Trabajo Práctico 1

Jairo Jiménez
Sergio De Raco

June 12, 2015

Objetivo

El objetivo del presente trabajo es analizar las característcas en las cuáles el algoritmo J48 es robusto. El algoritmo será aplicado sobre la base de datos Latinobarómetro [Lagos, 2012], la cual es una encuesta de percepción política, democrática y social de latinoamérica. Para poder observar las cualidades y deficiencias del algoritmo, este será sometido a diferentes tipos de pruebas para analizar su comportamiento.

Alcance

Los datos utilizados provienen de la última encuesta disponible sin suscripción (año 2011) de [Latinobarómetro]¡http://www.latinobarometro.org¿, un estudio de opinión pública que realiza anualmente alrededor de 20.000 entrevistas en 18 países de América Latina. Los datos se encuentran disponibles online, junto con el cuestionario y la metodología correspondientes. La encuesta se realiza en centros urbanos de los países de cobertura y contiene preguntas y datos acerca de la situación social, política y económica (actual, pasada y futura), distribución del ingreso, nivel educativo, percepciones acerca de la participación del Estado e información demográfica.

En este trabajo se seleccionaron variables (preguntas de la encuesta) que pudieran aportar a predecir la respuesta a la pregunta (P10ST): ¿Cree usted que el país está o no progresando? Para ello se seleccionaron los siguientes atributos:

Descripción	Variable	# Categ.
Autopercepción de clase social	S14	6
Estado civil	S15	4
Sexo	S16	2
Edad	S17	N
Propiedades y bienes	S28D	3
Satisfacción con la propia vida	P1ST	5
Situación económica actual	P3ST.A	5
Situación política actual	P3ST.B	5
Situación económica últimos 12 meses	P4ST	5
Percepción del nivel de (des)igualdad del ingreso	P12ST	5
Preferencia por la democracia	P13ST	4
Mejora en la democracia del país	P15N	4
Gobierno de pocos ó del pueblo	P19ST	3
Expectativa de mejora en el nivel educativo a futuro	P28NE	4
Nivel percibido de deomcracia en el país	P49ST.A	4
Orientación política declarada	P76ST	4
Fortaleza del Estado	P65ST	3
Situación económica personal y familiar	P6ST	6
Percepción de magnitud de futuras alzas de precios	S1NICC7	4

La base completa contiene 20.204 registros y 414 columnas, que representan las preguntas y datos de la encuesta. Para el trabajo se seleccionaron respuestas por países sudamericanos (Argentina, Bolivia, Brasil, Chile, Colombia, Ecuador, Paraguay, Perú, Uruguay y Venezuela), así se redujo a 11.614 observaciones para las 20 variables. En primer lugar, se procedió a transformar la variable de clase (P10ST) en una variable binaria que en el dataset aparece relativamente balanceada, con participaciones de 43.5 y 56.5 en cada categoría. Luego, en los atributos seleccionadso se agregaron categorías para evitar introducir un sesgo en la poda sobre atributos con menor cantidad de categorías.

Resultados esperados

Los resultados esperados para cada uno de los temas a trabajar en este trabajo son:

- Confianza: Sobreajuste a partir de algún valor de confianza, el performance en el grupo de training se aleja del performance en el grupo de test.
- Faltantes: Robustez cuando el árbol tiene pocos faltantes en todas las variables o faltantes sobre una sola variable. Sobreajuste cuando la imputación se hace teniendo en cuenta la clase y poco poder predictivo cuando la imputación se hace de manera global.
- Ruido: Poco poder predictivo del árbol en la medida que la cantidad de Ruido aumenta.

- Discretización: Robustez cuando la variable numérica no es buen discriminador. Sobre ajuste cuando la variablee numérica adquiere poder predictivo.

Metodología

Para la realización del presente trabajo, se utiliza el sofware de minería de datos Weka [Witten and Frank, 2005] mediante el lenguaje estadístico R [R Core Team, 2013] usando el paquete RWeka [Hornik et al., 2009]. Para cada punto se programaron funciones que llaman a distintos comandos de Weka desde R con el vector de parámetros de control adecuado.

1 Sobreajuste y poda

Como primer paso se generó una familia de datasets variando el factor de confianza como función de poda en el rango $[0, 0.5]$ a incrementos de 0.025. El mejor ajuste del árbol se observa cuando la confianza es de 0.1, luego comienza el sobreajuste del árbol, pues este deja baja la capacidad de pronosticar las nuevas instancias de manera correcta.
A medida que aumenta la confianza, aumentan el tamaño y el número de hojas Figura 1.

