MÁSTER EN SISTEMAS INTELIGENTES

Dpto. de Informática y Análisis Numérico

Descubrimiento de conocimiento útil para la detección del fracaso escolar. Aplicación a un caso real.

Proyecto de Fin de Máster

DIRECTOR: Sebastián Ventura Soto

AUTOR: Miguel Ángel Vallejo Gato

Córdoba, Diciembre de 2015

Resumen

El fracaso escolar es un problema que está presente en muchos ámbitos de la sociedad. Se produce cuando existe un fallo en una acción educativa (ya sea por parte del proceso de enseñanza propuesto por el centro o del proceso de aprendizaje del alumno).

La predicción temprana de alumnos en riesgo nos permite ejecutar medidas correctivas que permitan evitar el fracaso escolar de dichos alumnos (y por lo tanto el fracaso del centro).

El Grupo Santillana comercializa productos software que recogen información de los alumnos en distintos ámbitos (LMS, Test de conocimientos y gestión económica). Estos datos no han sido explotados y podrían ser una importante fuente de extracción de conocimiento.

En este proyecto se trata de generar un conjunto de datos sobre el cual se aplicarán técnicas de minería de datos para crear modelos predictivos que permitan detectar a los alumnos en riesgo de fracaso escolar. De esta forma el personal docente podrá actuar a tiempo para evitar esta situación.

Palabras Clave: EDM, minería de datos, data mining, clasificación, classification, fracaso escolar, school failure, abandono escolar, dropout, LMS, sistema para la gestión del aprendizaje, learning management system, Pleno, Datamart, modelos predictivos, predictive models.

Abstract

Educational failure is a problem that is very prevalent in many areas of our society. It is produced when there is failure in educational activities. This may be due to the process of the educational system or due to the student's learning process

Early detection of these at risks students allows us to execute corrective measures that will prevent educational failure, as well as the failure of that educational center.

Santillana Group commercializes software products that collect student's information in various areas (LMS, Knowledge testing and economy management). This data has not been researched and could be an important source of knowledge extraction.

This project attempts to generate a dataset on which data mining techniques will be applied to create predictive models to detect at risks students. This way the teachers may act accordingly in order to avoid this situation.

Keywords: EDM, minería de datos, data mining, clasificación, classification, fracaso escolar, school failure, abandono escolar, dropout, LMS, sistema para la gestión del aprendizaje, learning management system, Pleno, Datamart, modelos predictivos, predictive models.

Agradecimientos

Dedicado a la memoria de la Dra. Carmen del Rio Rincón a quien debo mi vocación investigadora.

En primer lugar, quiero agradecer su colaboración a mi director de proyecto, Sebastián Ventura Soto, por su aportación a este proyecto y por la orientación ofrecida. Por él me decidí a realizar este máster. Fue uno de mis mejores maestros hace más de 15 años y he tenido la suerte de tenerlo como mentor en este proyecto.

En el ámbito laboral debo agradecer:

A Cristóbal Romero quien me ha ayudado a encontrar soluciones a algunos problemas que me he encontrado en este proyecto.

A mis compañeros de trabajo, especialmente a Jose Antonio que siempre me ha prestado su ayuda cuando lo he necesitado de forma incondicional. Espero que podamos trabajar juntos por mucho tiempo.

A mis antiguos compañeros del Ayuntamiento de Montilla, especialmente Mari Sierra, Jose Antonio y Eva, con quienes he vivido momentos muy felices y a los que siempre consideraré mis amigos.

A mis antiguos compañeros del I.F.A.P.A. con quienes compartí 10 años de mi vida.

Por otro lado, quiero expresar mi agradecimiento a las personas que están a mi lado día a día ayudándome a superar las dificultades y que hacen que todo esfuerzo merezca la pena:

A mi querida mujer, Victoria. Juntos hemos conseguido este objetivo y es tanto de ella como mío. No podría haberlo hecho sin su ayuda y sacrificio.

A mi madre que siempre ha dedicado su vida a los demás, especialmente a su familia. No habría llegado hasta aquí sin ella.

A mi padre que siempre está cuando se le necesita y que me apoya incondicionalmente, incluso cuando tenemos opiniones distintas.

A mi abuela por seguir con nosotros para poder celebrar este día.

A mi hermano que nunca me ha fallado y con quien siempre puedo contar.

A mi sobrino Rafalín cuya alegría y entusiasmo me hace sentir vivo y feliz incluso en los momentos más duros.

A mis sobrinos Alejandro, Virginia, Miguel, Lorenzo, Lucía y Paula. Disfruto cada momento que compartimos. Desearía poder pasar más tiempo con ellos.

A mis amigos de siempre. Aunque no los vea mucho se que están ahí y puedo contar con ellos.

Índice general

Índice de figuras

Índice de tablas

1 Introducción

En este proyecto se trata de habilitar una importante fuente de datos educativos aún sin explotar, la cual tienen gran cantidad de información y están creciendo día a día.

El Grupo Santillana comercializa productos software en más de 15 países. Estos productos están siendo utilizados por más de 150.000 alumnos pertenecientes a 1100 colegios de todo el mundo. Actualmente se ha comercializado masivamente un sistema para la gestión del aprendizaje (LMS) y desde hace dos de años se está instaurando un software para la realización de tests de conocimiento. En estos momentos se está empezando a comercializar otros dos productos (sistema de gestión de alumnos y un nuevo proyecto del que aún se desconocen los detalles) por lo que en un futuro cercano contaremos con 4 fuentes de datos de las que se podrán extraer información muy variada de los mismos alumnos.

Estas fuentes de datos contienen información que aún no ha sido estudiada y podría ser de interés para distintos grupos:

- Investigadores: Contaríamos con un conjunto de datos de un tamaño considerable que se irá incrementando/actualizando cada año con un gran potencial para la extracción de conocimiento.
- Centros educativos: Se podrían utilizar estos datos para la creación de modelos predictivos que permitan detectar a los alumnos en riesgo para poder aplicar distintos protocolos de actuación que ayuden a mejorar su rendimiento escolar evitando el fracaso del alumno (y por lo tanto también del centro escolar). Esto puede repercutir directamente en su reputación y en su dotación económica.
- Grupo Santillana: Obviamente la comercialización de una aplicación que permita predecir en tiempo real el riesgo de fracaso escolar del alumno tendría importantes beneficios económicos para la empresa.

Una vez construido el dataset se realizarán una serie de experimentos para determinar si podemos extraer reglas de clasificación que permitan al personal docente detectar a los alumnos en riesgo. Se probarán varios algoritmos de estado del arte basados en reglas o árboles. Se buscará encontrar el algoritmo que dé mejores resultados y cuya interpretación por parte del personal docente sea sencilla.

Para determinar el algoritmo que dará mejores resultados se tendrán en cuenta varios factores:

- Precisión de la predicción
- Precisión de la predicción teniendo en cuenta sólo los alumnos en riesgo (normalmente dado por la especificidad).
- Detectar lo más temprano posible a los estudiantes que están en riesgo.
- Facilidad de interpretación del modelo predictivo.

1.1. Estructura del documento

El documento tendrá la siguiente estructura:

- Introducción: Breve descripción del proyecto. Contiene los siguientes puntos:
  - Descripción del problema. Identificamos las necesidades y las dificultades que nos encontramos para satisfacer dichas necesidades. Principalmente están relacionadas con la obtención e identificación de los campos a utilizar en el dataset.
  - Objetivos: Se establece el objetivo principal del proyecto y los subobjetivos que se deberán cumplir para la consecución del objetivo principal.
- Antecedentes: Revisamos el problema del fracaso escolar y la utilización de EDM en la predicción de los alumnos en riesgo de fracasar.
- Descripción del dataset: En este punto explicamos las bases de datos existentes, los pasos seguidos para la creación del dataset y finalmente definimos las variables utilizadas mediante la creación de un diccionario de datos.
- Experimentación: Definimos los experimentos a realizar, detallamos los algoritmos creados para su ejecución y realizamos las pruebas.
- Conclusiones: Se detallan las conclusiones obtenidas en la experimentación.
- Trabajos futuros: Se comentan posibles experimentos que se podrían realizar con el nuevo dataset.
- Bibliografía y Anexos: Mostramos los documentos consultados para la realización de este proyecto.

1.2. Descripción del problema

El fracaso escolar es un grave problema que puede condicionar la vida de muchas personas. Existen muchos motivos que pueden derivar en un menor rendimiento del alumno. Estos pueden depender del alumno, del centro formativo, del entorno sociocultural, etc.

Muchas veces el fracaso escolar se puede evitar si el problema se detecta a tiempo mediante la aplicación de distintas medidas correctoras.

SOLUCIÓN PROPUESTA:

Existen técnicas de EDM que nos permiten predecir qué alumnos están en riesgo de sufrir fracaso escolar en base a la información que tenemos de los mismos.

Se propone la creación de modelos predictivos para la detección del fracaso escolar basado en el conocimiento extraído de las bases de datos del Grupo Santillana.

Estos modelos deberán ser fácilmente interpretables por el personal docente para poder utilizarlos.

Para ello se deberá crear un conjunto de datos, determinar qué algoritmo de clasificación es el más apropiado (basándonos principalmente en la precisión y precocidad de la predicción) y facilitar las herramientas necesarias para que personal de desarrollo no experto en extracción de conocimiento puedan incorporarlo a las aplicaciones del grupo Santillana.

1.3. Objetivos

Objetivo: Poder predecir si un alumno va a suspender una determinada asignatura.

Para conseguir el objetivo se deben de cumplir una serie de subobjetivos:

1. Creación de un dataset.
2. Selección de varios colegios con datos suficientes para hacer un estudio.
3. Selección del mejor algoritmo para predecir con este dataset.
4. Facilitar la generación automática de modelos predictivos por parte del personal de desarrollo no experto en data mining.

2 Antecedentes

El fracaso escolar es un tema de gran relevancia que repercute en el desarrollo y bienestar de cualquier sociedad. La falta de acceso a una educación de calidad y un bajo nivel educativo de la población puede limitar el desarrollo de la misma. Debido a esto, se buscan soluciones a este problema en muchos ámbitos (gobiernos, centros educativos, etc).

En todos los centros educativos existen alumnos que pueden ser salvados del fracaso escolar aplicando diversos programas de recuperación como pueden ser orientación al alumno, apoyo familiar o asignación de tutores.

Es importante determinar cuales son estos alumnos dado que los recursos que se pueden dedicar a este tipo de acciones son limitados y es necesario orientarlo a los alumnos que realmente puedan aprovecharlos.

Evitar el fracaso escolar es una tarea compleja dado que es un problema que depende de muchos factores llegando a denominarse "el problema de las mil causas" [7].

Por otra parte, el fracaso escolar puede entenderse de distintas formas. Algunas de ellas serían:

- No obtener titulación. El alumno abandona los estudios antes de terminar [8][9].
- No pasar un curso. El alumno no supera los requisitos mínimos para pasar al siguiente curso y debe repetirlo [10][11].
- No aprobar una asignatura. El alumno no obtiene la nota mínima para finalizar la asignatura. Este es el caso que estudiaremos en nuestro proyecto [12].

EDM (Educational Data Mining) es una disciplina emergente que aplica técnicas de minería de datos (DM) sobre información extraída de los estudiantes y su entorno educativo con el fin de comprender mejor a los estudiantes y los métodos de aprendizaje [1][2][6].

Una de los principales objetivos de EDM es la mejora del rendimiento académico [5]. Esto se puede conseguir mediante la predicción temprana de alumnos en peligro de fracaso [13][8] [6] o detectando los factores que derivan en fracaso escolar para intentar evitarlos [14]. Nosotros nos centraremos en el primero.

El desarrollo de los sistemas de gestión de aprendizaje (LMS) ha facilitado la recopilación de datos relacionados con los estudiantes y sus hábitos de estudios [1]. En la actualidad hay un creciente interés en los factores que predicen el rendimiento de los estudiantes. Esto es aún más importante en la educación a distancia donde no hay un contacto directo con los alumnos y el número de estos suele ser superior, por lo que es más difícil conocer los hábitos y circunstancias de cada alumno. En la mayoría de los casos aplicamos EDM sobre conjuntos de datos obtenidos de un LMS. Romero y Ventura proponen un caso de estudio y tutorial sobre data mining aplicado a un curso de MOODLE que nos servirá de guía en este proyecto [4].

No existe un concenso sobre cual es el mejor método o algoritmo para predecir a los alumnos en riesgo. Existen muchos artículos al respecto, cada uno con su propia teoría, que en ocasiones son contrarias entre sí. Cada cual defiende su postura. Sin embargo, no se puede definir un algoritmo como el más adecuado por diversos motivos:

- Existen diversos niveles (universitario, primaria, secundaria, etc) con distinta distribución de los datos. La mayoría de los trabajos de investigación relacionados con este tema tratan el nivel de educación superior o universitaria existiendo pocos estudios sobre los niveles de educación básica y media [13][14].

- Se debe tener en cuenta el concepto de predicción temprana (normalmente es más importante predecir antes, aunque se pierda precisión). Sin embargo, no podemos hacerlo hasta que no se ha conseguido suficiente información para conseguir una buena clasificación y, por lo tanto, poder determinar que alumnos podrían estar en riesgo. Esto puede derivar en una situación en la que sea demasiado tarde para ayudar a estos alumnos a evitar el fracaso.

  Jiménez, Luna y Ventura tratan de predecir lo más temprano posible a los estudiantes que fracasaran en un centro escolar de educación secundaria en España. Para ello utilizan varios algoritmos de clasificación para poder ayudar cuanto antes a los alumnos en riesgo [15].

- Los datos suelen estar desbalanceados (mayor número de aprobados que de suspensos, es decir, el número de instancias de una clase es muy superior al número de instancias de la otra clase). Los algoritmos tradicionales de clasificación están diseñados para obtener la máxima precisión del modelo, pero cuando tenemos datos desbalanceados, la precisión no debe ser la variable a optimizar. Por ejemplo, en primaria donde el 99% de los alumnos aprueban, podríamos tener un modelo que acierte un 99% porque detecta a los alumnos que aprobarán, pero podría no estar acertando los que están en riesgo de fracaso (que son en realidad los que nos interesan).
  Este desbalanceo se dá aún más en las primeras etapas del aprendizaje (nuestro dominio de estudio) que en la etapa universitaria. Por lo tanto, la clase minoritaria puede ser fácilmente ignorada en los modelos. Para evitarlo tendremos que buscar algoritmos que sean capaces de clasificar de forma adecuada a la clase minoritaria (en nuestro caso los alumnos en riesgo de fracaso escolar) dándole más importancia a otras medidas (por ejemplo, la especificidad que nos indicaría el porcentaje de aciertos de la clase minoritaria).

3 Descripción del dataset

En este capítulo vamos a describir el proceso de creación del dataset y definiremos los campos obtenidos. Comenzaremos con una breve explicación de las bases de datos utilizadas.

3.1 Bases de datos existentes

Vamos a trabajar con dos fuentes de datos distintas:

- LMS: Contiene información del uso del sistema por parte del alumno (por ejemplo, accesos al sistema o a la aplicación y participación en foros) e información introducida en el sistema por el personal docente (por ejemplo, notas de actividades, control de asistencias y la más importante: la nota final). Utiliza MS SQL Server como sistema gestor de base de datos.
- PLENO: Contiene información sobre los test de conocimientos realizados por los alumnos. Guarda información como la nota del test y el tiempo que se tarda en responder. Utiliza MySQL como sistema gestor de base de datos.

Sobre estas bases de datos tenemos 3 vistas:

- DATAMART: Muestra el contenido de consultas realizadas sobre la base de datos del LMS. Montado en MS SQL Server.
- PLENO: Muestra el contenido de consultas realizadas sobre la base de datos PLENO. Montado en MySQL.
- EDMUCO: Muestra el contenido de consultas realizadas sobre las dos vistas anteriores, relacionando en algunos casos consultas de ambas vistas. Montado en MySQL.

No tenemos acceso a estas bases de datos. Tan solo tenemos acceso a las tres vistas.

Uno de los principales problemas que nos encontramos es que, Al utilizar dos sistemas gestores de bases de datos (mySQL y MS SQL Server) no es posible realizar consultas que relacionen tablas de las dos bases de datos.

A continuación, vamos a mostrar un gráfico para clarificar la relación entre las bases de datos y las vistas:

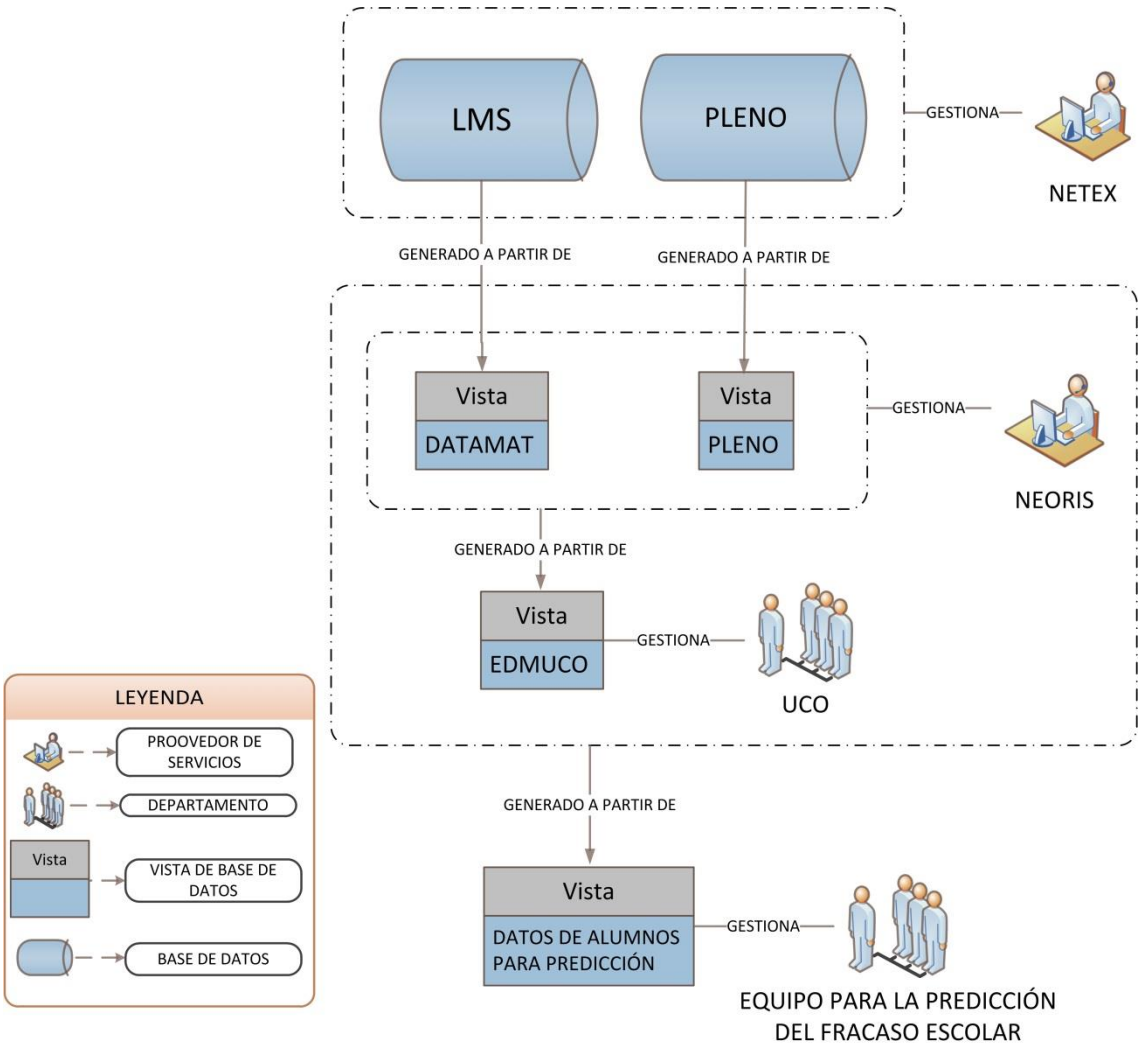

Figura 1. Origen de los datos. BBDD, vistas y gestores.

Para obtener mayor información sobre el funcionamiento de las bases de datos se aconseja consultar el documento "Definición funcional de los paquetes SSIS implementados en el proyecto SantillanaNeo de la plataforma Santillana" desarrollado por Neoris.

3.2 Diccionario de datos

En este apartado vamos a detallar los campos (variables) incluidos en el dataset utilizado para la predicción de las notas de clase.

3.2.1 Estructura base

La estructura base se hereda de Datamart. Es importante conocerla para poder entender el conjunto de datos y las relaciones entre las distintas bases de datos. Pleno tiene una estructura distinta, pero se le han incluido los campos de esta estructura base para poder relacionarlas.

