Detección de Perfiles de Rendimiento Académico en la Universidad Nacional del Este de Paraguay

La Red Martínez, David L.[1]

Bobadilla, Gabriela M.[2]

Delgado, Lourdes[2]

Ayala, Katia[2]

[1]Facultad Regional Resistencia, Universidad Tecnológica Nacional, Resistencia - Argentina
[2]Facultad Politécnica, Universidad Nacional del Este, Ciudad del Este – Paraguay

Resumen

La universidad enfrenta actualmente el desafío de mejorar su proceso de enseñanza-aprendizaje. Entre las variables que debe atender se encuentra el perfil de rendimiento académico de sus estudiantes. Estudios sobre este tema muestran que hay diversas formas de evaluar esta variable, que pueden variar según el contexto regional y la realidad social. Ante este problema, en este trabajo se propuso la detección temprana de perfiles de alumnos de la Universidad Nacional del Este, mediante técnicas de minería de datos. Se aplicaron las técnicas denominadas "agrupamiento" (*clustering*) y "árboles de decisión" (*decision trees*) sobre datos consolidados de 1801 alumnos. Estos datos fueron cargados en un almacén de datos. Se observó que las variables: "grado educacional de los padres" y "actitud hacia el estudio"; son algunas de las que más inciden en el rendimiento académico de los alumnos. Además, se aporta información que puede asistir a la toma de decisión.

Palabras clave: minería de datos, almacenes de datos, rendimiento académico.

Detection of Profiles of Academic Performance in the National University of the East of Paraguay

La Red Martínez, David L.[1]

Bobadilla, Gabriela M.[2]

Delgado, Lourdes[2]

Ayala, Katia[2]

[1]Facultad Regional Resistencia, Universidad Tecnológica Nacional, Resistencia - Argentina
[2]Facultad Politécnica, Universidad Nacional del Este, Ciudad del Este – Paraguay

Abstract

The university currently faces the challenge of improving its teaching-learning process. Among the variables it must address is the academic performance profile of its students. Studies on this topic show that there are various ways to evaluate this variable, which can vary according to the regional context and social reality. In view of this problem, this work proposes the early detection of student profiles at the Universidad Nacional del Este, through data mining techniques. Techniques "clustering" and "decision trees" were applied on consolidated data of 1801 students. These data were loaded in a data warehouse. It was observed that variables: "parents' educational grade" and "attitude towards study"; are some of the ones that more affect the academic performance of the students. In addition, information is provided that can assist in decision making.

Keywords: data mining, data warehouse, academic performance.

Introducción

La universidad enfrenta actualmente el desafío de mejorar su calidad académica enfocándose no solamente en el sistema de enseñanza-aprendizaje, sino contemplando otras variables, como la sistematización de procesos de evaluación permanentes (Briand et al., 1999). Entre estas variables, se destaca el estudio del perfil de rendimiento académico de los estudiantes.

Se define al rendimiento académico como la productividad del sujeto, matizado por sus actividades, rasgos y la percepción más o menos correcta de los cometidos asignados (Maletic et al., 2002).

Generalmente, al evaluar el rendimiento académico, se analizan en mayor o menor medida los elementos que influyen en el desempeño como ser, entre otros, factores socioeconómicos, la amplitud de programas de estudio, las metodologías de enseñanza y conocimientos previos del alumno (Marcus, 2003).

Se ha demostrado con varios estudios que el factor más relacionado con la calidad educativa es el propio alumno como coproductor, medido a través del nivel socioeconómico del hogar de donde proviene (Maradona & Calderón, 2007; Bobadilla Almada & la Red Martínez, 2015) y se ha evidenciado que la productividad del estudiante es mayor para las mujeres, para los estudiantes de menor edad y para quienes provienen de hogares con padres más educados (Porto, 2003).

También se ha mostrado el rendimiento entre las personas que trabajan y estudian y las que solamente estudian, encontrándose que no existen diferencias significativas en el rendimiento académico de los dos conjuntos (Reyes, 2004).

El problema de encontrar buenos predictores del rendimiento futuro de manera que se reduzca el fracaso académico en los programas de postgrado ha recibido una especial atención en EE. UU. (Wilson & Hardgrave, 1995), habiéndose encontrado que las técnicas de clasificación como el análisis discriminante o la regresión logística son más adecuadas que la regresión lineal múltiple a la hora de predecir el éxito/fracaso académico.

La diversidad de estudios sobre el rendimiento académico muestra que no existe una manera única para evaluarlo. Por ello, la determinación de grupos o clases de alumnos es un elemento para tener en cuenta para establecer las causas de los problemas relacionados al desempeño de estos. Más aún, los problemas pueden variar dependiendo

del contexto regional y la realidad social donde está inserto el alumno. Es decir, no existen herramientas que se puedan aplicar a todos los ámbitos y los resultados tampoco pueden ser extensibles para explicar todas las situaciones posibles. Esto denota claramente la necesidad de determinar perfiles en las instituciones educativas específicas adaptando las herramientas a cada situación particular.

Surge, entonces, la necesidad de implementar un mecanismo que permita determinar las características propias del estudiante, analizando la existencia de relaciones y patrones de comportamiento estudiantiles que posibilite la definición clara de perfiles de alumnos. Para ello una alternativa es utilizar técnicas de minería de datos para el modelado descriptivo (la Red Martínez & Podestá, 2014; la Red Martínez et al., 2015; Formia et al., 2013; Timarán Pereira, 2010).

A su vez, el modelado predictivo puede usarse para analizar una base de datos y determinar ciertas características esenciales acerca del conjunto de datos que permitan predecir el comportamiento de alguna variable (Connoly, 2005), en nuestro caso el rendimiento académico.

El presente trabajo identifica patrones característicos de los estudiantes de la UNE, según su rendimiento académico, considerando el nivel socioeconómico, como así también aspectos actitudinales e institucionales de los estudiantes universitarios, haciendo uso de técnicas de minería de datos (Data Mining, DM), (Hassanein & Elmelegy, 2014; Sadiku et al., 2015).

Metodología

Este trabajo consideró las etapas del proceso de descubrimiento de conocimiento en base de datos (Knowledge Discovery in Databases, KDD), donde la DM es parte significativa, por esto se implementó la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) (Chapman et al., 2000) (Figura 1), se aplicaron los modelos de minería de datos con técnicas de árbol de decisión y agrupamiento (cluster).