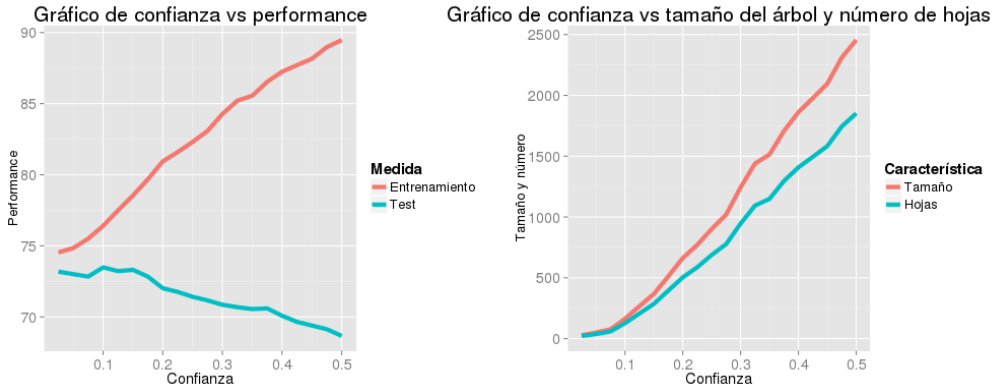

Figure 1: Gráficos de confianza

2 Tratamiento de datos faltantes

Para el análisis de datos faltantes, el algoritmo se sometió a 3 diferentes tipos de faltantes sobre el conjunto de datos. Primero se analizó el efecto de los datos faltantes sobre una sola columna, luego sobre varias columnas pero solamente

induciendo un faltante por fila y por último sobre varias columnas pero induciendo varios faltantes por fila. En los 3 casos, la cantidad de datos faltantes fue aumentando de 0 a 85%.
Para lograr lo propuesto anteriormente, se programaron una serie de funciones que incrementan en 2.5% el porcentaje de valores faltantes desde 0% hasta 85%, luego se imputaron los datos faltantes usando la moda (media) global o la moda (media) teniendo en cuenta la clase. Con cada set de datos, se calculó un nuevo árbol de clasificación y se tomaron los valores correspondientes para los análisis que se presentan a continuación.

2.1 Sobre una sola variable

En el caso de una variable, se seleccionó como candidata la variable que tiene mayor información de ganancia en el árbol (P15N).

