

Análisis predictivo de indicadores de desempeño por medio de la técnica de árboles de decisión j48 para las pruebas saber 11 Colombia 2020

Luis Felipe Vergara Serrato, David Santiago Mahecha Hernández, y cesar Yesid Barahona Rodríguez

Resumen - En Colombia el ICFES (Instituto Colombiano para la Evaluación de la Educación) realiza las pruebas saber 11 para medir las aptitudes de los estudiantes en diferentes áreas como: matemáticas, inglés, lectura crítica, sociales y ciencias naturales.

Se realizó un análisis predictivo de las pruebas saber por medio de las técnicas de minería de datos, para este caso de estudio se usará la herramienta WEKA ya que permite realizar una investigación a detalle de las pruebas saber 11 en Colombia del año 2020. Por otro lado, debido al estudio ya realizado donde, por medio de diferentes autores, se tiene como base para este estudio los indicadores de desempeño que pueden influir en los estudiantes, por medio de árboles de decisión (donde estos son una estructura representativa y analítica de todos los eventos que puede surgir de una decisión tomada), obtendremos resultados que serán analizados detalladamente, con esto poder demostrar si en la presente investigación estos indicadores influyen en el rendimiento académico de los estudiantes de secundaria.

I. INTRODUCCION

Colombia tiene las pruebas saber 11 el cual es considerado el examen de estado de la educación media, siendo una herramienta de medición estándar que evalúa la calidad de la educación formal.

Los indicadores de desempeño son instrumentos que proporcionan información cuantitativa y cualitativa de un objeto de estudio los cuales establecen una relación con diferentes variables, que, al ser comparados, se logra detectar fallas o avances del objeto de estudio [1].

Para esta investigación se realizó análisis predictivo consiste en el uso de técnicas estadísticas, con el fin de predecir tendencias futuras, además, se tiene un enfoque descriptivo para cuantificar las relaciones entre los datos usados, estudiando los beneficios pasados con base en los desempeños.

Los árboles de decisión son una técnica de clasificación que permiten examinar los resultados y determinar visualmente el flujo del modelo. Los resultados visuales ayudan a encontrar subgrupos y relaciones específicos que quizás no encontremos con las estadísticas tradicionales. Además, es una técnica estadística para la predicción; Para la elaboración de estos, se usa el algoritmo J48 el cual es una implementación libre en java del algoritmo C4.5, que utiliza el concepto de entropía de la información para la selección de variables que mejor clasifiquen al objeto estudiado [2].

II. MÉTODO

En la revisión bibliográfica donde se recolectaron los indicadores de desempeño académico, se utilizó la metodología implementada en el artículo de investigación de [3] donde se hace uso de herramientas y ecuaciones de búsqueda para las bases de datos de Scopus, ScienceDirect e IEEE.

Luego de realizar lo anterior y analizar las publicaciones que hacen parte del tema a desarrollar, se tomaron en cuenta los siguientes indicadores: el primer indicador es el estatus socioeconómico de los padres el cual gira entorno a las condiciones económicas como los ingresos familiares, como también la educación de los padres [4]; el capital social consta de la participación de los padres, es decir, si se encuentran los dos padres juntos [4]; la pandemia se tiene presente ya que se puede tener en cuenta las falencias con el acceso a internet o a computadores, [5]; también hace parte la desigualdad de oportunidades que se compone de el estrato socioeconómico como la profesión de los padres [6]; la composición social de la escuela se toma en cuenta si son colegios públicos o privados, entre otros indicadores [7].

Ahora bien, continuando con la metodología CRISP-DM (Tabla 1) ya que es el modelo de referencia más utilizada en el

desarrollo de proyectos enfocados en la minería de datos [8] se accedió a la base de datos del Icfes donde se consiguieron los resultados de las pruebas saber 11, en cuanto a su contenido nos ofrece información familiar, social, socioeconómica, etc., para poder visualizar este contenido de manera organizada se debe realizar el proceso de transformación limpieza de los datos la cual consiste en organizar, limpiar y eliminar algunas inconsistencias que pudiera tener una base de datos. En ocasiones, cuando se extraen datos de un sitio web estos contienen errores, los cuales deben ser corregidos para utilizar la información. La manera en la que realizamos la limpieza consta del uso de funciones para corregir el uso de caracteres especiales y eliminar datos erróneos o vacíos.

Fases CRISP-DM	Concepto
Comprensión del negocio	Comprensión de los objetivos de proyecto
Comprensión de los datos	Consta de la recopilación de datos inicial
Preparación de los datos	Transformación y la limpieza de datos
Modelado	Se selecciona y aplica las técnicas de modelado
Evaluación	Obtener una decisión sobre la aplicación de los resultados del proceso de análisis de datos.
Implementación	Generación de un informe

Tabla 1. Pasos metodología CRISP-DM.

1. Obtención de datos.



Fig.1 Obtención de datos

2. Transformación y limpieza



Fig.2 Transformación y limpieza de datos

3. Cargue de datos e implementación de los árboles de decisión

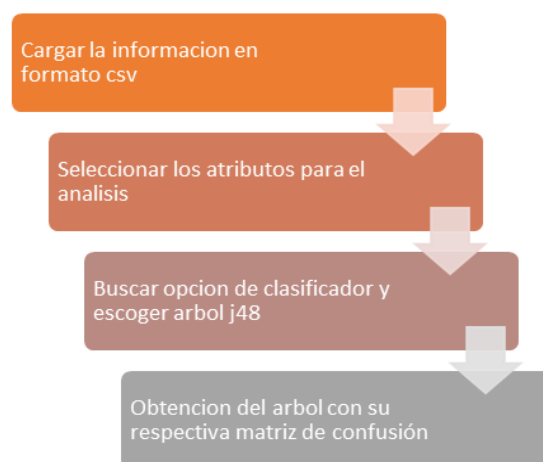


Fig.3 Cargue e implementación de los datos en los árboles de decisión

III. RESULTADOS Y DISCUSIÓN

1.