CRISP-DM es un método probado para orientar los trabajos de minería de datos.

Como metodología, incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas.

Como modelo de proceso, ofrece un resumen del ciclo de vida de minería de datos.

El modelo de CRISP-DM es flexible y se puede personalizar fácilmente. CRISP-DM permite crear un modelo de minería de datos que se adapte a necesidades concretas (Helberg, 2002).

Figura 1. Fases del proceso de modelado de la metodología CRISP-DM aplicado (fuente propia).

Obtención de datos

Se ha elaborado el instrumento de obtención de datos socioeconómicos y de actitud hacia el estudio (Figura 2), que se ha aplicado a las 6 unidades académicas y 1 escuela de la UNE durante los años 2018 y 2019, obteniéndose una muestra del 25,50% de los alumnos (1801 alumnos muestreados).

Luego se ha procedido a consolidar en un almacén de datos, los datos obtenidos con las notas de los alumnos, provistos por el sistema UNESYS de gestión académica de la UNE, habiéndose utilizado el ISW (InfoSphere Warehouse) 7.2 de IBM, software integrado para gestión de bases de datos, almacenes de datos, análisis multidimensional

y minería de datos descriptiva y predictiva; este software fue obtenido gratuitamente en el marco del Programa de Iniciativa Académica de dicha empresa.

Los datos han sido preparados integrando un almacén de datos o *data warehouse* (DW) para la aplicación de técnicas de minería de datos para la identificación de perfiles.

La estructura del DW consta de una tabla de hechos y varias tablas de dimensiones (Figura 3), que permiten hacer análisis multidimensional de los datos, estudiándolos desde distintas perspectivas, permitiendo también la integración de todas las tablas (Figura 4).

Figura 2. Formulario en línea para la obtención de los datos socioeconómicos (fuente propia).

Figura 3. Estructura global del almacén de datos utilizado (fuente propia).

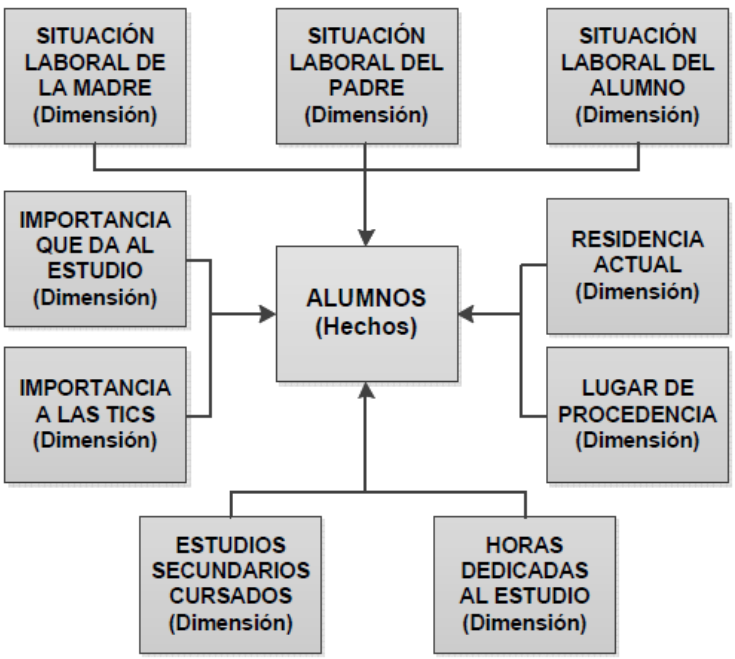
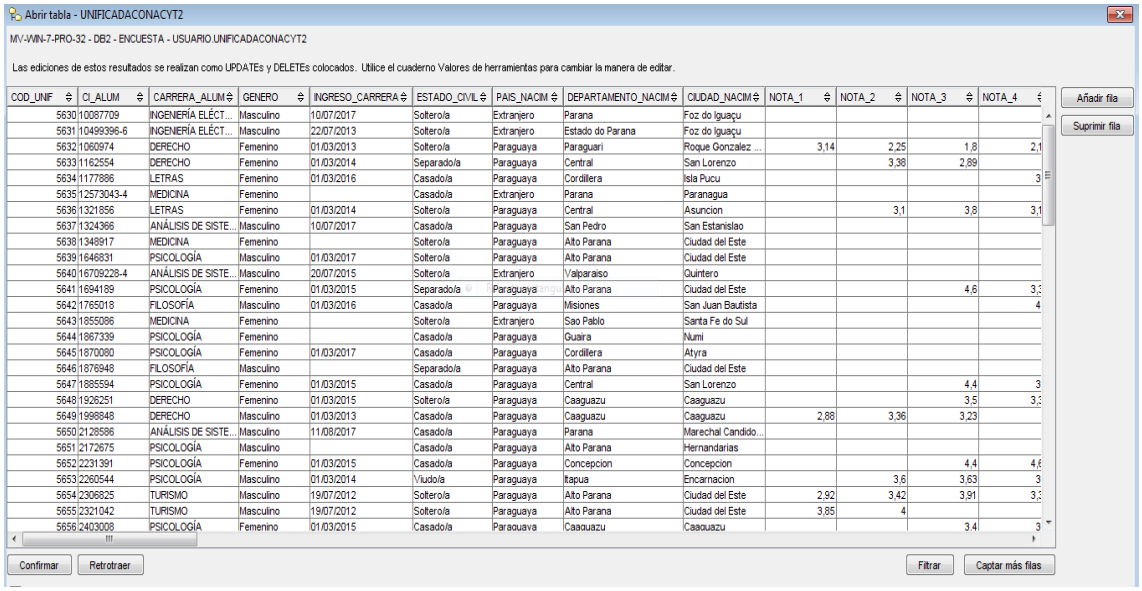


Figura 4. Visión parcial del almacén de datos preparado (fuente propia).



