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 realización de este trabajo se utilizó un conjunto de datos de las pruebas de estado saberpro que se realizan en Colombia a estudiantes universitarios ...

Keywords

árboles de decisión, J48, ICFES, saberpro

I. Introducción

A continuación se detallan las características de los atributos contemplados:

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)

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)

[?]

II. Diseño experimental

A. *Sobreajuste y poda*

texto

En la figura **??** se muestra la grafica el número de hojas en función de la función de poda.

En la figura **??** se muestra la grafica el performance en función de la función de poda.

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

B. *Faltantes*

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

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

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

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

C. *Ruido*

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

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

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

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

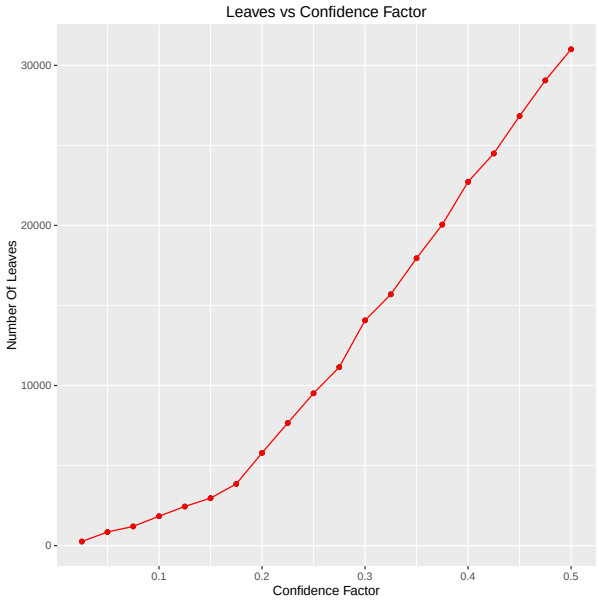