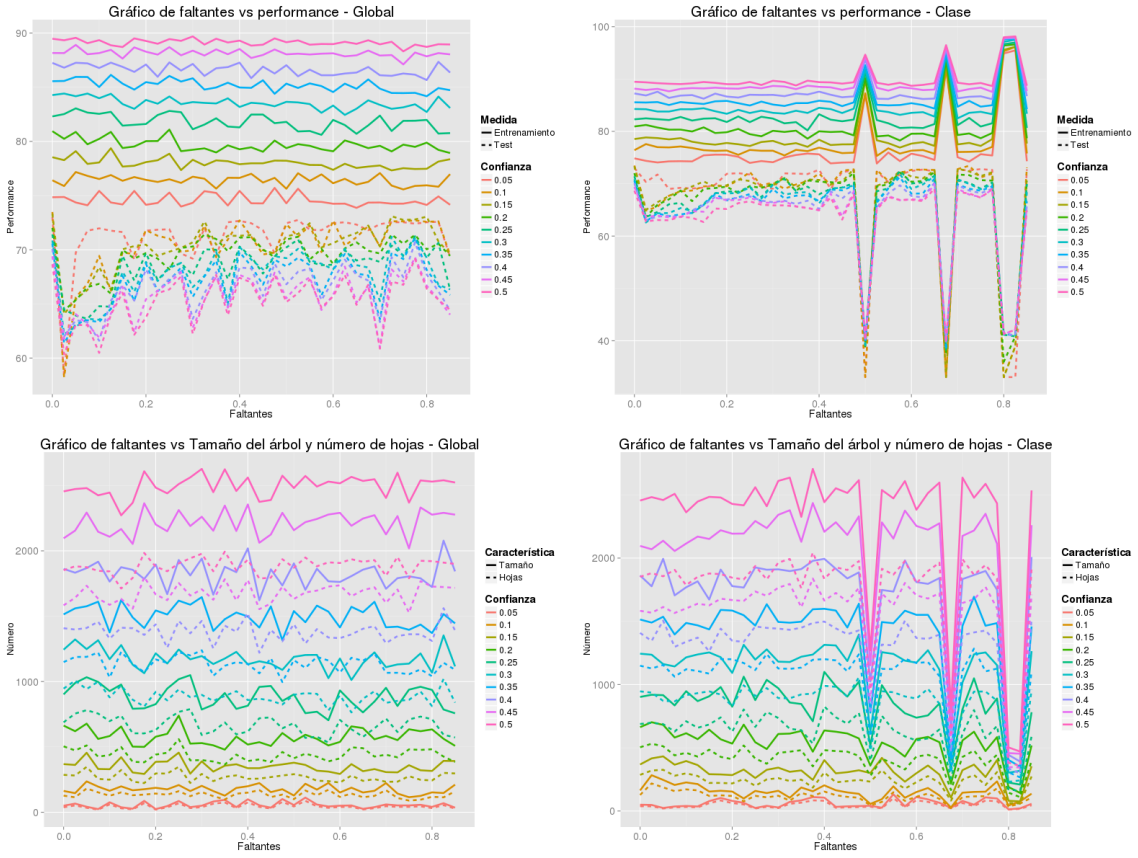

Figure 2: Gráficos de Faltantes en una sola variable

2.1.1 Imputación sin tener en cuenta la clase

Cuando la imputación se hace sin tener en cuenta la clase, el árbol no se ve afectado por la variable imputada en el conjunto de entrenamiento. Ahora bien, en el conjunto de test, se nota que el performance cae, pero empieza a recuperarse en la medida que la variable pierde poder discriminativo, hasta el punto en el que se estabiliza. En cuanto al tamaño del árbol, este no se ve afectado al ser solamente una variable la que presenta datos faltantes. Figura 2 Columna izquierda.

2.1.2 Imputación teniendo en cuenta la clase

La imputación teniendo en cuenta la clase, permite observar que existe un efecto dependiendo de en donde estén los faltantes, es decir, si los faltatantes están igualmente distribuidos entre las categorías de la clase, el performance y el tamaño del árbol se mantienen relativamente constantes, mientras que si el porcentaje de faltantes es mucho mayor para una de las categorías, se presentan sobreajustes, además de disminuciones drásticas en el tamaño del árbol. Figura 2 Columna derecha.

2.2 Sobre varias variables con un único faltante por fila

En este caso, se agregaron datos faltantes de manera aleatoria a las columnas, pero tengiendo en cuenta que solamente se admite un valor faltante por fila en el conjunto de datos. Figura 3