COD_UNIF	CI_ALUM	CARRERA_ALUM	GENERO	INGRESO_CARRERA	ESTADO_CIVIL	PAIS_NACIM	DEPARTAMENTO_NACIM	CIUDAD_NACIM	NOTA_1	NOTA_2	NOTA_3	NOTA_4
5630	10087709	INGENIERÍA ELÉCT...	Masculino	10/07/2017	Soltero/a	Extranjero	Parana	Foz do Iguaçu				
5631	10499396-6	INGENIERÍA ELÉCT...	Masculino	22/07/2013	Soltero/a	Extranjero	Estado do Parana	Foz do Iguaçu				
5632	1060974	DERECHO	Femenino	01/03/2013	Soltero/a	Paraguaya	Paraguari	Roque Gonzalez ...	3,14	2,25	1,8	2,1
5633	1162554	DERECHO	Femenino	01/03/2014	Separado/a	Paraguaya	Central	San Lorenzo		3,38	2,89	
5634	1177886	LETRAS	Femenino	01/03/2016	Casado/a	Paraguaya	Cordillera	Isla Pucu				3
5635	12573043-4	MEDICINA	Femenino		Casado/a	Extranjero	Parana	Paranagua				
5636	1321856	LETRAS	Femenino	01/03/2014	Soltero/a	Paraguaya	Central	Asuncion		3,1	3,8	3,1
5637	1324366	ANÁLISIS DE SISTE...	Masculino	10/07/2017	Casado/a	Paraguaya	San Pedro	San Estanislao				
5638	1348917	MEDICINA	Femenino		Soltero/a	Paraguaya	Alto Parana	Ciudad del Este				
5639	1646831	PSICOLOGÍA	Masculino	01/03/2017	Soltero/a	Paraguaya	Alto Parana	Ciudad del Este				
5640	16709228-4	ANÁLISIS DE SISTE...	Masculino	20/07/2015	Soltero/a	Extranjero	Valparaiso	Quintero				
5641	1694189	PSICOLOGÍA	Femenino	01/03/2015	Separado/a	Paraguaya	Alto Parana	Ciudad del Este			4,6	3,3
5642	1765018	FILOSOFÍA	Masculino	01/03/2016	Casado/a	Paraguaya	Misiones	San Juan Bautista				4
5643	1855086	MEDICINA	Femenino		Soltero/a	Extranjero	Sao Pablo	Santa Fe do Sul				
5644	1867339	PSICOLOGÍA	Femenino		Casado/a	Paraguaya	Guaira	Numi				
5645	1870080	PSICOLOGÍA	Femenino	01/03/2017	Casado/a	Paraguaya	Cordillera	Atyra				
5646	1876948	FILOSOFÍA	Masculino		Separado/a	Paraguaya	Alto Parana	Ciudad del Este				
5647	1885594	PSICOLOGÍA	Femenino	01/03/2015	Casado/a	Paraguaya	Central	San Lorenzo			4,4	3
5648	1926251	DERECHO	Femenino	01/03/2015	Soltero/a	Paraguaya	Caaguazu	Caaguazu			3,5	3,3
5649	1998848	DERECHO	Masculino	01/03/2013	Casado/a	Paraguaya	Caaguazu	Caaguazu	2,88	3,36	3,23	
5650	2128586	ANÁLISIS DE SISTE...	Masculino	11/08/2017	Casado/a	Paraguaya	Parana	Marechal Candido...				
5651	2172675	PSICOLOGÍA	Masculino		Casado/a	Paraguaya	Alto Parana	Hernandarias				
5652	2231391	PSICOLOGÍA	Femenino	01/03/2015	Casado/a	Paraguaya	Concepcion	Concepcion			4,4	4,6
5653	2260544	PSICOLOGÍA	Masculino	01/03/2014	Viudo/a	Paraguaya	Itapua	Encarnacion		3,6	3,63	3
5654	2306825	TURISMO	Masculino	19/07/2012	Soltero/a	Paraguaya	Alto Parana	Ciudad del Este	2,92	3,42	3,91	3,3
5655	2321042	TURISMO	Masculino	19/07/2012	Soltero/a	Paraguaya	Alto Parana	Ciudad del Este	3,85	4		
5656	2403008	PSICOLOGÍA	Femenino	01/03/2015	Casado/a	Paraguaya	Caaguazu	Caaguazu			3,4	3

Identificación de perfiles de rendimiento académico de los alumnos de la UNE, aplicando técnicas de minería de datos

Árboles de Decisión

Los árboles de decisión son algoritmos de aprendizaje supervisado (Han et al., 2012; Liu, 2011) que son muy populares cuando se encuentran relaciones entre los valores y atributos que describen el conjunto de datos.

Son bastante fáciles de tratar; el objetivo es encontrar los nodos que mejor clasifiquen los datos. Este proceso se realiza de forma recursiva y el algoritmo agrega un nodo al árbol cada vez que se encuentra una correlación significativa entre los atributos y las clases. Para establecer la conveniencia de utilizar un atributo como nodo de árbol, se utiliza la entropía como medida de obtención de información (Han et al., 2012; Liu, 2011). A lo largo del proceso de construcción del árbol, se calcula la entropía total del conjunto de datos y luego su entropía con respecto a cada atributo. El nodo con la mayor ganancia de información se elige en cada paso de la operación.

La ganancia de información se define como la diferencia entre la entropía total y la entropía del conjunto con respecto al atributo. Al ser un proceso recursivo, se exploran todas las posibles combinaciones de atributos para cada rama, que termina en un árbol completo que, en base a la entropía, explica claramente las relaciones entre los atributos.

Si bien la representación en árbol es útil a la hora de analizar datos, el mayor beneficio de este algoritmo es que permite determinar reglas de clasificación con dos parámetros asociados: soporte y confianza (Liu, 2011). El soporte mide la ocurrencia de la regla en el conjunto de datos, mientras que la confianza mide la confiabilidad de que, dados ciertos datos, se logra la clase predicha.

El software utilizado para árboles de decisión es parte del Intelligent Miner, componente del ISW ya mencionado y se describe en (Ballard et al., 2007), siendo un desarrollo de IBM en base a (Breiman et al., 1984).

Las principales opciones de configuración son las siguientes, para las cuales se han utilizado los valores por defecto en todos los casos:

Pureza máxima: Esta opción especifica la pureza máxima permitida de los nodos internos del árbol. Los valores permitidos son valores porcentuales entre 0 y 100; 100 es el valor predeterminado. Si se especifica un valor x <100, todos los nodos no hoja para los que más del x% de los registros de datos de entrenamiento asociados tienen un solo valor objetivo (o etiqueta de clase), se podan y se convierten en nodos hoja.

Profundidad máxima: Esta opción se utiliza para definir un límite superior para la profundidad del árbol. La profundidad de un nodo de árbol es el número de bordes entre este nodo y el nodo raíz del árbol. Si se especifica un valor entero x para la profundidad máxima, se podan todos los nodos de hojas con profundidades mayores que x. Por defecto, la profundidad máxima del árbol es ilimitada.