Figura 1. Number of leaves vs Confidence factor

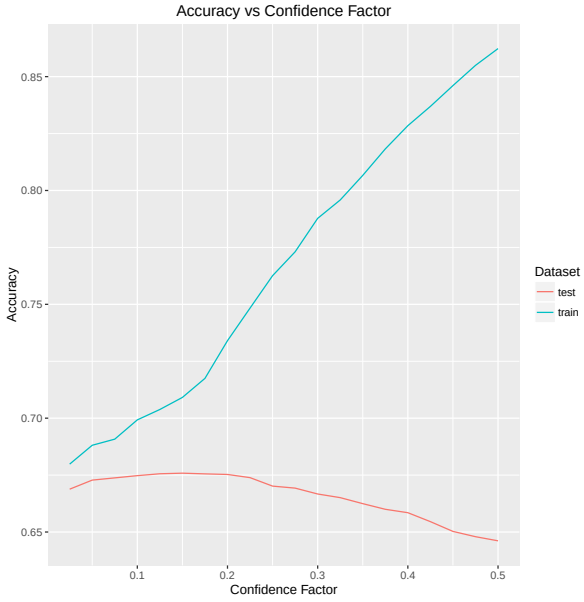

Figura 2. Accuracy vs Confidence factor

D. Discretización

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

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

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

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

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

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

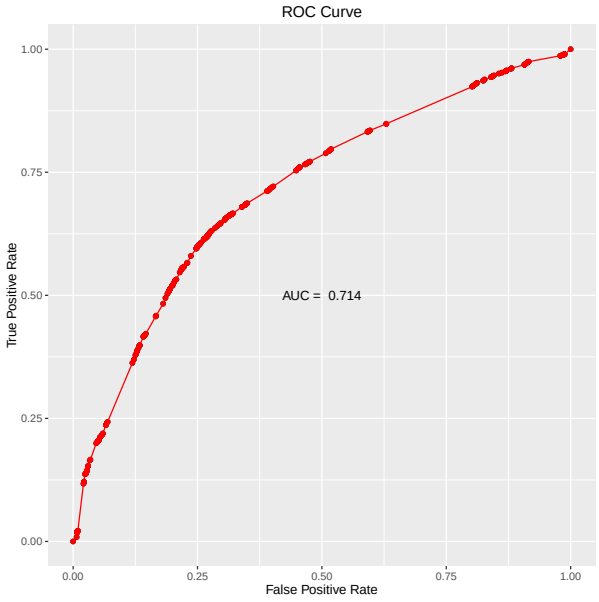

Figura 3. Curva ROC mejor árbol

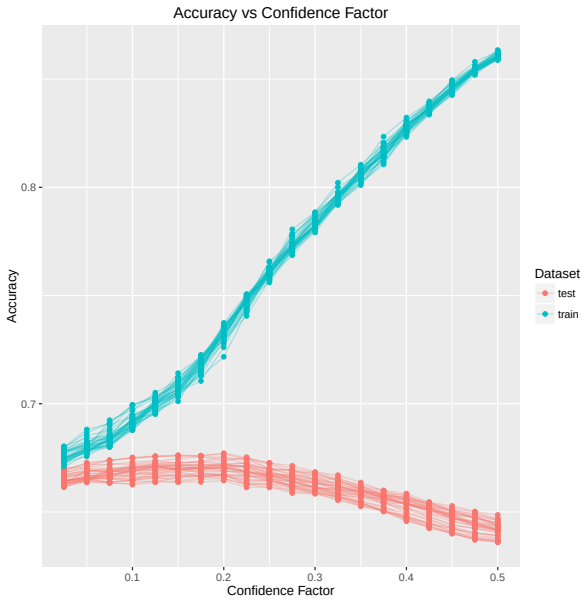

Figura 4. Accuracy vs Confidence factor with missing data

