Informe trabajo práctico número uno - Aprendizaje Automático

Omar Ernesto Cabrera Rosero
Universidad de Buenos Aires
Email: omarcabrera@udenar.edu.co

Jimmy Mateo Guerrero Restrepo
Universidad de Buenos Aires
Email: jimaguere@gmail.com

***Resumen*—En este trabajo práctico se analizan las particularidades de la utilización de algoritmos para la generación de árboles de decisión, para la aplicación se utilizó un conjunto de datos de las pruebas de estado de calidad de la educación superior Saber Pro en Colombia, según el decreto 3963 del 14 de octubre de 2009 [1], a estudiantes próximos a culminar los programas académicos de pregrado que ofrecen las instituciones de educación superior.**

Keywords*—*Árboles de decisión, J48, Saber Pro, sobreajuste

I. Introducción

El objetivo del presente informe es presentar los resultados del análisis del comportamiento del algoritmo de aprendizaje de árboles de decisión J48 en función del Confidence Factor (CF) y evaluar aspectos como el sobreajuste y su robustez ante variaciones en el conjunto de datos. La variaciones en el conjunto de datos fueron teniendo en cuenta: datos faltantes, la tolerancia al ruido y la discretización de atributos numéricos.

El conjunto de datos utilizado es de las pruebas de estado Saber Pro 2012-1 en Colombia, uno de los objetivos del examen de estado de calidad de la educación superior Saber Pro, según el decreto 3963 del 14 de octubre de 2009, Ministerio de Educación Nacional [1], es comprobar el grado de desarrollo de competencias de los estudiantes próximos a culminar los programas académicos de pregrado que ofrecen las instituciones de educación superior. El examen está compuesto por pruebas que evalúan competencias genéricas y específicas. De acuerdo a los lineamientos Saber Pro del Instituto colombiano para el Fomento de la Educación Superior [2], todos los estudiantes deben presentar los módulos de competencias genéricas sin importar el programa de formación que cursen, que incluye competencias de razonamiento cuantitativo, lectura crítica, escritura e inglés.

En la competencia de razonamiento cuantitativo se evalúan los desempeños relacionados con uso de lenguaje cuantitativo y solución de problemas [3]. En la competencia de lectura crítica se evalúan los desempeños asociados a lectura, pensamiento crítico y entendimiento interpersonal [3]. En escritura se evalúa la competencia para comunicar ideas por escrito referidas a un tema dado [2] [3]. En inglés se evalúa la competencia del estudiante para comunicarse efectivamente en inglés.

El conjunto de datos pertenece a los datos de las pruebas Saber Pro 2012-1, la cual cuenta con 94 variables y 97.068 registros, a este conjunto de datos se le realizó un tratamiento de transformación de variables para reducir la dimesión, con esto se obtuvo un nuevo conjunto de datos con 31 variables y 96.775 registros.

Las variables que se usaron para el análisis representan, información personal del estudiante como lo muestra la tabla I, información familiar del estudiante como lo muestra la tabla II, información de la institución que cursa el estudiante como lo muestra la tabla III y la información socio-económica del estudiante como lo muestra la tabla IV.

La elección del conjunto de datos se fundó en las características enunciadas en [4] con respecto a la clase de problemas que son apropiados para trabajar con árboles de decisión particularmente el hecho de que cada atributo toma un número pequeño de valores posibles.

II. Diseño experimental

El diseño experimental se lo realizó usando el algoritmo J48, el cual es una implementación de [5]

Tabla I. Información personal estudiante

Atributo/Clase	Nombre	Tipo	Descripción	Estadística
Clase	mod_razona_cuantitativo	Cualitativa Nominal	Nivel asignado al modulo de Razonamiento Cuantitativo.	mode = BAJO LA MEDIA (48757), least = SOBRE LA MEDIA (48018)
Atributo	estu_genero	Cualitativa Nominal	Género alumno.	mode = F – Femenino(40084),least= F – Masculino(56691)
Atributo	estu_edad	Cuantitativa	Edad alumno al momento de tomar la prueba.	Min=9.00, 1st Qu=22, Median=24, Mean=26.03, 3rd Qu=28, Max=74.
Atributo	estu_estado_civil	Cualitativa Nominal	Estado civil alumno.	mode = Soltero(a)(77732), least = Viudo(a)(163)
Atributo	estu_hogar_actual	Cualitativa Nominal	Su hogar actual.	mode = Es el habitual-permanente(79298), least = Es temporal por razones de estudio u otra razón(17477)
Atributo	estu_sn_cabeza_fmlia	Cualitativa Nominal	Es cabeza de familia.	mode = No(80380), least = Si(16395)
Atributo	estu_grupo_referencia	Cualitativa Nominal	Nombre del grupo de referencia al que pertenece el programa académico del evaluado.	mode = CIENCIAS ECONOMICAS Y ADMINISTRATIVAS(26557), least = ARTES - DISEÑO - COMUNICACION(30)
Atributo	estu_pje_creditos	Cualitativa ordinal	Porcentaje de créditos cursados y aprobados.	mode = MAS DE 90 %(46506), least = MENOS DEL 75 %(2883)
Atributo	estu_titulo_bto	Cualitativa Nominal	Título de bachiller obtenido.	mode = Académico(73955), least = Técnico(4267)
Atributo	estu_financiacion_matricula	Cualitativa Nominal	Fuente de los recursos con que canceló la Matrícula.	mode = PADRES(38622), least =PROPIO, BECA O SUBSIDIO(232)
Atributo	estu_estrato	Cualitativa ordinal	Estrato socioeconómico de la vivienda donde reside actualmente su hogar habitual o permanente según el recibo del servicio de energía Eléctrica?	mode = Estrato3(36274), least= Vive en una zona rural donde no hay estratificación socioeconómica(112)
Atributo	estu_trabaja	Cualitativa Nominal	Si el alumno usted actualmente?	mode NO(42914), least = SI, POR SER PRACTICA OBLIGATORIA DEL PROGRAMA(7300)
Atributo	estu_metodo_prgm	Cualitativa Nominal	Metodología del programa académico que pertenece el evaluado.	mode = PRESENCIAL(84059), least = SEMIPRESENCIAL(3)
Atributo	estu_area_conoc	Cualitativa Nominal	Nombre del área de conocimiento a la que pertenece el programa académico del evaluado.	mode = ECONOMIA, ADMINISTRACION, CONTADURIA Y AFINES(27034), least = AGRONOMIA VETERINARIA Y AFINES(1523)
Atributo	num_estu_zona	Cualitativa ordinal	Nivel estudiantes por zona	mode = Media(56900), least=Baja(6408)

