Algoritmos para predecir el éxito académico en las pruebas saber pro

Samuel Meneses Universidad Eafit Colombia smenesesd@eafit.edu.co	Neller Pellegrino Universidad Eafit Colombia npellegrib@eafit.edu.co	Miguel Correa Universidad Eafit Colombia macorream@eafit.edu.co	Mauricio Toro Universidad Eafit Colombia mtorobe@eafit.edu.co

RESUMEN

La participación de los jóvenes colombianos en las Pruebas Saber Pro es de suma importancia para sus vidas como profesionales, Unos se quieren superar y hay otros que se quedan en el nivel que siempre han estado. Hay muchas variables que pueden verse reflejadas en los resultados de dicha prueba, como lo es el género de estudiante, su esfuerzo en aprender cada vez más y el estado socioeconómico de los padres. El problema que se va a solucionar en este proyecto es saber la cantidad de estudiantes que van a estar en un nivel superior dependiendo de sus notas. Es muy importante resolver este problema porque así vemos como los estudiantes Colombianos van saliendo adelante en sus estudios y en sus notas, para poder ver que hay talento Colombiano en cada universidad local.

¿Cuál es el algoritmo propuesto? ¿Qué resultados obtuvieron? ¿Cuáles son las conclusiones de este trabajo? El resumen debe tener como máximo **200 palabras**. (*En este semestre, usted debe resumir aquí los tiempos de ejecución, el consumo de memoria, la exactitud, la precisión y la sensibilidad*)

Palabras clave

Árboles de decisión, aprendizaje automático, éxito académico, predicción de los resultados de los exámenes

1. INTRODUCCIÓN

Las Pruebas Saber Pro como su nombre lo dice son unas pruebas que todo estudiante terminando su pregrado debe de hacer, estas pruebas se hacen para saber su nivel personal y posicionar también el nivel de educación de su universidad. La estructura del examen consta de 5 módulos los cuales son lectura crítica, razonamiento cuantitativo, competencias ciudadanas, comunicación escrita e inglés. Esta prueba consta de dos sesiones, una de 4 horas y media y la otra sesión de 4 horas

1.1. Problema

El problema que está planteado en este proyecto es predecir que estudiantes les va a ir bien en las pruebas. Este problema si llega a solucionarse de manera casi inmediata en Colombia, cada estudiante que vaya a presentar esta prueba tendría un conocimiento claro de lo que podría sacar. Colombia es uno de los países con más bajos niveles académicos tiene, según la Ocde. Las cifras no serían las mismas solucionando este problema ya que las universidades de todos los niveles estarían entrenadas y cada estudiante sabría que sacar en promedio para ayudarse a sí mismo y al nivel de la educación de Colombia.

1.2 Solución

En este trabajo, nos centramos en los árboles de decisión porque proporcionan una gran explicabilidad. Evitamos los métodos de caja negra como las redes neuronales, las máquinas de soporte vectorial y los bosques aleatorios porque carecen de explicabilidad

1.3 Estructura del artículo

En lo que sigue, en la sección 2, presentamos el trabajo relacionado con el problema. Más adelante, en la sección 3, presentamos los conjuntos de datos y métodos utilizados en esta investigación. En la sección 4, presentamos el diseño del algoritmo. Después, en la sección 5, presentamos los resultados. Finalmente, en la sección 6, discutimos los resultados y proponemos algunas direcciones de trabajo futuras.

2. TRABAJOS RELACIONADOS

2.1 Minería de datos: Lo que los estudiantes solucionan al momento de hacer este proyecto es averiguar las causas de las personas que entran a la universidad y se retirar de la carrera antes del momento de graduarse. Lo que ellos hicieron fue una minería de datos ayudados del minero de datos (weka) y los árboles de decisión. Lo que pudieron sacar en conclusión de las personas que desertaban era por tres razones, 1: la edad de la persona, 2: los ingresos familiares y 3: el nivel de inglés.

2.2 Pruebas saber 11º Es parecido lo que estos señores hacen en este artículo, solo que estos hablan del desempeño académico en las pruebas saber 11º en el año 2015 y 2016 en Colombia, Como el ejemplo anterior también usaron la minería de datos en (weka) y arboles de decisión. Lo que ellos lograron fue deducir tras toda esta búsqueda que los mejores resultados fueron los estudiantes que estudiaron en un colegio de categoría media o alta.