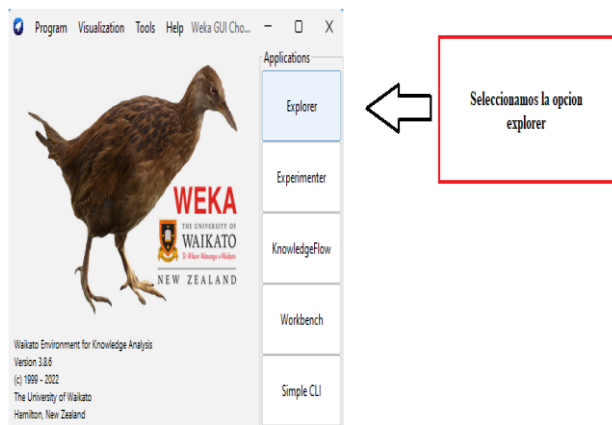


Fig.4 Inicio Weka.

2.



Fig.5 Selección del archivo.

3.

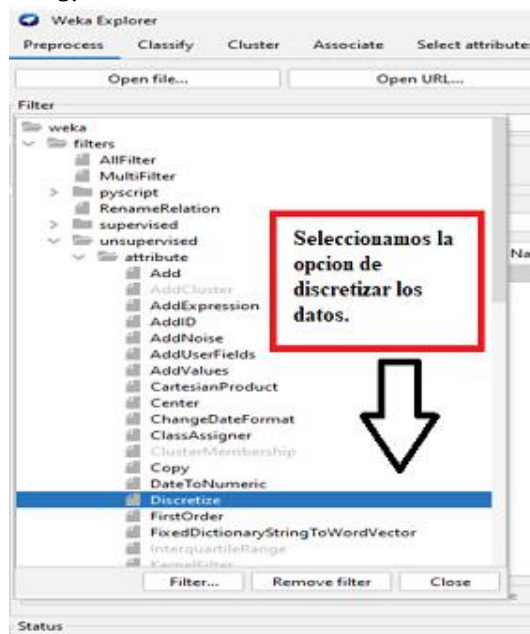


Fig.6 Selección del filtro.

El filtro de discretización es necesario cuando se quiere hacer una clasificación sobre un atributo numérico, en nuestro caso se usó para que el atributo de puntaje global se dividiera en rangos.

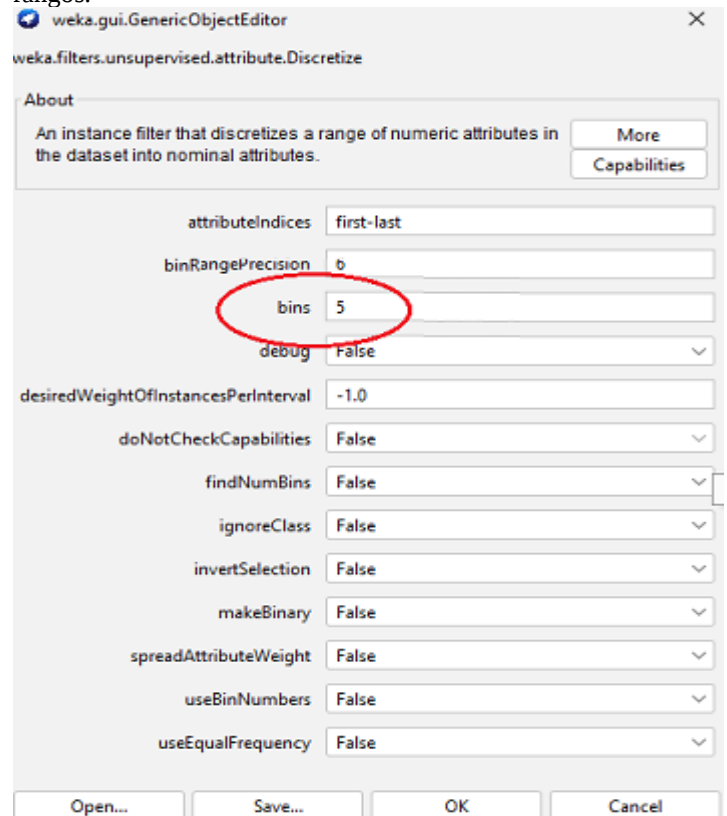


Fig.7 Selección de bis.

En la opción de bins podemos escoger en cuantos rangos vamos a dividir los atributos numéricos.

4.

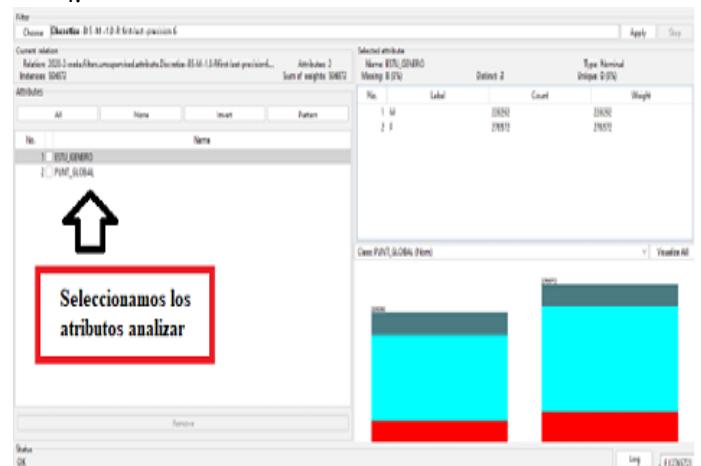


Fig.8 Selección de atributos.