desarrollado por Ross Quinlan. Éste a su vez es una extensión del algoritmo ID3, del mismo autor.

En los diseños experimentales se dividieron en dos dos partes, un 80 % de las instancias se utilizaron para el entrenamiento del árbol, mientras que el 20 % restante se utilizó para testearlo. La división se hizo de forma aleatoria, manteniendo la misma proporción de clases en cada uno de los subconjuntos de datos.

Para la implementación de los experimentos se utilizó la biblioteca 'RWeka' [6], una implementación de los algoritmos de WEKA para lenguaje de programación R, el código fuente de la implementación puede ser consultada en el repositorio[1]

[1] Repositorio tp1AA https://github.com/poldrosky/tp1AA

A. Sobreajuste y poda

Se conoce como sobreajuste u overfitting, al efecto que consiste en pegarse mucho a los datos de entrenamiento, dicho en otros términos, sobre-entrenar el árbol, perjudicando la performance sobre el conjunto de validación. [4] define el overfitting como: "A hypothesis overfits the training examples if some other hypothesis that fits training examples less well actually performs better over the entire distribution of instances."

Metodología utilizada: Se ejecutaron corridas del algoritmo J48 variando la función de poda. Para ello, se utilizó como parámetro el ConfidenceFactor (CF). Se iteró desde 2,5 % a 50 %, con intervalos de 2,5 %. A menor porcentaje de CF, el algoritmo incurre en un mayor nivel de poda.

Resultados esperados: En función de lo enun-

Tabla II. Información familiar estudiante

Atributo	Nombre	Tipo	Descripción	Estadística
Atributo	fami_num_pers_cargo	Cuantitativa	Tiene personas a cargo (cuando es cabeza de familia).	mode = No(68472), least = Si(28303)
Atributo	fami_nivel_educa_padres	Cualitativa Nominal	Nivel educativo de los padres.	mode = SECUNDARIA (BACHILLERATO) COMPLETA(19899),least = NINGUNO(661)
Atributo	fami_ocup_madre	Cualitativa Nominal	Cuál es actualmente la ocupación de su madre? (o última si Falleció?).	mode = Hogar r(41120), least = Empleado-con cargo-como-director(a)(1487)
Atributo	fami_ocup_padre	Cualitativa Nominal	Cuál es actualmente la ocupación de su padre? (o última si Falleció?)	mode = trabajador por cuenta propia(23955), Least = Hogar(1943)
Atributo	fami_nivel_sisben	Cualitativa ordinal	Su familia está clasificada en el nivel 1, 2 ó 3 del SISBEN?	mode = No está clasificada por el SISBEN(54353), least = Está clasificada en otro nivel(804)
Atributo	fami_ing_fmliar_mensual	Cualitativa ordinal	Cuál es el total de ingresos mensuales de su hogar habitual o permanente (por trabajo u otros conceptos) en salarios mínimos:SM-?	mode = DOS SALARIOS(30151), least = SIETE SALARIOS(4033)

Tabla III. Información institución estudiante

Atributo	Nombre	Tipo	Descripción	Estadística
Atributo	inst_tipo	Cualitativa Nominal	Tipo institución	mode = PRIVADA(58025), least = REGIMEN ESPECIAL(47)
Atributo	inst_caracter_academico	Cualitativa Nominal	Carácter Académico.	mode = ACADEMICO(73955) ,least = ESCUELA TECNOLOGICA(4267)
Atributo	inst_acreditada	Cualitativa Nominal	Institución alumno acreditada?	mode = INSTITUCION NO ACREDITADA(79807), least = INSTITUCION ACREDITADA(16968)
Atributo	inst_programa_zona	Cualitativa Nominal	Zona del programa de estudio del alumno.	mode = BOGOTA(33467) , least = MARINILLA(2)
Atributo	num_instituciones_zona	Cualitativa ordinal	Nivel instituciones por zona	mode = Alta(49946), least = Baja (19903)