Los datos están agrupados en varios niveles. En primer lugar. vamos a mostrarlos para facilitar el seguimiento de la división de los datos:

Figura 2. Niveles de agrupación de los datos

La primera división que nos encontramos es el campo Empresa. Una empresa es una entidad que comercializa el software en un determinado país. En un mismo país pueden operar más de una empresa cada una con un sistema diferente (por ejemplo, Uno y Compartir). Algunas de las empresas que podemos encontrar son las siguientes: *UNO México, Santillana Compartir México, UNO Colombia, Santillana Compartir Colombia…*

Un ciclo es un periodo de tiempo que identifica una campaña escolar. Una empresa puede tener varios ciclos, pero un ciclo es único para una empresa. En la siguiente tabla podemos ver un ejemplo:

Empresa	CicloId	Ciclo
UNO Mexico	20	2012/2013
	23	2013/2014
	24	2014/2015
Santillana Compartir Mexico	29	2012/2013
	65	2013/2014
	66	2014/2015
UNO Colombia	34	A-2013
	55	A-2014
	56	A-2015
	70	B-2013
	71	B-2014
	72	B-2015
Santillana Compartir Colombia	50	A-2013
	60	A-2014
	61	A-2015
	77	B-2013
	78	B-2014
	79	B-2015
...	...	...

Tabla 1. Ejemplo relación Empresa-Ciclo

A continuación, se encuentran los colegios. En realidad, no es un nivel inferior sino que estarían la mismo nivel que los ciclos. Cada colegio pertenece a una sola empresa pero la relación con los ciclos es N:N (es decir, un colegio tiene varios ciclos y un ciclo afecta a varios colegios). Esto se puede observar en el siguiente ejemplo:

Empresa	CicloId	Ciclo	Colegio
UNO Mexico	20	2012/2013	Colegio1
	23	2013/2014	Colegio1
			Colegio2
	24	2014/2015	Colegio1
			Colegio2
Santillana Compartir Mexico	29	2012/2013	Colegio3
			Colegio4
	65	2013/2014	Colegio3
			Colegio4
	66	2014/2015	Colegio3
			Colegio4
UNO Colombia	34	A-2013	Colegio5
	55	A-2014	Colegio5
	56	A-2015	Colegio5
	70	B-2013	Colegio6
	71	B-2014	Colegio6
	72	B-2015	Colegio6
Santillana Compartir Colombia	50	A-2013	Colegio7
			Colegio8
	60	A-2014	Colegio7
			Colegio8
	61	A-2015	Colegio7
			Colegio8
	77	B-2013	Colegio9
	78	B-2014	Colegio9
	79	B-2015	Colegio9
…	…	…	…

Tabla 2. Ejemplo relación Empresa-Ciclo-Colegio

Cada colegio puede tener diferentes niveles en cada ciclo. Algunos de los niveles que podemos encontrar pueden ser preescolar, primaria y secundaria. Podemos ver un ejemplo en la siguiente tabla.

Empresa	CicloId	Ciclo	Colegio	Nivel
UNO México	20	2012/2013	Colegio1	Primaria
				Secundaria
	23	2013/2014	Colegio1	Primaria
				Secundaria
			Colegio2	Primaria
	24	2014/2015	Colegio1	Primaria
				Secundaria
			Colegio2	Primaria
Santillana Compartir México	29	2012/2013	Colegio3	Primaria
			Colegio4	Primaria
	65	2013/2014	Colegio3	Primaria
			Colegio4	Primaria
	66	2014/2015	Colegio3	Primaria
			Colegio4	Primaria
UNO Colombia	34	A-2013	Colegio5	Primaria
				Secundaria
	55	A-2014	Colegio5	Primaria
				Secundaria
	56	A-2015	Colegio5	Primaria
				Secundaria
	70	B-2013	Colegio6	Primaria
	71	B-2014	Colegio6	Primaria
	72	B-2015	Colegio6	Primaria
Santillana Compartir Colombia	50	A-2013	Colegio7	Primaria
			Colegio8	Primaria
	60	A-2014	Colegio7	Primaria
			Colegio8	Primaria
	61	A-2015	Colegio7	Primaria
			Colegio8	Primaria
	77	B-2013	Colegio9	Primaria
	78	B-2014	Colegio9	Primaria
	79	B-2015	Colegio9	Primaria
…	…	…	…	

Tabla 3. Ejemplo relación Empresa-Ciclo-Colegio-Nivel

Finalmente, dentro de cada nivel pueden existir distintos grados. La siguiente tabla muestra un ejemplo:

Empresa	CicloId	Ciclo	Colegio	Nivel	Grado
UNO México	20	2012/2013	Colegio1	Primaria	1º Primaria
					2º Primaria
					…
				Secundaria	1º Secundaria
					2º Secundaria
					…
	23	2013/2014	Colegio1	Primaria	1º Primaria
					…
				Secundaria	1º Secundaria
					…
			Colegio2	Primaria	1º Primaria
					…
	24	2014/2015	Colegio1	Primaria	1º Primaria
					…
				Secundaria	1º Secundaria
					…
			Colegio2	Primaria	1º Primaria
					…
…	…	…	…	…	…

Tabla 4. Ejemplo relación Empresa-Ciclo-Colegio-Nivel-Grado

Por debajo de grado existen 2 niveles más (clase y materia). No profundizaremos en ellos ya que serán sustituidos por la variable "Area" como se explicará en puntos posteriores.

3.2.2 Consideraciones

Existen una serie de consideraciones a tener en cuenta para comprender mejor el conjunto de datos:

- Los datos de las pruebas que se han incluido de Pleno corresponden con las pruebas finalizadas (tenemos fecha de inicio y fecha de finalización de la prueba). Las pruebas que un alumno ha empezado, pero no ha finalizado no se han tenido en cuenta.

- Cada instancia podrá ser identificada unívocamente por los campos AlumnoID, ClaseID, CicloId, Subciclo e IdPrueba. En caso de que el alumno no haya realizado ninguna prueba, cada instancia estará identificada unívocamente por los campos AlumnosID, ClaseID, CicloId y Subciclo (IdPrueba no tendrá valor).

- Al introducir una instancia por cada prueba realizada podemos estudiar la posible relación entre las notas de las distintas pruebas con la nota final. Esto puede ser interesante para detectar distintos factores dentro de las pruebas.

- Los nombres de los campos comenzarán por un acrónimo que indicará la procedencia del campo. Estos acrónimos serán:
  - PL: Si vienen de PLENO.
  - LMS: Si vienen de datamart (LMS).

- Para la predicción temprana al inicio del curso (cuando aún no se tienen datos de ese ciclo) vamos a generar varios campos basados en las notas de cursos anteriores. Se ha decidido que las áreas más influyentes serán matemáticas, lenguaje y la de la clase a estudiar (por ejemplo, si la clase corresponde al área de las Ciencias Naturales, esta área será más influyente que otras). Estos campos son: "CicloAnterior", "NotaCicloAntArea", "NotaCicloAntLeng", "NotaCicloAntMate", "MismoNivelCicloAnt" y "Repetidor".

3.2.3 Campos (variables) utilizados para predicción

En la siguiente tabla se muestra un resumen de las variables utilizadas y su origen para facilitar la visión global de la tabla de datos. Posteriormente se detallarán cada uno de los campos.

Atributo	**Origen**
LMSEmpresaId	Datamart – Múltiples tablas
LMSEmpresa	Datamart – Múltiples tablas
LMSColegioID	Datamart – Múltiples tablas
LMSNombreColegio	Datamart – Múltiples tablas
LMSCicloId	Datamart – Múltiples tablas
LMSCiclo	Datamart – Múltiples tablas
LMSSubciclo	EDMUCO – EDMCicloSubciclo
LMSNivelId	Datamart – Múltiples tablas
LMSNivel	Datamart – Múltiples tablas
LMSGradoId	Datamart – Múltiples tablas
LMSGrado	Datamart – Múltiples tablas
LMSClaseId	Datamart – Múltiples tablas
LMSClase	Datamart – Múltiples tablas
LMSArea	Datamart – Múltiples tablas
LMSAlumnoId	Datamart – Múltiples tablas
LMSNSesionesTotalClases	Datamart - tablas “FACTSesionLogin_H” y “Sesion_Antigua”
LMSNAccesosAPPTotalClases	Datamart - tabla “FACTAccesosAPP”
LMSNAsistencias	Datamart - tabla “FACTAsistenciasNAEDM”
LMSPAsistencias	Datamart - tabla “FACTAsistenciasNAEDM”
LMSNRetardos	Datamart - tabla “FACTRetardosEDM”
LMSPRetardos	Datamart - tabla “FACTFaltasEDM”
LMSNFaltas	Datamart - tabla “FACTFaltasEDM”
LMSPFaltas	Datamart - tabla “FACTFaltasEDM”
LMSNFaltasJust	Datamart - tabla “FACTFaltasJustificadasEDM”
LMSPFaltasJust	Datamart - tabla “FACTFaltasJustificadasEDM”
LMSMensajesEnviados	Datamart - tabla “FACTMensajes_H”
LMSMensajesRecibidos	Datamart - tabla “FACTMensajes_H”
LMSNTareas	Datamart - tabla “FACTNotasActividadEDM”
LMSNotaTareas	Datamart - tabla “FACTNotasActividadEDM”
LMSPTareas	Datamart - tabla “FACTNotasActividadEDM”
LMSPNotasTareas	Datamart - tabla “FACTNotasActividadEDM”
LMSNExamenes	Datamart - tabla “FACTNotasActividadEDM”
LMSNotaExamenes	Datamart - tabla “FACTNotasActividadEDM”
LMSPExamenes	Datamart - tabla “FACTNotasActividadEDM”
LMSPNotasExamenes	Datamart - tabla “FACTNotasActividadEDM”
LMSNPostForos	Datamart - tabla “FACTForosPost”
LMSCicloAnterior	Datamart - tabla “FACTForosPost”
LMSNotaCicloAntArea	Datamart – Múltiples tablas
LMSNotaCicloAntLeng	Datamart – Múltiples tablas
LMSNotaCicloAntMate	Datamart – Múltiples tablas
LMSMismoNivelCicloAnt	Datamart – Múltiples tablas
LMSRepetidor	Datamart – Múltiples tablas
PLPruebaId	Pleno – Múltiples tablas
PLTiempoRespuesta	Datamart - tabla “preguntas_prueba_contestada”
PLTestSuperado	Datamart - tablas “prueba_contestada” y “pruebas”
PLPRespuestasCorrectas	Datamart - tabla “prueba_contestada”
PLPNecesarioParaLogro	Datamart - tabla “pruebas”
Nota	Datamart - tabla “FACTNotasClaseEDM”

Tabla 5. Origen de los atributos

3.2.3.1 LMSEmpresaId

- **Definición**: Identificador de la empresa que comercializa la aplicación. Cada país está abastecido por una o más empresas (normalmente una o dos) y cada empresa opera en un único país. El funcionamiento puede variar entre dos empresas.
- **Origen**: Datamart – Múltiples tablas.
- **Uso**: Selección.

3.2.3.2 LMSEmpresa

- **Definición**: Nombre de la empresa que comercializa la aplicación.
- **Origen**: Datamart – Múltiples tablas.
- **Uso**: Selección.

3.2.3.3 LMSColegioID

- **Definición**: Identificador numérico unívoco de un colegio.
- **Origen**: Datamart – Múltiples tablas.
- **Uso**: Selección.

3.2.3.4 LMSNombreColegio

- **Definición**: Nombre de un colegio.
- **Origen**: Datamart – Múltiples tablas.
- **Uso**: Selección.

3.2.3.5 LMSCicloId

- **Definición**: Identificador de un periodo de tiempo entre dos fechas equivalente a un curso escolar.
- **Origen**: Datamart – Múltiples tablas.
- **Uso**: Selección.

3.2.3.6 LMSCiclo

- **Definición**: Nombre asignado al periodo de tiempo entre dos fechas equivalente a un curso escolar.
- **Origen**: Datamart – Múltiples tablas.
- **Uso**: Selección.

3.2.3.7 LMSSubciclo

- **Definición**: Distintas particiones de un ciclo. En principio se creará un subciclo denominado Subciclo0 que indica el inicio del curso y comprende tan sólo ese día. A continuación, se dividirá el ciclo en 6 subciclos en el que cada subciclo contendrá los subciclos anteriores más un intervalo de tiempo de igual tamaño que el primer subciclo (mismo número de días con una posible variación mínima debida el redondeo). Por lo tanto, siendo "tInicio" la fecha de inicio del ciclo, "tFin" la fecha en que termina el ciclo y "t" el número de días comprendido entre "tInicio" y "tFin" dividido entre el número de subciclos ((tFin-tInicio)/6), los subciclos comprenderán los siguientes intervalos de tiempo:
  - Subciclo0: [tInicio, tInicio].
  - Subciclo1: [tInicio, tInicio + t].
  - Subciclo2: [tInicio, tInicio + 2t].
  - Subciclo3: [tInicio, tInicio + 3t].
  - Subciclo4: [tInicio, tInicio + 4t].
  - Subciclo5: [tInicio, tInicio + 5t].
  - Subciclo6: [tInicio, tFin] (coincide con el ciclo).

- **Origen**: EDMUCO – EDMCicloSubciclo.
- **Uso**: Selección.

3.2.3.8 LMSNivelId

- **Definición**: Identificador del nivel educativo. No está relacionado con el campo NombreNivel (Pleno). Estos identificadores pueden variar dependiendo de la empresa distribuidora (mismo nombre de nivel pero distinto NivelId).
- **Origen**: Datamart – Múltiples tablas.
- **Uso**: Selección.

3.2.3.9 LMSNivel

- **Definición**: Nombre del nivel educativo identificado en el campo NivelId. No está relacionado con el campo NombreNivel (Pleno). Ejemplos de niveles educativos son bachillerato, preescolar o primaria.
- **Origen**: Datamart – Múltiples tablas.
- **Uso**: Selección.

3.2.3.10 LMSGradoId

- **Definición**: Identificador del grado. Subdivide un nivel. Estos identificadores pueden variar dependiendo de la empresa distribuidora (mismo nombre de grado, pero distinto GradoId).
- **Origen**: Datamart – Múltiples tablas.
- **Uso**: Selección.

3.2.3.11 LMSGrado

- **Definición**: Nombre del grado. Ejemplos de grados son 1º de bachillerato, kinder o 5º de primaria.
- **Origen**: Datamart – Múltiples tablas.
- **Uso**: Selección.

3.2.3.12 LMSClaseId

- **Definición**: Identificador numérico unívoco de la clase. Una clase es un grupo de alumnos que cursa una o varias materias comunes.
- **Origen**: Datamart – Múltiples tablas.
- **Uso**: Selección.

3.2.3.13 LMSClase

- **Definición**: Nombre de la clase. Ejemplos de clase son Ciencias Sociales 8º B o Educación Física 6º A.
- **Origen**: Datamart – Múltiples tablas.
- **Uso**: Selección.

3.2.3.14 LMSArea

- **Definición**: Area temática a la que corresponden los contenidos impartidos en una clase. Ejemplos de Area pueden ser "Matemáticas", "Lenguaje" o "Ciencias Naturales". Se obtiene mediante la ejecución del script "AgrupacionAreas.R" mapeando los nombres de las clases y los nombres de las materias que se imparten en dicha clase. Se utilizará para agrupar clases con contenidos similares.
- **Origen**: Datamart – Múltiples tablas.
- **Uso**: Selección o predicción. Dependerá del nivel de profundidad en el que hagamos la predicción (por nivel, por grado o por área). Si profundizamos a nivel de área se utilizará para selección. En otro caso se utilizará como campo en la predicción.
- **Relación**: Cada clase estará vinculada a un área (aunque podría estar relacionada con más de una por los contenidos. El algoritmo seleccionará la más importante).

3.2.3.15 LMSAlumnoId

- **Definición**: Identificador numérico unívoco del alumno.
- **Origen**: Datamart – Múltiples tablas.
- **Uso**: Selección.

3.2.3.16 LMSNSesionesTotalClases

- **Definición**: Número de sesiones abiertas (accesos al LMS) por el usuario a lo largo del subciclo.
- **Origen**: Datamart - tablas "FACTSesionLogin_H" y "Sesion_Antigua".
- **Uso**: Predicción
- **Relación**: Un valor por cada Alumno/Ciclo/Subciclo.
- **Inconvenientes**
    - Al no tenerlo dividido por clases, en cada clase tendremos la suma de las sesiones abiertas por el usuario en todas las clases.

3.2.3.17 LMSNAccesosAPPTotalClases

- **Definición**: Número accesos a la aplicación (App) realizados por el usuario a lo largo del subciclo.
- **Origen**: Datamart - tabla "FACTAccesosAPP".
- **Uso**: Predicción
- **Relación**: Un valor por cada Alumno/Ciclo/Subciclo.
- **Inconvenientes**
  - Al no tenerlo dividido por clases, en cada clase tendremos la suma de los accesos establecidos por el usuario en todas las clases.

3.2.3.18 LMSNAsistencias

- **Definición**: Indica el número de días que el alumno ha asistido a clase.
- **Origen**: Datamart - tabla "FACTAsistenciasNAEDM".
- **Uso**: Predicción.
- **Relación**: Un valor por cada Alumno/Clase/Ciclo/Subciclo.

3.2.3.19 LMSPAsistencias

- **Definición**: Indica el porcentaje de días que el alumno ha asistido a clase. Se calcula dividiendo el número de días que ha asistido entre el número de días que debería haber asistido (calculado como la suma de asistencias, retardos, faltas y faltas justificadas).
- **Origen**: Datamart - tabla "FACTAsistenciasNAEDM".
- **Uso**: Predicción.
- **Relación**: Un valor por cada Alumno/Clase/Ciclo/Subciclo.

3.2.3.20 LMSNRetardos

- **Definición**: Indica el número de días que el alumno ha asistido a clase.
- **Origen**: Datamart - tabla "FACTRetardosEDM".
- **Uso**: Predicción.
- **Relación**: Un valor por cada Alumno/Clase/Ciclo/Subciclo.

3.2.3.21 LMSPRetardos

- **Definición**: Indica el porcentaje de días que el alumno ha llegado tarde a clase. Se calcula dividiendo el número de días que llego tarde entre el número de días que debería haber asistido (calculado como la suma de asistencias, retardos, faltas y faltas justificadas).
- **Origen**: Datamart - tabla "FACTRetardosEDM".
- **Uso**: Predicción.
- **Relación**: Un valor por cada Alumno/Clase/Ciclo/Subciclo.

3.2.3.22 LMSNFaltas

- **Definición**: Indica el número de días que el alumno ha faltado a clase sin entregar un justificante.
- **Origen**: Datamart - tabla "FACTFaltasEDM".
- **Uso**: Predicción.
- **Relación**: Un valor por cada Alumno/Clase/Ciclo/Subciclo.

3.2.3.23 LMSPFaltas

- **Definición**: Indica el porcentaje de días que el alumno ha faltado a clase sin entregar un justificante. Se calcula dividiendo el número de días que falto sin justificar entre el número de días que debería haber asistido (calculado como la suma de asistencias, retardos, faltas y faltas justificadas).
- **Origen**: Datamart - tabla "FACTFaltasEDM".
- **Uso**: Predicción.
- **Relación**: Un valor por cada Alumno/Clase/Ciclo/Subciclo.

3.2.3.24 LMSNFaltasJust

- **Definición**: Indica el número de días que el alumno ha faltado a clase justificando la falta.
- **Origen**: Datamart - tabla "FACTFaltasJustificadasEDM".
- **Uso**: Predicción.

3.2.3.25 LMSPFaltasJust

- **Definición**: Indica el porcentaje de días que el alumno ha faltado a clase entregando un justificante. Se calcula dividiendo el número de días que falto de forma justificada entre el número de días que debería haber asistido (calculado como la suma de asistencias, retardos, faltas y faltas justificadas).
- **Origen**: Datamart - tabla "FACTFaltasJustificadasEDM".
- **Uso**: Predicción.
- **Relación**: Un valor por cada Alumno/Clase/Ciclo/Subciclo.

3.2.3.26 LMSMensajesEnviados

- **Definición**: Indica el número de mensajes enviados por un alumno a lo largo de un ciclo.
- **Origen**: Datamart - tabla "FACTMensajes_H".
- **Uso**: Predicción.
- **Relación**: Un valor por cada Alumno/Ciclo.
- **Inconvenientes**
  - Al no tenerlo dividido por clases, en cada clase tendremos la suma de los mensajes enviados por el usuario en todas las clases.
  - Al no estar dividido en subciclos, tan solo tendremos un valor para el último subciclo (que coincide con el ciclo). El problema se debe a que la tabla que utilizamos para obtener este campo presenta los valores finales agrupados (es decir, el número total mensajes enviados al final de curso). Si fuera posible acceder a su origen y obtener la fecha en que se envían, se podría solucionar este problema y dividirlo en subciclos.
  - Actualmente hay un error en la consulta por lo que no se puede incluir este campo.

3.2.3.27 LMSMensajesRecibidos

- **Definición**: Indica el número de mensajes recibidos por un alumno a lo largo de un ciclo.
- **Origen**: Datamart - tabla "FACTMensajes_H".
- **Uso**: Predicción.
- **Relación**: Un valor por cada Alumno/Ciclo.
- **Inconvenientes**
  - Al no tenerlo dividido por clases, en cada clase tendremos la suma de los mensajes recibidos por el usuario en todas las clases.
  - Al no estar dividido en subciclos, tan solo tendremos un valor para el último subciclo (que coincide con el ciclo). El problema se debe a que la tabla que utilizamos para obtener este campo presenta los valores finales agrupados (es decir, el número total mensajes recibidos al final de curso). Si fuera posible acceder a su origen y obtener la fecha en que se reciben, se podría solucionar este problema y dividirlo en subciclos.
  - Actualmente hay un error en la consulta por lo que no se puede incluir este campo.