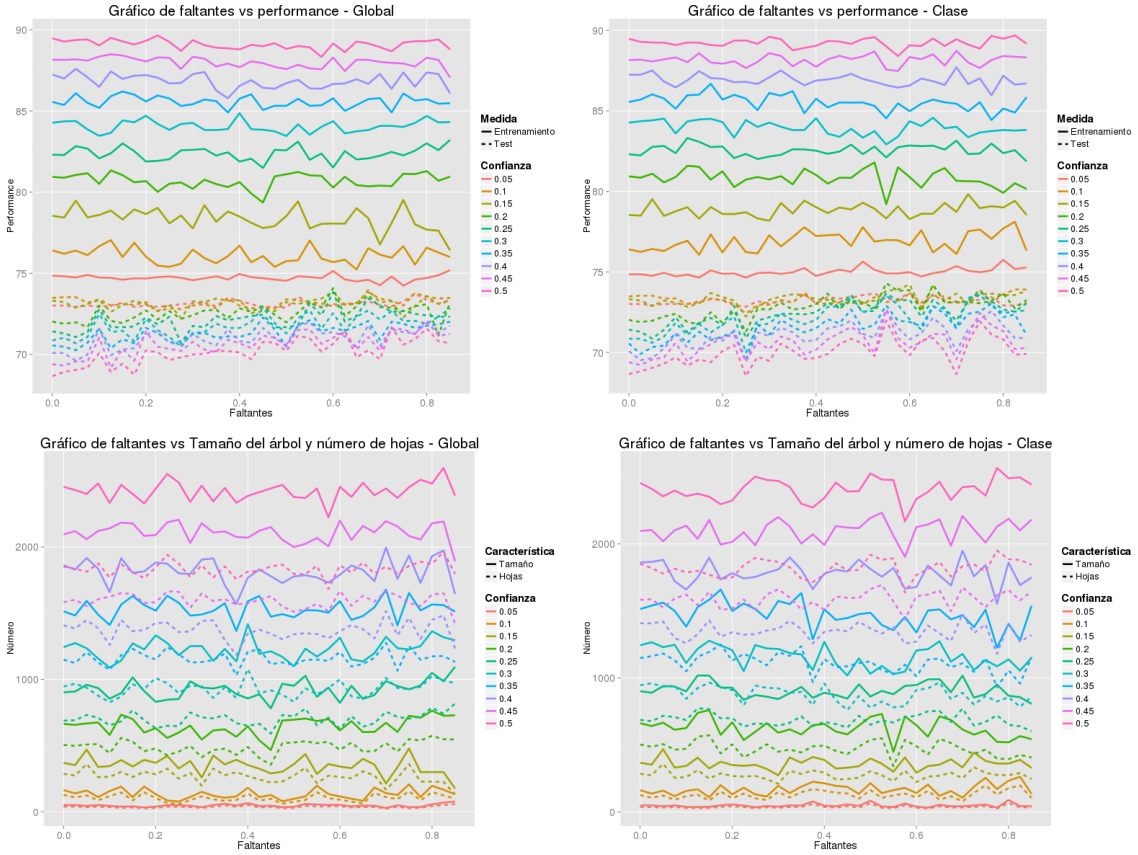

Figure 3: Gráficos de Faltantes en múltiples variables con un solo faltante por fila

En ninguna de las metodologías de imputación se notan efectos importantes sobre el performance o el tamaño del árbol y el número de hojas a medida que aumenta el porcentaje de faltantes. Este análisis puede considerarse como un caso particular del caso propuesto en la siguiente sección.

2.3 Sobre varias variables con múltiples faltantes por fila

Para el último análisis, se agregaron datos faltantes sobre el total de entradas de el conjunto de datos (total de filas por total de columnas). Figura 4

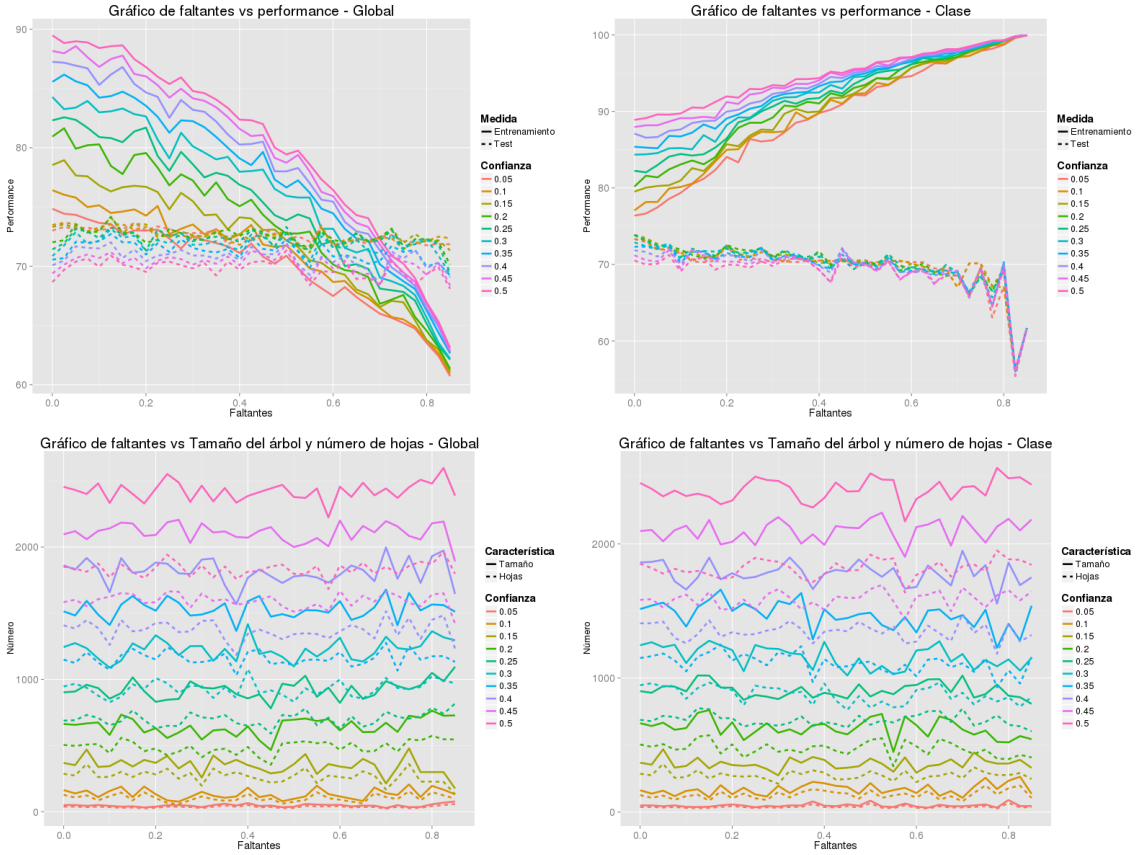

Figure 4: Gráficos de Faltantes en múltiples variables con múltiples faltantes por fila