Tabla IV. Información socioeconómica estudiante

Atributo	Nombre	Tipo	Descripción	Estadística
Atributo	eco_condicion_vivienda	Cualitativa ordinal	Condición económica vivienda.	mode = BUENA(78857), least = REGULAR(2721)
Atributo	eco_condicion_hogar	Cualitativa ordinal	Condición económica hogar.	mode = CONDICION VIVIENDA BUENA(53131), least = CONDICION VIVIENDA MALA(9139)
Atributo	eco_condicion_transporte	Cualitativa ordinal	Condición económica de transporte.	mode = CONDICION TRANSPORTE PUBLICO(63499), least =CONDICION TRANSPORTE PARTICULAR(33276)
Atributo	eco_condicion_tic	Cualitativa ordinal	Condición tecnológica hogar.	mode = CONDICION HOGAR BUENA(85270), least = CONDICION HOGAR MALA(4706)
Atributo	eco_condicion_vive	Cualitativa ordinal	Condición hacinamiento vivienda.	mode = SIN HACINAMIENTO(93333), least = HACINAMIENTO CRITICO(445)

ciado en la descripción precedente, se espera que a medida que crezca el tamaño del árbol, medido en función de la cantidad de nodos, crezca monótonamente la performance sobre el conjunto de entrenamiento, y crezca y luego disminuya sobre el de validación. Asi mismo, por cómo opera el CF, el tamaño del árbol debería aumentar a medida que crece su valor.

Análisis de los resultados: Los resultados obtenidos una vez realizado el experimento se pueden resumir en las figuras presentadas a continuación:

La figura 1 se observa que a medida que se incrementa el valor del CF la cantidad de nodos también crece, mostrando una clara relación positiva entre el CF y el tamaño del árbol; en ese sentido cabe recordar que “The default confidence value is set at 25 % and works reasonably well in most cases; possibly it should be altered to a lower value, which causes more drastic pruning” [7]; es decir que un valor bajo del CF implica una poda muy grande y por otro lado un valor muy alto señalaría que el árbol no sufriría poda alguna. Es

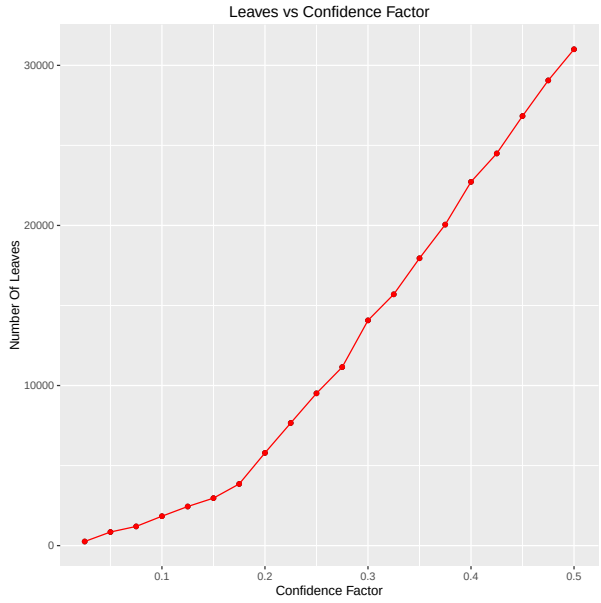


Figura 1. Leaves vs CF

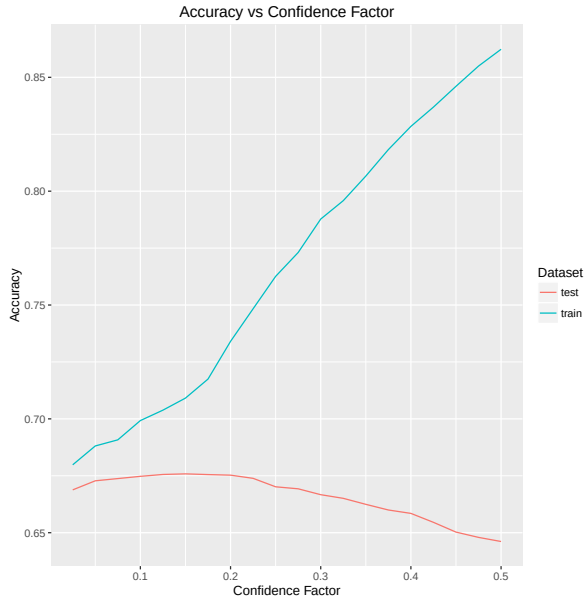


Figura 2. Accuracy vs CF

razonable entonces que en el gráfico una poda muy grande esté asociada a una poca cantidad de nodos y por el contrario una poda pequeña se vincule a una cantidad grande de nodos, debido a que se dejó crecer el árbol sin ninguna restricción; sin embargo esta figura, tomada de manera aislada, no dice absolutamente nada acerca de la performance de la técnica predictiva sobre el conjunto de entrenamiento y de validación.

La figura 2 es la que muestra la relación entre la performance y el CF. Es importante recordar que un CF mayor significa un menor nivel de poda y, teniendo en cuenta la relación puesta en evidencia en el gráfico anterior, un incremento del tamaño del árbol. De este gráfico es importante destacar que a medida que el árbol es más grande la performance sobre el conjunto de entrenamiento es mayor, en tanto que en el caso del conjunto de validación, el rendimiento también crece hasta un cierto nivel, para luego disminuir y finalmente estabilizarse manteniéndose casi constante. Todo lo mencionado hace suponer que si se toman valores muy grande de CF se colapsaría en el fenómeno conocido como overfitting, donde el árbol de decisión clasifica muy bien las instancias pertenecientes al conjunto de entrenamiento, sin embargo no llega a tener una predicción adecuada sobre nuevas instancias no conocidas.