3.2.3.28 LMSNTareas

- **Definición**: Indica el número de actividades del LMS realizadas por el alumno a lo largo de un subciclo en una clase.
- **Origen**: Datamart - tabla "FACTNotasActividadEDM".
- **Uso**: Predicción.
- **Relación**: Un valor por cada Alumno/Clase/Ciclo/Subciclo.
- **Inconvenientes**
  - No contamos con la fecha de entrega de las actividades. Contamos con la fecha límite de entrega por lo que se está utilizando ésta para estimar el subciclo en el que se realiza la tarea.

3.2.3.29 LMSNotaTareas

- **Definición**: Indica la media de las notas obtenidas en las actividades del LMS realizadas a lo largo del subciclo en una clase. (SumaDeNotas/NTareas)
- **Origen**: Datamart - tabla "FACTNotasActividadEDM".
- **Uso**: Predicción.
- **Relación**: Un valor por cada Alumno/Clase/Ciclo/Subciclo.
- **Inconvenientes**
  - No contamos con la fecha de entrega de las actividades. Contamos con la fecha límite de entrega por lo que se está utilizando esta para estimar el subciclo donde se realiza la tarea.

3.2.3.30 LMSPTareas

- **Definición**: Indica el porcentaje de tareas realizadas (NTareas/NTareasMax) respecto al número de tareas que debería haber realizado. Se debe tener en cuenta que las tareas no realizadas podrían ser optativas. NTareasMax es el número total de actividades que puede realizar un alumno. Al no contar con este dato utilizamos para su estimación el número máximo de tareas que ha realizado un alumno en esa clase y subciclo (por ejemplo, si este alumno ha realizado 8 tareas y otro alumno en la misma clase y subciclo ha realizado 10 tareas tendremos: PTareas=(8/10)*100).
- **Origen**: Datamart - tabla "FACTNotasActividadEDM".
- **Uso**: Predicción.
- **Relación**: Un valor por cada Alumno/Clase/Ciclo/Subciclo.
- **Inconvenientes**
  - No contamos con la fecha de entrega de las actividades. Contamos con la fecha límite de entrega por lo que se está utilizando ésta para estimar el subciclo en el que se realiza la tarea.
  - No contamos con número máximo de tareas. Lo estimamos calculando el número máximo de tareas realizadas por un alumno en dicha clase y subciclo.

3.2.3.31 LMSPNotasTareas

- **Definición**: Indica la media de las notas obtenidas en las actividades del LMS realizadas a lo largo del subciclo en una clase dividida entre el total de tareas que podría haber realizado. (SumaDeNotas/PTareas)
- **Origen**: Datamart - tabla "FACTNotasActividadEDM".
- **Uso**: Predicción.
- **Relación**: Un valor por cada Alumno/Clase/Ciclo/Subciclo.
- **Inconvenientes**
    - No contamos con la fecha de entrega de las actividades. Contamos con la fecha límite de entrega por lo que se está utilizando ésta para estimar el subciclo en el que se realiza la tarea.

3.2.3.32 LMSNExamenes

- **Definición**: Indica el número de exámenes realizados por el alumno a lo largo de un subciclo en una clase.
- **Origen**: Datamart - tabla "FACTNotasActividadEDM".
- **Uso**: Predicción.
- **Relación**: Un valor por cada Alumno/Clase/Ciclo/Subciclo.
- **Inconvenientes**
    - No contamos con la fecha de entrega de los exámenes. Contamos con la fecha límite de entrega por lo que se está utilizando ésta para estimar el subciclo en el que se realiza el examen.

3.2.3.33 LMSNotaExamenes

- **Definición**: Indica la media de las notas obtenidas en los exámenes realizados a lo largo del subciclo en una clase. (SumaDeNotas/NExámenes)
- **Origen**: Datamart - tabla "FACTNotasActividadEDM".
- **Uso**: Predicción.
- **Relación**: Un valor por cada Alumno/Clase/Ciclo/Subciclo.
- **Inconvenientes**
    - No contamos con la fecha de entrega de los exámenes. Contamos con la fecha límite de entrega por lo que se está utilizando ésta para estimar el subciclo en el que se realiza el examen.

3.2.3.34 LMSPExamenes

- **Definición**: Indica el porcentaje de exámenes realizados (NExamenes/NExamenesMax) respecto al número de exámenes que debería haber realizado. Se debe tener en cuenta que algunos exámenes no realizadas podrían ser optativos. NExamenesMax es el número total de exámenes que puede realizar un alumno. Al no contar con este dato utilizamos para su estimación el número máximo de exámenes que ha realizado un alumno en esa clase y subciclo (por ejemplo, si este alumno ha realizado 8 exámenes y otro alumno en la misma clase y subciclo ha realizado 10 exámenes tendremos: PExamenes = (8/10) * 100).
- **Origen**: Datamart - tabla "FACTNotasActividadEDM".
- **Uso**: Predicción.
- **Relación**: Un valor por cada Alumno/Clase/Ciclo/Subciclo.
- **Inconvenientes**
  - No contamos con la fecha de entrega de los exámenes. Contamos con la fecha límite de entrega por lo que se está utilizando ésta para estimar el subciclo en el que se realiza el examen.
  - No contamos con número máximo de exámenes. Lo estimamos calculando el número máximo de exámenes realizados por un alumno en dicha clase y subciclo.

3.2.3.35 LMSPNotasExamenes

- **Definición**: Indica la media de las notas obtenidas en los exámenes realizados a lo largo del subciclo en una clase dividida entre el total de exámenes que podría haber realizado. (SumaDeNotas/PExamenes)
- **Origen**: Datamart - tabla "FACTNotasActividadEDM".
- **Uso**: Predicción.
- **Relación**: Un valor por cada Alumno/Clase/Ciclo/Subciclo.
- **Inconvenientes**
  - No contamos con la fecha de entrega de los exámenes. Contamos con la fecha límite de entrega por lo que se está utilizando ésta para estimar el subciclo en el que se realiza el examen.

3.2.3.36 LMSNPostForos

- **Definición**: Indica el número de mensajes publicados en los foros por el alumno en una determinada clase y subciclo.
- **Origen**: Datamart - tabla "FACTForosPost".
- **Uso**: Predicción.
- **Relación**: Un valor por cada Alumno/Clase/Ciclo/Subciclo.

3.2.3.37 LMSCicloAnterior

- **Definición**: Indica el identificador del ciclo que realizó el alumno en el curso anterior.
- **Origen**: Datamart - tabla "FACTForosPost".
- **Uso**: Cálculo de los campos "NotaCicloAntArea", "NotaCicloAntLeng", "NotaCicloAntMate", "MismoNivelCicloAnt" y "Repetidor".
- **Relación**: Un valor por cada Alumno/Clase/Ciclo.

3.2.3.38 LMSNotaCicloAntArea

- **Definición**: Indica la nota que obtuvo el alumno en el ciclo anterior en la misma área (por ejemplo, en una instancia de una clase con contenidos de matemáticas nos daría la nota que tuvo el alumno en una clase con estos contenidos en el ciclo anterior).
- **Origen**: Datamart – Múltiples tablas.
- **Uso**: predicción.
- **Relación**: Un valor por cada Alumno/Clase/Ciclo.

3.2.3.39 LMSNotaCicloAntLeng

- **Definición**: Indica la nota que obtuvo el alumno en el ciclo anterior en una clase correspondiente al área Lenguaje.
- **Origen**: Datamart – Múltiples tablas.
- **Uso**: predicción.
- **Relación**: Un valor por cada Alumno/Clase/Ciclo.

3.2.3.40 LMSNotaCicloAntMate

- **Definición**: Indica la nota que obtuvo el alumno en el ciclo anterior en una clase correspondiente al área Matemáticas.
- **Origen**: Datamart – Múltiples tablas.
- **Uso**: predicción.
- **Relación**: Un valor por cada Alumno/Clase/Ciclo.

3.2.3.41 LMSMismoNivelCicloAnt

- **Definición**: Indica si el alumno está en el mismo nivel que en el ciclo anterior. Por ejemplo, si este año está en secundario indicaría si el ciclo anterior estaba en secundaria o viene de primaria. Esto resulta útil debido a que un cambio de nivel puede influir en el rendimiento del alumno.
- **Origen**: Datamart – Múltiples tablas.
- **Uso**: predicción.
- **Relación**: Un valor por cada Alumno/Clase/Ciclo.

3.2.3.42 LMSRepetidor

- **Definición**: Indica si el alumno está cursando el mismo grado que el ciclo anterior (está repitiendo curso).
- **Origen**: Datamart – Múltiples tablas.
- **Uso**: predicción.
- **Relación**: Un valor por cada Alumno/Clase/Ciclo.

3.2.3.43 PLPruebaId

- **Definición**: Identificador numérico unívoco de una prueba de Pleno.
- **Origen**: Pleno – Múltiples tablas.
- **Uso**: Selección y predicción (podría utilizarse en la predicción para ver qué pruebas son relevantes en la nota final, cuales pueden dar resultados no esperados, etc).

3.2.3.44 PLTiempoRespuesta

- **Definición**: Tiempo que tarda un alumno en responder a una prueba. Se calcula sumando los tiempos que tarda el alumno en responder a cada pregunta.
- **Origen**: Datamart - tabla “preguntas_prueba_contestada”.
- **Uso**: Predicción.
- **Relación**: Un valor por cada prueba. Cada prueba está vinculado a Alumno/Clase/Subciclo.

3.2.3.45 PLTestSuperado

- **Definición**: Indica si la prueba ha sido superada (es decir, el porcentaje de respuestas acertadas en la prueba es mayor o igual al porcentaje necesario para lograr superarlo). Solo se han tenido en cuenta las pruebas finalizadas. Las pruebas que se han dejado a medias no se han incluido en la tabla.
- **Origen**: Datamart - tablas "prueba_contestada" y "pruebas".
- **Uso**: Predicción.
- **Relación**: Un valor por cada prueba. Cada prueba está vinculado a Alumno/Clase/Subciclo.

3.2.3.46 PLPRespuestasCorrectas

- **Definición**: Porcentaje de preguntas que se han respondido correctamente en una prueba.
- **Origen**: Datamart - tabla "prueba_contestada".
- **Uso**: Predicción.
- **Relación**: Un valor por cada prueba. Cada prueba está vinculado a Alumno/Clase/Subciclo.

3.2.3.47 PLPNecesarioParaLogro

- **Definición**: Porcentaje necesario para superar una determinada prueba.
- **Origen**: Datamart - tabla "pruebas".
- **Uso**: Predicción – Pendiente de evaluar. Podría eliminarse en versiones futuras.
- **Relación**: Un valor por cada prueba. Cada prueba está vinculado a Alumno/Clase/Subciclo.

3.2.3.48 Nota

- **Definición**: Nota final del curso. Este es el objetivo de la predicción.
- **Origen**: Datamart - tabla "FACTNotasClaseEDM".
- **Uso**: Predicción (Objetivo).
- **Relación**: Una nota por Alumno/Clase/Ciclo. Se repetirá la nota de un ciclo a lo largo de los subciclos.

3.3 Proceso de obtención del dataset

Para la creación del dataset vamos a realizar una serie de pasos:

1.- Creación del dataset base con los datos obtenidos del LMS y del PLENO.
2.- Creación de la variable Area.
3.- Creación de datos históricos.
4.- Transformación de datos.

Para ello utilizaremos una serie de scripts y funciones desarrolladas con el lenguaje de programación R.

3.3.1 Creación del dataset base con datos de LMS y PLENO

El siguiente algoritmo creará un data.frame con las variables que se pueden extraer directamente de las bases de datos. Esta tarea la hemos dividido en 14 pasos:

1. Cargar las librerías necesarias
2. Creación del dataset
3. Añadir los nombres correspondientes a los distintos identificadores
4. Creación la tabla subciclos
5. Añadir Accesos a la aplicación
6. Añadir mensajes enviados y recibidos
7. Añadir asistencias
8. Añadir retardos
9. Añadir faltas
10. Añadir faltas justificadas
11. Calculo de los porcentajes de asistencias, retardos y faltas
12. Añadir actividades de clase y exámenes
13. Añadir número de sesiones
14. Añadir participación en foros
15. Añadir campos de la base de datos de cuestionarios (Pleno)
16. Guardar los resultados en un archivo

3.3.1.1 Cargar las librerias necesarias

Tendremos que cargar las librerías RODBC y RMySQL para poder acceder a las bases de datos en MS SQL SERVER y MySQL respectivamente.

También utilizaremos las librerías dplyr y data.table para manipular de forma más eficiente los data frames.

3.3.1.2 Creación del dataset

Creamos el dataset a partir de la tabla que contiene las notas finales con las claves que identificarán cada registro (EmpresaId, CicloId, ColegioID, NivelId, GradoId, AlumnoId, ClaseId) y la variable objetivo (Nota).

También añadiremos la variable que contiene el nombre de la clase (clase). El resto de nombres correspondientes a los otros identificadores no se encuentran en esta tabla y se extraen en el siguiente punto.

3.3.1.3 Añadir los nombres correspondientes a los distintos identificadores

Estos estarán en distintas tablas de EDMUCO. Tendrán que extraerse individualmente y posteriormente fusionar los resultados con el dataset.

3.3.1.4 Creación la tabla subciclos

Cada ciclo lo vamos a dividir en 6 partes de igual duración a los que denominaremos subciclos. A continuación, crearemos una tabla con los ciclos existentes y sus subciclos correspondientes que denominaremos EDMCicloSubciclo.

La razón por la que se ha elegido hacer 6 divisiones es porque 6 es múltiplo de 2 y de 3, por lo que podremos cambiar de un modo de 6 divisiones a un modo de 2 o 3 divisiones sin tener que hacer ningún cambio.

Para entenderlo mejor definiremos lo siguiente:

- t=DuracionCiclo/Número de subciclos (sin contar el cero).
- SubcicloX es un intervalo de tiempo que empieza en Subciclo0 y termina en Subciclo0 + X*t donde X es el número del subciclo (Ej, fecha fin de Subciclo2 = Subciclo0 + 2*t).

Por lo tanto, siendo t un intervalo de tiempo cuya fórmula

Intervalo para 6 divisiones	6 divisiones	3 divisiones	2 divisiones
t=0	Subciclo0	Subciclo0	Subciclo0
t	Subciclo1		
t*2	Subciclo2	Subciclo1	
t*3	Subciclo3		Subciclo1
t*4	Subciclo4	Subciclo2	
t*5	Subciclo5		
t*6	Subciclo6	Subciclo3	Subciclo2

Tabla 6. Posibles divisiones de un ciclo en subciclos.

Si observamos la tabla, vemos que podemos cambiar de 6 divisiones a 3 omitiendo los subciclos 1, 3 y 5 y tomando el subciclo 2 como el 1, el subciclo 4 como el 2 y el subciclo 6 como el 3. De igual forma, si queremos cambiar a 2 divisiones lo haremos omitiendo los subciclos 1,2,4 y 5 y tomando el subciclo 3 como el 1 y el subciclo 6 como el 2.

A continuación, crearemos una tabla con los ciclos existentes y sus subciclos correspondientes que denominaremos EDMCicloSubciclo.

3.3.1.5 Añadir Accesos a la aplicación

Vamos a incluir en el dataset el número de accesos que realiza cada alumno a la aplicación del LMS. Esta información se guardará en el campo "NAccesosAPPTotalClases".

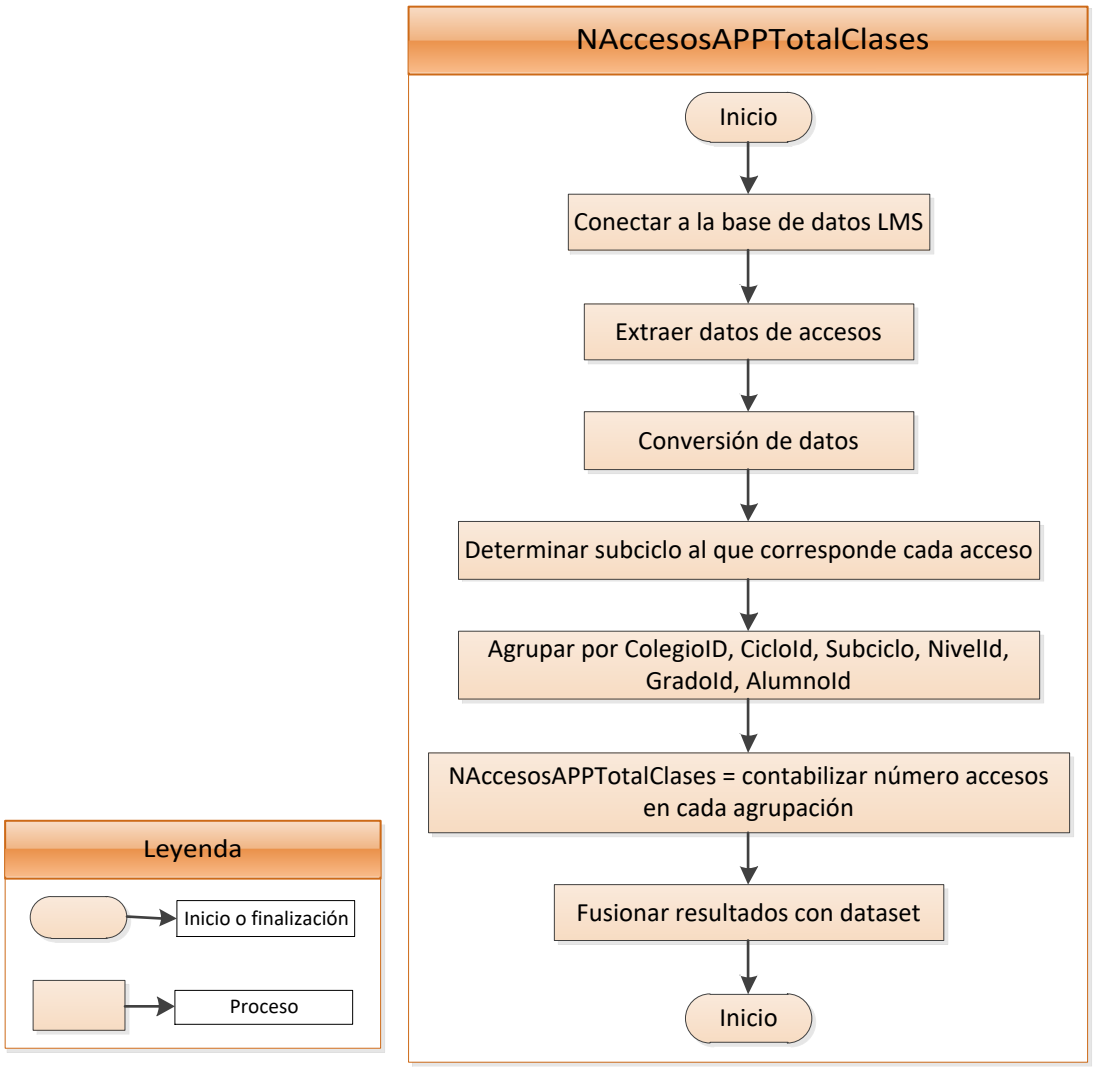

Figura 3. Proceso de creación del atributo "NAccesosAPPTotalClases".

3.3.1.6 Añadir mensajes enviados y recibidos

Vamos a incluir en el dataset el número mensajes que cada alumno ha enviado y recibido. Esta información se guardará en los campos "MensajesEnviados" y "MensajesRecibidos".