2.3 Descubrimiento de patrones de desempeño académico en las competencias genéricas. Es un tema idéntico al de este proyecto habla de las pruebas saber pro, esta investigación al

hacen en las pruebas del segundo semestre del año 2011, fueron averiguando los desempeños en cada módulo de competencia. Para implementar esto usaron los datos ya registrados en la página oficial de los icfes, y muchas tablas para empezar a comprar.

2.4 Modelos predictivos y técnicas de minería de datos Lo que ellos quieren hacer es un análisis de diferentes campos en las diferentes carreras de la FACENA de la UNNE, los campos son, la prueba de diagnóstico de matemáticas y las condiciones socioeconómicas de los alumnos. Los algoritmos que usaron fue las técnicas de minería clásicas y métodos simbólicos o inteligentes. Los resultados logrados fueron evidentes para saber que la universidad cuenta con un nivel alto de matemáticas y en la parte socioeconómica la universidad implementa becas para los estudiantes que no pueden afrontar el costo de su carrera.

3. MATERIALES Y MÉTODOS

En esta sección se explica cómo se recopilaron y procesaron los datos y, después, cómo se consideraron diferentes alternativas de solución para elegir un algoritmo de árbol de decisión.

3.1 Recopilación y procesamiento de datos

Obtuvimos datos del *Instituto Colombiano de Fomento de la Educación Superior* (ICFES), que están disponibles en línea en ftp.icfes.gov.co. Estos datos incluyen resultados anonimizados de Saber 11 y Saber Pro. Se obtuvieron los resultados de Saber 11 de todos los gradudados de escuelas secundarias colombianas, de 2008 a 2014, y los resultados de Saber Pro de todos los graduados de pregrados colombianos, de 2012 a 2018. Hubo 864.000 registros para Saber 11 y 430.000 para Saber Pro. Tanto Saber 11 como Saber Pro, incluyeron, no sólo las puntuaciones sino también datos socioeconómicos de los estudiantes, recogidos por el ICFES, antes de la prueba.

En el siguiente paso, ambos conjuntos de datos se fusionaron usando el identificador único asignado a cada estudiante. Por lo tanto, se creó un nuevo conjunto de datos que incluía a los estudiantes que hicieron ambos exámenes estandarizados. El tamaño de este nuevo conjunto de datos es de 212.010 estudiantes. Después, la variable predictora binaria se definió de la siguiente manera: ¿El puntaje del estudiante en el Saber Pro es mayor que el promedio nacional del período en que presentó el examen?

Se descubrió que los conjuntos de datos no estaban equilibrados. Había 95.741 estudiantes por encima de la media y 101.332 por debajo de la media. Realizamos un submuestreo para equilibrar el conjunto de datos en una proporción de 50%-50%. Después del submuestreo, el conjunto final de datos tenía 191.412 estudiantes.

Por último, para analizar la eficiencia y las tasas de aprendizaje de nuestra implementación, creamos al azar subconjuntos del conjunto de datos principal, como se muestra en la Tabla 1. Cada conjunto de datos se dividió en un 70% para entrenamiento y un 30% para validación. Los conjuntos de datos están disponibles en https://github.com/mauriciotoro/ST0245-Eafit/tree/master/proyecto/datasets .

	Conjunto de datos 1	Conjunto de datos 2	Conjunto de datos 3	Conjunto de datos 4	Conjunto de datos 5
Entrenamiento	15,000	45,000	75,000	105,000	135,000
Validación	5,000	15,000	25,000	35,000	45,000

Tabla 1. Número de estudiantes en cada conjunto de datos utilizados para el entrenamiento y la validación.

3.2 Alternativas de algoritmos de árbol de decisión

En lo que sigue, presentamos diferentes algoritmos usardos para construir automáticamente un árbol de decisión binario. *(En este semestre, ejemplos de tales algoritmos son ID3, C4.5 y CART).*

3.2.1 Arbol CART (Classification and regession trees)

Cart es un algoritmo que genera arboles de clasificación y de regresión, lo que hace este algoritmo es dividir el nodo del árbol en dos ramas exactas y solo permite crear arboles de valores binarios; este árbol permite resolver problemas de clasificación y de regresión como lo dicho anteriormente.