2.3.1 Imputación sin tener en cuenta la clase

Cuando la cantidad de datos faltantes es inducida sobre la matriz completa, se observa que a medida que aumenta el porcentaje de faltantes, el ajuste del árbol en entrenamiento decae para todos los valores de confianza aunque el comportamiento en el grupo de test es relativamente constante. En cuanto al tamaño del árbol y el número de hojas, estos permanecen relativamente constantes para cada nivel de confianza. Figura 4 Columna izquierda.

2.3.2 Imputación teniendo en cuenta la clase

En este caso, a medida que aumenta el porcentaje de valores faltantes, se observa ue el ajuste en los datos de entrenamiento aumenta rápidamente, en contraste con el ajuste sobre los datos de test, sobre los cuales el decrecimiento es mucho

más lento. El tamaño del árbol y el número de hojas se mantienen relativamente constantes nuevamente. Figura 4 Columna derecha.

3 Tolerancia al ruido

Para observar la tolerancia del algoritmo J-48 al ruido, se generó ruido entre el 0% y el 35% aumentando de a 0.025%. Este ruido fue generado de manera aleatoria para cada una de las categorías de la variable de clase.

Se puede observar que el Figura 5

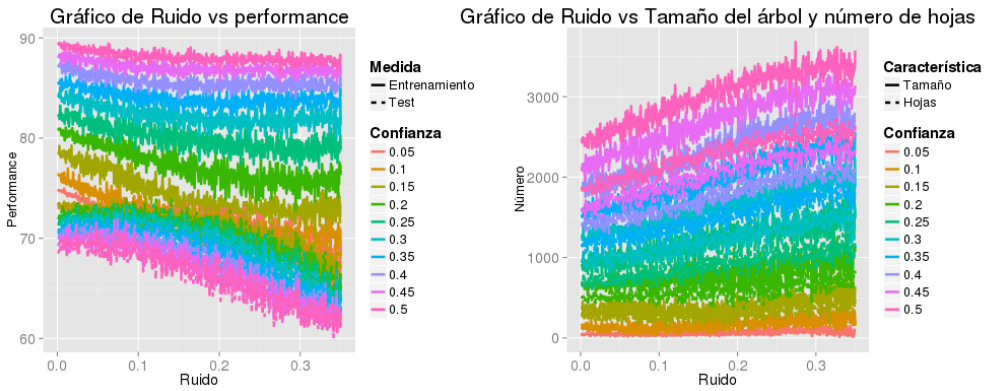

Figure 5: Gráficos de Ruido

4 Discretización de atributos numéricos

La única variable numérica considerada es la EDAD ($S17$), que se procedió a discretizar según tres criterios para el diseño de las categorías: a) de igual densidad, b) de igual frecuencia, y c) supervisado, en base a la clase.

Si la variable a discretizar resultara relevante en términos de ganancia de información, se esperaría encontrar que a mayor cantidad de bins se incremente la performance del conjunto de entrenamiento. Para el caso, la variable EDAD evidencia poco aporte de información, independientemente del criterio de discretización utilizado.

Para el ejercicio se programó una función que selecciona entre los distintos criterios y luego, para cada bin del intervalo $[1, 20]$, se arma un dataset con la variable discretizada que se particiona en subconjuntos de entrenamiento (80%) y test (20%). Luego, para la familia de combinaciones de factor de confianza en el intervalo $[0.05, 0.5]$ a pasos de 0.05, se corre llama al J48 sobre el primer dataset para extraer el modelo a utilizar con el conjunto de test y de ambas corridas se extrae la información que se precisa: tamaño del árbol, cantidad de hojas y performance.

5 Discretización de atributos numéricos

La única variable numérica considerada es la EDAD ($S17$), que se procedió a discretizar según tres criterios para el diseño de las categorías: a) de igual densidad, b) de igual frecuencia, y c) supervisado, en base a la clase.

Si la variable a discretizar resultara relevante en términos de ganancia de información, se esperaría encontrar que a mayor cantidad de bins se incremente la performance del conjunto de entrenamiento. Para el caso, la variable EDAD evidencia poco aporte de información, independientemente del criterio de discretización utilizado.