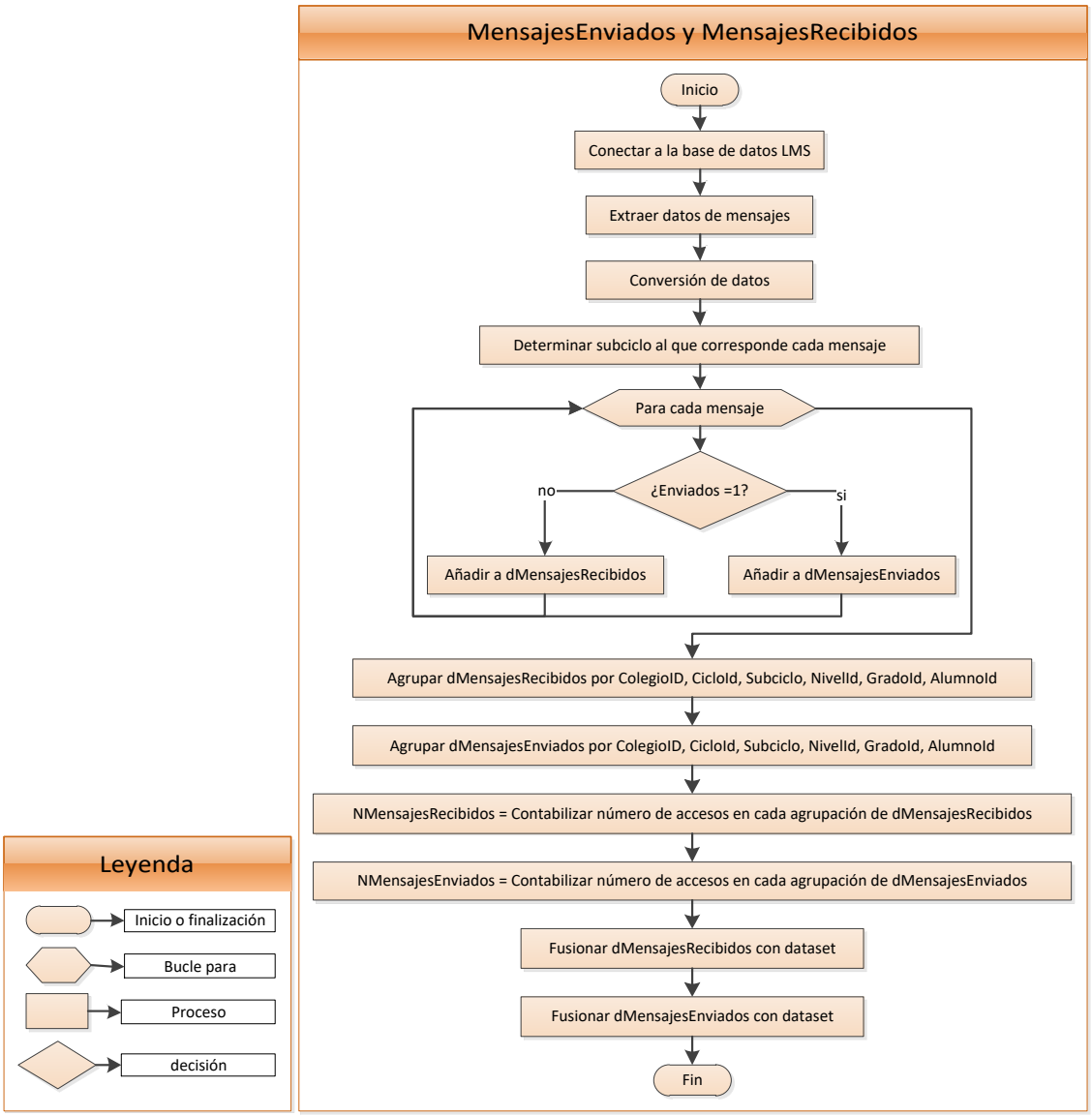

Figura 4. Proceso de creación de los atributos "MensajesEnviados" y "MensajesRecibidos".

3.3.1.7 Añadir asistencias

Vamos a incluir en el dataset el número de asistencias a clase de cada alumno. Esta información se guardará en el campo "NAsistencias".

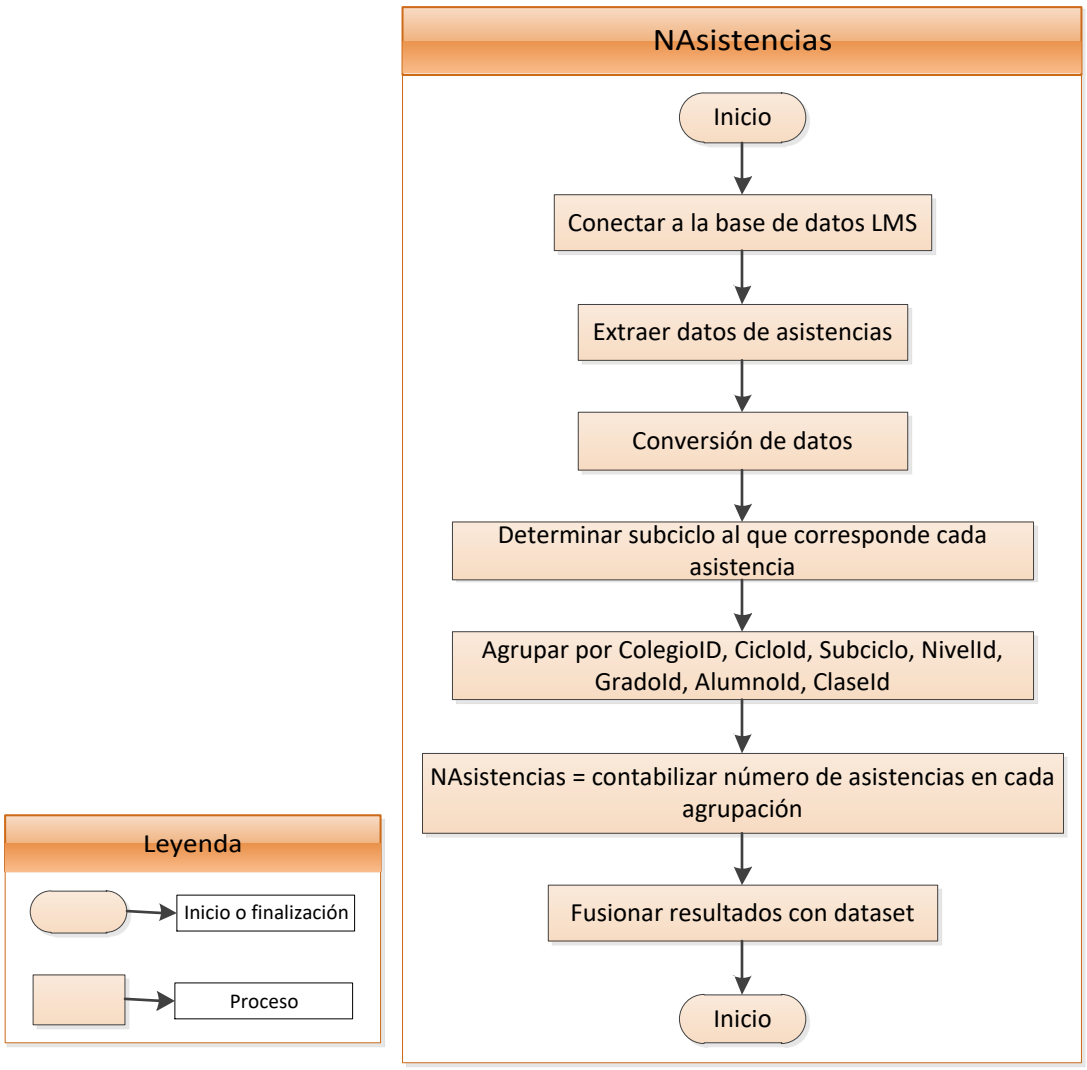

Figura 5. Proceso de creación del atributo "NAsistencias".

3.3.1.8 Añadir retardos

Vamos a incluir en el dataset el número de veces que un alumno llega tarde a clase. Esta información se guardará en el campo "NRetardos".

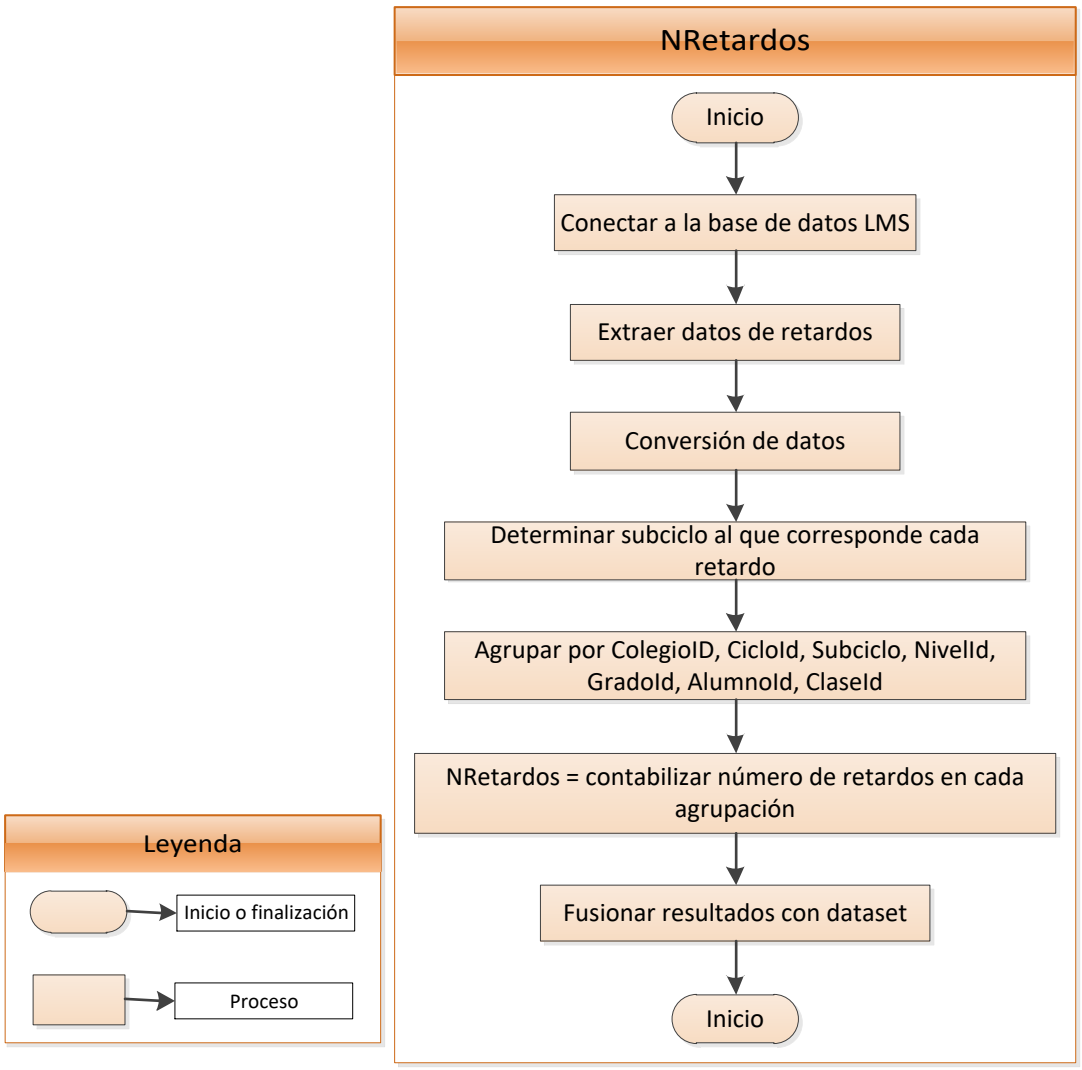

Figura 6. Proceso de creación del atributo "NRetardos".

3.3.1.9 Añadir faltas

Vamos a incluir en el dataset el número de ausencias de cada alumno. Esta información se guardará en el campo "NFaltas".

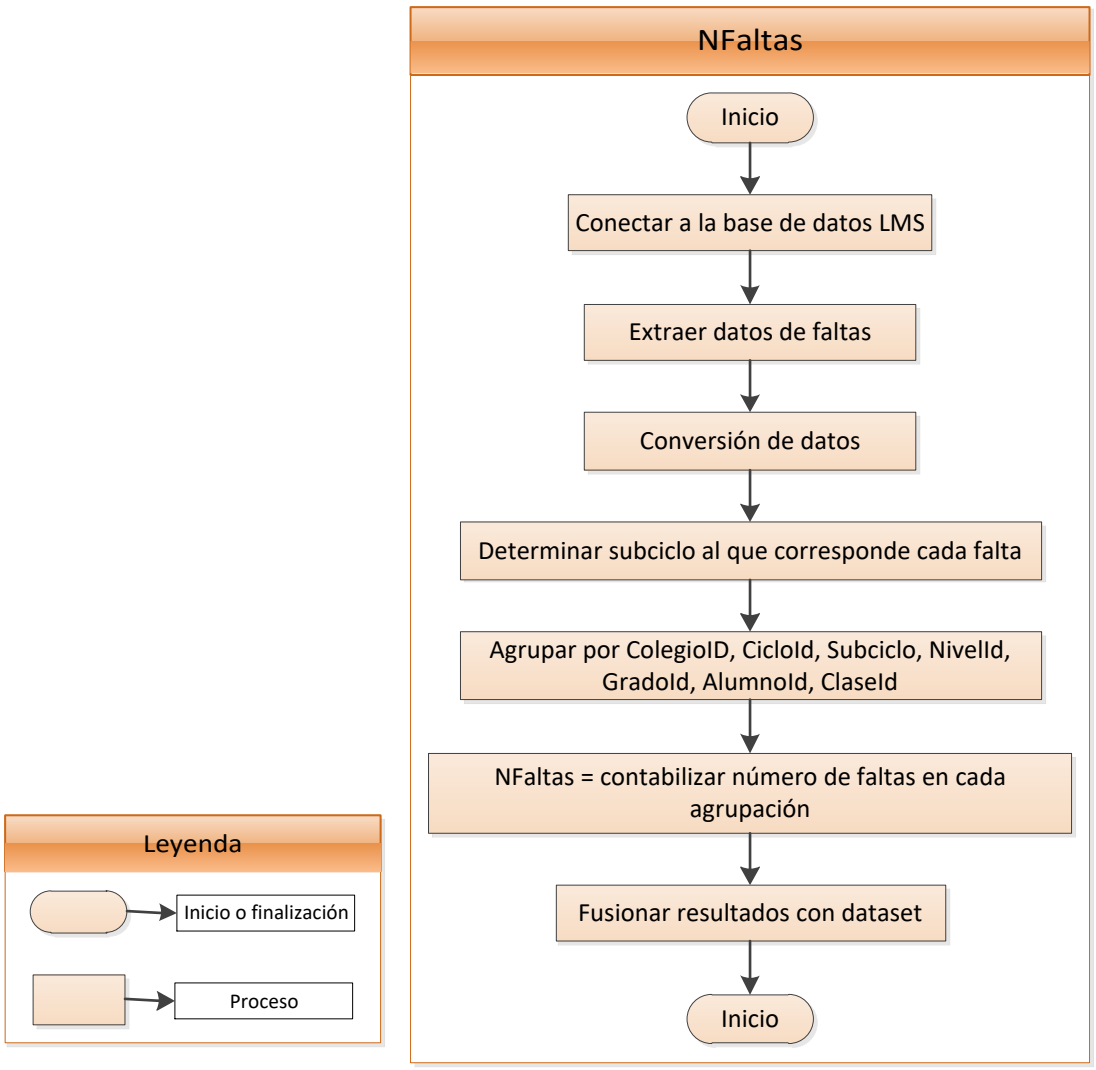

Figura 7. Proceso de creación del atributo "NFaltas".

3.3.1.10 Añadir faltas justificadas

Vamos a incluir en el dataset el número de ausencias justificadas de cada alumno. Esta información se guardará en el campo "NFaltasJust".

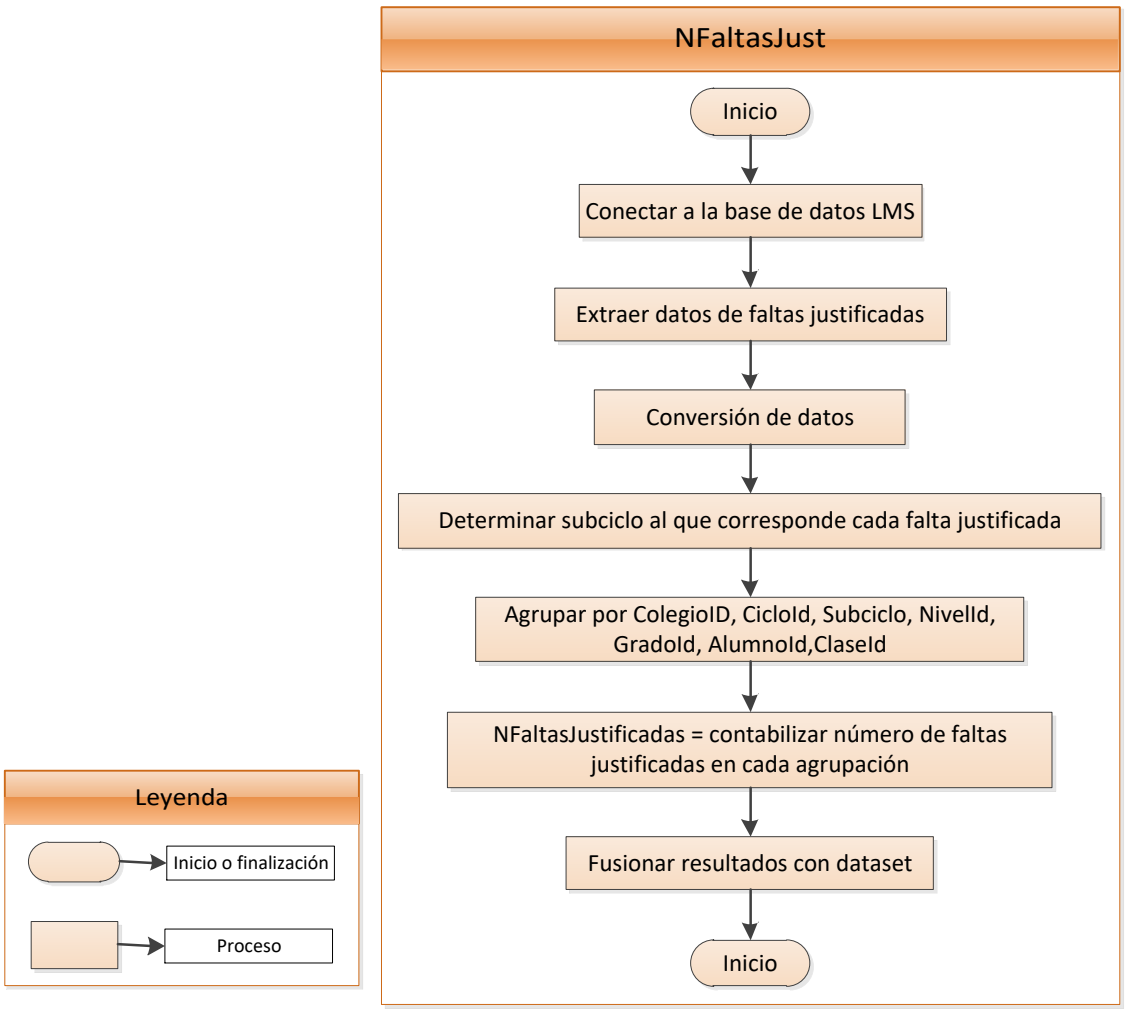

Figura 8. Proceso de creación del atributo "NFaltasJust".

3.3.1.11 Cálculo de los porcentajes de asistencias, retardos y faltas

Vamos a incluir en el dataset los porcentajes de asistencias, retardos, faltas y faltas justificadas.

Para ello vamos a dividir cada una de las variables por la suma de todas ellas. En lugar de un diagrama utilizaremos Pseudocódigo ya que tan sólo se realizan asignaciones de valores.

PSEUDOCÓDIGO:

```
#Calculamos los porcentajes de asistencias

Sumatorio <- NAsistencias + NFaltas + NFaltasJust + NRetardos

NAsistencias <- (NAsistencias / Sumatorio) * 100
PFaltas <- (NFaltas / Sumatorio) * 100
NFaltasJust <- (NFaltasJust / Sumatorio) * 100
NRetardos <- (NRetardos / Sumatorio) *100
```

3.3.1.12 Añadir actividades de clase y exámenes

Vamos a incluir en el dataset los resultados de las actividades de clase y los examenes. Para ello introduciremos 8 variables:

- NTareas: Número de tareas realizadas por el alumno.
- NotaTareas: Nota media de las tareas realizadas por el alumno.
- PTareas: Porcentaje de tareas realizadas por el alumno con respecto al número de tareas que debería haber realizado. Se calcula aplicando la fórmula: (NTareas/NTareasMax) * 100
- PNotasTareas: Indica la nota media de las actividades respecto al número total de actividades que debería haber realizado. Las actividades no realizadas reciben una nota de cero puntos. Se calcula aplicando la fórmula: (NotaTareas/NTareasMax) * NTareas
- NExamenes: Número de examenes realizados por el alumno.
- NotaExamenes: Nota media de los examenes realizados por el alumno.
- PExamenes: Porcentaje de examenes realizados por el alumno con respecto al número de examenes que debería haber realizado. Se calcula aplicando la fórmula: (examenes$NExamenes/examenes$NExamenesMax) * 100
- PNotasExamenes: Indica la nota media de los examenes respecto al número total de examenes que debería haber realizado. Los examenes no realizados reciben una nota de cero puntos. Se calcula aplicando la fórmula: (examenes$NotaExamenes/examenes$NExamenesMax) *examenes$NExamenes

Dado que no tenemos el número total de tareas, tendremos que estimarlo. Vamos a suponer que el número de tareas a realizar es igual al número máximo de tareas que ha realizado un alumno en esa clase, es decir, tomamos el número de tareas que han realizado cada alumno y nos quedamos con el valor mayor. Esto mismo también se aplicará al número máximo de examenes.