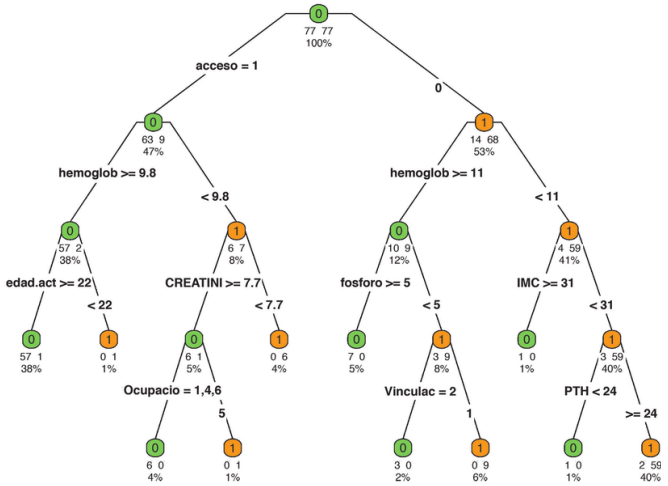
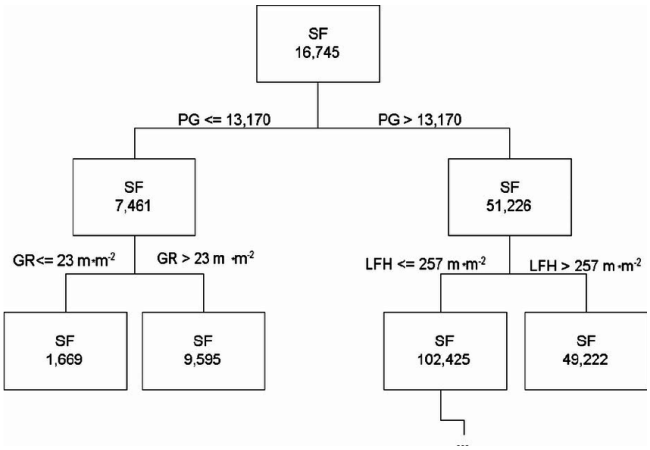


Imagen 1.

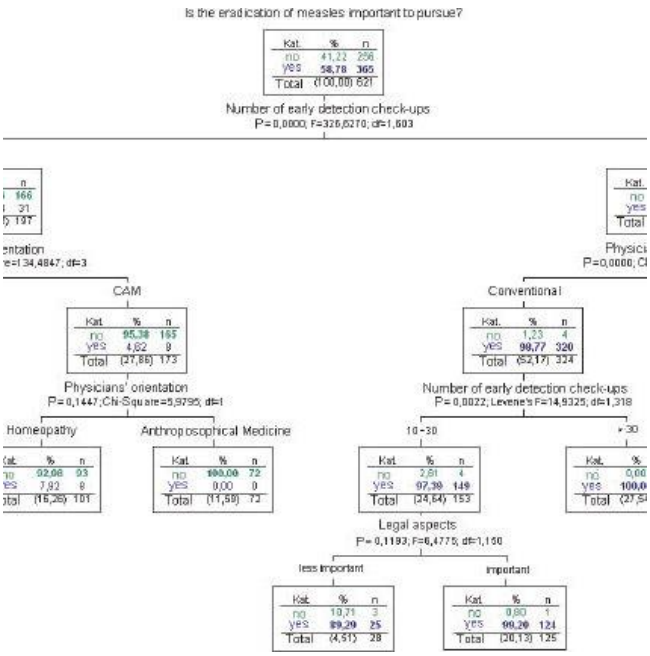
3.2.2 Árbol CHAID (CHi-square Automatic Interaction Detection)

Este árbol tiene un antecedente por allá en la época de los 70s e inicios de los 80s ya que se usaba el AID, y ya después del AID surgió el CHAID, se llama así porque el objetivo de este árbol es guiarse en las interacciones entre las variables y en la clasificación.



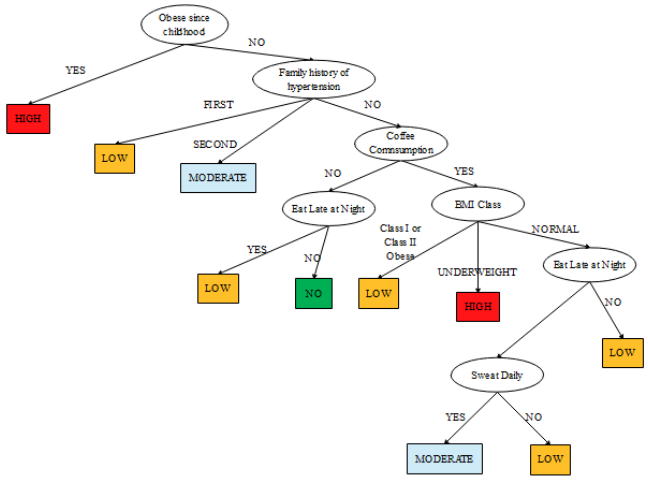
3.2.3 Árbol QUEST (Quick, Unbiased, Efficient Statistical Tree)