Conclusión: Los resultados obtenidos se alinean con los esperados. Para el conjunto de datos utilizado en el experimento, el árbol crecerá a medida que los niveles de poda se restrinjan mediante la elección del valor del CF. Este crecimiento del árbol influirá positivamente sobre la performance en el conjunto de entrenamiento; sin embargo en el caso de los datos de validación el crecimiento del árbol tendrá un efecto positivo sobre su performance hasta un cierto punto, siendo perjudicial una vez superado ese límite. Por lo que parece recomendable elegir un valor de CF de entre 15 % y 18 %, que maximizaría los niveles de predicción sobre los datos no conocidos. Asimismo esto se traducirá en árboles con lenguajes de hipótesis no tan expresivos, pero que explicarán mejor los hechos; los cuales según Occam son las teorías preferibles en condiciones similares. “Occam’s Razor shaves philosophical hairs off a theory”

En la figura 3 se muestra la grafica de la curva ROC para el mejor árbol,

B. Tratamiento de datos faltantes

Descripción: En la práctica, es común encontrarse y tener que trabajar con datasets que contienen datos nulos o incompletos. Existen distintos métodos para tratar con ellos, en esta sección se exploraron dos. Los cuales tratan de rellenar los datos faltantes con el valor modal del atributo, y se diferencian en que uno toma en cuenta el valor de la clase para el individuo o registro analizado, mientras que el otro no tiene en cuenta el valor de la clase.

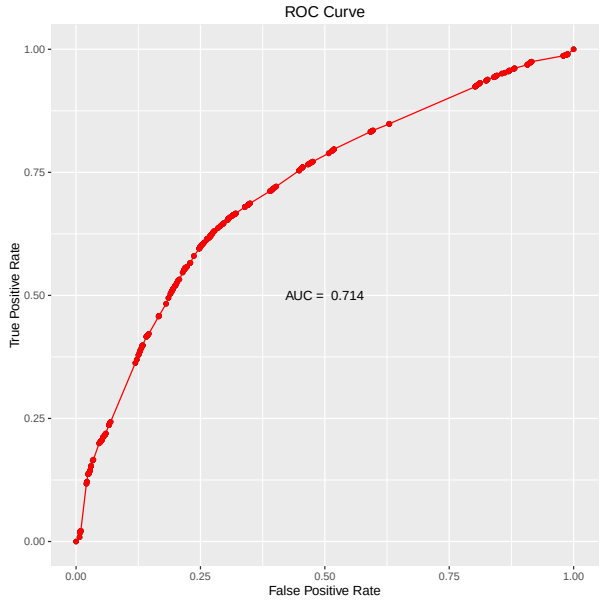


Figura 3. Curva ROC mejor árbol

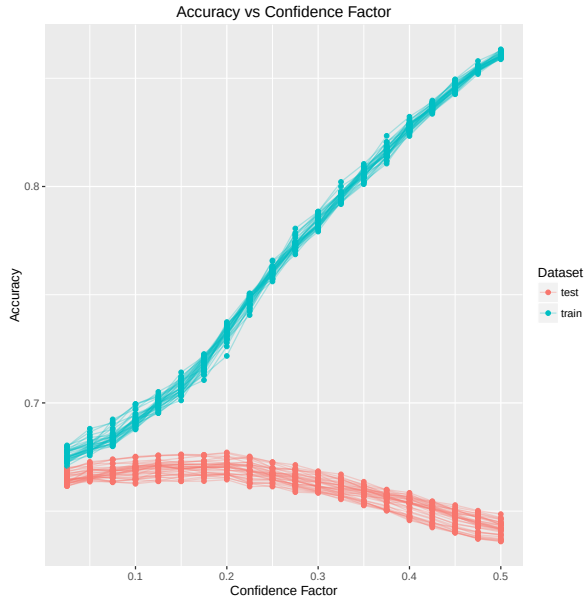


Figura 4. Accuracy vs CF with missing data

Metodología utilizada: Para el tratamiento de datos faltantes se indujo valores nulos al 80 % del dataset en los atributos con mayor GainRatio, se preservó el 20 % de los datos para validación. A continuación se detallan las características de los atributos contemplados:

Tabla V. Atributos con mayor GainRatio

Atributo	GainRatio
inst_acreditada	0.0436292021
estu_metodo_prgm	0.0295357038

El porcentaje de imputación de datos faltantes fue variando de 0 a 0.85 con incrementos de 0.025 generando 36 datasets con datos faltantes. En cada generación de datos faltantes, se relleno con la moda los datos nulos del atributo inst_acreditada y con la modaclase para los valores nulos del atributo estu_metodo_prgm. Después se construyeron modelos con las estrategias de relleno de datos faltantes anteriormente descritas variando el CF (Confidence Factor) de 0 a 0.5 y finalmente se evaluaron los mismos con el set de validación.

Resultados esperados: La performance sobre el set de validación no debería verse sensiblemente afectada, al menos al introducir proporciones bajas o medias de datos faltantes, dada la robustez del algoritmo J48 para lidiar con esta problemática.

Con respecto al tamaño del árbol, es de esperar que el mismo aumente a medida que aumenta la función de poda y no por el porcentaje inducido de datos faltantes.

Análisis de los resultados: En la figura 4 Accuracy vs Confidence factor with missing data, se observa un patrón de comportamiento en las curvas de entrenamiento y validación para la accuracy, dicho patron es similar a las curvas generadas en las corridas donde no se imputaron datos faltantes, esto nos dice que el algoritmo J48 presenta resistencia y robustez a datos faltantes ya que la performance se ve afectada más por la función de poda que por la imputación de datos faltantes.