5.

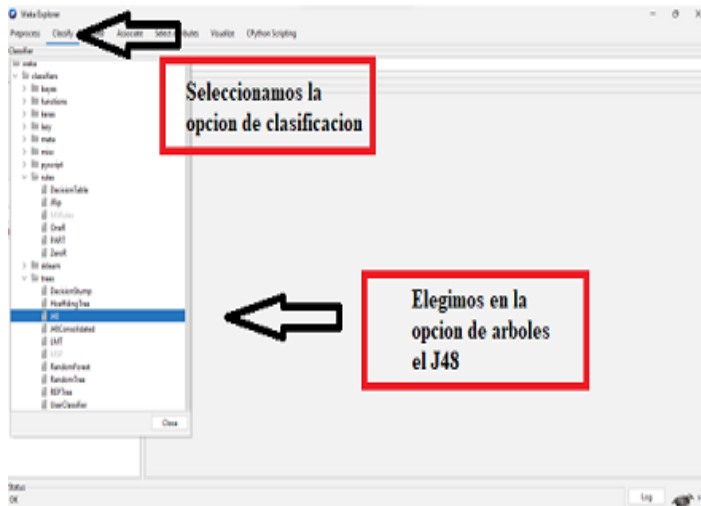


Fig.9 Selección del clasificador.

Se hizo uso de los árboles de decisión j48 que además de su gran reputación, permite la selección del atributo principal el cual da lugar a la construcción de dicho árbol buscando y dividiendo en subcategorías o subconjuntos donde haya relación con los demás atributos seleccionados, además, tienen como ventaja que sus resultados son fáciles de interpretar, se calculan rápido y se puede aplicar a variables continuas o categóricas [9].

6.

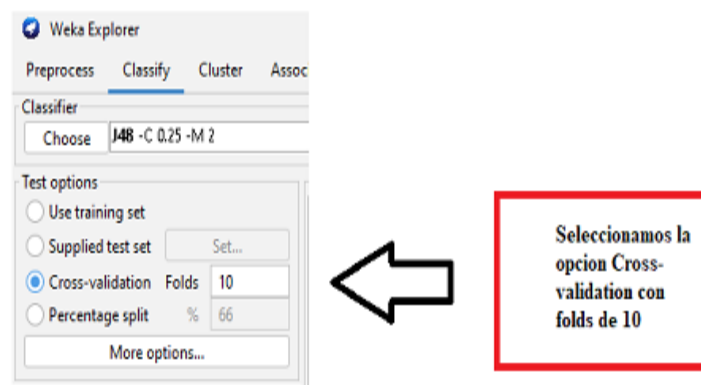


Fig.10 Selección de la prueba.

“Esta opción permite hacer una validación cruzada de K iteraciones. Los datos de la muestra se dividen en K subconjuntos. En cada una de las K iteraciones, el subconjunto K se utiliza como datos de prueba y el resto (K-1), se utiliza como datos de entrenamiento. Una vez se termina el proceso de validación cruzada, se realiza la media aritmética de los resultados de cada iteración para obtener un único resultado. Este método resulta preciso puesto que evalúa K combinaciones de prueba y entrenamiento, aunque desde el punto de vista computacional resulta más lento mientras más iteraciones se añadan. La elección del número de iteraciones debe depender de la medida del conjunto de datos. El programa utiliza por defecto 10 iteraciones.” [10]

7.

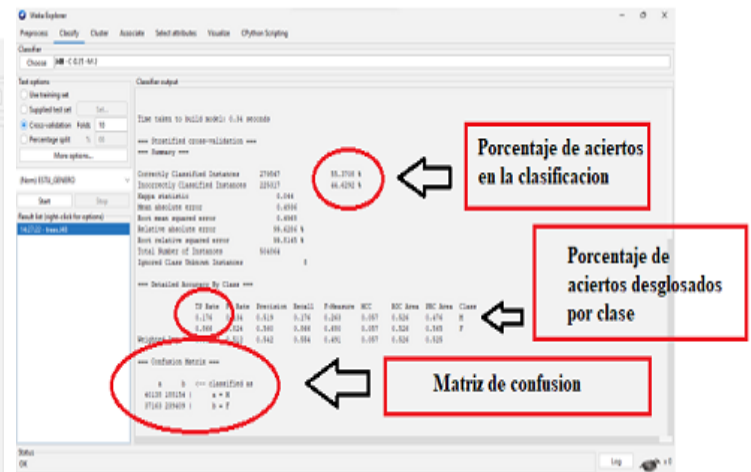


Fig.11 Información del clasificador.

Análisis de los indicadores:

- **Genero:**
✓ **2018:**

Correctly Classified Instances	303544	55.1964 %
Incorrectly Classified Instances	246390	44.8036 %
Kappa statistic	0.0654	
Mean absolute error	0.4939	
Root mean squared error	0.4969	
Relative absolute error	99.4415 %	
Root relative squared error	99.7207 %	
Total Number of Instances	549934	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.283	0.220	0.522	0.283	0.367	0.073	0.533	0.489	M
	0.780	0.717	0.562	0.780	0.653	0.073	0.533	0.561	F
Weighted Avg.	0.552	0.489	0.544	0.552	0.522	0.073	0.533	0.528	

=== Confusion Matrix ===

a	b	<-- classified as
71349 181079		a = M
65311 232195		b = F

Fig.12 Informe suministrado del árbol de decisión J48 indicador género 2018.