Su nombre se refiere a un árbol rápido (quick), eficiente e insesgado. Este árbol o algoritmo intenta ahorrarte tiempo comprando el tiempo de los dos árboles anteriores toman para generar el algoritmo. Este método no tiene una división exacta, lo que hace es seleccionar la mejor forma de segmentar los datos y ahí decide la división propia de esta.



3.2.4 C4.5

Este algoritmo Construye arboles de decisión, fue una mejora del algoritmo ID3 que se desarrolló en 1993. Este árbol se construye mediante la estrategia de depht-first



4. DISEÑO DE LOS ALGORITMOS

4.1 Estructura de los datos

La estructura de datos que vamos a utilizar para hacer la predicción es el árbol de decisión binario, para ver de esta manera la probabilidad de los estudiantes en sacar un buen resultado, como ya sabemos un árbol de decisión ayudan a realizar elecciones correctas entre distintas posibilidades; entonces el árbol de decisión binaria son estructuras de datos que tienen como valor una función boolean

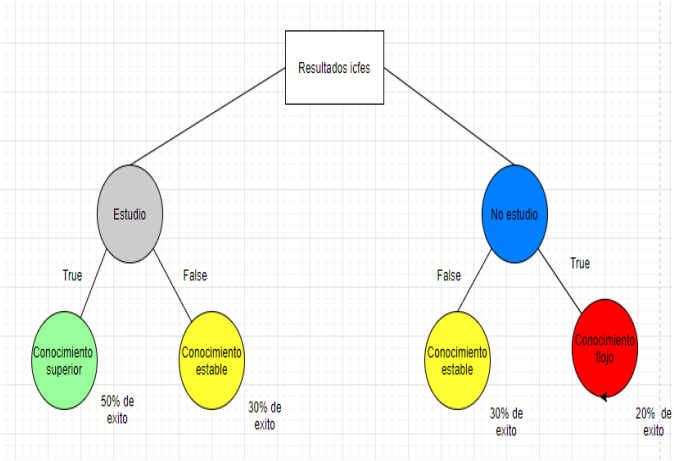


Figura 1: En este árbol de decisión binario podemos ver que se basa en el resultado del icfes, entonces de este centro saca dos opciones si estudio o si no estudio, de no estudio salen otras dos opciones las cuales si de verdad no estudio tendrá un conocimiento flojo del tema y tendrá un 20% de éxito o menos, la otra opción si es falso tiene un conocimiento estable y tendrá aproximadamente un 30% de éxito en la prueba; pasando a la otra opción de si sí estudio, volvemos a tener dos opciones, si es falso tendrá un conocimiento estable con un aproximado de 30% de éxito y si es verdadero el resultado este tendrá un conocimiento superior con un 50% de éxito en la prueba.

4.2 Algoritmos

Explica el diseño del algoritmo para resolver el problema y haz una figura. No uses figuras de Internet, haz las tuyas propias. *(En este semestre, un algoritmo debe ser un algoritmo para entrenar un algoritmo de árbol de decisión como ID3, C4.5, CART y el segundo algoritmo debe ser un algoritmo para clasificar los nuevos datos utilizando dicho árbol).*

4.2.1 Entrenamiento del modelo

Explique, brevemente, cómo entrenó a la modelo: Esto equivale a explicar cómo su algoritmo construye automáticamente un árbol de decisión binario.

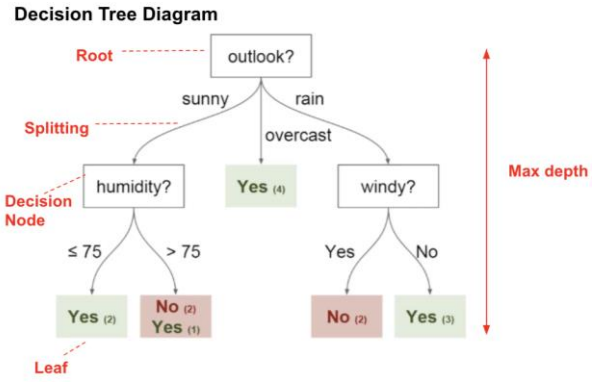


Figura 2: Entrenamiento de un árbol de decisión binario usando *(En este semestre, uno podría ser CART, ID3, C4.5... por favor, elija)*. En este ejemplo, mostramos un modelo para predecir si se debe jugar al golf o no, según el clima.