En la figura 5 Leaves vs missing percentage, se observa claramente que el tamaño del árbol incrementa debido al aumento de la función de poda (confidence factor), respecto al porcentaje de datos faltantes no se mira que este afecte en el tamaño del arbol.

En la figura 6 Accuracy vs missing percentage,se observa un comportamiento similar para los diferentes porcentajes de datos faltantes imputados, la performance se comporta de igual forma manteniendo la variación debido a la función de poda independientemente del porcentaje de datos faltantes.

Los anteriores supuestos los podemos confirmar claramente en la figura 7 Missing Percentage vs Confidence Factor, donde se puede mirar que a valores

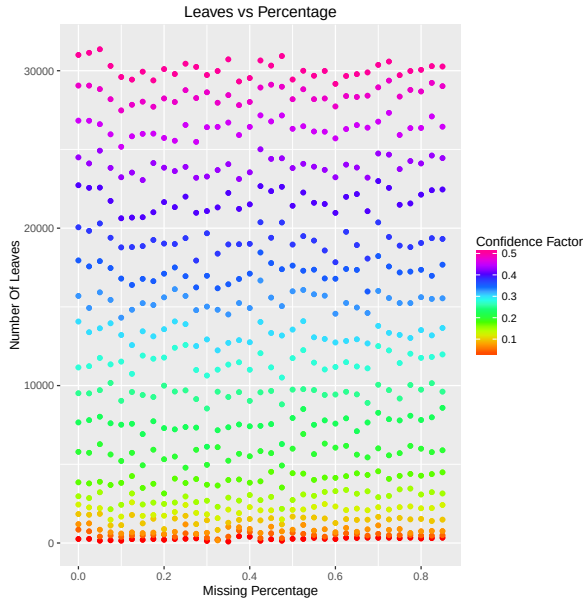


Figura 5. Leaves vs missing percentage

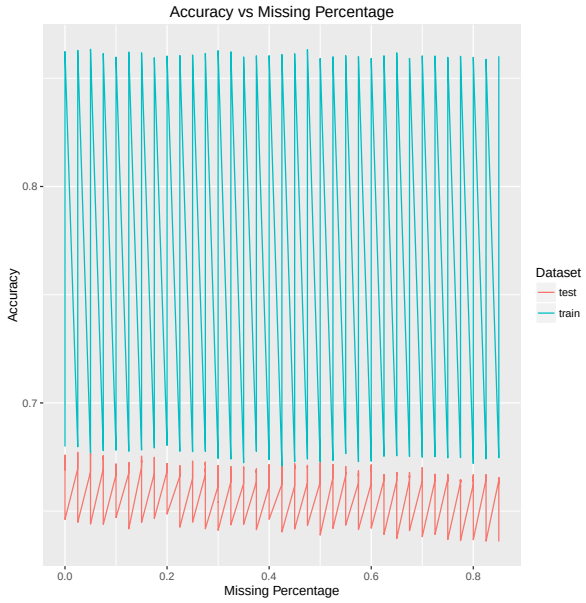


Figura 6. Accuracy vs missing percentage

bajos CF hay un alto performance en el set de datos de validación, y a medida que aumenta el valor de CF (eje x), también aumenta el valor de la performance del set de entrenamiento (tamaño de la burbuja) así como baja la performance del set de testing (color de la burbuja) por este motivo el arbol se sobre ajusta al conjunto de entrenamiento. Esta lectura es posible realizarla según el porcentaje de datos faltantes (de 0 a 0,85, en el eje Y; Cada serie de burbujas horizontales corresponden al mismo experimento).

Conclusión: Los resultados se presentaron, en mayor o menor medida, en línea con lo esperado. La performance no se vio sensiblemente deteriorada al igual que el tamaño del árbol, aliniandose a los resultados esperados. por tanto el algoritmo de árboles J48 es robusto a datos faltantes ya que el rendimiento y tamaño del árbol observados en los gráficos se ven afectados por la función de poda y no por la variación de datos faltantes.

En la figura 7 Missing percentage vs Confidence factor,

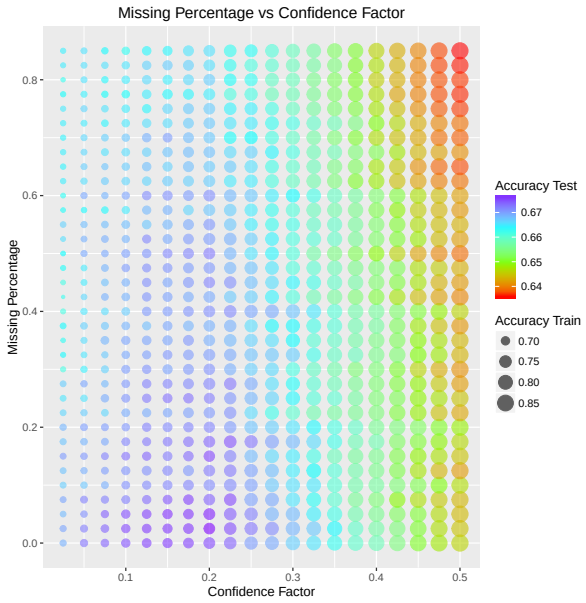


Figura 7. Missing percentage vs CF

C. Tratamiento de ruido

En la figura 8 se muestra la precisión del árbol en función del CF

En la figura 9 se muestra el número de hojas en función de porcentaje de datos con ruido.

En la figura 10 se muestra la precisión del árbol en función del porcentaje.

En la figura 11 se muestra el porcentaje de ruido en función del CF.