Número mínimo de registros: Esta opción se utiliza para definir un límite inferior para el número de registros en un solo nodo de árbol. Si se especifica un valor entero x para número mínimo de registros, todos los nodos que contienen menos de x registros se podan y fusionan en sus nodos principales. De forma predeterminada, el número mínimo de registros por nodo de árbol es cinco.

Aplicada a esta investigación, la técnica de Árboles de Decisión identifica perfiles descriptivos que se observan en la Tabla 1, clasificados por el atributo *Situación Final*, el cual se obtiene a partir de los promedios generales de hasta los cinco promedios totales del programa cursado por el alumno ya sea en forma semestral o anual, clasificados en 4 rangos [0-1,99], [2-2,99], [3-3,99] y [4-5].

Se puntualiza los siguientes resultados generales para todos los intervalos. En cuanto a los aspectos sociales, los alumnos en su mayoría han nacido y residen en el departamento de Alto Paraná, viven con familiares, son trabajadores del ámbito informal,

no poseen seguro social y provienen de colegios públicos. En cuanto a su actitud hacia el estudio, le dan mayor importancia que a la diversión, su motivación es aprender integralmente y aprobar, en cuanto al uso de las Tecnologías de la Información y Comunicación (TIC) opinan que facilitan el proceso de enseñanza.

En los resultados particulares por intervalos se destacan:

Intervalo [4-5]: corresponde a alumnos en su mayoría de género femenino, el último estudio cursado de los padres es universitario, el número de horas semanales dedicadas al estudio es más de 10 horas.

Intervalo [3-3,99]: corresponde a alumnos que en su mayoría son de género femenino, el último estudio cursado de los padres es el nivel medio concluido, el número de horas semanales dedicadas al estudio es menos de 10 horas.

Intervalo [2-2,99]: integrado por alumnos que en su mayoría son de género masculino, el último estudio cursado de los padres es el nivel medio concluido, el número de horas semanales dedicadas al estudio es menos de 10 horas.

Intervalo [0-1,99] integrado en igual proporción por alumnos de género femenino y masculino, el último estudio cursado por el padre es la primaria concluida y de la madre el nivel medio concluido, el número de horas semanales dedicadas al estudio es más de 10 horas.

Tabla 1: Resumen de resultados en porcentaje, aplicando técnicas de árboles de decisión (fuente propia).

	Tamaño en Porcentaje	**26,32%**		**51,19%**		**20,38%**		**2,11%**	
	Tamaño absoluto	**474**		**922**		**367**		**38**	
	Situación Final	**Promedio [4-5]**		**Promedio [3-3,99]**		**Promedio [2-2,99]**		**Promedio [0-1,99]**	
	Campo	**Valor modal**	**Frecuencia Modal %**	**Valor modal**	**Frecuencia Modal %**	**Valor modal**	**Frecuencia Modal %**	**Valor modal**	**Frecuencia Modal %**
DATOS GENERALES	GÉNERO	Femenino	60,55	Femenino	55,75	Masculino	56,13	Femenino	50,00
	ESTADO CIVIL	Soltero/a	92,19	Soltero/a	93,93	Soltero/a	94,55	Soltero/a	94,74
	DEPARTAMENTO NACIMIENTO	Alto Paraná	75,74	Alto Paraná	76,90	Alto Paraná	79,29	Alto Paraná	81,58
	CIUDAD DE RESIDENCIA	Ciudad del Este	69,62	Ciudad del Este	69,41	Ciudad del Este	64,31	Ciudad del Este	50,00
	TIPO DE RESIDENCIA	Con familiares	83,76	Con familiares	83,84	Con familiares	87,47	Con familiares	92,11

SITUACIÓN SOCIAL DEL ALUMNO	SITUACIÓN LABORAL	Trabaja	48,95	Trabaja	52,71	Trabaja	51,50	Trabaja	44,74
	SEGURO SOCIAL	No	59,70	No	62,04	No	65,12	No	55,26
	DEPENDENCIA COLEGIO SECUNDARIO	Pública	60,13	Pública	62,26	Pública	62,13	Pública	71,05
SITUACIÓN SOCIAL DE LOS PADRES	ÚLTIMOS ESTUDIOS DEL PADRE	Universitario Concluido	26,16	Nivel Medio Concluido	27,44	Nivel Medio Concluido	27,52	Primaria Concluida	23,68
	ÚLTIMOS ESTUDIOS DE LA MADRE	Universitario Concluido	29,54	Nivel Medio Concluido	23,54	Nivel Medio Concluido	25,61	Nivel Medio Concluido	26,32
	SITUACIÓN LABORAL DEL PADRE	Trabaja	78,69	Trabaja	75,16	Trabaja	82,56	Trabaja	65,79
	SITUACIÓN LABORAL DE LA MADRE	Trabaja	56,75	Trabaja	59,00	Trabaja	55,31	Trabaja	55,26
ACTITUD HACIA EL ESTUDIO	IMPORTANCIA AL ESTUDIO	Más que a la diversión	63,71	Más que a la diversión	57,16	Más que a la diversión	58,86	Más que a la diversión	68,42
	HORAS SEMANALES DE ESTUDIO	Más de 10	54,64	Menos de 10	53,90	Menos de 10	53,95	Más de 10	60,53
	MOTIVACIÓN PARA ESTUDIAR	Aprender integralmente y Aprobar	52,53	Aprender integralmente y Aprobar	49,13	Aprender integralmente y Aprobar	50,95	Aprender integralmente y Aprobar	39,47
	UTILIDAD DE LAS TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIÓN	Facilitan el proceso de enseñanza	49,37	Facilitan el proceso de enseñanza	50,87	Facilitan el proceso de enseñanza	55,86	Facilitan el proceso de enseñanza	52,63

Agrupamientos (Clusters)

El método utilizado para encontrar agrupamientos (*clusters*) es el de agrupación demográfica para la agrupación de datos con valores de atributos similares. La idea principal es el uso de métricas de distancia entre los valores de los atributos. La agrupación demográfica compara cada instancia con los grupos previamente establecidos y la asigna a la que maximiza la valoración de la similitud. Este es un proceso iterativo y realiza una gran cantidad de rastreo de datos con el objetivo de minimizar el error de asignación.

