca JFreeChart.

estadística del conjunto de entrenamiento podemos ver que el Hospital San Roque contribuye con mucho menos instancias a la distribución total de la variable, contrario a lo esperado por nosotros. *Ver detalle en el grafico 1*.

- Con un confidence factor de **0.15** el árbol crece un nivel mas y los casos positivos empiezan a dividirse por rangos de edad. Aquí podemos ver que los casos positivos solo resultan en dos hojas; los rangos de edad que van de 6 a 11 y de 12 a 23 meses, teniendo el ultimo la mayor cantidad de casos. En bibliografía leída al respecto hemos encontrado la confirmación de que esa franja es la mas afectada por esta enfermedad.

- Con un confidence factor de **0.45**, sigue clasificando los casos positivos en los dos intervalos de edad, pero esta vez, realiza un split por estación del año, confirmando que el invierno es la época con las mayor cantidad de casos positivos. El árbol generado es muy similar al anterior, pero para el caso de Hospital de Clínicas podemos ver que avanzo un paso mas en el espacio de hipótesis detectando que para el H.C (Hospital de clínicas) todo los casos positivos que tenia, en los 47 casos que no fueron clasificados por el factor de poda en el árbol anterior. Podemos ver que exploto el nodo del Hospital de clínicas con cuatro niveles, primero por el mes, día de la semana, luego por edad y sexo. Y solo en los meses de invierno, manteniendo la misma clasificación de positivos en los rangos de edad como había hecho en el paso anterior para el Hospital San Roque. Los casos de rotavirus que se encuentran en los meses bajos fueron separados en masculino y femenino, clasificando los casos positivos en los masculinos.

- No encontramos ninguna explicación a la hipótesis de división por día de la semana. Lo único que tratamos de hacer fue ver si estaba correlacionada con Rotavirus. Hicimos un test de hipótesis a través de la tabla de contingencia, y vimos que el coeficiente de Pearson nos dio 0.0382. Con lo cual, no podemos rechazar que sean independientes. Si aplicamos el mismo test para el caso del hospitales, pudimos ver que no son independientes por que el coeficiente de Pearson es <0.0001. Hay una fuerte correspondencia entre estas dos variables.

- Con confidence factor **0.45**, podemos ver el fenomeno de sobreajuste. El arbol empieza a hacerse muy especifico. El conjunto de entrenamiento llega al 85% de performance, pero el conjunto de validacion baja al 84% su performance. El conjunto de hipotesis generado a partir del conjunto de entrenamiento pierde prediccion sobre el conjunto de validacion.

- Con un confidence factor de **0.475** se sigue dando el fenómeno de sobreajuste. El conjunto de entrenamiento mejora en clasificacion (alcanza el 85.5%), pero el conjunto de validacion vuelve a perder performance(81%). La brecha entre la performance de ambos conjuntos parece crecer aún más. Es claro que la explicacion a esto viene dado por el fenomeno antes mencionado

- Con un confidence factor de **0.51** ocurre algo atipico. El conjunto de entrenamiento mejora en clasificacion(88%), pero a diferencia de lo esperado, el conjunto de validacion tambien aumenta la performance (82%). ¿Cual es la explicacion a esta situacion? La cantidad de casos clasificados correctamente mejoro en 6 y dentro de esos 6 hay uno que mejoro en el conjunto de validacion. Por eso sucede este fenomero.

- A partir del confidence factor 0.51 el arbol se mantiene inalterable.

Discretización:

Keywords: Variables numéricas en el dataset. Frecuencias. Densidad. Bines. Tamaño y performance de los arboles generados. Gráfico.

El dataset original tenía una única variable numérica. Era la edad con una variabilidad de 3 días a 73 meses.

Para hacer el procesamiento de la discretización se utilizó los mismos confidence factors que en el punto de anterior. Pero solo nos hemos quedado con los árboles que no se repiten.
Para cada uno de esos confidence factors, se hizo un conjunto de discretizaciones que variaban la cantidad de bines de 1 a 20. El detalle da cada uno puede verse en el programa que acompaña en trabajo.

El tamaño del árbol crece a medida que aumenta el confidence factor y la cantidad de bines. Algo razonablemente esperable. Pero para confidence factor pequeños se puede observar picos de creciemiento debido al split que hizo el algoritmo de la variable edad. Encuentra una discretización tal que determina hacer el crecimiento por esta variable. Esto puede ser por que la concentración de casos positivos pueden encontrarse en un intervalo.

También se puede ver que en los casos dónde hay picos de crecimiento la performance decrece, tanto para el conjunto de entrenamiento como para el conjunto de validación.
Para ambos casos, igual frecuencia e igual densidad, pero sobre todo para los de igual frecuencia, se da el fenómeno de sobreajuste en los valores más altos de confidence factor.

Metodologías

La preparación del archivo en formato .arff empezó siendo una tarea manual, pero debido a la cantidad de cambios que iba sufriendo a medida que avanzaba el desarrollo del TP, optamos por automatizar esta tarea. La opción elegida fué hacer un programa java que tome como entrada el archivo .csv y generrar el archivo .arff con los ajustes que ibamos determinando. Ajustes que iban desde la discretización de alguna variable hasta la corrección en el tipo de datos que Weka necesitaba.
Esto hizo que fuera más práctico ir incorporando cada uno de estos procesamientos automatizados al código de weka. Haciendo uso de los códigos fuentes de Weka y siguiendo algunas referencias de la wiki pudimos hacer la resolución del trabajo haciendo una extensión de la herramienta, y así evitar la exportación de la información generada para los tratamientos posteriores.
Aprovechando el trabajo realizado en java, agregamos unas bibliotecas que simplifican el trabajo con gráficos. Sobre todo para poder usar las imágenes generadas a partir de ellos.

Bibliografia

- Machine Learning, Tom Mitchell, McGraw Hill, 1997.
- Weka Wikispaces
- Rotavirus infection in the Paraguayan population from 2004 to 2005: High incidence of rotavirus strains with short electropherotype in children and adults

Referencias

1. Machine Learning, Tom Mitchell, McGraw Hill, 1997.

2. S.R.: Hospital San Roque, http://www.sanroque.com.py/
3. H.C.: Hospital de Clinicas
4. I.P.S.: Instituto de Prevision Social, http://www.ips.gov.py
5. I.P.N.: Instituto Privado del Ninio, http://www.ipn.com.py