Para el ejercicio se programó una función que selecciona entre los distintos criterios y luego, para cada bin del intervalo $[1, 20]$, se arma un dataset con la variable discretizada que se particiona en subconjuntos de entrenamiento (80%) y test (20%). Luego, para la familia de combinaciones de factor de confianza en el intervalo $[0.05, 0.5]$ a pasos de 0.05, se corre llama al J48 sobre el primer dataset para extraer el modelo a utilizar con el conjunto de test y de ambas corridas se extrae la información que se precisa: tamaño del árbol, cantidad de hojas y performance.

Adicionalmente, dado que se encuentra que la variable a discretizar aporta escasa o nula información en la determinación de la clase para el conjunto de variables seleccionadas, se realiza el ejercicio de discretización con la clase, la variable EDAD y una variable equiparable a esta en términos de poder de discriminación ($S15$, ESTADO CIVIL) para observar el comportamiento del algoritmo.

5.1 Todas las variables

Cuando se incluye el conjunto completo de variables seleccionadas, en todos los casos se observa para distintos factores de confianza una mejora la performance del conjunto de entrenamiento y test a la vez para el intervalo de valores del factor de confianza especificado. No obstante, para cantidades mayores de niveles (bins) se observa incluso una caída en la performance de entrenamiento y test en las estrategias que no consideran la clase. Estos resultados indicarían que la variable EDAD no aporta mayor información que el resto de las variables consideradas.