La calidad del conglomerado o agrupamiento se mide no sólo internamente (menor número de reasignaciones y menor error de distancia cuadrática media, etc.) (Liu, 2011), sino también con medidas intragrupo (Arbelaitz et al., 2013).

La agrupación demográfica se basa en la distribución. Proporciona una agrupación rápida y natural de bases de datos muy grandes. Los clústeres se caracterizan por la

distribución de valor de sus miembros. Determina automáticamente el número de clústeres que se generarán.

Normalmente, los datos demográficos contienen muchas variables categóricas. La función de minería funciona bien con conjuntos de datos que constan de este tipo de variables. También puede utilizar variables numéricas. El algoritmo de agrupación demográfica trata las variables numéricas asignando similitudes de acuerdo con la diferencia numérica de los valores.

La agrupación demográfica es un proceso iterativo sobre los datos de entrada. Cada registro de entrada se lee sucesivamente. Se calcula la similitud de cada registro con cada uno de los clústeres existentes actualmente. Si la mayor similitud calculada está por encima de un umbral determinado, el registro se agrega al grupo correspondiente. Las características de este grupo cambian en consecuencia. Si la similitud calculada no está por encima del umbral, o si no hay ningún grupo (que es el caso inicialmente), se crea un nuevo grupo que contiene solo el registro. Puede especificar el número máximo de clústeres, así como el umbral de similitud.

La agrupación demográfica utiliza el criterio estadístico de Condorcet (MATH 1340, 2010; Young, 1988; Laffond et al., 1995) para gestionar la asignación de registros a las agrupaciones y la creación de nuevas agrupaciones. El criterio de Condorcet evalúa qué tan homogéneo es cada grupo descubierto (en el sentido de que los registros que contiene son similares) y qué tan heterogéneos son los grupos descubiertos entre sí. El proceso iterativo de descubrimiento de agrupaciones se detiene después de dos o más pasadas sobre los datos de entrada si la mejora del resultado de agrupación según el criterio de Condorcet no justifica una nueva pasada (IBM, 2020).

Aplicada a esta investigación, se obtuvo que los agrupamientos 7, 3 y 9 son los de mayor tamaño y se muestran en la Tabla 2; los datos generales y sociales de los alumnos presentan resultados generales como: los promedios de las cinco notas del programa cursado por el alumno se encuentran en el rango de [3-3,99], en su mayoría de género femenino, de estado civil soltera/o, nacidos y fijan residencia en el departamento de Alto Paraná, viven con familiares, no poseen seguro social, provienen de colegios públicos.

En los resultados particulares se destacan: en el cluster 7, se agrupan alumnos que en su mayoría no trabajan, en los clusters 3 y 9 los alumnos agrupados son trabajadores en actividades no relacionadas a sus estudios.

En todos los casos, lo que se muestra son los *valores preponderantes* de las variables consideradas.

Tabla 2: Resumen de resultados en porcentaje, de los clusters con mayor tamaño, respecto de datos generales y sociales de los alumnos (fuente propia).

		Clúster 7		**Clúster 3**		**Clúster 9**	
	Tamaño	**33%**		**23,9%**		**11,3%**	
	Tamaño absoluto	**587**		**431**		**204**	
	Campo	**Valor modal**	**Frec. Modal %**	**Valor modal**	**Frec. Modal %**	**Valor modal**	**Frec. Modal %**
	SITUACIÓN FINAL	Promedio [3-3,99]	54,17	Promedio [3-3,99]	48,23	Promedio [3-3,99]	76,96
DATOS GENERALES	GÉNERO	Femenino	50,43	Femenino	54,05	Femenino	74,51
	ESTADO CIVIL	Soltero/a	98,81	Soltero/a	87,11	Soltero/a	94,12
	DEPARTAMENTO NACIMIENTO	Alto Paraná	90,46	Alto Paraná	76,51	Alto Paraná	77,45
	CIUDAD RESIDENCIA	Ciudad del Este	78,19	Ciudad del Este	64,03	Ciudad del Este	60,78
SITUACIÓN SOCIAL DEL ALUMNO	TIPO RESIDENCIA	Con familiares	95,91	Con familiares	82,74	Con familiares	87,25
	SEGURO SOCIAL	No	60,99	No	62,79	No	64,71
	SITUACIÓN LABORAL ALUMNO	No Trabaja	65,59	Trabaja	48,02	Trabaja	87,75
	ACTIVIDAD ECONÓMICA ALUMNO	No Trabaja	65,59	Otro	34,51	Otro	32,84
	HS TRABAJADAS ALUMNO	No Trabaja	65,93	De 36 o más	34,72	De 36 o más	32,84
	RELACIÓN TRABAJO CON EL ESTUDIO	No Trabaja	68,65	No relacionada	41,58	No relacionada	50,00
	DEPENDENCIA COLEGIO SECUNDARIO	Pública	53,49	Pública	64,45	Pública	76,47

En los clusters 10, 4 y 1, visualizados en la Tabla 3, se observa en los datos generales y sociales de los alumnos, que en su mayoría son de estado civil soltera/o, nacidos y con residencia en el departamento de Alto Paraná, viven con familiares, son alumnos trabajadores en actividades no relacionadas a sus estudios, trabajan 36 o más horas, provienen de colegios públicos.

En cuanto a los resultados particulares se hace notar: en el cluster 10, alumnos con promedio final en el intervalo [3-3,99]; en el cluster 4, promedio final en el intervalo [2-

2,99] y en el cluster 1, promedio final en el intervalo [4-5]; en estos dos últimos, la mayoría de los alumnos no posee seguro social.

Tabla 3: Resumen de resultados en porcentaje, de agrupación de los clusters de tamañao mendiano con los datos generales y sociales de los alumnos (fuente propia).