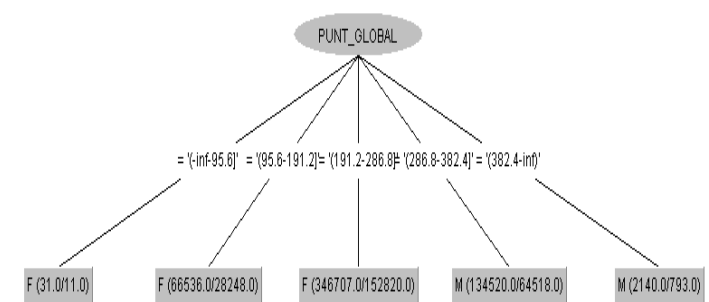


Fig.13 Árbol de decisión J48 indicador genero 2018.

✓ 2020:

Classifier output

Time taken to build model: 0.34 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	279547	55.3708 %
Incorrectly Classified Instances	225317	44.6292 %
Kappa statistic	0.044	
Mean absolute error	0.4936	
Root mean squared error	0.4968	
Relative absolute error	99.6286 %	
Root relative squared error	99.8145 %	
Total Number of Instances	504864	
Ignored Class Unknown Instances	8	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.176	0.134	0.519	0.176	0.263	0.057	0.526	0.476	M
	0.866	0.824	0.560	0.866	0.680	0.057	0.526	0.565	F
Weighted Avg.	0.554	0.512	0.542	0.554	0.491	0.057	0.526	0.525	

=== Confusion Matrix ===

a	b	c	d	e	<-- classified as
40138	188154				a = M
37163	239409				b = F

Fig.14 Informe suministrado del árbol de decisión J48 indicador género 2020.

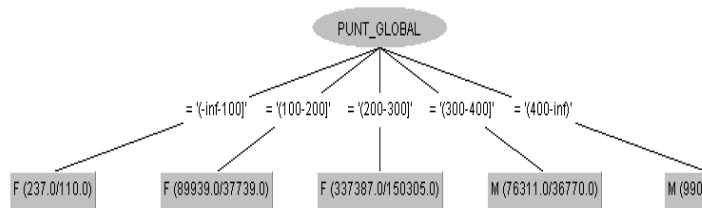


Fig.15 Árbol de decisión J48 indicador genero 2020.

Análisis: Podemos observar que en ambos años el género femenino obtiene puntajes de 300 hacia abajo, mientras que el género masculino obtiene resultados de 300 hacia arriba. En este caso vemos que el porcentaje de acierto que nos genera esta clasificación está en aproximadamente el 55%.

• Educación del padre:

✓ 2018:

Correctly Classified Instances	351252	63.8717 %
Incorrectly Classified Instances	196602	36.1283 %
Kappa statistic	0.0387	
Mean absolute error	0.2013	
Root mean squared error	0.3167	
Relative absolute error	95.3036 %	
Root relative squared error	97.469 %	
Total Number of Instances	549934	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.000	0.000	?	0.000	?	?	0.599	0.000	'(-inf-95.6)'
	0.000	0.000	?	0.000	?	?	0.620	0.177	'(95.6-191.2)'
	0.995	0.965	0.637	0.995	0.777	0.113	0.596	0.691	'(191.2-286.8)'
	0.047	0.006	0.713	0.047	0.089	0.140	0.676	0.414	'(286.8-382.4)'
	0.000	0.000	?	0.000	?	?	0.819	0.026	'(382.4-inf)'
Weighted Avg.	0.639	0.610	?	0.639	?	?	0.619	0.558	

=== Confusion Matrix ===

a	b	c	d	e	<-- classified as
0	0	31	0	0	a = '(-inf-95.6)'
0	0	66318	218	0	b = '(95.6-191.2)'
0	0	344871	1836	0	c = '(191.2-286.8)'
0	0	128139	6381	0	d = '(286.8-382.4)'
0	0	1627	513	0	e = '(382.4-inf)'

Fig.16 Informe suministrado del árbol de decisión J48 indicador educación del padre 2018.

FAMI_EDUCACIONPADRE = Ninguno: '(191.2-286.8)' (20094.78/7536.91)

FAMI_EDUCACIONPADRE = Primaria incompleta: '(191.2-286.8)' (110193.57/31863.46)

FAMI_EDUCACIONPADRE = Tecnica o tecnologica completa: '(191.2-286.8)' (37205.54/16548.77)

FAMI_EDUCACIONPADRE = Primaria completa: '(191.2-286.8)' (61616.5/19058.33)

FAMI_EDUCACIONPADRE = Secundaria (Bachillerato) completa: '(191.2-286.8)' (122404.4/44165.37)

FAMI_EDUCACIONPADRE = Tecnica o tecnologica incompleta: '(191.2-286.8)' (11050.79/4473.52)

FAMI_EDUCACIONPADRE = No sabe: '(191.2-286.8)' (35203.01/13322.75)

FAMI_EDUCACIONPADRE = Secundaria (Bachillerato) incompleta: '(191.2-286.8)' (77909.36/25208.17)

FAMI_EDUCACIONPADRE = Educacion profesional completa: '(191.2-286.8)' (48599.53/25816.68)

FAMI_EDUCACIONPADRE = Educacion profesional incompleta: '(191.2-286.8)' (10386.85/5560.26)

FAMI_EDUCACIONPADRE = Postgrado: '(286.8-382.4)' (9566.8/3113.03)

FAMI_EDUCACIONPADRE = No Aplica: '(191.2-286.8)' (5702.88/2326.6)

Fig.17 Árbol de decisión J48 indicador educación del padre 2018.