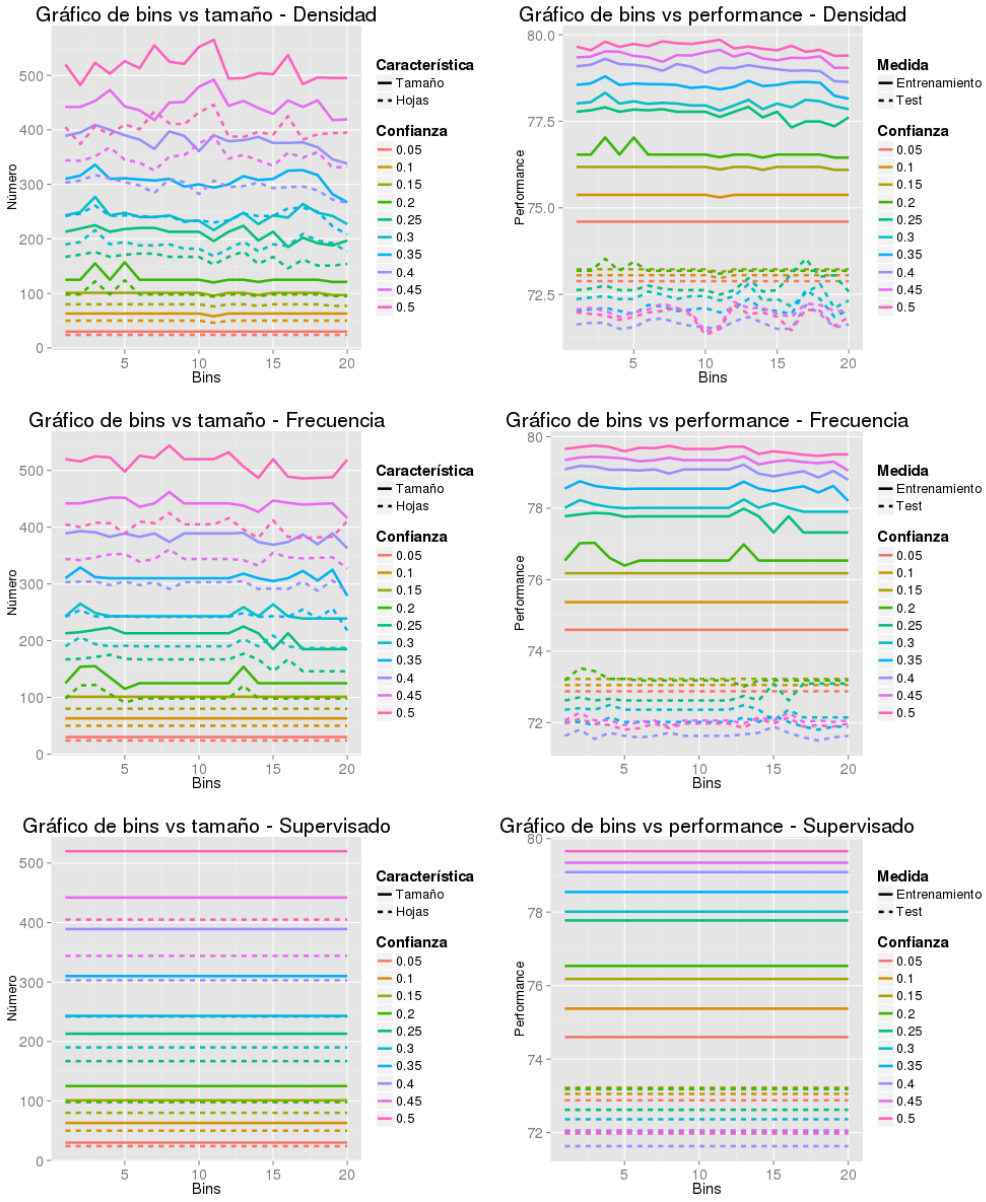

Figure 6: Gráficos de Ruido

5.1.1 Discretización por densidad

En este caso, como se observa en la Figura 6 se encontró que la variable discretizada prácticamente no incide en la performance (gráficos a la derecha) ni en en el tamaño del árbol (gráficos a la izquierda), lo cual estaría indicando que

es un atributo sin mayor poder de discriminación.

5.1.2 Discretización por frecuencias

En el caso de la discretización por frecuencias se observa una respuesta similar, con signos de independencia tanto del tamaño del árbol como de la performance respecto de la cantidad de bins de discretización. En comparación con el criterio anterior, aparentemente no introduce mejoras a la performance en el conjunto de entrenamiento.

5.1.3 Método supervisado

Para este criterio, dado que no puede elegirse la cantidad de bines sino que es una salida de la corrida algoritmo, corresponde observar cuántos niveles utiliza el algoritmo en la discretización. Con el dataset utilizado el J48 entrega una sola categoría (bin) que para la variable de clase no aporta nueva información.

5.2 Modelo reducido de variables con bajo poder discriminativo

En este apartado se realiza el ejercicio de discretización de la variable EDAD para un conjunto reducido de variables que incluye a la clase, la EDAD y el ESTADO CIVIL.