		Cluster 10		Cluster 4		Cluster 1	
	Tamaño	**6,2%**		**5,5%**		**5,4%**	
	Tamaño absoluto	**111**		**99**		**97**	
	Campo	**Valor modal**	**Frec. Modal %**	**Valor modal**	**Frec. Modal %**	**Valor modal**	**Frec. Modal %**
	SITUACIÓN FINAL	Promedio [3-3,99]	90,99	Promedio [2-2,99]	95,96	Promedio [4-5]	97,94
DATOS GENERALES	GÉNERO	Masculino	56,76	Masculino	70,71	Femenino	52,58
	ESTADO CIVIL	Soltero/a	86,49	Soltero/a	91,92	Soltero/a	85,57
	DEPARTAMENTO NACIMIENTO	Alto Paraná	71,17	Alto Paraná	69,70	Alto Paraná	49,48
	CIUDAD RESIDENCIA	Ciudad del Este	63,06	Ciudad del Este	53,54	Ciudad del Este	62,89
SITUACIÓN SOCIAL DEL ALUMNO	TIPO RESIDENCIA	Con familiares	69,37	Con familiares	82,83	Con familiares	73,20
	SEGURO SOCIAL	Sí	43,24	No	58,59	No	58,76
	SITUACIÓN LABORAL ALUMNO	Trabaja	65,77	Trabaja	47,47	Trabaja	64,95
	ACTIVIDAD ECONÓMICA ALUMNO	Otro	32,43	Otro	45,45	Otro	26,80
	HS TRABAJADAS ALUMNO	De 36 o mas	32,43	De 36 o mas	45,45	De 36 o mas	26,80
	RELACIÓN TRABAJO CON EL ESTUDIO	No relacionada	34,23	No relacionada	52,53	No relacionada	35,05
	DEPENDENCIA COLEGIO SECUNDARIO	Pública	51,35	Pública	60,61	Pública	71,13

En la Tabla 4 se visualizan los clusters 7, 3 y 9, con los datos sociales de los padres y actitud hacia el estudio de los alumnos. El cluster 7 agrupa mayoritariamente a alumnos cuyo último estudio de su padre y madre es universitario y dedican más de 10 horas al estudio. Los clusters 3 y 9 agrupan mayoritariamente a alumnos donde el último estudio de su padre y madre es el nivel medio y dedican al estudio menos de 10 horas.

Los clusters en forma general presentan que los padres son trabajadores, la actividad económica de los mismos es informal, trabajan más de 36 horas. La importancia que dan al estudio los alumnos es más que a la diversión, en cuanto a lo que opinan sobre la TICs, expresan que facilitan el proceso de enseñanza y la motivación para estudiar es

aprender integralmente y aprobar. Como ya se ha aclarado, los valores mostrados para las distintas variables son los valores preponderantes.

Tabla 4: Resumen de resultados en porcentaje, de agrupación de los clusters de mayor tamaño con los datos sociales de los padres y la actitud hacia el estudio de los alumnos (fuente propia).

		Cluster 7		**Cluster 3**		**Cluster 9**	
	Tamaño	**33%**		**23,9%**		**11,3%**	
	Tamaño absoluto	**587**		**431**		**204**	
	Campo	**Valor modal**	**Frec. Modal %**	**Valor modal**	**Frec. Modal %**	**Valor modal**	**Frec. Modal %**
	SITUACIÓN FINAL	Promedio [3-3,99]	54,17	Promedio [3-3,99]	48,23	Promedio [3-3,99]	76,96
SITUACIÓN SOCIAL PADRE	ÚLTIMOS ESTUDIOS PADRE	Universitario Concluido	31,69	Nivel Medio Concluido	25,78	Nivel Medio Concluido	29,90
	SITUACIÓN LABORAL PADRE	Trabaja	86,88	Trabaja	75,26	Trabaja	79,90
	ACTIVIDAD ECONÓMICA PADRE	Otro	17,21	Otro	21,00	Otro	26,47
	HS TRABAJADAS PADRE	De 36 o más	54,51	De 36 o más	41,37	De 36 o más	39,71
SITUACIÓN SOCIAL MADRE	ÚLTIMOS ESTUDIOS MADRE	Universitario Concluido	44,29	Nivel Medio Concluido	22,45	Nivel Medio Concluido	22,55
	SITUACIÓN LABORAL MADRE	Trabaja	95,57	No Trabaja	88,57	Trabaja	100,00
	ACTIVIDAD ECONÓMICA MADRE	Otro	16,87	No Trabaja	88,57	Otro	32,35
	HS TRABAJADAS MADRE	De 36 o más	57,07	No Trabaja	89,19	De 36 o más	44,61
ACTITUD HACIA EL ESTUDIO	HORAS ESTUDIO	Más de 10	62,35	Menos de 10	48,23	Menos de 10	88,73
	IMPORTANCIA ESTUDIO	Más que a la diversión	68,99	Más que a la diversión	53,43	Más que a la diversión	46,08
	UTILIDAD DE LAS TECNOLOGIAS DE LA INFORMACIÓN Y COMUNICAC IÓN	Facilitan el proceso de enseñanza	55,71	Facilitan el proceso de enseñanza	48,86	Facilitan el proceso de enseñanza	53,92
	MOTIVACIÓN PARA ESTUDIAR	Aprender integralmente y Aprobar	63,03	Aprender integralment e y Aprobar	45,95	Aprender integralmente y Aprobar	35,78

En la Tabla 5, se visualizan los clusters 10, 4 y 1 con los datos sociales de los padres y actitud hacia el estudio de los alumnos.

En los resultados particulares se destacan: el cluster 10 agrupa alumnos que en su mayoría lograron promedios finales en el intervalo [3-3,99], el último estudio de los padres es el nivel medio concluido, el padre trabaja en forma informal y se dedica más de 36 horas, las madres en su mayoría no trabajan, en cuanto a la actitud de los alumnos hacia el estudio, dedican menos de 10 horas semanales al mismo, su motivación para

estudiar es aprender íntegramente y aprobar, la importancia que dan al estudio es más que al trabajo y opinan que las TICs facilitan el proceso de enseñanza.

El cluster 4 agrupa a alumnos que en su mayoría tienen promedios finales en el intervalo [2-2,99], el último estudio de los padres es universitario y de las madres es el nivel medio concluido, los padres son trabajadores en forma informal y las madres en su mayoría no trabajan, la actitud de los alumnos hacia el estudio es que dedican menos de 10 horas semanales al mismo, su motivación para estudiar es aprender íntegramente y aprobar, la importancia que dan al estudio es más que la diversión y opinan que las TICs facilitan el proceso de enseñanza.

El cluster 1 agrupa mayoritariamente a alumnos cuyos promedios finales se encuentran en el intervalo [4-5], el último estudio de los padres es primaria no concluida y de las madres es primaria concluida, los padres son trabajadores informales y las madres en su mayoría no trabajan, la actitud de los alumnos hacia el estudio es que le dedican más de 10 horas semanales al mismo, su motivación para estudiar es aprender íntegramente y aprobar, la importancia que dan al estudio es más que la diversión y opinan respecto de las TICs que será imprescindible su dominio para el ejercicio profesional.