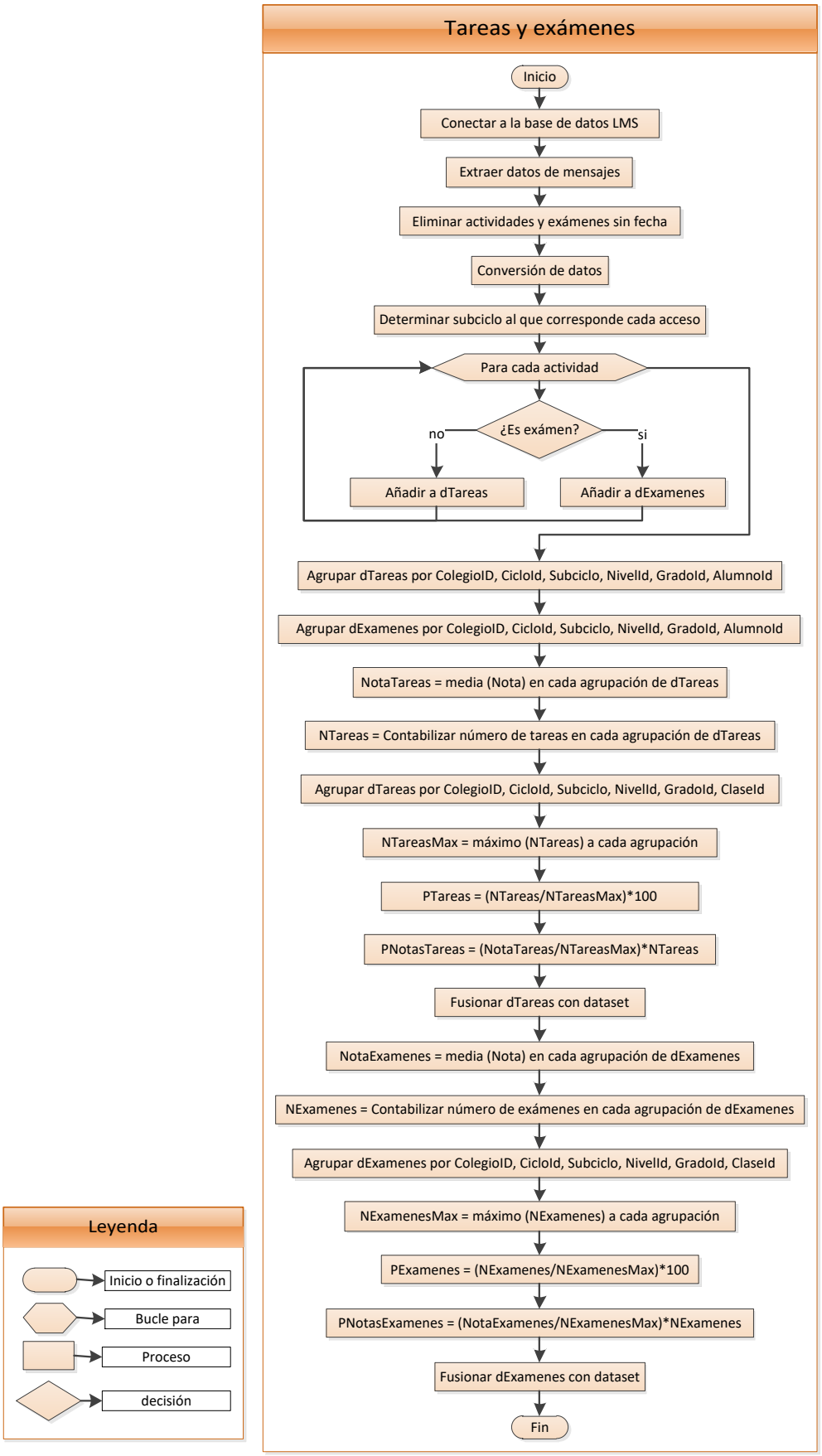

Figura 9. Proceso de creación de los atributos "NTareas", "NotaTareas", "PTareas", "PNotasTareas", "NExamenes", "NotaExamenes", "PExamenes" y "PNotasExamenes".

3.3.1.13 Añadir número de sesiones

Vamos a incluir en el dataset el número de sesiones abiertas en el sistema LMS por el alumno. Esta información se guardará en el campo "NSesionesTotalClases". Estos datos se extraeran de dos tablas distintas.

- FACTSesionLogin_H: Datos actuales de sesiones.
- Sesion_Antigua: Histórico de sesiones.

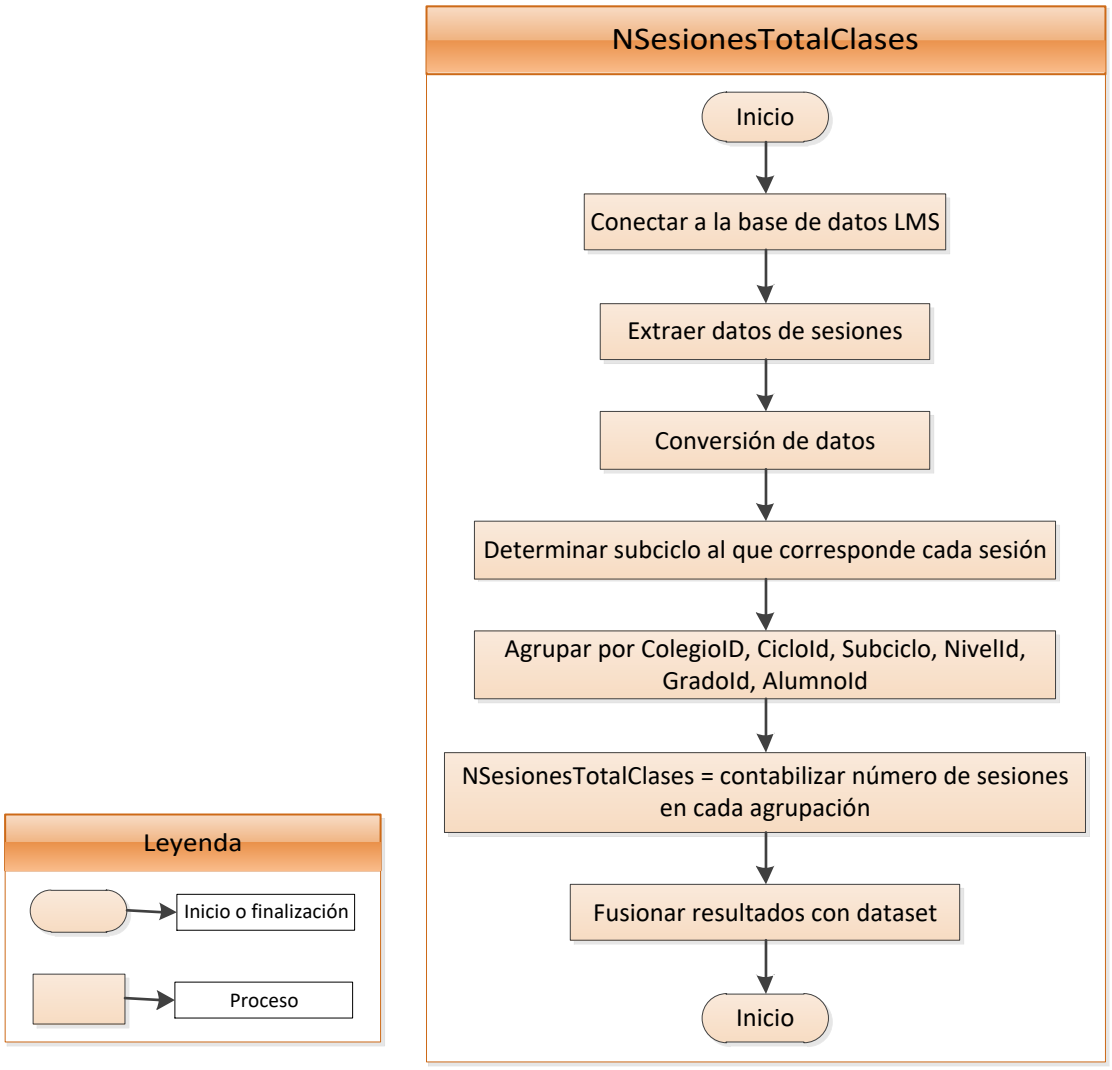

Figura 10. Proceso de creación del atributo "NSesionesTotalClases".

3.3.1.14 Añadir participación en foros

Vamos a incluir en el dataset el número de participaciones en foros de cada alumno. Esta información se guardará en el campo "NPostForos".

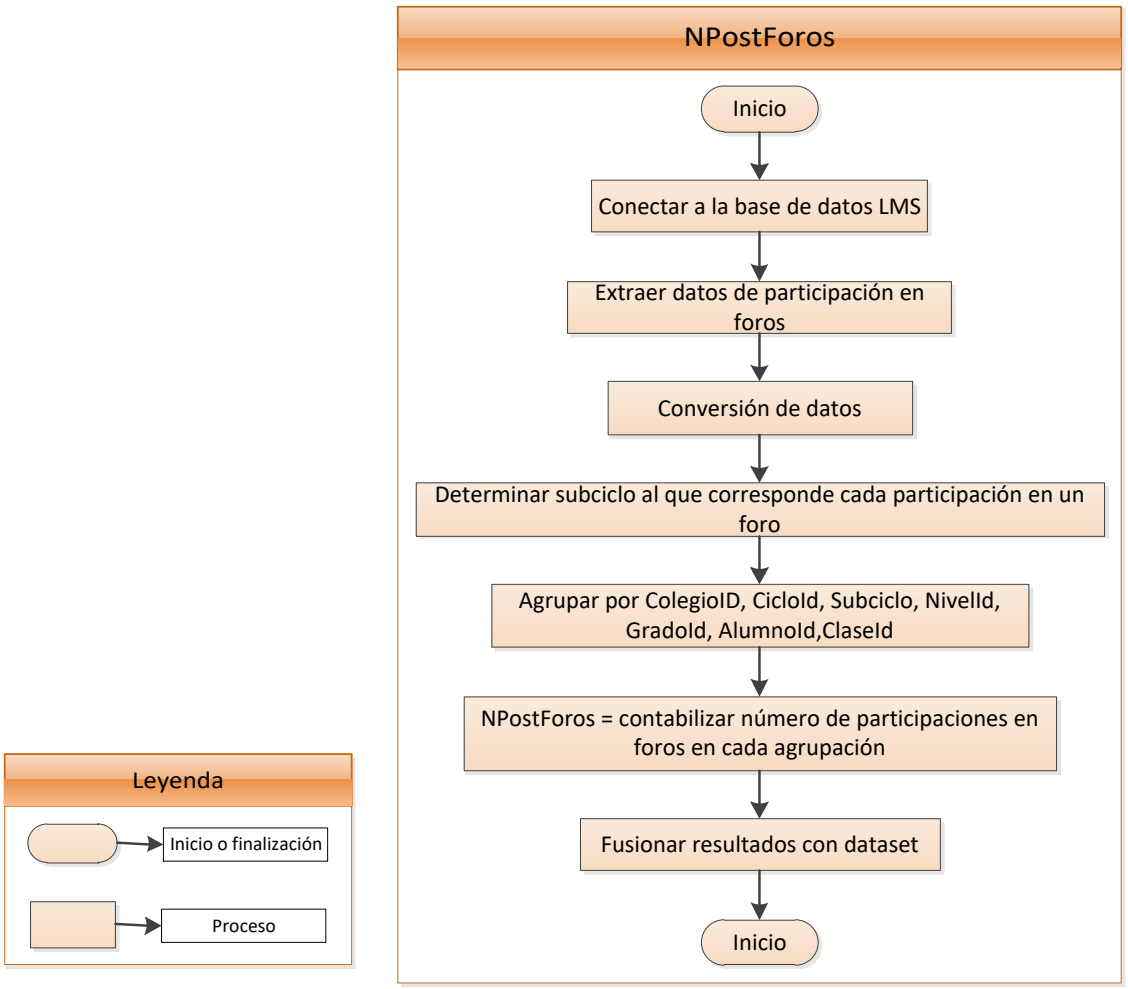

Figura 11. Proceso de creación del atributo "NPostForos".

3.3.1.15 Añadir campos de la base de datos de cuestionarios (Pleno)

Vamos a incluir en el dataset la información contenida en la base de datos de test (PLENO).

Los campos a introducir son:

- PruebaId: Identificador de la prueba.
- TiempoRespuesta: Tiempo que tarda el alumno en terminar el test.
- TestSuperado: Indica si el alumno aprobó el test.
- PRespuestasCorrectas: Porcentaje de respuestas que ha acertado el alumno.
- PNecesarioParaLogro: Porcentaje de respuestas mínimo que el alumno debe acertar para superar el test.

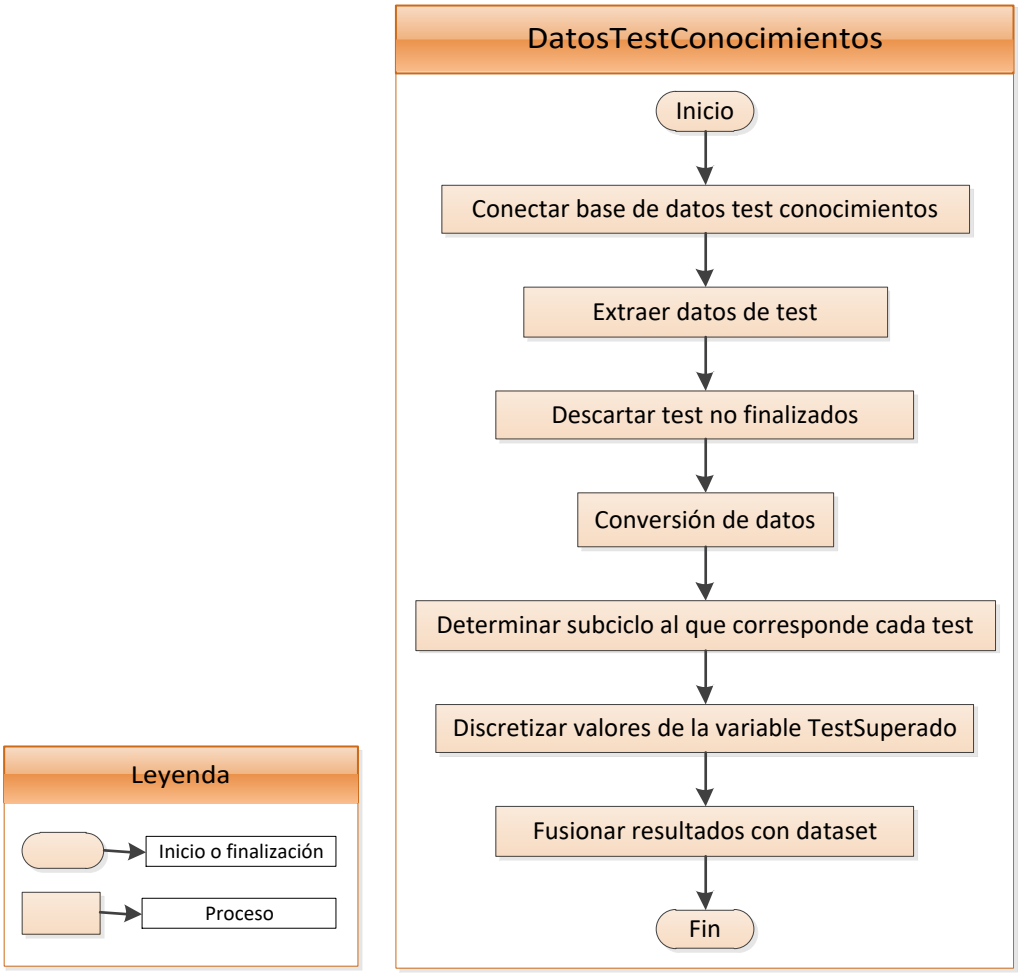

Figura 12. Proceso de creación de los atributos relacionados con los tests de conocimientos.

3.3.1.16 Guardar los resultados en un archivo

Finalmente guardamos los resultados en un fichero RData de manera que podamos recuperar el dataset fácilmente.

3.3.2 Creación de la variable Area

Uno de los principales problemas del dataset es que la nota final (que es el objetivo que deseamos predecir) está vinculada a la clase.

Una clase es un conjunto de alumnos que cuentan con unas materias en común.

Cada alumno tiene una nota final para cada clase a la que pertenece (o en la que está matriculado).

El problema de la clase es que puede tener un nombre asignado por el docente.

En muchos casos este nombre será identificativo (por ejemplo "matemáticas primero B") pero en otros casos no será identificativo (por ejemplo "clase de la profesora Mari Carmen") en cuyo caso no se podrá utilizar para predicción. Además, aunque el nombre sea identificativo de los contenidos impartidos, puede cambiar de un curso para otro (de un cilclo para otro) por lo que, en ocasiones, no podemos utilizar las notas de un curso en el curso siguiente.

Por otro lado, cada clase tiene una o más materias que indican los contenidos de la clase. Ejemplos de materias son matemáticas o lenguaje.

La nota final de la clase se calcula en base a las notas individuales de cada materia, las cuales tienen asignadas unos pesos (los cuales no están a nuestra disposición por lo que no podremos usar estas notas).

A pesar de la existencia de materias comunes creadas por la empresa, las materias también pueden ser creadas por el docente (materias privadas) por lo que no coincidirán.

Para solventar este problema vamos a crear una variable denominada Area que nos permitirá unificar alumnos que estén estudiando los mismos contenidos. Hay que tener en cuenta que Area estará entre los niveles grado y clase (un nivel superior a clase). A cada ClaseId se le asignara un nombre de área. Las clases a las que no podamos asignar un área serán descartadas.

En resumen, con esta variable conseguimos:

- Agrupar varias clases similares en un nivel superior (por ejemplo "matemáticas primero A", "matemáticas primero B" y "matemáticas primero C").
  De esta forma conseguimos conjuntos de datos con más registros para la creación del modelo predictivo al juntar todos los alumnos de la misma área para un determinado nivel/grado (en lugar de utilizar los alumnos de una única clase).
- Vincular datos entre varios ciclos, aunque cambie el nombre de la clase (si los contenidos son los mismos se podrán relacionar).
  Esto será imprescindible para la creación de datos históricos como la nota que obtuvo en esta área en el ciclo anterior. Al trabajar con clases no podríamos hacerlo ya que el nombre de la clase puede cambiar de un ciclo a otro.

Para la asignación del área vamos a realizar un mapeo basado en los nombres de las clases. Una vez terminado lo realizaremos en base a los nombres de las materias.

Al utilizar las materias en segundo lugar, estas sobreescribirán los valores previos obtenidos con los nombres de las clases con lo que le damos prioridad a los nombres de las materias sobre los nombres de las clases.

De igual forma, el orden en que se asignan los nombres de las areas ha sido establecido para dar prioridad a determinadas areas en caso de duda.

Las clases con más de una materia han sido omitidas (por ejemplo "clase de Manuel" que tiene como materias "Matemátias", "Lengua", "Ciencias Sociales" y "Ciencias Naturales").

Vamos a crear las siguientes áreas:

1. Matematicas
2. Ciencias Sociales
3. Ciencias Naturales
4. Lenguaje
5. Ingles
6. Educacion Artistica
7. Educación Física
8. Educación Religiosa
9. Quimica
10. Fisica
11. Tecnología e informatica
12. Frances
13. Filosofia
14. Formacion Civica y Etica
15. Geografia
16. Historia

Este proceso se realizará en varios pasos:

1. Cargar librerías necesarias
2. Obtención de nombres de clases y materias
3. Mapeo de nombres de clases
4. Mapeo de nombres de materias
5. Mapeo de casos especiales por país
6. Eliminación de registros duplicados y clases multiáreas.
7. Añadir variable Area al dataset

3.3.2.1 Cargar librerías necesarias

Utilizaremos RODBC para el acceso a las bases de datos en MS SQL Server y dplyr para manejar de forma más eficiente las variables tipo data.frame.

3.3.2.2 Obtención de nombres de clases y materias

En este punto nos conectamos a la base de datos del LMS para extraer las materias de la tabla de notas de exámenes y actividades.

3.3.2.3 Mapeo de nombres de clases

En primer lugar, haremos un mapeo de los nombres de las clases a los nombres de las áreas que hemos creado. Para ello utilizaremos, dentro del lenguaje R, PERL y el comando grep.

A continuación, mostramos una tabla con los mapeos realizados por nombre de clase:

AREA	MAPEO POR NOMBRE DE CLASE								
Matematicas	matem	stad	lgebr	ometr					
Ciencias Sociales	sociales								
Ciencias Naturales	naturale	biolog							
Lenguaje	"Lengua" y no "extranje[oa]"	Lect[ou]ra	Ortograf[ií]a	Español	Espanhol				
Ingles	English	Ingl[eé]s	Listen	Speak	Comprehe	writ[ei]	read	grammar	spell
Educacion Artistica	art[ií]stic[oa]	arte	^arte	M[úu]sica	Danza				
Fisica	"F[íi]sica" y no "Educaci[óo]n"								
Educación Fisica	Educaci[oó]n F[ií]sica	e.*[.] fisica$	F[ií]sico	Deporte					
Etica	[eé]tica	^[eé]tica							
Educación Religiosa	Religi[oó]	catequ[ie]s	e.* fe$						
Quimica	Qu[ií]m	emist							
Tecnologia e Informatica	tecno	inform[aá]t	computaci[oó]n						
Frances	Franc[eé]								
Filosofica	Filosof								
Geografia	Geograf								
Historia	Historia								

Tabla 7. Mapeo de áreas por nombre de clase.

3.3.2.4 Mapeo de nombres de materias

Realizaremos el mapeo de los nombres de las materias a los nombres de las áreas que hemos creado. Para ello volvemos a utilizar, dentro del lenguaje R, PERL y el comando grep.