D. Discretización de datos numéricos

Descripción: La discretización consistió en convertir atributos numéricos en categóricos, las estrategias de discretización que se utilizaron fueron: por

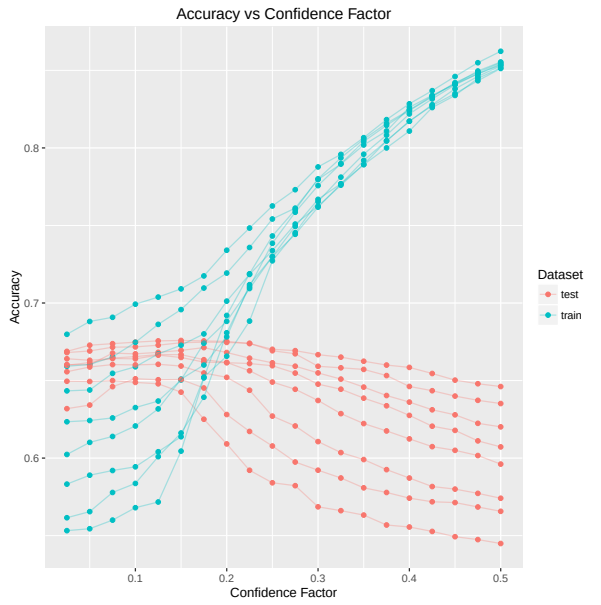


Figura 8. Accuracy vs CF with noise data

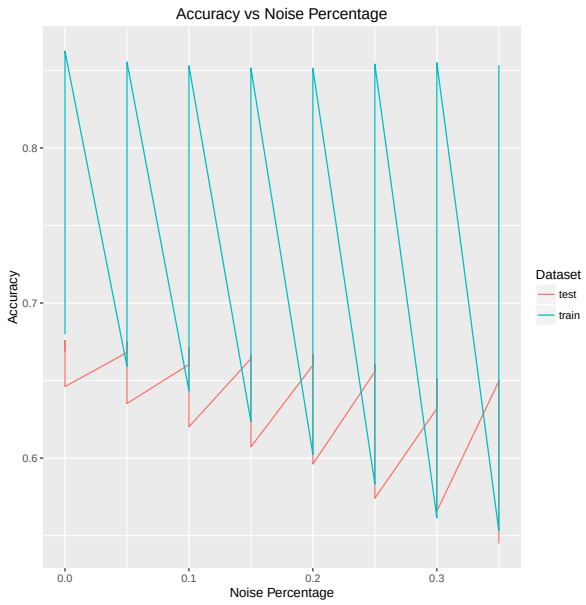


Figura 10. Accuracy vs noise percentage

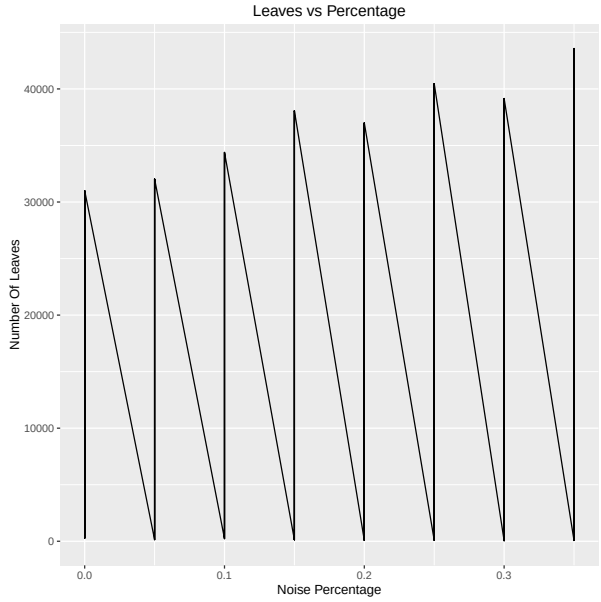


Figura 9. Leaves vs noice percentage

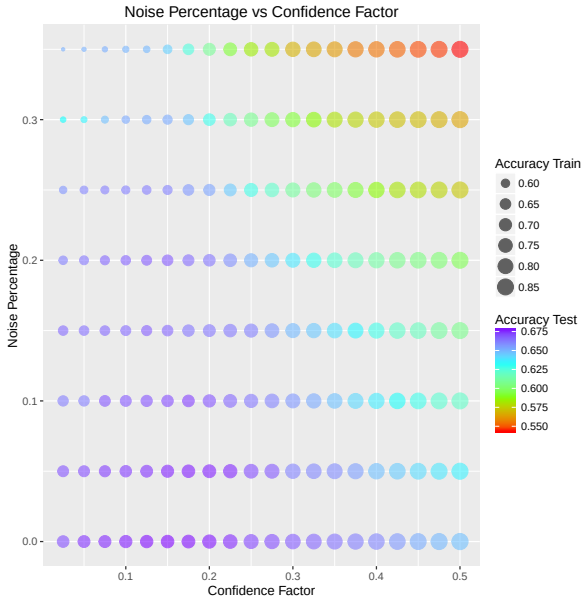


Figura 11. Noise percentage vs CF

frecuencia (generan bins en cantidades similares de instancias en cada uno), densidad (con bins del mismo ancho), y supervisado (utiliza un modelo de discretización que toma en cuenta la clase en el armado de los bins)

Metodología utilizada: para este experimento, se crearon conjuntos de datos para cada una de las tres estrategias de discretización previamente comentadas, para crear los bins por frecuencia y densidad se utilizo la biblioteca 'arules', y para la discretización supervisada se uso la biblioteca 'RWeka'.