4.2.2 Algoritmo de prueba

Explique, brevemente, cómo probó el modelo: Esto equivale a explicar cómo su algoritmo clasifica los nuevos datos después de que se construya el árbol.

4.3 Análisis de la complejidad de los algoritmos

Explique en sus propias palabras el análisis para el peor caso usando la notación O. ¿Cómo calculó tales complejidades.

Algoritmo	La complejidad del tiempo
Entrenar el árbol de decisión	O(N2*M2)
Validar el árbol de decisión	O(N3*M*2N)

Tabla 2: Complejidad temporal de los algoritmos de entrenamiento y prueba. *(Por favor, explique qué significan N y M en este problema.)*

Algoritmo	Complejidad de memoria
Entrenar el árbol de decisión	O(N*M*2N)
Validar el árbol de decisión	O(1)

Tabla 3: Complejidad de memoria de los algoritmos de entrenamiento y prueba. *(Por favor, explique qué significan N y M en este problema.)*

4.4 Criterios de diseño del algoritmo

Explica por qué el algoritmo fue diseñado de esa manera. Use un criterio objetivo. Los criterios objetivos se basan en la

eficiencia, que se mide en términos de tiempo y consumo de memoria. Ejemplos de criterios no objetivos son: "Estaba enfermo", "fue la primera estructura de datos que encontré en Internet", "lo hice el último día antes del plazo", etc. Recuerde: Este es el 40% de la calificación del proyecto.

5. RESULTADOS

5.1 Evaluación del modelo

En esta sección, presentamos algunas métricas para evaluar el modelo. La precisión es la relación entre el número de predicciones correctas y el número total de datos de entrada. Precisión. es la proporción de estudiantes exitosos identificados correctamente por el modelo y estudiantes exitosos identificados por el modelo. Por último, Sensibilidad es la proporción de estudiantes exitosos identificados correctamente por el modelo y estudiantes exitosos en el conjunto de datos.

5.1.1 Evaluación del modelo en entrenamiento

A continuación presentamos las métricas de evaluación de los conjuntos de datos de entrenamiento en la Tabla 3.

	Conjunto de datos 1	***Conjunto de datos 2***	***...Conjunto de datos n***
Exactitud	0.7	0.75	0.9
Precisión	0.7	0.75	0.9
Sensibilidad	0.7	0.75	0.9

Tabla 3. Evaluación del modelo con los conjuntos de datos de entrenamiento.

5.1.2 Evaluación de los conjuntos de datos de validación

A continuación presentamos las métricas de evaluación para los conjuntos de datos de validación en la Tabla 4.

	Conjunto de datos 1	***Conjunto de datos 2***	***...Conjunto de datos n***
Exactitud	0.5	0.55	0.7
Precisión	0.5	0.55	0.7
Sensibilidad	0.5	0.55	0.8

Tabla 4. Evaluación del modelo con los conjuntos de datos de validación.

5.2 Tiempos de ejecución

Calcular el tiempo de ejecución de cada conjunto de datos en Github. Medir el tiempo de ejecución 100 veces, para cada conjunto de datos, e informar del tiempo medio de ejecución para cada conjunto de datos.

	Conjunto de datos 1	***Conjunto de datos 2***	***...Conjunto de datos n***
Tiempo de entrenamiento	10.2 s	20.4 s	5.1 s
Tiempo de validación	1.1 s	1.3 s	3.3 s

Tabla 5: Tiempo de ejecución del algoritmo *(Por favor, escriba el nombre del algoritmo, C4.5, ID3)* para diferentes conjuntos de datos.

5.3 Consumo de memoria

Presentamos el consumo de memoria del árbol de decisión binario, para diferentes conjuntos de datos, en la Tabla 6.

	Conjunto de datos 1	***Conjunto de datos 2***	***...Conjunto de datos n***
Consumo de memoria	10 MB	20 MB	5 MB

Tabla 6: Consumo de memoria del árbol de decisión binario para diferentes conjuntos de datos.