✓ 2020:

Correctly Classified Instances	339435	67.2319 %
Incorrectly Classified Instances	165437	32.7681 %
Kappa statistic	0.0405	
Mean absolute error	0.1918	
Root mean squared error	0.3096	
Relative absolute error	96.1439 %	
Root relative squared error	98.004 %	
Total Number of Instances	504872	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.000	0.000	?	0.000	?	?	0.769	0.001	'(-inf-100)'
	0.000	0.000	?	0.000	?	?	0.634	0.264	'(100-200)'
	0.990	0.964	0.674	0.990	0.802	0.090	0.557	0.697	'(200-300)'
	0.072	0.009	0.577	0.072	0.129	0.165	0.698	0.306	'(300-400)'
	0.000	0.000	?	0.000	?	?	0.782	0.009	'(400-inf)'
Weighted Avg.	0.672	0.645	?	0.672	?	?	0.592	0.559	

=== Confusion Matrix ===

a	b	c	d	e	<-- classified as
0	0	237	0	0	a = '(-inf-100)'
0	0	89583	357	0	b = '(100-200)'
0	0	333913	3480	0	c = '(200-300)'
0	0	70790	5522	0	d = '(300-400)'
0	0	781	209	0	e = '(400-inf)'

Fig.18 Informe suministrado del árbol de decisión J48 indicador educación del padre 2020.

FAMI_EDUCACIONPADRE = Secundaria Bachillerato completa: '(200-300)' (123499.56/37026.87)

FAMI_EDUCACIONPADRE = Postgrado: '(300-400)' (9820.68/4266.76)

FAMI_EDUCACIONPADRE = Tecnica o tecnologica completa: '(200-300)' (33946.42/11286.21)

FAMI_EDUCACIONPADRE = Primaria completa: '(200-300)' (47143.98/14403.87)

FAMI_EDUCACIONPADRE = Primaria incompleta: '(200-300)' (94520.96/29191.96)

FAMI_EDUCACIONPADRE = Ninguno: '(200-300)' (21347.25/9385.25)

FAMI_EDUCACIONPADRE = Educacion profesional incompleta: '(200-300)' (10501.19/4152.98)

FAMI_EDUCACIONPADRE = No sabe: '(200-300)' (30660.88/9747.93)

FAMI_EDUCACIONPADRE = Educacion profesional completa: '(200-300)' (47971.27/20257.75)

FAMI_EDUCACIONPADRE = No Aplica: '(200-300)' (7939.27/2617.59)

FAMI_EDUCACIONPADRE = Secundaria Bachillerato incompleta: '(200-300)' (67029.62/19940.2)

FAMI_EDUCACIONPADRE = Tecnica o tecnologica incompleta: '(200-300)' (10490.92/3286.89)

Fig.19 Árbol de decisión J48 indicador educación del padre 2020.

Análisis: En ambos años podemos observar que el puntaje en casi todas las opciones es el mismo, menos cuando el padre realiza un postgrado, notamos que esto influye positivamente ya que en el 2018 los estudiantes están en el rango de (286.8-

382.4) en comparación al resto que todos están en el rango de (191.2-286.8) y en el año 2020, observamos que los estudiantes están en el rango de (300-400), mientras que el resto está en (200-300). En esta clasificación se obtuvo un porcentaje de acierto en el 2018 del 63%, mientras que en el 2020 un porcentaje de 67%.

• **Educación de la madre:**

✓ **2018:**

Correctly Classified Instances	351211	63.8642 %
Incorrectly Classified Instances	198723	36.1358 %
Kappa statistic	0.0408	
Mean absolute error	0.2003	
Root mean squared error	0.3159	
Relative absolute error	94.8501 %	
Root relative squared error	97.2222 %	
Total Number of Instances	549934	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.000	?	0.000	?	?	0.632	0.000	'(-inf-95.6)'
0.000	0.000	?	0.000	?	?	0.641	0.186	'(95.6-191.2)'
0.993	0.962	0.638	0.993	0.777	0.111	0.600	0.691	'(191.2-286.8)'
0.052	0.008	0.683	0.052	0.097	0.140	0.679	0.414	'(286.8-382.4)'
0.000	0.000	?	0.000	?	?	0.829	0.024	'(382.4-inf)'
Weighted Avg.	0.639	0.608	?	0.639	?	?	0.625	0.560

=== Confusion Matrix ===

a	b	c	d	e	<-- classified as
0	0	31	0	0	a = '(-inf-95.6)'
0	0	66286	250	0	b = '(95.6-191.2)'
0	0	344221	2486	0	c = '(191.2-286.8)'
0	0	127530	6990	0	d = '(286.8-382.4)'
0	0	1635	505	0	e = '(382.4-inf)'