Para la discretización por frecuencia y densidad se varió la cantidad de bins entre 1 y 20. Cabe resaltar que elegir 1 bin implica prescindir del atributo discretizado a la hora de construir el árbol, ya que el mismo tomaría un solo valor.

Para la discretización supervisada, no es posible

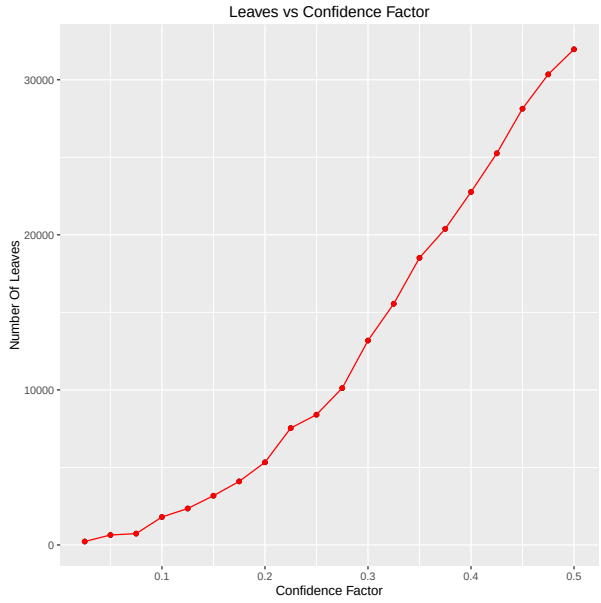


Figura 12. Leaves vs CF

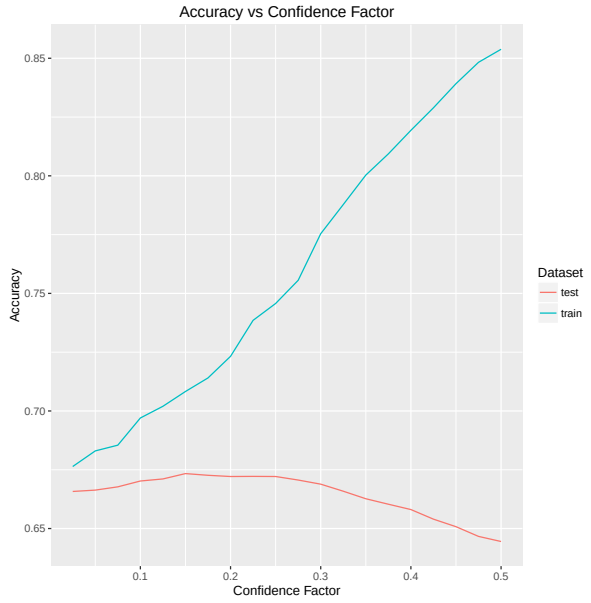


Figura 13. Accuracy vs CF

seleccionar la cantidad de bins de salida ya que la misma es determinada por el propio filtro utilizando un criterio de entropía sobre ganancia de información.

Se hicieron corridas sobre las tres familias utilizando el algoritmo J48 variando el CF de forma análoga a lo hecho en la Sección II.

Con las salidas, se analizó, para los casos de discretización no supervisada, la relación entre la cantidad de bins, la cantidad de nodos del árbol de clasificación y su performance. Asimismo, se compararon las distintas estrategias entre sí, y contra los resultados de la Sección II.

Análisis de los resultados: Los resultados obtenidos una vez realizado el experimento se pueden resumir en las figuras presentadas a continuación.

Las figuras 12, 13 y 14 hace referencia a la discretización supervisada, se puede dar cuenta que son muy similares a la figuras de la subseccion sobreajuste y poda en la cual no hay discretización.

En la figura 13 se muestra la grafica de la curva ROC para el mejor árbol,

En la figura 14 se muestra la grafica de la curva ROC para el mejor árbol,

En la figura 15 se observa que hay un nivel de asociación positivo entre los bins utilizados en la estrategia de discretización y los nodos del árbol de