A continuación, mostramos una tabla con los mapeos realizados por nombre de materia:

AREA	MAPEO POR NOMBRE DE MATERIA								
Matematicas	matem								
Ciencias Sociales	sociales								
Ciencias Naturales	naturale	biolog							
Lenguaje	"Lengua" y no "extranje[oa]"	Lect[ou]ra	Ortograf[iíÍ]a	Español	Espanhol				
Ingles	English	Ingl[eé]s	Listen	Speak	Comprehe	writ[ei]	read	grammar	spell
Educacion Artistica	art[iíÍ]stic[oa]	arte	^arte	M[úu]sic	Danza				
Fisica	"F[íi]sica" y no "Educaci[óo]n"								
Educación Fisica	Educaci[oó]n F[iíÍ]sica	e.*[.] fisica$	ed.* fisica$	F[iíÍ]sico	Deporte				
Etica	[eé]tica	^[eé]tica							
Educación Religiosa	Religi[oó]	catequ[ie]s	e.* fe$						
Quimica	Qu[íi]m								
Tecnologia e Informatica	tecno	inform[aá]t	computaci[oó]n						
Frances	Franc[eé]								
Filosofica	Filosof								
Geografia	Geograf								
Historia	Historia								

Tabla 8. Mapeo de áreas por nombre de materia.

3.3.2.5 Mapeo de casos especiales por país

Determinados países poseen materias únicas que deben ser tratadas de una forma diferente.

La tabla de mapeo es la siguiente:

PAIS	AREA	MAPEO ESPECIAL POR PAISES	
Mexico	Ingles	Lengua Extranjera	mat_Español
Ecuador	Educacion Fisica	Cultura F[iíÍ]sica	

Tabla 9. Excepciones de mapeo de áreas por países.

3.3.2.6 Eliminación de registros duplicados y clases multiáreas.

Las clases con más de un área deben ser eliminadas. Estas se producen cuando una clase es multimateria (por ejemplo, una clase donde se impartan contenidos de lenguaje y matemáticas).

Para ello debemos calcular la frecuencia de cada agrupación EmpresaId, ColegioID, CicloId y ClaseId. Todos los que tengan una frecuencia superior a uno deberán ser eliminados.

3.3.2.7 Añadir variable "Area"" al dataset

Finalmente cargamos el dataset guardado en el punto 3.3.1.16 y lo fusionamos con las áreas. Una vez terminado volvemos a guardar el nuevo dataset en un fichero RData para facilitar su recuperación en pasos posteriores.

3.3.3 Creación de datos históricos

En este punto vamos a introducir el concepto de ciclo anterior. En la base de datos no se puede saber cual es el ciclo en el que estuvo el alumno el año lectivo anterior. Existen ciclos distintos para cada país y en cada país puede haber varios ciclos que corresponden a un periodo lectivo determinado (por ejemplo 2 ciclos que comprenden desde septiembre de 2013 a junio de 2014 para el mismo país).

Este problema se intenta solucionar mediante el cálculo de una nueva variable que se añadirá al dataset denominada cicloAnterior.

Este campo se calcula mirando los alumnos que tienen al menos 2 ciclos y buscando para cada instancia si existe una instancia del mismo alumno cuya fecha de inicio del ciclo corresponda al año anterior (año - 1).

Una vez obtenida esta variable podemos calcular otras variables que nos serán muy útiles para la predicción temprana ya que serán los únicos que tendrán valores al inicio del curso (antes de empezar las clases). Estos campos son:

- NotaCicloAntArea: Nota que obtuvo en la misma área en el año anterior.
- NotaCicloAntMate: Nota que obtuvo en matemáticas en el año anterior.
- NotaCicloAntLeng: Nota que obtuvo en lenguaje en el año anterior.
- MismoNivelCicloAnt: Indica si el alumno ha cambiado de nivel (por ejemplo, de primaria a secundaria).
- Repetidor: Indica si está cursando el mismo grado que el año anterior (mismo curso).

Se podrían considerar la inclusión de la nota que ha obtenido en el ciclo anterior en otras áreas, pero estimamos que las más influyentes en las notas de las distintas áreas son matemáticas y lenguaje. Por norma general, los alumnos con problemas en estas materias suelen arrastrar problemas en el resto de materias.

También se debe añadir la nota que obtuvo en la misma área porque es obvio que ésta va a ser la más influyente en dicha área.

Finalmente es interesante ver si ha repetido curso o si ha experimentado un cambio de nivel (por ejemplo, el paso de primaria a secundaria, lo cual implica un incremento importante de la exigencia al alumno).

Se seguirán los siguientes pasos:

1. Cargar librerías necesarias
2. Obtención de los ciclos anteriores
3. Creación de las variables históricas.

3.3.3.1 Cargar librerías necesarias

Cargaremos la librería RODBC para el acceso a las bases de datos en MS SQL Server, las librerías dplyr y data.table para facilitar el manejo de los data.frames y la librería lubridate para facilitar el uso de variables tipo fecha.

3.3.3.2 Obtención de los ciclos anteriores

Para calcular el ciclo anterior de cada alumno/ciclo, buscamos el mayor ciclo cursado por ese alumno que cumple ser menor que el ciclo del que estamos calculando el ciclo anterior.

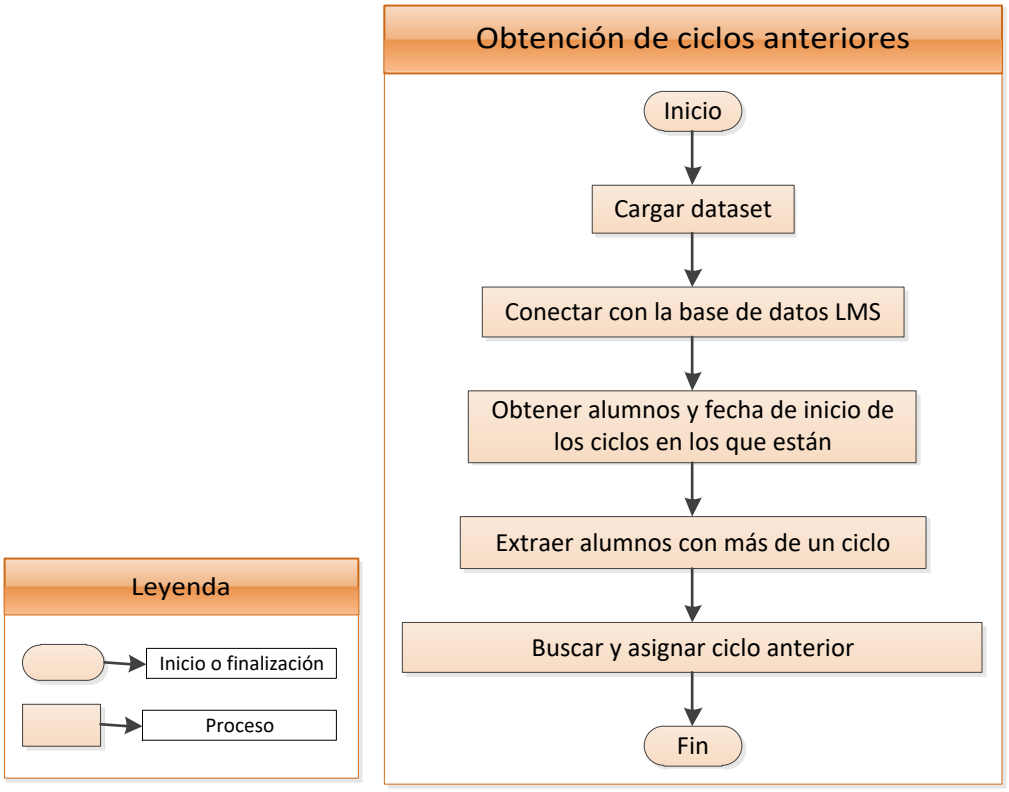

Figura 13. Proceso de obtención de ciclos anteriores.

3.3.3.3 Creación de las variables históricas.

Crearemos variables cuyo valor derivan del ciclo anterior y por lo tanto estarán disponibles al inicio del curso (en el subciclo 0).

Estas variables son:

- NotaCicloAntArea: Nota que obtuvo en la misma área en el ciclo anterior.
- NotaCicloAntLeng: Nota que obtuvo en lenguaje en el ciclo anterior.
- NotaCicloAntMate: Nota que obtuvo en matemáticas en el ciclo anterior.
- MismoNivelCicloAnt: Indicara si el alumno a cambiado de nivel (por ejemplo, pasar de primaria a secundaria).
- Repetidor: Indica si el alumno esta repitiendo curso (está cursando el mismo grado que en el ciclo anterior).

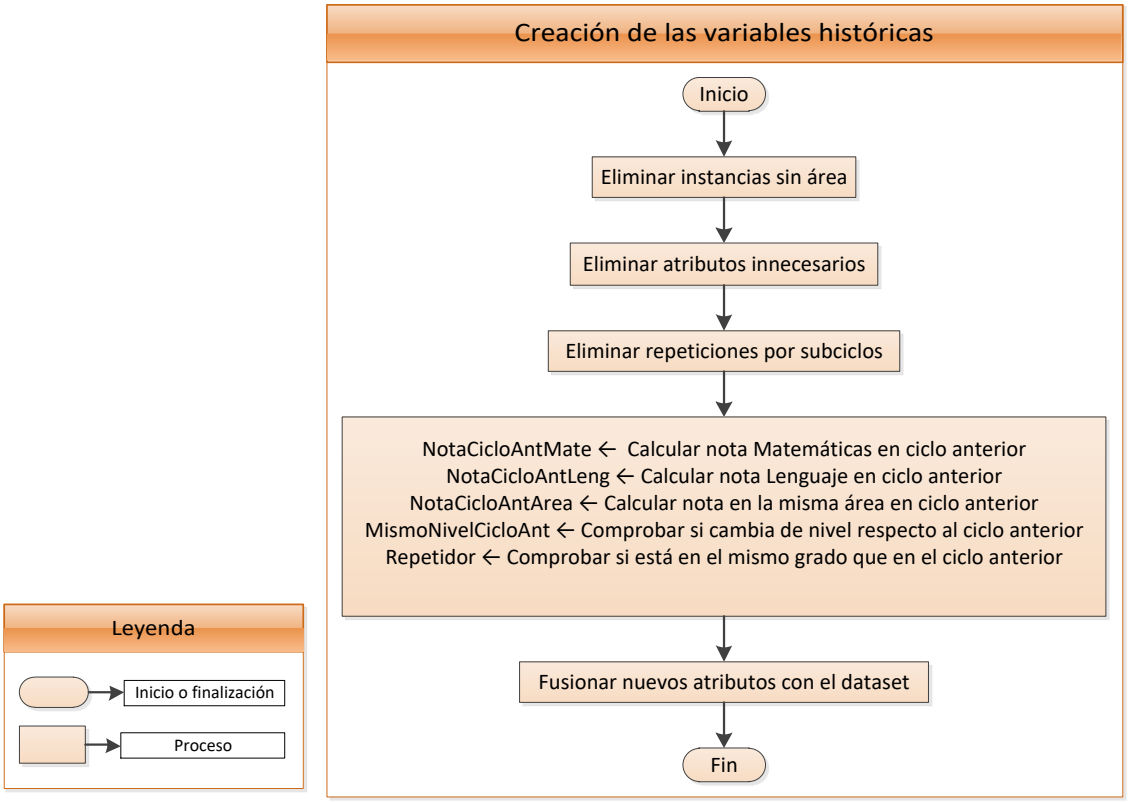

Figura 14. Proceso de creación de las variables históricas.

3.3.4 Transformación de datos

Vamos a realizar una serie de transformaciones en los datos para evitar problemas a la hora de ejecutar algoritmos de clasificación. Estas transformaciones serán de tipo de datos y de sistema de codificación (para evitar tildes y caracteres problemáticos)

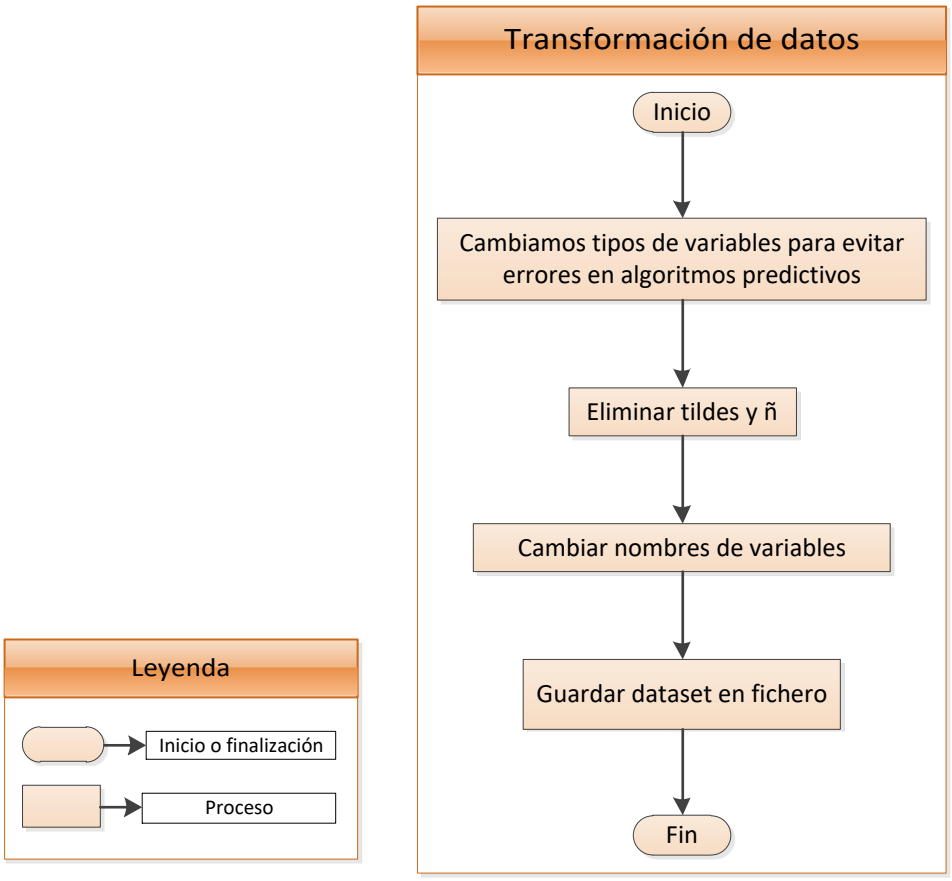

Figura 15. Proceso de transformación de datos.

4 Experimentación

En este capítulo vamos a exponer los experimentos realizados. Vamos a dividirlo en tres apartados:

- Definición del experimento: Indicaremos la motivación y definiremos las pruebas que vamos a realizar y cómo las realizaremos.
- Configuración de la experimentación: Se detallarán los algoritmos creados para la ejecución de los experimentos y los subconjuntos de datos generados a partir del dataset y que serán objeto de estudio.
- Pruebas: Se mostrarán los resultados de las pruebas realizadas

4.1 Definición del experimento

El estudio que vamos ha realizar tendrá dos objetivos fundamentales:

- Determinar el algoritmo predictivo que da mejores resultados sobre el conjunto de datos.
- Comprobar que la creación de la variable "Area" para sustituir la variable "clase" nos permite realizar una buena predicción de la nota final de clase.

Para ello vamos a realizar el siguiente experimento (utilizando el lenguaje de programación R) con cada subconjunto de datos propuesto en el punto 4.2.2:

1. Dividimos el conjunto de datos en 7 (uno por subciclo) y aplicamos lo siguiente:
    a. Realizaremos 10 ejecuciones de cada algoritmo sobre el subconjunto de datos. En cada una de las ejecuciones se aplicarán los siguientes pasos:
        i. Dividiremos el dataset (subconjunto del original) en 2 conjuntos de datos de forma aleatoria, asignando el 70% de las instancias a trainData el 30% restante a testData.
        ii. Creamos el modelo utilizando trainData.
        iii. Probamos el modelo con testData, generamos la matriz de confusión y calculamos varios estadísticos.
        iv. Generamos el árbol (si es aplicable).
        v. Guardamos toda la información en ficheros. Estos serán:
            1. ListadoReglas.csv: Contiene las reglas obtenidas
            2. MatrizConfusion.csv: Contiene la matriz de confusión.
            3. Modelo.pdf: Dibujo del árbol obtenido.
            4. Modelo.RData: Fichero R que contiene todas las variables necesarias (modelo, conjuntos de datos, etc) para reproducir la prueba.
            5. OverallStatistics.csv: Contiene distintos estadísticos como la precisión y el índice Kappa.
            6. StatisticsByClass.csv: Contiene distintos estadísticos de clase como la sensibilidad y la especificidad.
            7. VariablesDisponibles: Contiene las variables (atributos) utilizados en el modelo.

b. Extraemos la información de cada ejecución y creamos archivos con resultados globales:
    i. MejorIteracion.txt: Guarda la iteración (de las 10) que ha obtenido mejores resultados. Para determinarlo se ha seleccionado la iteración con mayor sensibilidad y en caso de empate se utiliza la precisión.
    ii. OverallStatisticsTotales.csv: Agrupa los resultados de "OverallStatistics.csv" de todas las iteraciones.
    iii. OverallStatisticsTotalesSummary: Almacena un resumen del fichero "OverallStatisticsTotales.csv". Muestra, para cada estadístico, los siguientes valores: mínimo, mediana, cuartiles (incluye media) y máximo.
    iv. StatisticsByClassTotales.csv: Agrupa los resultados de "StatisticsByClass.csv" de todas las iteraciones.
    v. StatisticsByClassTotales Summary: Almacena un resumen del fichero "StatisticsByClassTotales.csv". Muestra, para cada estadístico, los siguientes valores: mínimo, mediana, cuartiles (incluye media) y máximo.

2. Generamos el fichero "Resultados.csv" que contiene las medias de la precisión, especificidad y sensibilidad obtenidas en los pasos anteriores para cada combinación Subciclo/Algoritmo.
3. Generamos 3 gráficas comparando los resultados del punto anterior (una para precisión, otra para especificidad y otra para sensibilidad) y las guardamos en el fichero "ComparativaAlgoritmos.pdf".

En base a lo mostrado en las gráficas obtenidas y almacenadas en el fichero "ComparativaAlgoritmos.pdf" podremos determinar qué algoritmo nos interesa más y si los resultados son buenos.

Para la realización de las pruebas hemos escogido los siguientes algoritmos:

- **ZeroR**: Es el método de clasificación más simple. Se basa únicamente en el target (variable objetivo de la predicción) sin tener en cuenta el resto de predictores por lo que tan solo predice la clase mayoritaria. No hay poder predictivo real en ZeroR pero es útil para determinar la línea mínima que otros métodos de clasificación deben superar para demostrar su rendimiento.
- **JRip (RIPPER)**: Este algoritmo está basado en un aprendiz de reglas proposicionales que optimiza IREP mediante la repetición de una poda incremental para la reducción de error. Devuelve un conjunto de reglas.
- **J48 (C4.5):** Algoritmo que genera un árbol de decisión. Es una extensión del algoritmo ID3. En cada nodo realiza una división de los datos por el atributo que aporta una mayor ganancia de información. Se va dividiendo recursivamente el problema en sublistas más pequeñas.
- **C50:** Versión del algoritmo C4.5 que mejora los siguientes aspectos:
  - Importante aumento de la velocidad.
  - Uso más eficiente de la memoria.
  - Resultados más fáciles de interpretar (arboles de decision más pequeños).
  - Soporta boosting, lo cual mejora los árboles consiguiendo mayor precisión.
  - Permite la asignación de pesos.
  - Posee una opción para la selección automatic de atributos para eliminar aquellos que no son útiles.
- **RPart (CART):** Algoritmo basado en un particionamiento recursivo. Una vez generado el árbol se realiza una poda del mismo. Tiene una alta tolerancia a outliners.
- **NaiveBayes:** Clasificador probabilístico bayesiano que asume independencia entre todas las características que componen cada patrón de la muestra [3]. Construye, en base a los datos de entrenamiento, un modelo probabilístico que otorga distintos pesos y nos da como resultado la probabilidad de pertenencia a una clase. Al no ser fácilmente interpretable solo se utilizará para comparar los resultados con el resto de algoritmos, pero no será seleccionable.

4.2 Configuración de la experimentación

4.2.1 Algoritmos creados para ejecutar las pruebas

Vamos a crear varios niveles de profundidad para la generación de modelos.

1.- GenerarModelosColeNivelGradoArea: Le indicamos varios colegios, niveles, grados y áreas. Crea la estructura de directorios necesaria para cada combinación, realiza los cálculos necesarios y guarda los resultados en cada carpeta. Este es el nivel más alto.

Llama a la función GenerarModeloColeNivelGradoAreaAlgoritmoSubciclo.

2.- GenerarModeloColeNivelGradoAreaAlgoritmoSubciclo: Dado un dataset (normalmente enviado desde la función GenerarModelosColeNivelGradoArea) genera n modelos por cada par algoritmo/subciclo.

Llama a la función GenerarModeloColeNivelGradoAreaAlgoritmoSubcicloIteracion.