Fig.20 Informe suministrado del árbol de decisión J48 indicador educación de la madre 2018.

```
FAMI_EDUCACIONMADRE = Primaria incompleta: '(191.2-286.8)' (86678.46/25384.87)
FAMI_EDUCACIONMADRE = Secundaria (Bachillerato) incompleta: '(191.2-286.8)' (84163.67/25707.07)
FAMI_EDUCACIONMADRE = Tecnica o tecnologica incompleta: '(191.2-286.8)' (16432.16/6406.77)
FAMI_EDUCACIONMADRE = Secundaria (Bachillerato) completa: '(191.2-286.8)' (142672.86/49359.76)
FAMI_EDUCACIONMADRE = No sabe: '(191.2-286.8)' (9780.21/4282.32)
FAMI_EDUCACIONMADRE = Tecnica o tecnologica completa: '(191.2-286.8)' (55092.37/23639.73)
FAMI_EDUCACIONMADRE = Primaria completa: '(191.2-286.8)' (61710.12/18740.34)
FAMI_EDUCACIONMADRE = Educacion profesional completa: '(191.2-286.8)' (55816.99/29649.54)
FAMI_EDUCACIONMADRE = Ninguno: '(191.2-286.8)' (12894.57/5062.85)
FAMI_EDUCACIONMADRE = Educacion profesional incompleta: '(191.2-286.8)' (12836.86/6516.43)
FAMI_EDUCACIONMADRE = Postgrado: '(286.8-382.4)' (10934.47/3861.66)
FAMI_EDUCACIONMADRE = No Aplica: '(191.2-286.8)' (921.27/463.64)
```

Fig.21 Árbol de decisión J48 indicador educación de la madre 2018

✓ **2020:**

Correctly Classified Instances	339236	67.1925 %
Incorrectly Classified Instances	165636	32.8075 %
Kappa statistic	0.0442	
Mean absolute error	0.1909	
Root mean squared error	0.3088	
Relative absolute error	95.6948 %	
Root relative squared error	97.7668 %	
Total Number of Instances	504872	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.000	?	0.000	?	?	0.683	0.001	'(-inf-100)'
0.000	0.000	?	0.000	?	?	0.651	0.275	'(100-200)'
0.987	0.959	0.675	0.987	0.801	0.090	0.566	0.704	'(200-300)'
0.081	0.012	0.552	0.081	0.142	0.169	0.711	0.311	'(300-400)'
0.000	0.000	?	0.000	?	?	0.791	0.009	'(400-inf)'
Weighted Avg.	0.672	0.643	?	0.672	?	?	0.604	0.566

=== Confusion Matrix ===

a	b	c	d	e	<-- classified as
0	0	237	0	0	a = '(-inf-100)'
0	0	89488	452	0	b = '(100-200)'
0	0	333029	4364	0	c = '(200-300)'
0	0	70105	6207	0	d = '(300-400)'
0	0	774	216	0	e = '(400-inf)'

Fig.22 Informe suministrado del árbol de decisión J48 indicador educación de la madre 2020.

```
FAMI_EDUCACIONMADRE = Secundaria Bachillerato completa: '(200-300)' (141424.7/41352.66)
FAMI_EDUCACIONMADRE = Educacion profesional completa: '(200-300)' (57890.35/24006.71)
FAMI_EDUCACIONMADRE = Tecnica o tecnologica incompleta: '(200-300)' (15615.11/4471.07)
FAMI_EDUCACIONMADRE = Secundaria Bachillerato incompleta: '(200-300)' (70052.22/20229.6)
FAMI_EDUCACIONMADRE = Primaria completa: '(200-300)' (47249.1/15049.31)
FAMI_EDUCACIONMADRE = Primaria incompleta: '(200-300)' (73245.51/24172.13)
FAMI_EDUCACIONMADRE = Postgrado: '(300-400)' (11547.45/5301.52)
FAMI_EDUCACIONMADRE = Ninguno: '(200-300)' (12702.3/5915.4)
FAMI_EDUCACIONMADRE = Educacion profesional incompleta: '(200-300)' (12967.38/5003.1)
FAMI_EDUCACIONMADRE = Tecnica o tecnologica completa: '(200-300)' (51873.63/16182.42)
FAMI_EDUCACIONMADRE = No Aplica: '(200-300)' (1202.11/562.25)
FAMI_EDUCACIONMADRE = No sabe: '(200-300)' (9102.13/3541.72)
```

Fig.23 Árbol de decisión J48 indicador educación de la madre 2020.

Análisis: Al igual que el indicador anterior, notamos el mismo caso, el cual es que influye positivamente el hecho de que la madre tenga una educación de postgrado con los siguientes resultados, en el 2018 se obtiene un rango de (286.8-382.4) en comparación al resto que todos están en el rango de (191.2-286.8) y en el año 2020, observamos que los estudiantes están en el rango de (300-400), mientras que el resto está en (200-300). En esta clasificación se obtuvo un porcentaje de acierto en el 2018 del 63%, mientras que en el 2020 un porcentaje de 67%.

- **Tiene internet:**
✓ **2018:**

```

Correctly Classified Instances   313877       61.051 %
Incorrectly Classified Instances 200246       38.949 %
Kappa statistic                 0.1011
Mean absolute error             0.4549
Root mean squared error        0.4769
Relative absolute error         93.7471 %
Root relative squared error     96.8235 %
Total Number of Instances      514123
Ignored Class Unknown Instances 35911
  
```

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Si	0.164	0.074	0.611	0.164	0.258	0.141	0.605	0.462
No	0.926	0.836	0.610	0.926	0.736	0.141	0.630	0.664
Weighted Avg.	0.611	0.520	0.611	0.611	0.538	0.141	0.620	0.580

=== Confusion Matrix ===

```

a      b  <-- classified as
34898 178071 |      a = No
22175 278979 |      b = Si
  
```

Fig.24 Informe suministrado del árbol de decisión J48 indicador el estudiante tiene internet 2018.