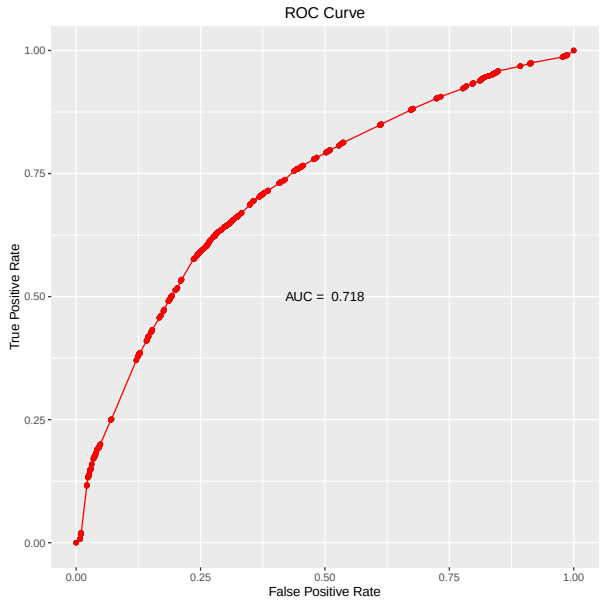
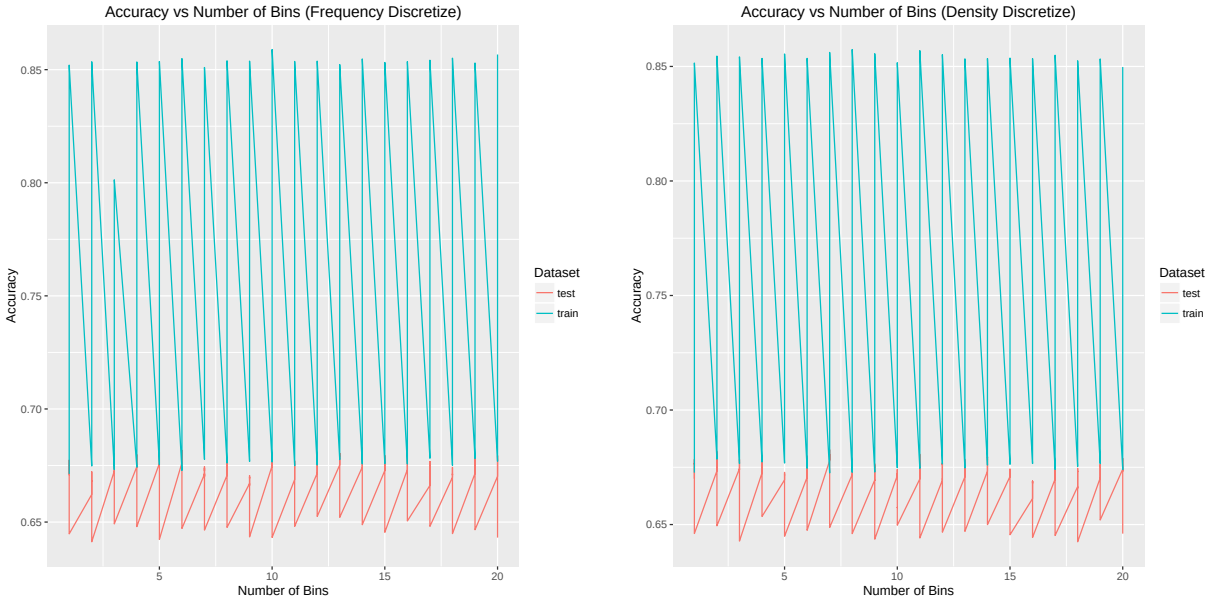


Figura 14. Curva ROC mejor árbol

clasificación, en línea con lo esperado. Las líneas, que representan la estrategia empleada se cruzan en varios tramos de la representación, lo que no permite determinar si la asociación enunciada es más fuerte en una o en otra.

En la figura 16 muetran la precisión del árbol en función del CF para la discretización por frecuencia y densidad respectivamente, se observa que las precisiones se encuentran en la discretización por densidad



(a) Frequency discretize

(b) Density discretize

Figura 16. Accuracy vs CF (Discretización)

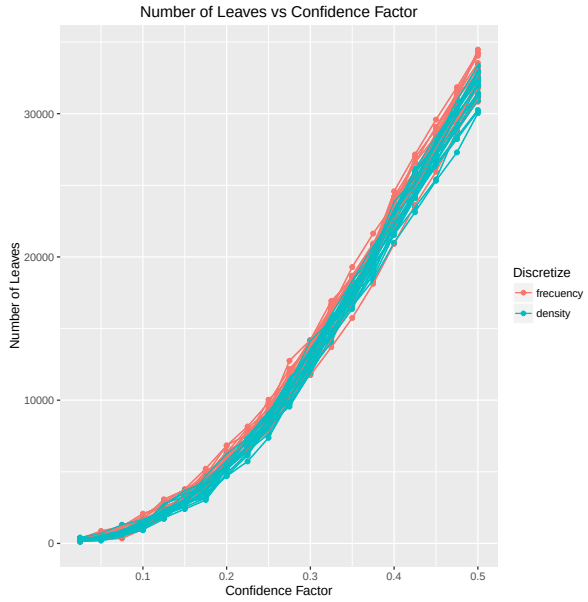


Figura 15. Leaves vs missing percentage with discretize

pero la diferencia es muy mínima.

III. Conclusiones

Referencias

[1] MEN, "Decreto 3963 de octubre 14 de 2009, por el cual se reglamenta el examen de estado de calidad de la educación superior," http://www.mineducacion.gov.co/1621/article-205955.html, Ministerio de Educación Nacional (MEN), 2009, accessed: 2014-08-20.

[2] ICFES, "Lineamientos saber pro," http://aprendeenlinea.udea.edu.co/lms/moodle/file.php/532/Lineamientos_SABER_PRO_2011_2_30_08_1_.pdf, 2011, accessed: 2014-08-20.

[3] ——, "Examen saber pro, junio de 2012–i. módulos de competencias genéricas y especificas disponibles. evaluación de la calidad de la educaciónsuperior. bogotá," http://www.icfes.gov.co/examenes/component/docman/doc_download/151-saber-pro-modulos-de-competencias-genericas-y-especificas-disp Itemid=f, 2012, accessed: 2014-08-20.

[4] T. Mitchell, *Machine Learning*, ser. McGraw-Hill International Editions. McGraw-Hill, 1997. [Online]. Available: https://books.google.com.ar/books?id=EoYBngEACAAJ

[5] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[6] K. Hornik, C. Buchta, and A. Zeileis, "Open-source machine learning: R meets Weka," *Computational Statistics*, vol. 24, no. 2, pp. 225–232, 2009.

[7] I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*, ser. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science, 2005. [Online]. Available: https://books.google.com.ar/books?id=QTnOcZJzlUoC