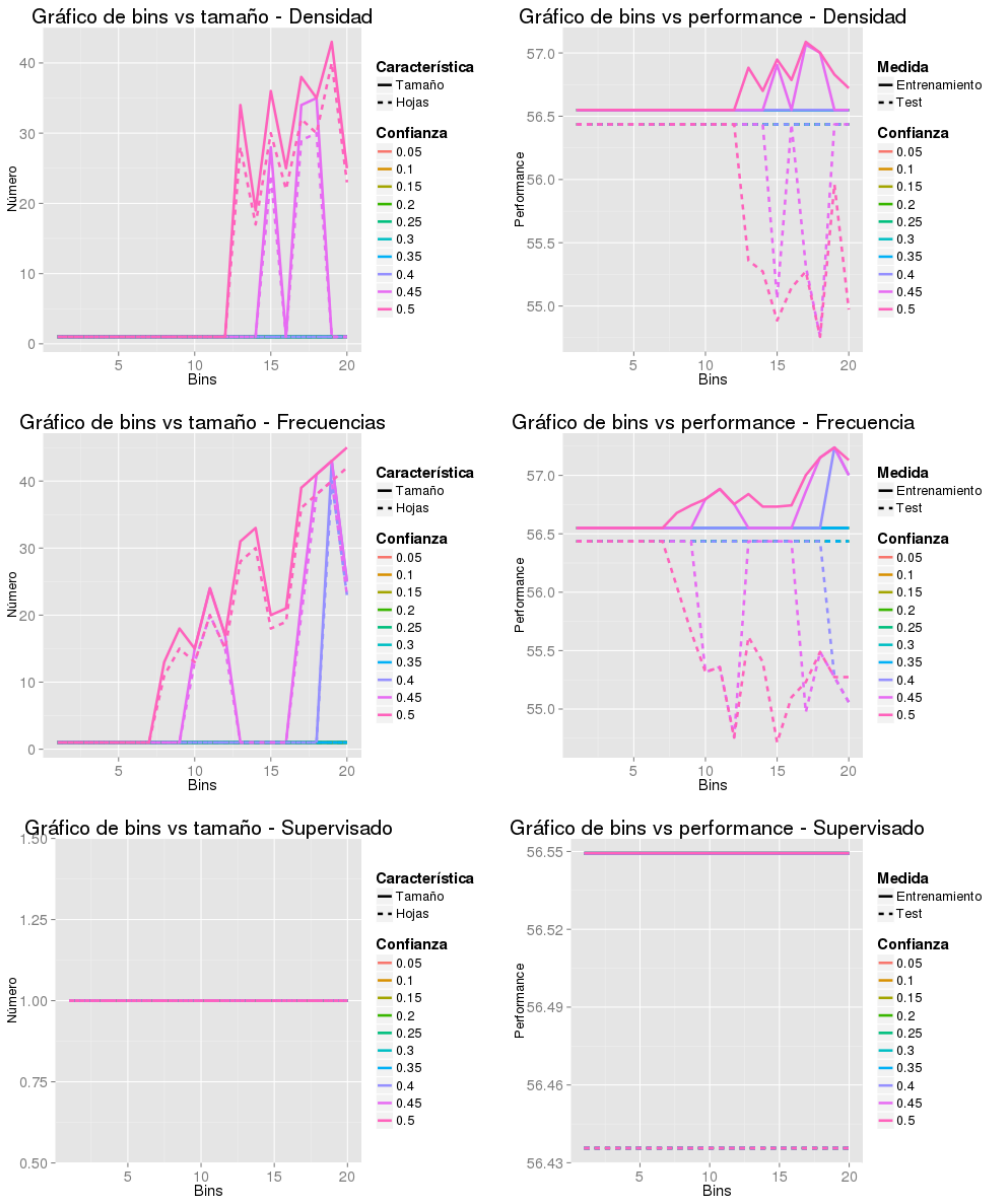

Figure 7: Categorización tenien

Esta última variable presenta un poder de discriminación similar al de la EDAD por lo que aparecen mejoras en la performance de entrenamiento a medida que se incrementan los niveles de la variable discretizada. Sin embargo, a la vez se observa un empeoramiento aún mayor en la preformance de test (ver

Figura 7). Por otro lado se observa la presencia de overfitting para mayor cantidad de bins para los criterios que no consideran la clase. El caso de la estrategia supervisada es representativo de la ausencia de nuevos aportes de información de esta variable en la determinación de la clase. En resumen, podríamos concluir que la variable numérica EDAD podría dejarse fuera de futuros modelos de predicción de la clase elegida ya que su contribución es prácticamente nula.

Trabajo futuro

Para poder verificar las propiedades del método de árboles de clasificación para su aplicación en problemas de la vida real, sería conveniente mezclar las diferentes pruebas hechas en este trabajo. Por ejemplo, surge como pregunta natural: ¿qué sucede cuando el árbol es sometido a ruido y a faltantes al mismo tiempo? Además de ello, sería conveniente intentar otros métodos para imputación de datos categóricos que sesguen menos los datos, como por ejemplo el método de imputación Hot Deck, el cual, busca individuos similares al individuo con el dato faltante para "donarle" el dato que le hace falta. [Rodgers, 1984].

Desde el punto de vista del dataset seleccionado, Latinobarómetro 2011, en este trabajo se observó que algunas de las variables incluidas en el análisis aportaban escasa información nueva por lo que el algoritmo no las mostraba en el árbol. La selección de variables fue realizada en base a un criterio de sentido común, sin ningún tipo de análisis estadístico de las mismas que ordenara de alguna forma su significación en la explicación de la variable de clase. Una mejora a este trabajo podría consistir en realizar un análisis de métodos factoriales para la selección de variables de modo de elegir aquellas más cercanas a la variable de clase. Por otro lado, una extensión natural sería utilizar los árboles generados para predecir la clase en datos de otros períodos, disponibles online, y evaluar la utilidad de los modelos entrenados y la estabilidad de los resultados de las medidas de tamaño y performance comparativamente.

Conclusiones

Con respecto a los datos faltantes, el árbol tiene una tolerancia similar a la mayoría de métodos estadísticos, en donde se admite al rededor de un 20% de datos faltantes.
En cuanto a la discretización de variables, en el ejercicio se observa claramente que la variable numérica seleccionada EDAD no aporta elementos determinantes en la predicción del valor de la clase. Al revisar su comportamiento discretizándola en un conjunto más reducido de variables con performance similar (con poca ganancia de información), en ningún momento forma parte del árbol, es decir que el algoritmo detecta el bajo poder de discriminación presente en la misma y la descarta lo cual resulta un comportamiento esperable.

References

[Hornik et al., 2009] Hornik, K., Buchta, C., and Zeileis, A. (2009). Open-source machine learning: R meets Weka. *Computational Statistics*, 24(2):225–232.

[Lagos, 2012] Lagos, M. (2012). Latinobarometro. opinion publica latinoamericana. Web.

[R Core Team, 2013] R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Rodgers, 1984] Rodgers, W. (1984). An evaluation of statistical matching. *Journal of Business and Economic Statistics*, page 91–102.

[Witten and Frank, 2005] Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2nd edition.