Tabla 5: Resumen de resultados en porcentaje, de agrupación de los clusters de tamañao medio con los datos sociales de los padres y la actitud hacia el estudio de los alumnos (fuente propia).

		Cluster 10		**Cluster 4**		**Cluster 1**	
	Tamaño	**6,2%**		**5,5%**		**5,4%**	
	Tamaño absoluto	**111**		**99**		**97**	
	Campo	**Valor modal**	**Frec. Modal %**	**Valor modal**	**Frec. Modal %**	**Valor modal**	**Frec. Modal %**
	SITUACIÓN FINAL	Promedio [3-3,99]	90,99	Promedio [2-2,99]	95,96	Promedio [4-5]	97,94
SITUACIÓN SOCIAL PADRE	ÚLTIMOS ESTUDIOS PADRE	Nivel Medio Concluido	27,93	Universitario Concluido	23,23	Primaria No Concluida	18,56
	SITUACIÓN LABORAL PADRE	Trabaja	52,25	Trabaja	68,69	Trabaja	64,95
	ACTIVIDAD ECONÓMICA PADRE	Otro	27,03	Otro	26,26	Otro	26,80
	HS TRABAJADAS PADRE	De 36 o más	47,75	De 36 o más	40,40	De 36 o más	37,11
	ÚLTIMOS ESTUDIOS MADRE	Nivel Medio Concluido	29,73	Nivel Medio Concluido	29,29	Primaria Concluida	26,80

SITUACIÓN SOCIAL MADRE	SITUACIÓN LABORAL MADRE	No Trabaja	79,28	No Trabaja	58,59	No Trabaja	60,82
	ACTIVIDAD ECONÓMICA MADRE	No Trabaja	79,28	No Trabaja	58,59	No Trabaja	60,82
	HS TRABAJADAS MADRE	No Trabaja	79,28	No Trabaja	58,59	No Trabaja	67,01
ACTITUD HACIA EL ESTUDIO	HORAS ESTUDIO	Menos de 10	77,48	Menos de 10	54,55	Más de 10	53,61
	IMPORTANCIA ESTUDIO	Más que el trabajo	55,86	Más que a la diversión	61,62	Más que a la diversión	58,76
	UTILIDAD DE LAS TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIÓN	Facilitan el proceso de enseñanza	45,95	Facilitan el proceso de enseñanza	55,56	Será imprescindible su dominio para el ejercicio profesional	43,30
	MOTIVACIÓN PARA ESTUDIAR	Aprender integralmente y Aprobar	36,04	Aprender integralmente y Aprobar	53,54	Aprender integralmente y Aprobar	42,27

Conclusiones

Los resultados obtenidos con la aplicación de las técnicas de minería de datos para identificar perfiles descriptivos de los alumnos de la UNE (clusters y árboles de decisión), han evidenciado las características de las clases representativas del rendimiento académico de los alumnos, considerando los promedios generales de hasta los cinco promedios totales del programa cursado por el alumno, clasificados en los rangos [0-1,99], [2-2,99], [3-3,99] y [4-5].

Los árboles de decisión han identificado las características sociales generales, tales como que los alumnos en su mayoría han nacido y residen en el departamento de Alto Paraná, viven con familiares, son trabajadores del ámbito informal, no poseen seguro social y provienen de colegios públicos. En cuanto al perfil de rendimiento se ha determinado que está relacionado con el nivel de estudio de los padres y la actitud hacia el estudio de los alumnos, específicamente la cantidad de horas otorgadas al estudio.

En el intervalo [4-5] se ha observado como características más relevantes que los alumnos en su mayoría son de género femenino, no trabajan, el último estudio cursado de los padres es universitario y el número de horas semanales dedicadas al estudio es de más de 10 horas.

En el intervalo [3-3,99] se ha observado que los alumnos se caracterizan en su mayoría por ser de género femenino, el último estudio cursado de los padres es el nivel medio concluido y el número de horas semanales dedicadas al estudio es de menos de 10 horas.

El intervalo [2-2,99] corresponde a alumnos que en su mayoría son de género masculino, el último estudio cursado de los padres es el nivel medio concluido, el número de horas semanales dedicadas al estudio es de menos de 10 horas.

El intervalo [0-1,99] indica igual proporción de género femenino y masculino, el último estudio cursado por los padres es la primaria concluida y de las madres el nivel medio concluido, el número de horas semanales dedicadas al estudio es de más de 10 horas.

Los clusters de mayor tamaño corresponden a alumnos mayoritariamente con promedios generales en el rango de [3-3,99], en su mayoría de género femenino, de estado civil soltera/o, nacidos y con residencia en el departamento de Alto Paraná, que viven con familiares, no poseen seguro social y provienen de colegios públicos.

Los clusters de menor tamaño corresponden a alumno cuyos rangos de notas mayoritariamente se encuentran en [2-2,99] y [4-5], éstos también presentan características sociales comunes, como que residen en el departamento de Alto Paraná, viven con familiares, no poseen seguro social y provienen de colegios públicos, ahora bien, las principales diferencias entre los clusters están relacionadas con el nivel de estudio de los padres y la actitud hacia el estudio de los alumnos, específicamente la cantidad de horas otorgadas al estudio.

Se destaca en el intervalo [4-5] el nivel de formación de los padres y la opinión de los alumnos hacia el uso de las TICs, considerando que será imprescindible su dominio para el ejercicio profesional.

Los resultados obtenidos han permitido caracterizar los distintos perfiles de los alumnos de la UNE especialmente en función de su rendimiento académico, lo cual permitirá el desarrollo de modelos predictivos de dicho rendimiento (la Red Martínez et al., 2016; 2017; 2018), tendientes a detectar tempranamente perfiles asociados a bajo rendimiento académico, lo cual permitirá tomar decisiones y realizar acciones tendientes a mejorar dicho rendimiento, contribuyendo a disminuir la deserción estudiantil.

Agradecimientos

Este trabajo es sostenido por el Proyecto de Investigación "Estudio del rendimiento académico y determinación temprana de perfiles de alumnos en la Universidad Nacional del Este de Paraguay, aplicando técnicas de minería de datos", CONACYT, PINV – 15 - 559. El software ISW (InfoSphere Warehouse) de gestión de DW y de DM fue obtenido de IBM a través de la Iniciativa Académica de dicha empresa.