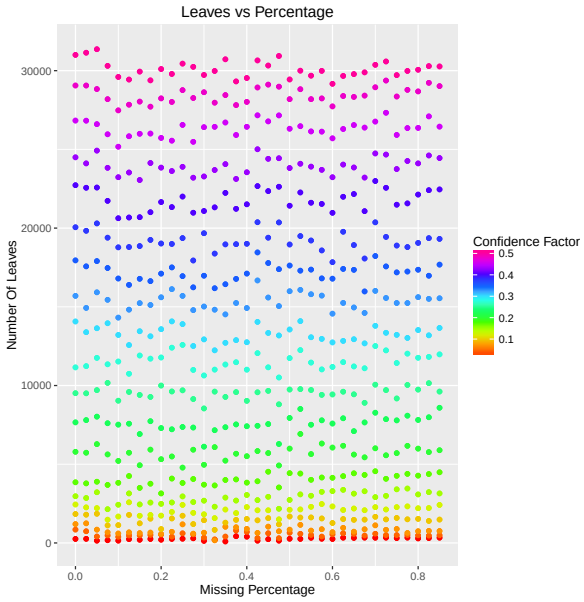

Figura 5. Leaves vs missing percentage

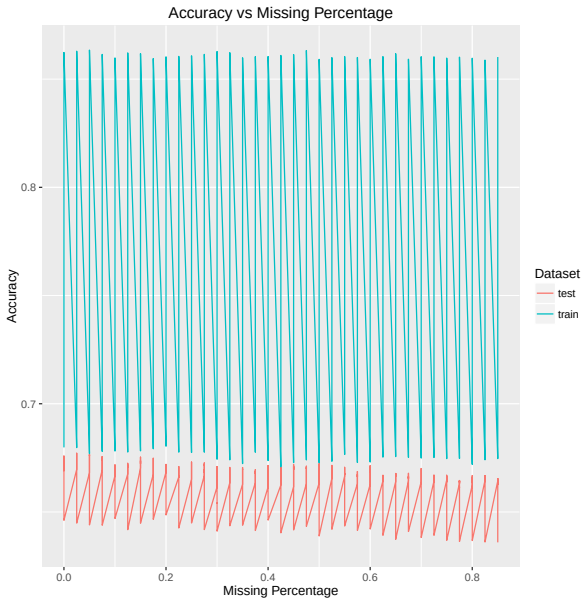

Figura 6. Accuracy vs missing percentage

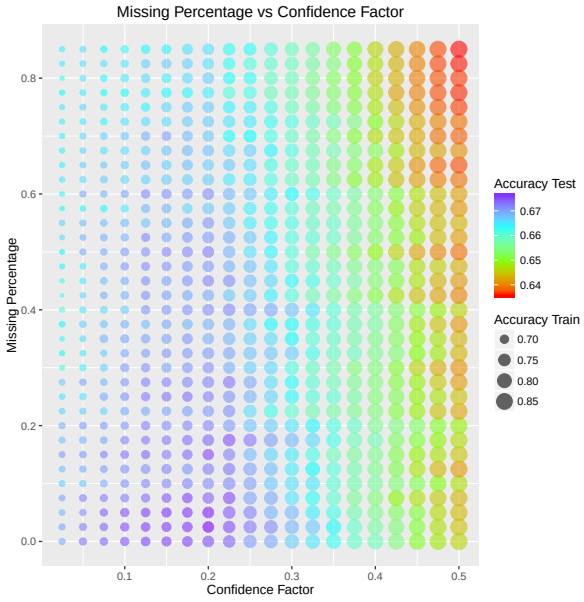

Figura 7. Missing percentage vs Confidence factor

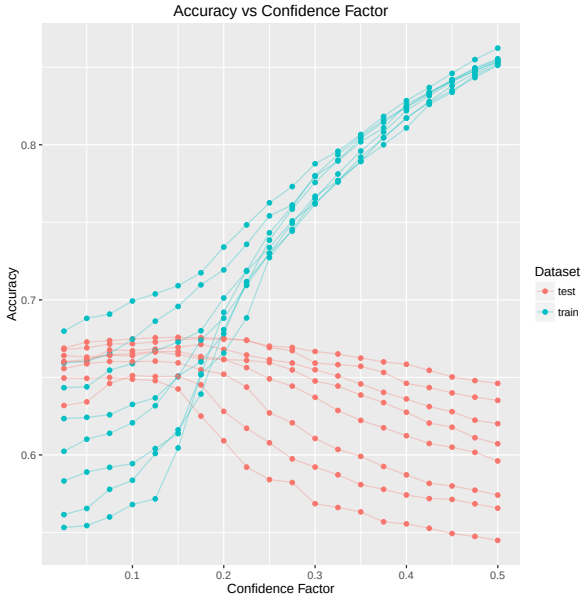

Figura 8. Accuracy vs Confidence factor with noise data

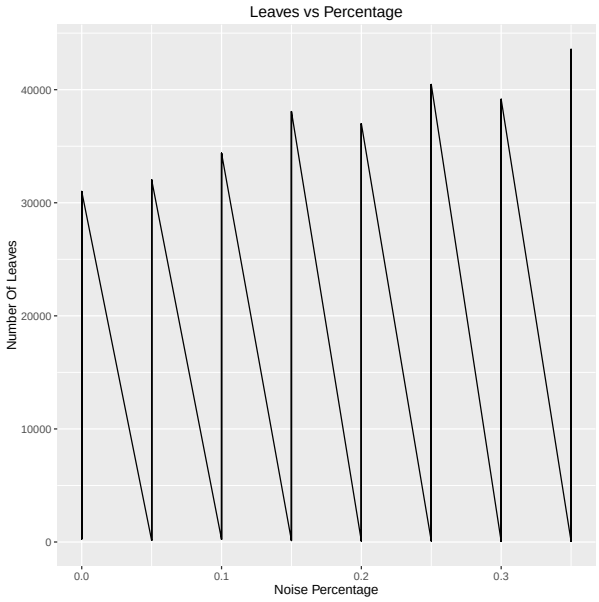

Figura 9. Leaves vs noice percentage

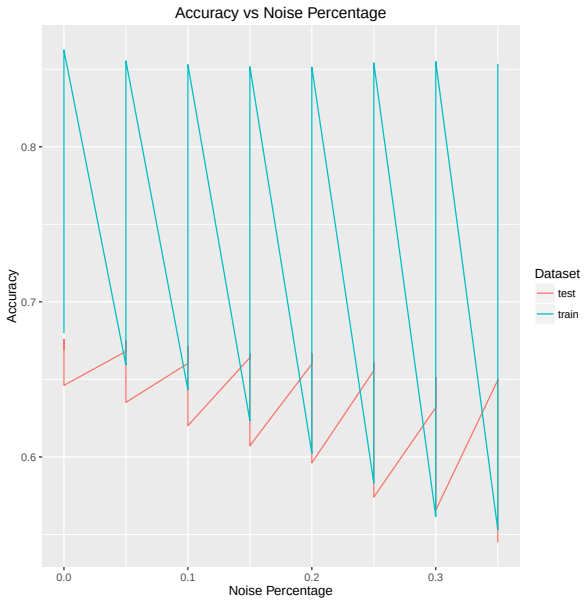

Figura 10. Accuracy vs noise percentage

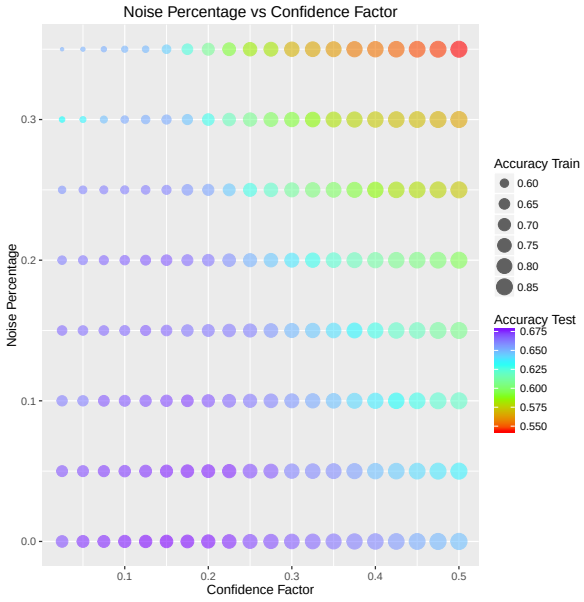

Figura 11. Noise percentage vs Confidence factor