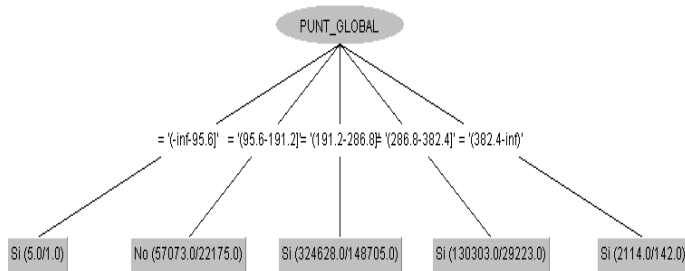


Fig.25 Árbol de decisión J48 indicador el estudiante tiene internet 2018.

- ✓ **2020:**

```

Correctly Classified Instances   345073       70.25 %
Incorrectly Classified Instances 146134       29.75 %
Kappa statistic                 0.2
Mean absolute error             0.3927
Root mean squared error        0.4431
Relative absolute error         93.1256 %
Root relative squared error     96.5019 %
Total Number of Instances      491207
Ignored Class Unknown Instances 13665
  
```

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
Si	0.877	0.701	0.743	0.877	0.804	0.213	0.632	0.775	Si
No	0.299	0.123	0.513	0.299	0.378	0.213	0.633	0.400	No
Weighted Avg.	0.703	0.526	0.674	0.703	0.676	0.213	0.632	0.662	

=== Confusion Matrix ===

```

a      b  <-- classified as
300638 42187 |      a = Si
103947 44435 |      b = No
  
```

Fig.26 Informe suministrado del árbol de decisión J48 indicador el estudiante tiene internet 2020.

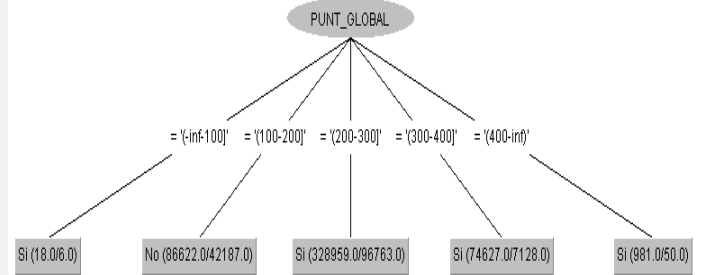


Fig.27 Árbol de decisión J48 indicador el estudiante tiene Internet 2020.

Análisis: En este indicador vemos claramente en ambos años que el hecho de tener internet influye positivamente en los resultados. En el 2018 obtuvimos un 63% de acierto en la clasificación, mientras que en el 2020 un 70%.

- **Tiene computador:**
✓ **2018:**

```

Correctly Classified Instances   320048       60.3329 %
Incorrectly Classified Instances 210422       39.6671 %
Kappa statistic                 0.1001
Mean absolute error             0.4589
Root mean squared error        0.479
Relative absolute error         94.027 %
Root relative squared error     96.968 %
Total Number of Instances      530470
Ignored Class Unknown Instances 19464
  
```

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
Si	0.922	0.831	0.602	0.922	0.728	0.139	0.621	0.664	Si
No	0.169	0.078	0.612	0.169	0.265	0.139	0.612	0.491	No
Weighted Avg.	0.603	0.513	0.606	0.603	0.532	0.139	0.617	0.591	

=== Confusion Matrix ===

```

a      b  <-- classified as
282190 24022 |      a = Si
186400 37858 |      b = No
  
```

Fig.28 Informe suministrado del árbol de decisión J48 indicador estudiante tiene computador 2018.

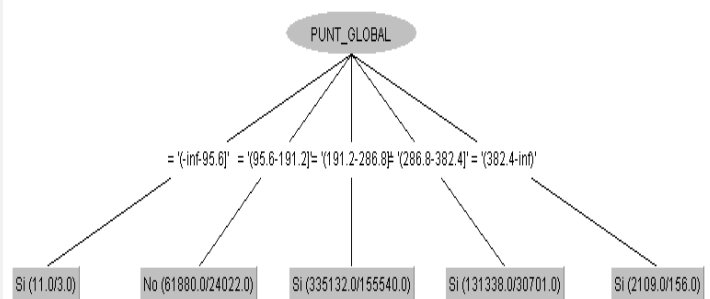


Fig.29 Árbol de decisión J48 indicador el estudiante tiene computador 2018.

✓ 2020:

Correctly Classified Instances	310428	64.0483 %
Incorrectly Classified Instances	174250	35.9517 %
Kappa statistic	0.1669	
Mean absolute error	0.4439	
Root mean squared error	0.4711	
Relative absolute error	92.9315 %	
Root relative squared error	96.4014 %	
Total Number of Instances	484678	
Ignored Class Unknown Instances	20194	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.884	0.733	0.649	0.884	0.749	0.194	0.625	0.692	Si
	0.267	0.116	0.599	0.267	0.369	0.194	0.623	0.477	No
Weighted Avg.	0.640	0.490	0.630	0.640	0.599	0.194	0.624	0.607	

=== Confusion Matrix ===

a	b	<-- classified as
259387	34123	a = Si
140127	51041	b = No

Fig.30 Informe suministrado del árbol de decisión J48 indicador estudiante tiene computador 2020.