Referencias Bibliográficas

Arbelaitz , O., Gurrutxaga, I., Muguerza , J., Pérez , J., & Perona, I. (2013). An extensive comparative study of cluster validity indices. Obtenido de https://doi.org/10.1016/J.PATCOG.2012.07.021

Ballard, C., Rollins, J., Ramos, J., & Perkins. (2007). Dynamic Warehousing: Data Mining Made Easy". IBM Information Management Software. En *IBM Information Managemet Software.*

Bobadilla Almada, G., & la Red Martínez, D. L. (2015). Estudio del rendimiento académico y determinación de perfiles de alumnos de la Facultad Politécnica de la Universidad Nacional del Este Paraguay. *FPUNE Scientifc*, 43-48.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Predicting materials properties and behavior using classification and regression trees. *Taylor & Francis Group*. doi:https://doi.org/10.1201/9781315139470

Briand, L., Daly, J., & Wüst, J. (1999). A unified framework for coupling measurement in object oriented systems. *IEEE Transactions on Software*, 25.

Chapman, P., Clinton, J., Kerber, R., & Khabaza, T. (2000). *CRISP-DM 1.0: Guía de minería de datos paso a paso.* Obtenido de https://www.semanticscholar.org/paper/CRISP-DM-1.0%3A-Step-by-step-data-mining-guide-Chapman-Clinton/54bad20bbc7938991bf34f86dde0babfbd2d5a72

Connoly, T. B. (2005). Sistemas de bases de Datos. *Un enfoque práctico para diseño, implementación y gestión*(4ta Edición).

Formia, S., Lanzarini, L., & Hasperué, W. (2013). Characterization of University Drop-Out at UNRN Using Data Mining. A study case. *CACIC.* Mar del Plata, Buenos Aires, Argentina.

Han J, Kamber M, & Pei J. (2012). *Data mining concepts and techniques* (Third ed.). (M. Waltham, & K. Morgan , Edits.) Obtenido de https://www.amazon.de/Data-Mining-Concepts-Techniques-

Management/dp/0123814790/ref=tmm_hrd_title_0?ie=UTF8&qid=1366039033&sr=1-1

Hassanein, W., & Elmelegy, A. (2014). Clustering Algorithms for Categorical Data Using Concepts of Significance and Dependence of Attributes. *European Scientific Journal*. Obtenido de https://doi.org/10.19044/esj.2014.v10n3p%p

Helberg, C. (2002). *Data Mining with Confidence.* SPSS Inc. doi:ISBN 1-56827-287-1.

IBM. (2020). *IBM Knowledge Center*. Obtenido de https://www.ibm.com/support/knowledgecenter/es/SSEPGG_10.5.0/com.ibm.im.model.doc/c_distribution_based_clustering.html

la Red Martínez , D., & Podestá, C. E. (2014). Contributions from Data Mining to Study Academic Performance of Students of a Tertiary Instute;. *American Journal of Educational Research, Volume 02*(9), 713-726.

la Red Martínez, D., Giovannini, M., Báez , M., Molinas, Torre, J., & Yaccuzzi, N. (2017). Academic performance problems: A predictive data mining-based model. *Academia Journal of Educational Research, 5*, págs. 61-75.

la Red Martínez, D., Karanik, M., & Giovannini, M. (2018). Academic achievement profiles: An intelligent predictive model based on data mining. *Academia Journal of Educational Research, 6*, págs. 279-289.

la Red Martínez, D., Karanik, M., Giovannini, M., & Scappini, R. (2016). Towards to a Predictive Model of Academic Performance Using Data Mining in the UTN-FRReTowards to a Predictive Model of Academic Performance Using Data Mining in the UTN-FRRe. *Journal of Systemics, Cybernetics and Informatics, 14*(2), págs. 36-41.

la Red Martínez, D., Karanik, M., Giovannini, M., Báez, M., & Pinto, N. (2015). Academic Performance Profiles: A Descriptive Model Based on Data Mining. *European Scientific Journal (ESJ)*, 17-38.

Laffond G, Laslier JF, & Le Breton M. (s.f.). Condorcet choice correspondences: A set-theoretical comparison. *Mathematical Social Sciences* (30), 23-25.

Liu B. (2011). *Web Data Mining. Computer Knowledge and Technology.* Academic Berlin, Heidelberg: Springer Berlin Heidelberg. Obtenido de https://doi.org/10.1007/978-3-642-19460-3

Maletic, J., Collard, M., & Marcus, A. (2002). Source Code Files as Structured Documents. *in Proceedings 10th IEEE International Workshop onProgram Comprehension (IWPC'02)*, 289-292.

Maradona, G., & Calderón, M. I. (2007). Una apliaplicacióndel enfoque de la función de producción en educación. *Revista de Economía y Estadística, XLII*.

Marcus, A. (2003). *Semantic Driven Program Analysis. Kent State University.* Kent. USA: OH.

MATH 1340. Mathematics and Politics: Condorcet's Method and Condorcet Winners. (2010). Obtenido de Cornell University. .

Porto, A. y. (2003). *Características y rendimiento de estudiantes universitarios. El casode la Facultad de Ciencias Económicas de la UniUniversidad.* Universidad Nacional de La Plata.

Reyes R, S. L. (enero-junio de 2004). El Bajo Rendimiento Académico de los Estudiantes Universitarios. Una Aproximación a sus Causas. *Theorethikos, Año VI*(N° 18).

Sadiku, M., Shadare, A., & Musa, S. (2015). Data Mining: a Brief Introduction. *European Scientific Journal, 21*. Obtenido de http://eujournal.org/index.php/esj/article/view/6017

Timarán Pereira, R. (2010). Una lectura sobre deserción universitaria en estudiantes de pregrado desde la perspectiva de la minería de datos. *Revista Científica Guillermo de Ockham., 8*(1), 121-130.

Wilson, R. L., & Hardgrave, B. C. (1995). Predicting graduate student success in an MBA program: Regression versus classification. *Educational and Psychological Measurement*(55), 186-195.

Young, H. (1988). Condorcet's theory of voting. *American Political Science, 82*(4), 1231-1244. doi:https://doi.org/10.2307/1961757