Para medir el consumo de memoria, debería usar un generador de perfiles (*profiler*). Uno muy bueno para Java es VisualVM, desarrollado por Oracle, http://docs.oracle.com/javase/7/docs/technotes/guides/visualvm/profiler.html. Para Python, use C-profiler.

6. DISCUSIÓN DE LOS RESULTADOS

Explique los resultados obtenidos. ¿Son la precisión, exactitud y sensibilidad apropiadas para este problema? ¿El modelo está sobreajustado? ¿Es el consumo de memoria y el consumo de tiempo sib apropiados? *(En este semestre, de acuerdo con los resultados, ¿se puede aplicar esto para dar becas o para ayudar a los estudiantes con baja probabilidad de éxito? ¿Para qué es mejor?)*

6.1 Trabajos futuros

Respuesta, ¿qué le gustaría mejorar en el futuro? ¿Cómo le gustaría mejorar su algoritmo y su implementación? ¿Qué hay de usar un bosque aleatorio?

AGRADECIMIENTOS

Identifique el tipo de agradecimiento que quiere escribir: Para una persona o para una institución. Considere las siguientes pautas: 1. El nombre del profesor no se menciona porque es un autor. 2. No debe mencionar sitios web de autores de artículos que no haya contactado. 3. Debe

mencionar estudiantes y profesores de otros cursos que le hayan ayudado.

Como ejemplo: Esta investigación fue apoyada parcialmente por [Nombre de la Fundación, Donante].

Agradecemos la asistencia con [técnica particular, metodología] a [nombre apellido, cargo, nombre de la institución] por los comentarios que mejoraron enormemente el manuscrito.

REFERENCIAS

1.Icfes. Acerca del examen. https://www.icfes.gov.co/acerca-del-examensaber-pro

2.Sergio Valero, Alejandro Salvador, Marcela Garcia. Mineria de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. https://d1wqtxts1xzle7.cloudfront.net/34203825/e1 .pdf?1405405774=&response-contentdisposition=inline%3B+filename%3DMineria_de_datos_prediccion_de_la_deserc.pdf&Expires=1597604666&Signature=bq92Es~7rNPqCYpvnVvg47u96f12Y71compLlT8~5HYxYFCIf5DRuQpt6GaIgeUkWsrOi6Usnfo~1d7gSYoJIalUpPZ5XZVrWtRBzVcwJVBi3R0kRGElod3RMt6TX5Ct7dLGBVhZm95O25v-CfrbHnjlWoVZGKPJKcPh~l4oECjkCm14TjLJ16LWLc07RK6zfnH8OJJ5yGRDTruuZfBNmzDEvrmKnnWu5EoXLE5bY GK~hZd4ohx8vKi0gPrIAMGkRqBxKJPmiU2eRqImxQ7~AZLery8RFLcwCHoH22qlWb4ilKU2p4 AlqWo-lhnilQEGRSo60bOpnxmvsefwuA__&Key-PairId=APKAJLOHF5GGSLRBV4ZA

3. Timarán-Pereira, R., Caicedo-Zambrano, J., & HidalgoTroya, A. (2019). Árboles de decisiones para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas saber 11°. Rev.investig.desarro.innov., 9 (2), 363-378. doi: 10.19053/20278306.v9.n2.2019.9184

4. Timarán-Pereira, S. R., Hernández-Arteaga, I., CaicedoZambrano, S. J., Hidalgo-Troya, A. y AlvaradoPérez, J. C. (2016). Descubrimiento de patrones de desempeño académico en las competencias genéricas. En Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional (pp. 101-150). Bogotá: Ediciones Universidad Cooperativa de Colombia. doi: http://dx.doi.org/10.16925/9789587600490

5. Porcel, Eduardo; Dapozo, Gladys; López, María V. Departamento de Informática. Facultad de Ciencias Exactas y Naturales y Agrimensura Universidad Nacional del Nordeste. 9 de Julio Nº 1449. CP 3400. Corrientes. Argentina. http://sedici.unlp.edu.ar/bitstream/handle/10915/19846/Documento_completo.pdf?sequence=1&isAllowed=y

6. Ing. Bruno Lopez. Inteligencia artifical. http://www.itnuevolaredo.edu.mx/takeyas/apuntes/Inteligencia%20Artificial/Apuntes/tareas_alumnos/C4.5/C4.5(2005-II-B).pdf

7. Francisco Parra. Estadistica y machine learning. https://bookdown.org/content/2274/portada.html