3.- GenerarModeloColeNivelGradoAreaAlgoritmoSubcicloIteracion: Esta función nos permite generar modelos a partir de un dataset. Podrá utilizar uno o varios de los siguientes algoritmos (la definición de los mismos se puede ver en el apartado anterior “Definición del experimento”):

- Rpart
- C5.0
- JRip
- J48
- NaiveBayes

Se realizarán n ejecuciones de cada algoritmo solicitado sobre el conjunto de datos. Los resultados se guardarán en distintas carpetas. Los modelos generados también se guardarán por lo que podrán ser recuperados posteriormente.

Para la creación de los modelos, el conjunto de datos se dividirá en dos datasets. Uno para entrenamiento que contendrá el 70% de los datos y otro para test con un 30%.

Para simplificar el código se han creado funciones que son necesarias para la correcta ejecución del mismo y se detallan en el apartado “otras funciones”.

4.2.1.1 GenerarModeloColeNivelGradoArea

A esta función le pasaremos una lista de colegios, niveles, grados y areas y creará modelos para cada combinación de los mismos.

Se creará una carpeta de cada colegio, dentro de esta una por cada nivel, cada nivel contendrá carpetas para cada grado e igualmente cada grado tendrá una carpeta por cada area.

Si alguna combinación no tiene datos no se creará la carpeta.

Finalmente, dentro de cada combinación se realizará una llamada al siguiente nivel de profundidad mediante el uso de la función "GenerarModeloColeNivelGradoAreaAlgoritmoSubciclo" la cual se encargará de crear los modelos.

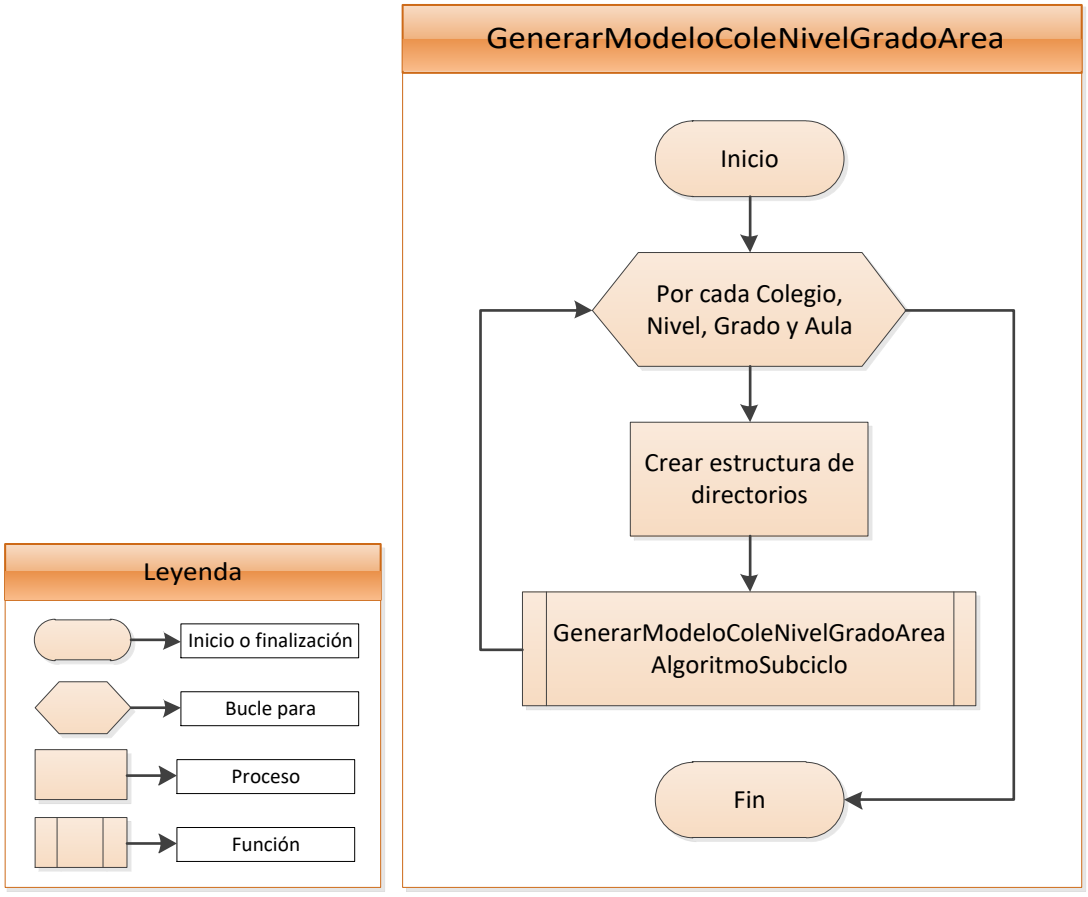

Figura 16. Proceso GenerarModeloColeNivelGradoArea.

RESUMEN:

- Definición: Genera los modelos para los colegios indicados y guarda los resultados en el directorio de trabajo actual (creando la estructura de directorios necesaria).
- Entrada:
  - dataset: data.frame que contiene los datos.
  - colegios: Vector que contiene los índices de los colegios a procesar.
  - niveles: Vector que contiene los índices de los niveles a procesar de cada colegio.
  - grados: Vector que contiene los índices de los grados a procesar de cada nivel.
  - areas: Vector que contiene los nombres de las áreas a procesar de cada grado.
  - algoritmosProbados: Algoritmos que vamos a lanzar en el experimento.
  - repeticiones: Número de iteraciones (modelos construidos) para cada Colegio/Nivel/Grado/Area/Subciclo
  - notaCorte: Nota mínima para aprobar.
  - umbralNota: Se le sumará a la nota de corte para indicar que los alumnos por debajo de esa nota están en peligro de suspender la asignatura. Por ejemplo, si la nota de corte es 50 y el umbral 10, todos los que tengan una nota inferior a 60 se consideran en peligro.
- Salida: Ficheros con los resultados de los distintos modelos.
- Llamadas a funciones: GenerarModeloColeNivelGradoAreaAlgoritmoSubciclo.
- Observaciones: Se pueden procesar colegios que tengas distintos niveles o grados. Si no existe la combinación Colegio/nivel/grado/area se omitirá.

4.2.1.2 GenerarModeloColeNivelGradoAreaAlgoritmoSubciclo

Dado un dataset (que normalmente proviene de una combinación Colegio/Nivel/Grado/Area solicitada por la función GenerarModeloColeNielGradoArea) crea una carpeta por algoritmo que se deba probar y dentro de estas crea una carpeta por cada subciclo (del 0 al 6).

Dentro de cada combinación algoritmo/subciclo se realizará una llamada al siguiente nivel de profundidad mediante el uso de la función "GenerarModeloColeNivelGradoAreaAlgoritmoSubcicloIteracion" la cual se encargará de crear los modelos.

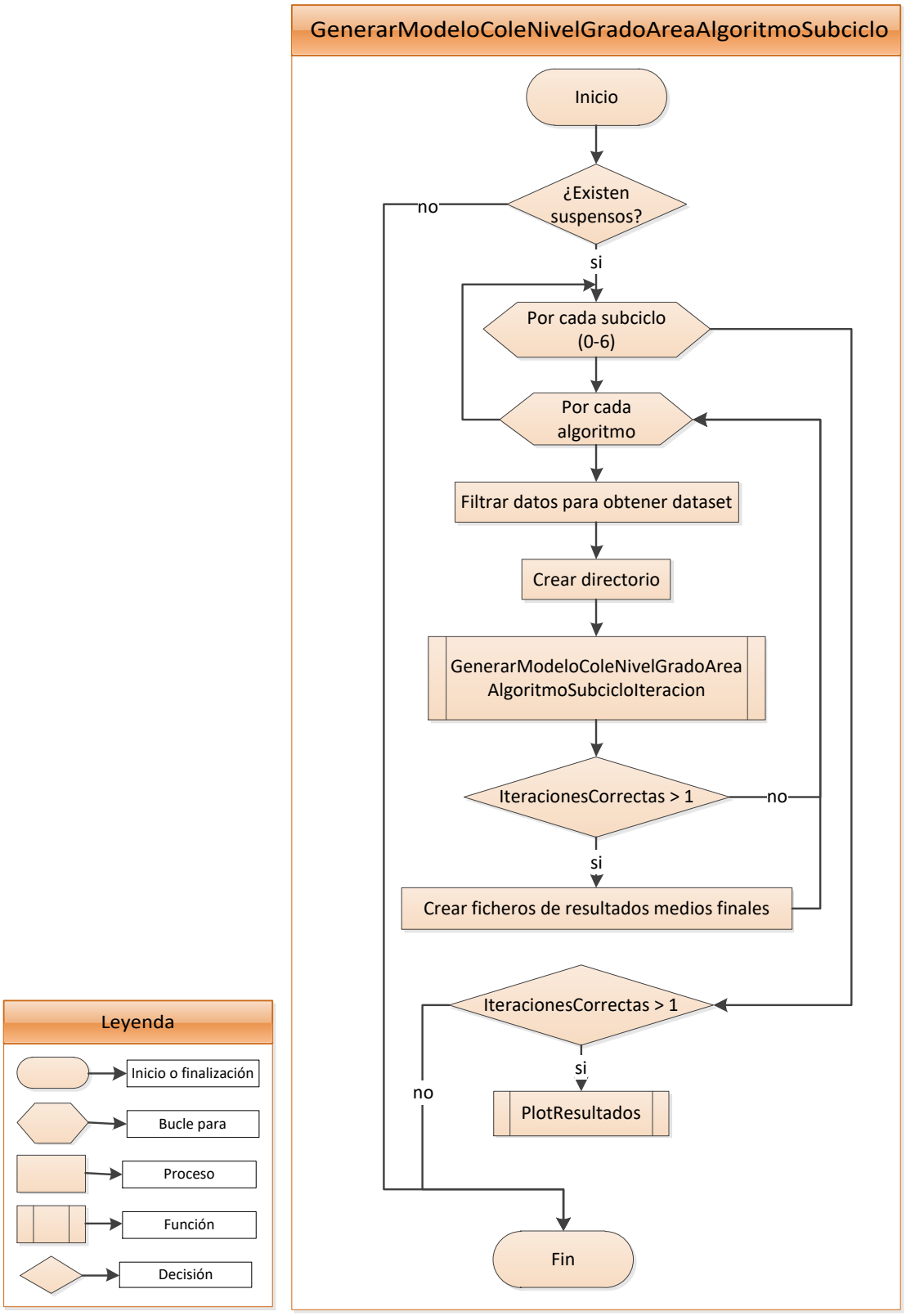

Figura 17. Proceso GenerarModeloColeNivelGradoAreaAlgoritmoSubciclo.

RESUMEN:

- Definición: Dado un dataset (normalmente enviado desde la función GenerarModeloColeNivelGradoArea) genera n modelos (donde n es el número de repeticiones) por cada par algoritmo/subciclo.
- Entrada:
  - datosCompletos: data.frame que contiene los datos.
  - algoritmosProbados: Algoritmos que vamos a lanzar en el experimento.
  - minSplit: Número mínimo de observaciones que deben existir en un nodo para intentar partirlo en dos ramas.
  - minBucket: Número mínimo de observaciones que debe haber en un nodo hoja. repeticiones: Número de iteraciones (modelos construidos) para cada Colegio/Nivel/Grado/Area/Subciclo.
  - notaCorte: Nota mínima para aprobar.
  - umbralNota: Se le sumará a la nota de corte para indicar que los alumnos por debajo de esa nota están en peligro de suspender la asignatura. Por ejemplo, si la nota de corte es 50 y el umbral 10, todos los que tengan una nota inferior a 60 se consideran en peligro.
- Salida: Ficheros con los resultados de los distintos modelos.
- Llamadas a funciones:
  - GenerarModeloColeNivelGradoAreaAlgoritmoSubcicloIteracion
  - PlotResultados

4.2.1.3 GenerarModeloColeNivelGradoAreaAlgoritmoSubcicloIteracion

Esta función recibe un dataset y crea modelos predictivos utilizando un algoritmo especificado en la llamada.

Realizará por defecto 10 modelos. Cada uno de ellos lo guardará en una carpeta diferente. Guardará en ficheros el árbol generado (si es aplicable) y varios estadísticos generados al crear el modelo (precisión, sensibilidad, especificidad...)

El modelo, dataset y las variables más importantes se guardarán en un fichero RData, las estadísticas se guardarán en ficheros .csv y los árboles se guardarán en formato pdf. De esta forma se podrá recuperar los resultados para facilitar la labor de los desarrolladores de software que podrán incluirlo fácilmente en sus aplicaciones.

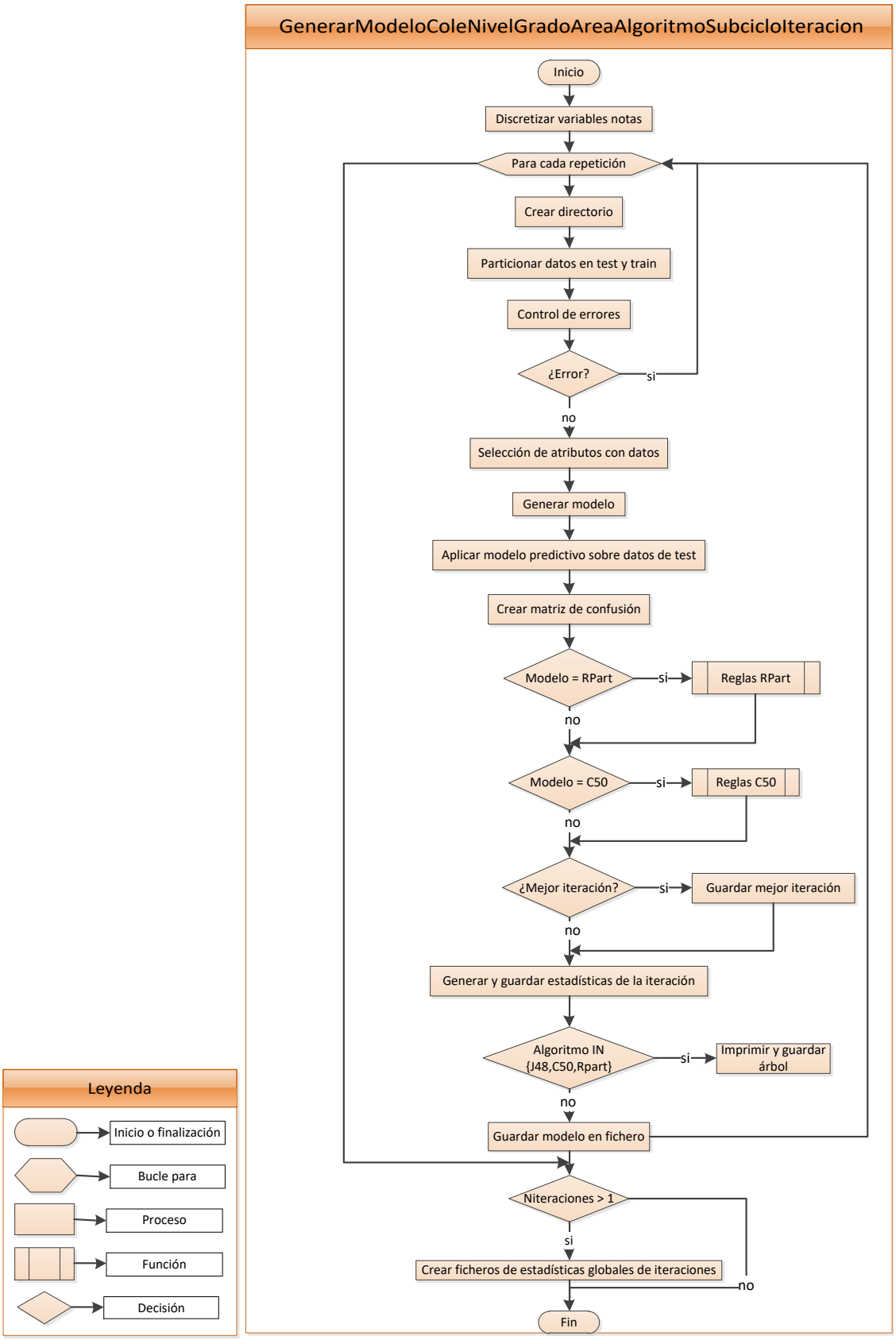

Figura 18. Proceso GenerarModeloColeNivelGradoAreaAlgoritmoSubcicloIteracion.

RESUMEN:

- Definición: Genera modelos utilizando los datos pasados como parámetro y seleccionando el campo Nota como target.
- Entrada:
  - datos: data.frame que contiene los datos.
  - split: Número mínimo de observaciones que deben existir en un nodo para intentar partirlo en dos ramas.
  - bucket: Número mínimo de observaciones que debe haber en un nodo hoja.
  - notaCorte: Nota mínima para aprobar.
  - repeticiones: Numero de iteraciones (modelos construidos) para cada Colegio/Nivel/Grado/Area/Subciclo
  - semilla: valor utilizado para reproducir los resultados. Si no se quieren obtener los mismos resultados con el mismo conjunto de datos se le deberá asignar un valor aleatorio.
  - tipo: Algoritmo de clasificación que se aplicará a los datos para la generación del modelo.
  - umbralNota: Se le sumará a la nota de corte para indicar que los alumnos por debajo de esa nota están en peligro de suspender la asignatura. Por ejemplo, si la nota de corte es 50 y el umbral 10, todos los que tengan una nota inferior a 60 se consideran en peligro.
  - nombreCarpeta: Nombre del directorio donde se almacenarán los ficheros generados con los resultados.
- Salida: Ficheros con los resultados de los modelos generados en las distintas repeticiones.

4.2.1.4 Otras funciones

En este apartado expondremos algunas funciones menores necesarias para el correcto funcionamiento de las anteriores.

4.2.1.4.1 ReglasRPart

RESUMEN:

- Definición: Extrae las reglas de un modelo creado con el algoritmo RPart.
- Entrada:
  - modelo: Modelo creado con RPart.
- Salida: Devuelve un dataframe con 4 columnas: Antecedente, Consecuente, Cobertura y Comentarios (Los comentarios estarán vacíos, pero se habilita para que el usuario pueda añadirlos).

4.2.1.4.2 ObtenerReglasC50

RESUMEN:

- Definición: Extrae las reglas de un modelo creado con el algoritmo C5.0.
- Entrada:
  - C50.model: Modelo creado con el algoritmo C5.0 y la opción rules=TRUE.
- Salida: Devuelve un dataframe con 4 columnas: Antecedente, Consecuente, Cobertura y Comentarios (Los comentarios estarán vacíos, pero se habilita para que el usuario pueda añadirlos).

4.2.1.4.3 PlotResultados

RESUMEN:

- Definición: Imprime comparativas de los modelos obtenidos con los distintos algoritmos con respecto a la precisión, especificidad y sensibilidad.
- Entrada:
  - resultados: data.frame que contiene los resultados de las tres medidas para cada subciclo y algoritmo.
  - ZeroR: Precisión que se alcanzaría si el modelo predictivo dijera que el resultado de la predicción es siempre la clase dominante.
- Salida: Imprime una imagen con 3 gráficas (una por cada medida).

4.2.2 Subconjuntos de datos que se utilizarán

En este apartado vamos a seleccionar varios subconjuntos de datos sobre los que realizaremos los experimentos.

Para preservar la confidencialidad de los datos, utilizaremos el identificador del colegio en lugar del nombre del mismo.

Las características deseadas en estos subconjuntos son:

- Cada subconjunto será de un colegio, en un nivel, grado y área específico. Se buscará diversificar estos atributos, pero no es la prioridad.
- Se le dará prioridad a los que tengan mayor número de instancias.
- Buscaremos los conjuntos con menor número de datos perdidos u omitidos. Para ello debemos establecer una medida del porcentaje de alumnos con datos en cada atributo.
- El conjunto debe presentar alumnos aprobados y alumnos en riesgo. En la medida de lo posible buscaremos conjuntos en los que la clase minoritaria tenga suficiente representación para evitar el fallo de todos los algoritmos predictivos.

En última instancia el equipo experto es el que tiene la última palabra para seleccionar el subconjunto que determine apropiado para el experimento.

Para ayudar a la selección crearemos un fichero excel con las distintas agrupaciones que se pueden hacer (colegio-nivel-grado-area), indicando algunos valores que nos servirán de guía para realizar la selección. Estos valores son los siguientes:

Los campos que identifican al subconjunto son:

- ColegioID: Identificador del colegio.
- NivelId: Identificador del nivel.
- GradoId: Identificador del grado.
- Area: Identificador del área.

Los atributos que nos muestran las características del subconjunto son:

- NCiclos: Número de ciclos que tiene.
- NumeroAlumnos: Número de instancias del subconjunto.
- Aprobados: Número de instancias con resultado aprobado.
- PAprobados: Porcentaje de instancias con resultado aprobado.
- Suspensos: Número de instancias con resultado suspenso.
- PSuspensos: Porcentaje de instancias con resultado susplenso.
- NSesionesTotalClases: Número de instancias con sesiones.