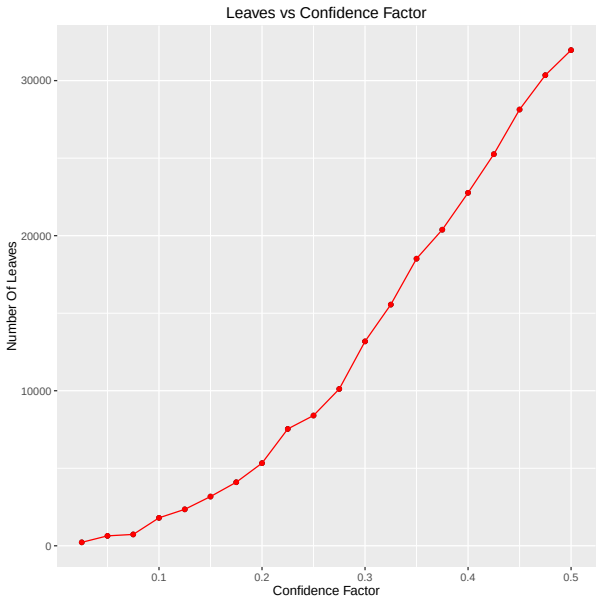

Figura 12. Number of leaves vs Confidence factor with supervised discretize

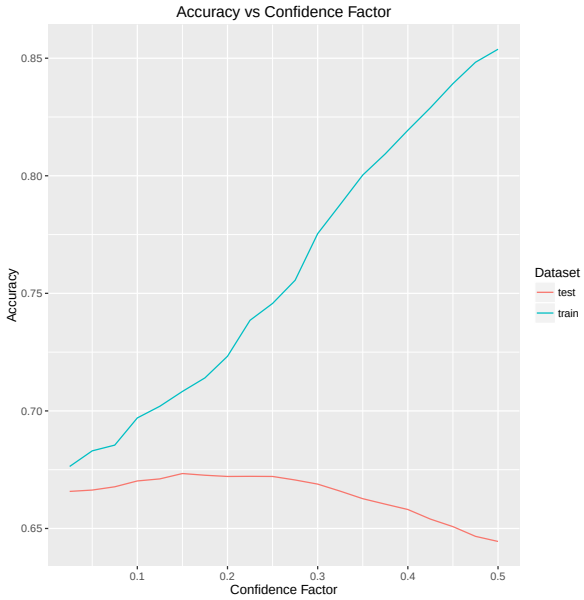

Figura 13. Accuracy vs Confidence factor with supervised discretized

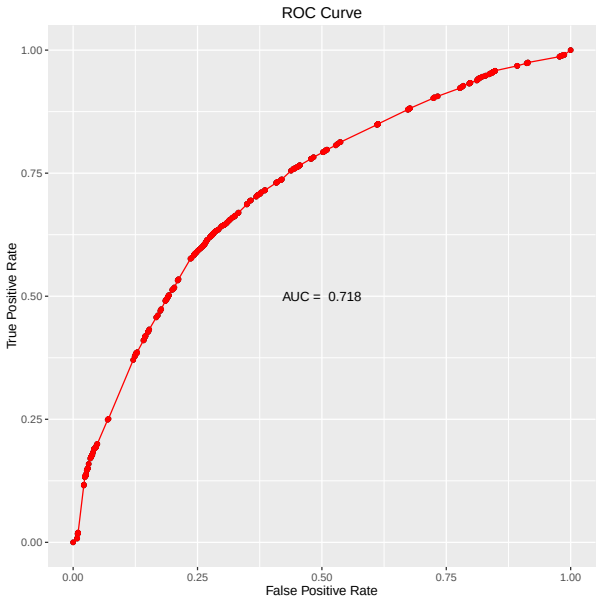

Figura 14. Curva ROC mejor árbol with supervised discretized

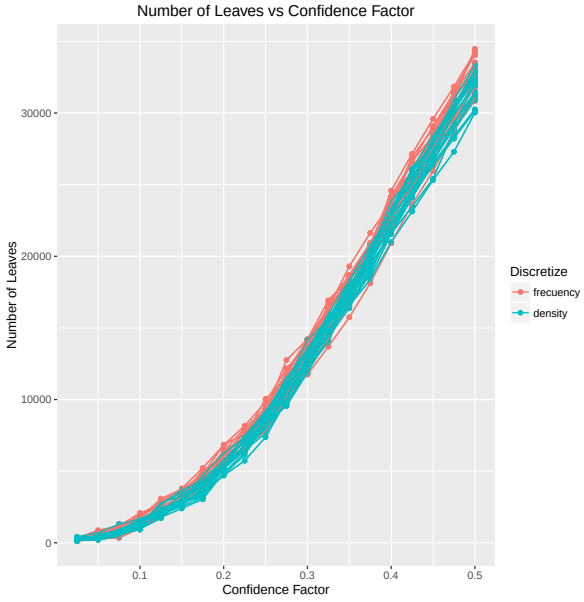

Figura 15. Leaves vs missing percentage with discretize

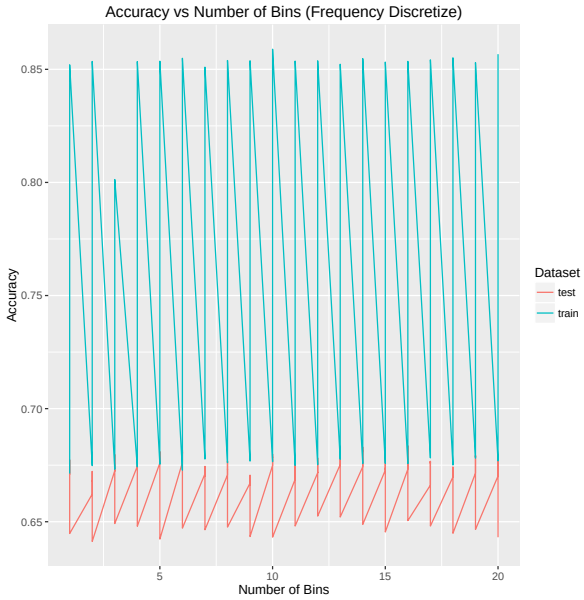

Figura 16. Accuracy vs Number of bins (Frequency discretize)

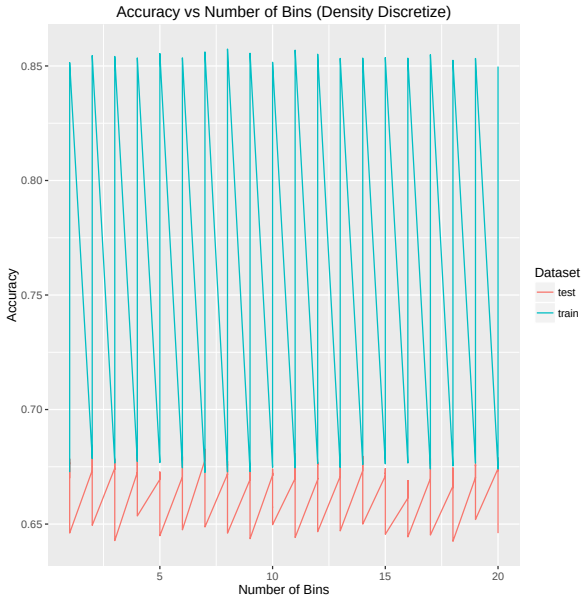

Figura 17. Accuracy vs Number of bins (Density discretize)