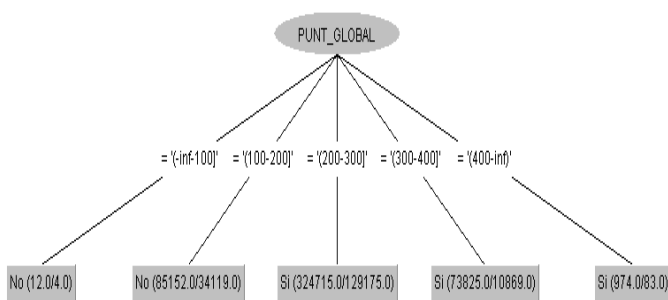


Fig.31 Árbol de decisión J48 indicador el estudiante tiene computador 2020.

Análisis: En este indicador vemos claramente en ambos años que el hecho de tener computador influye positivamente en los resultados. En el 2018 obtuvimos un 60% de acierto en la clasificación, mientras que en el 2020 un 64%.

IV. CONCLUSIONES

Nuestro análisis predictivo de los resultados conseguidos con el árbol de decisión J48 para determinar si los indicadores analizados influyen en el desempeño académico de los estudiantes colombianos en las pruebas de estado para la educación media saber 11° en los años 2018 y 2020, señalando que estos si tienen influencia no solo con el análisis realizado con los datos del ICFES sino también con el respaldo teórico. En nuestro primer indicador notamos una brecha de genero bastante marcada donde el género masculino tiende a obtener mejores resultados en estas pruebas que el género femenino, En el segundo y tercer indicador el cual es la educación alcanzada por los padres notamos que en ambos casos donde alguno de los 2 tuviese un postgrado tiende a que los estudiantes obtengan un mejor rendimiento en las pruebas y los 2 últimos indicadores los cuales son si el estudiante cuenta con internet y si cuenta con computador, vemos que en ambos casos influye positivamente el hecho de obtener alguno de estos 2 estos.

Por otro lado, en el análisis del COVID el cual se conforma por tener acceso a internet y tener computador en una comparación con el año donde empieza todo siendo 2020 con uno anterior como 2018 se evidencia que no hay gran influencia en cuanto a los resultados obtenidos.

REFERENCIAS

- [1] Lima, "MINISTERIO DE ECONOMÍA Y FINANZAS DIRECCIÓN GENERAL DEL PRESUPUESTO PÚBLICO Instructivo para la Formulación de Indicadores de Desempeño," 2010.
- [2] Vanesa Berlanga Silvente and María José Rubio Hurtado, "Cómo aplicar árboles de decisión en SPSS.," 2013, Accessed: Aug. 20, 2022. [Online]. Available: <http://diposit.ub.edu/dspace/bitstream/2445/43762/1/618361.pdf>
- [3] Mahecha Hernandez David Santiago and Vergara Serrato Luis Felipe, "Revisión literaria de indicadores que influyen en la educación mediante técnicas de minería de datos".
- [4] O. v. Kotomina and A. I. Sazhina, "The Influence of Family Factors on the Academic Performance of Schoolchildren and University Students: Review of Foreign Studies," Education and Self Development, vol. 16, no. 4, pp. 74–92, 2021, doi: 10.26907/esd.16.4.07.
- [5] Y. Wang, M. Xia, W. Guo, F. Xu, and Y. Zhao, "Academic performance under COVID-19: The role of online learning readiness and emotional competence," Current Psychology, 2022, doi: 10.1007/s12144-022-02699-7.
- [6] J. A. Sarmiento Espinel, A. C. Silva Arias, and E. van Gamen, "Evolution of the inequality of educational opportunities from secondary education to university," International Journal of Educational Development, vol. 66, pp. 193–202, Apr. 2019, doi: 10.1016/j.ijedudev.2018.09.006.
- [7] K. A. Adebayo, N. Ntokozo, and N. Z. Grace, "Availability of Educational Resources and Student Academic Performances in South Africa," Universal Journal of Educational Research, vol. 8, no. 8, pp. 3768–3781, Aug. 2020, doi: 10.13189/ujer.2020.080858.
- [8] J. H. Orallo, *Introducción a la minería*. PRENTICE HALL, 2008.
- [9] Cardona Carlos and Garcia Nancy, "Análisis de datos mediante el algoritmo de clasificación J48, sobre un cluster en la nube de AWS," Nov. 2016. <https://revistas.udistrital.edu.co/index.php/REDES/article/view/11985/12615> (accessed Nov. 13, 2022).
- [10] Rodríguez Manzanares, C. (2018). Análisis de los métodos de predicción aplicados a los desvíos en el sistema eléctrico ibérico. (Trabajo Fin de Máster Inédito). Universidad de Sevilla, Sevilla.

Agradecimientos

Este trabajo se desarrolló gracias al apoyo de la Universidad de Cundinamarca y el grupo de Investigación GISTFA (Grupo de Investigación de Sistemas y Tecnología de Facatativá), Colombia

Conflicto de Intereses

Los autores consideran que no existen conflictos de intereses.

Contribución de los autores

Luis Felipe Vergara Serrato. ORCID: <https://orcid.org/0000-0003-2679-8375>

Diseño y desarrollo de los experimentos. Redacción del manuscrito.

David Santiago Mahecha Hernández: <https://orcid.org/0000-0002-2801-6333>

Diseño y desarrollo de los experimentos. Redacción del manuscrito.

Cesar Yesid Barahona Rodríguez. ORCID: <https://orcid.org/0000-0001-7673-7381>

Diseño y desarrollo de los experimentos. Redacción del manuscrito.