- PNSesionesTotalClases: Porcentaje de instancias con sesiones.
- NAccesosAPPTotalClases: Número de instancias con acceso a la aplicación.
- PNAccesosAPPTotalClases: Porcentaje de instancias con acceso a la aplicación.
- NAsitencias: Número de instancias con asistencias.
- PNAsitencias: Porcentaje de instancias con asistencias.
- NRetardos: Número de instancias con retardos.
- PNRetardos: Porcentaje de instancias con retardos.
- NFaltas: Número de instancias con faltas.
- PNFaltas: Porcentaje de instancias con faltas.
- NFaltasJust: Número de instancias con faltas justificadas.
- PNFaltasJust: Porcentaje de instancias con faltas justificadas.
- NExamenes: Número de instancias con exámenes realizados.
- PNExamenes: Porcentaje de instancias con exámenes realizados.
- NTareas: Número de instancias con tareas realizadas.
- PNTareas: Porcentaje de instancias con tareas realizadas.
- NTest: Número de instancias con test de Pleno realizados.
- PNTest: Porcentaje de instancias con test de Pleno realizados.
- NotaCicloAntArea: Número de instancias con nota en el ciclo anterior en la misma área.
- PNotaCicloAntArea: Porcentaje de instancias con nota en el ciclo anterior en la misma área.
- NotaCicloAntLeng: Número de instancias con nota de Lenguaje en el ciclo anterior.
- PNotaCicloAntLeng: Porcentaje de instancias con nota de Lenguaje en el ciclo anterior.
- NotaCicloAntMate: Número de instancias con nota de Matemáticas en el ciclo anterior.
- PNotaCicloAntMate: Porcentaje de instancias con nota de Matemáticas en el ciclo anterior.
- MismoNivelCicloAntTrue: Número de instancias que continúan en el mismo nivel que en el ciclo anterior.
- PMismoNivelCicloAntTrue: Porcentaje de instancias que continúan en el mismo nivel que en el ciclo anterior.
- MismoNivelCicloAntFalse: Número de instancias que han cambiado de nivel con respecto al ciclo anterior (por ejemplo, han cambiado de primaria a secundaria).
- PMismoNivelCicloAntFalse: Porcentaje de instancias que han cambiado de nivel con respecto al ciclo anterior (por ejemplo, han cambiado de primaria a secundaria).

- RepetidorTrue: Número de instancias que contienen alumnos que están repitiendo curso.
- PRepetidorTrue: Porcentaje de instancias que contienen alumnos que están repitiendo curso.
- RepetidorFalse: Número de instancias que contienen alumnos que podemos confirmar que no están repitiendo curso.
- PRepetidorFalse: Porcentaje de instancias que contienen alumnos que podemos confirmar que no están repitiendo curso.
- ColConDatos: Número de columnas con datos. Indica el número de campos no vacíos de los que acabamos de definir, es decir, cuantos de ellos tienen realmente datos. A pesar de que no se tiene en cuenta la importancia de cada atributo, nos interesa maximizar este número ya que indica que tenemos una mayor variedad de información (la cantidad viene determinada por el número de instancias y los porcentajes de instancias con datos en cada atributo).

Para realizar la exploración debemos determinar un subciclo. Hemos seleccionado el intermedio (subciclo 3) debido a que, de esta forma, obtenemos un compromiso entre el número de datos que puede tener y la precocidad.

Una vez obtenidos los resultados, omitiremos las filas que no tengan suspensos o no tengan áreas y lo ordenaremos (en orden descendente) por número de ciclos, número de columnas con datos y número de suspensos (NCiclos, ColConDatos y Suspensos).

SUBCONJUNTOS SELECCIONADOS:

RESUMEN DE SUBCONJUNTOS SELECCIONADOS			
ColegioID	NivelId	GradoId	Area
2100	24	50	Matematicas
1047	18	25	Lenguaje
2336	18	27	Tecnologia e Informatica
2202	36	103	Geografia
2053	24	50	Ciencias Sociales
2030	22	44	Matematicas
2030	22	43	Ingles
3170	23	48	Ciencias Naturales

Figura 19. Resumen de subconjuntos seleccionados

Atributo	Exper. 1	Exper. 2	Exper. 3	Exper. 4
ColegioID	2100	1047	2336	2202
NivelId	24	18	18	36
GradoId	50	25	27	103
Area	Matematicas	Lenguaje	Tecnologia e Informatica	Geografia
NCiclos	2	3	2	2
NumeroAlumnos	438	152	98	120
Aprobados	317	147	97	101
PAprobados	0,7237	0,9671	0,9898	0,8417
Suspensos	121	5	1	19
PSuspensos	0,2763	0,0329	0,0102	0,1583
NSesionesTotalClases	409	98	97	119
PNSesionesTotalClases	0,9338	0,6447	0,9898	0,9917
NAccesosAPPTotalClases	0	4	0	0
PNAccesosAPPTotalClases	0	0,0263	0	0
NAsitencias	0	51	98	34
PNAsitencias	0	0,3355	1	0,2833
NRetardos	0	1	12	0
PNRetardos	0	0,0066	0,1224	0
NFaltas	0	34	48	2
PNFaltas	0	0,2237	0,4898	0,0167
NFaltasJust	0	10	1	0
PNFaltasJust	0	0,0658	0,0102	0
NExamenes	61	0	52	71
PNExamenes	0,1393	0	0,5306	0,5917
NTareas	104	152	52	87
PNTareas	0,2374	1	0,5306	0,725
NTest	68	0	0	0
PNTest	0,1553	0	0	0
NotaCicloAntArea	215	1	3	0
PNotaCicloAntArea	0,4909	0,0066	0,0306	0
NotaCicloAntLeng	215	1	3	46
PNotaCicloAntLeng	0,4909	0,0066	0,0306	0,3833
NotaCicloAntMate	215	1	2	46
PNotaCicloAntMate	0,4909	0,0066	0,0204	0,3833
MismoNivelCicloAntTrue	0	0	0	0
PMismoNivelCicloAntTrue	0	0	0	0
MismoNivelCicloAntFalse	215	74	3	46
PMismoNivelCicloAntFalse	0,4909	0,4868	0,0306	0,3833
RepetidorTrue	0	0	0	0
PRepetidorTrue	0	0	0	0
RepetidorFalse	215	74	3	46
PRepetidorFalse	0,4909	0,4868	0,0306	0,3833
ColConDatos	19	25	25	19

Figura 20. Características de los subconjuntos seleccionados 1.

Atributo	Exper. 5	Exper. 6	Exper. 7	Exper. 8
ColegioID	2053	2030	2030	3170
NivelId	24	22	22	23
GradoId	50	44	43	48
Area	Ciencias Sociales	Matematicas	Ingles	Ciencias Naturales
NCiclos	3	3	3	2
NumeroAlumnos	745	235	215	173
Aprobados	737	232	211	171
PAprobados	0,9893	0,9872	0,9814	0,9884
Suspensos	8	3	4	2
PSuspensos	0,0107	0,0128	0,0186	0,0116
NSesionesTotalClases	163	164	188	173
PNSesionesTotalClases	0,2188	0,6979	0,8744	1
NAccesosAPPTotalClases	1	1	3	0
PNAccesosAPPTotalClases	0,0013	0,0043	0,014	0
NAsitencias	112	235	189	85
PNAsitencias	0,1503	1	0,8791	0,4913
NRetardos	1	15	1	1
PNRetardos	0,0013	0,0638	0,0047	0,0058
NFaltas	9	76	41	15
PNFaltas	0,0121	0,3234	0,1907	0,0867
NFaltasJust	1	31	8	0
PNFaltasJust	0,0013	0,1319	0,0372	0
NExamenes	44	71	25	0
PNExamenes	0,0591	0,3021	0,1163	0
NTareas	516	164	188	130
PNTareas	0,6926	0,6979	0,8744	0,7514
NTest	0	0	0	0
PNTest	0	0	0	0
NotaCicloAntArea	203	105	88	40
PNotaCicloAntArea	0,2725	0,4468	0,4093	0,2312
NotaCicloAntLeng	200	105	88	42
PNotaCicloAntLeng	0,2685	0,4468	0,4093	0,2428
NotaCicloAntMate	203	105	88	44
PNotaCicloAntMate	0,2725	0,4468	0,4093	0,2543
MismoNivelCicloAntTrue	0	0	0	0
PMismoNivelCicloAntTrue	0	0	0	0
MismoNivelCicloAntFalse	204	105	88	44
PMismoNivelCicloAntFalse	0,2738	0,4468	0,4093	0,2543
RepetidorTrue	0	0	0	0
PRepetidorTrue	0	0	0	0
RepetidorFalse	204	105	88	44
PRepetidorFalse	0,2738	0,4468	0,4093	0,2543
ColConDatos	27	27	27	21

Figura 21. Características de los subconjuntos seleccionados 2.

4.3 Pruebas

Vamos a centrarnos en los resultados finales para simplificar la lectura de las pruebas.

Los gráficos que muestran la precisión tendrán dos características especiales:

- En el eje Y tan sólo mostraremos el intervalor y = [0.5,1] omitiendo la mitad inferior. Esto se debe a que en ningún caso nos va a interesar una predicción inferior al 50%.
- Se trazará una línea negra paralela al eje X en y=ZeroR donde ZeroR es la predicción que conseguimos basándonos únicamente en el target (variable objetivo de la predicción) sin tener en cuenta el resto de predictores (predice la clase mayoritaria). Cualquier predicción por debajo de este valor no será de utilidad.

4.3.1 Experimento 1

RESUMEN DEL DATASET	
PAIS:	COLOMBIA
COLEGIOID:	2100
NIVEL:	MEDIA (24)
GRADO:	DÉCIMO MEDIA (50)
AREA:	MATEMATICAS
Nº DE INSTANCIAS:	3204
% ALUMNOS EN RIESGO:	38.79% (1243 INSTANCIAS)

Tabla 10. Resumen del dataset del experimento 1.

Resultados (medias de las 10 ejecuciones)

SUBCICLO 0			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,8290	0,0884	0,9873
C50	0,8235	0,0000	1,0000
J48	0,8235	0,0000	1,0000
JRip	0,8235	0,0000	1,0000
NaiveBayes	0,8023	0,2387	0,9226

SUBCICLO 1			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,7503	0,6044	0,8467
C50	0,6005	0,0000	1,0000
J48	0,6005	0,0000	1,0000
JRip	0,6005	0,0000	1,0000
NaiveBayes	0,5636	0,1682	0,8277

SUBCICLO 2			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,6944	0,3486	0,9130
C50	0,6109	0,0000	1,0000
J48	0,6109	0,0000	1,0000
JRip	0,6109	0,0000	1,0000
NaiveBayes	0,6100	0,3266	0,7899

SUBCICLO 3			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,7159	0,4884	0,8680
C50	0,6455	0,5489	0,7101
J48	0,5992	0,0000	1,0000
JRip	0,5992	0,0000	1,0000
NaiveBayes	0,5841	0,3467	0,7426

SUBCICLO 4			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,7053	0,5778	0,7914
C50	0,6439	0,5567	0,7027
J48	0,5992	0,0000	1,0000
JRip	0,5992	0,0000	1,0000
NaiveBayes	0,5015	0,3362	0,6120

SUBCICLO 5			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,8732	0,7425	0,9512
C50	0,8444	0,5921	0,9938
J48	0,6288	0,0000	1,0000
JRip	0,6288	0,0000	1,0000
NaiveBayes	0,8190	0,7358	0,8690

SUBCICLO 6			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9876	0,9794	0,9928
C50	0,9830	0,9578	0,9979
J48	0,6288	0,0000	1,0000
JRip	0,6288	0,0000	1,0000
NaiveBayes	0,9098	0,9490	0,8870

Tabla 11. Resultados del experimento 1.

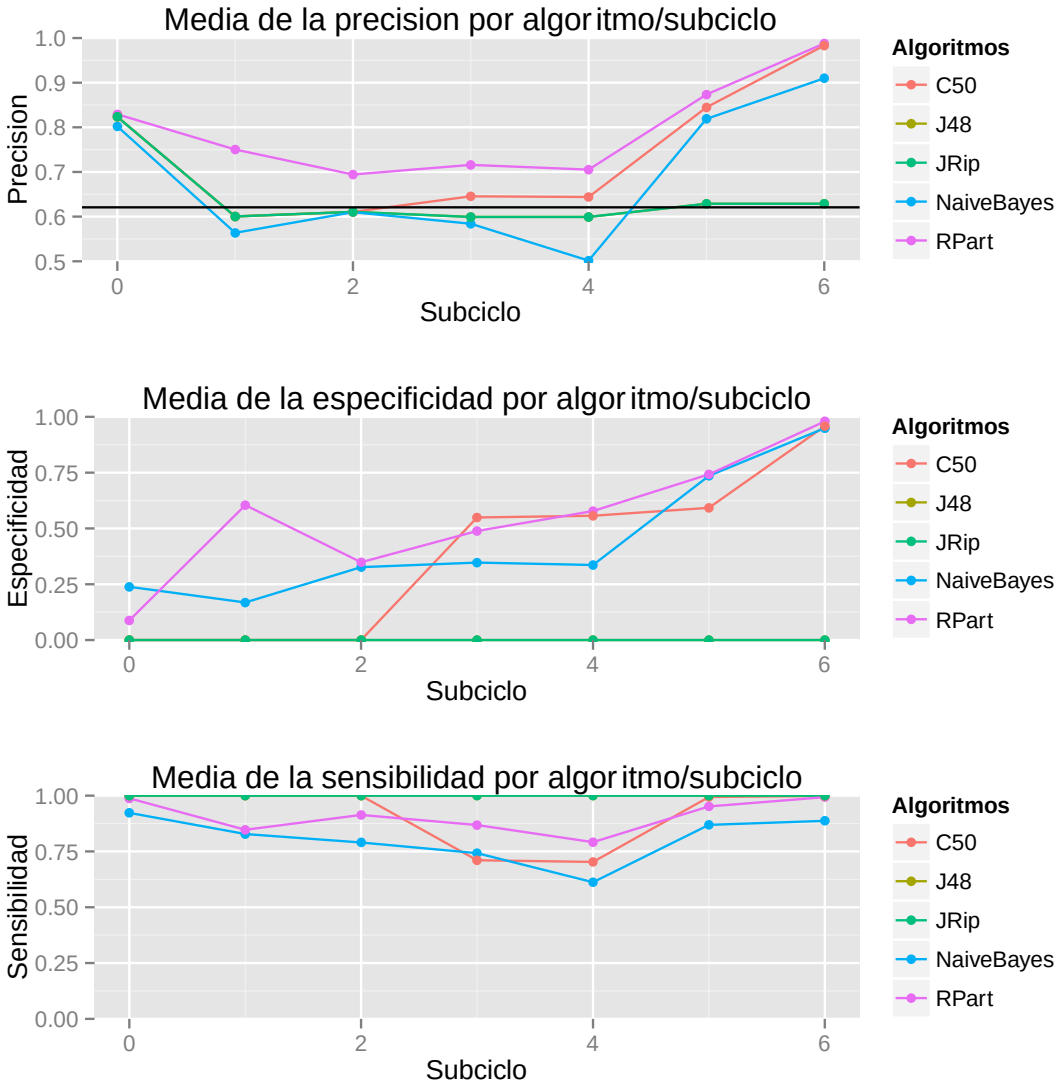

Figura 22. Comparativa de algoritmos del experimento 1.

4.3.2 Experimento 2

RESUMEN DEL DATASET	
PAIS:	MÉJICO
COLEGIOID:	1047
NIVEL:	SECUNDARIA (18)
GRADO:	PRIMERO SECUNDARIA (25)
AREA:	LENGUAJE
Nº DE INSTANCIAS:	1064
% ALUMNOS EN RIESGO:	17.76% (189 INSTANCIAS)

Tabla 12. Resumen del dataset del experimento 2.

Resultados (medias de las 10 ejecuciones)

SUBCICLO 0			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,8484	NA	1,0000
C50	0,8484	NA	1,0000
J48	0,8484	NA	1,0000
JRip	0,8484	NA	1,0000
NaiveBayes	0,8576	NA	1,0000

SUBCICLO 1			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9120	0,6895	0,9629
C50	0,8896	0,3271	1,0000
J48	0,8379	0,0000	1,0000
JRip	0,8379	0,0000	1,0000
NaiveBayes	0,8217	0,5571	0,8760

SUBCICLO 2			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9152	0,4837	1,0000
C50	0,9109	0,4417	1,0000
J48	0,8413	0,0000	1,0000
JRip	0,8413	0,0000	1,0000
NaiveBayes	0,8882	0,6709	0,9338

SUBCICLO 3			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9391	0,7395	0,9791
C50	0,9152	0,4716	1,0000
J48	0,8413	0,0000	1,0000
JRip	0,8413	0,0000	1,0000
NaiveBayes	0,8935	0,7086	0,9317

SUBCICLO 4			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9152	0,4716	1,0000
C50	0,9457	0,6786	0,9949
J48	0,8413	0,0000	1,0000
JRip	0,8413	0,0000	1,0000
NaiveBayes	0,8819	0,7131	0,9147

SUBCICLO 5			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9696	0,8051	1,0000
C50	0,9696	0,8051	1,0000
J48	0,8413	0,0000	1,0000
JRip	0,8413	0,0000	1,0000
NaiveBayes	0,9035	0,7129	0,9437

SUBCICLO 6			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9717	0,8482	1,0000
C50	0,9391	0,6291	1,0000
J48	0,8413	0,0000	1,0000
JRip	0,8413	0,0000	1,0000
NaiveBayes	0,8991	0,7706	0,9265

Tabla 13. Resultados del experimento 2.

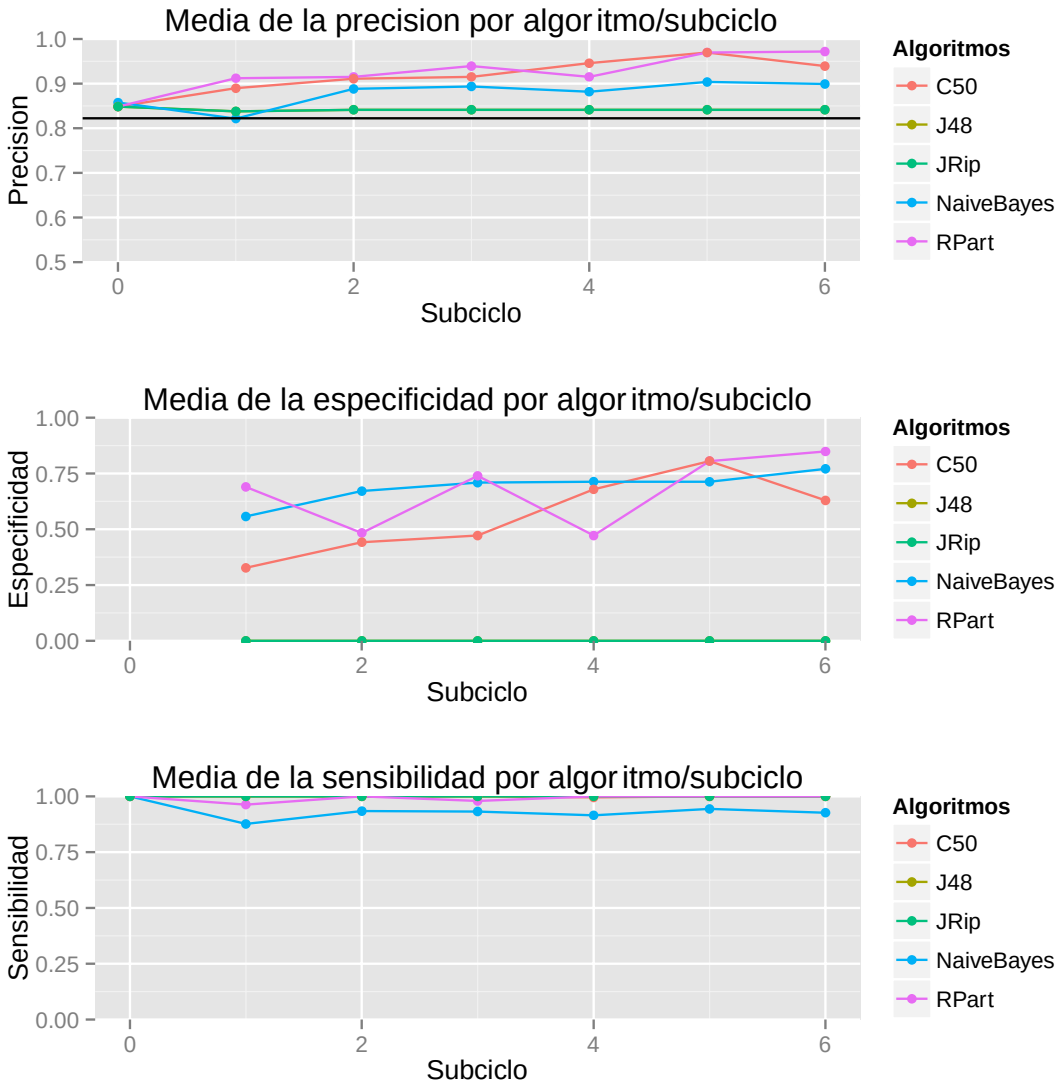

Figura 23. Comparativa de algoritmos del experimento 2.

4.3.3 Experimento 3

RESUMEN DEL DATASET	
PAIS:	MÉJICO
COLEGIOID:	2336
NIVEL:	SECUNDARIA (18)
GRADO:	TERCERO SECUNDARIA (27)
AREA:	TECNOLOGIA E INFORMATICA
Nº DE INSTANCIAS:	686
% ALUMNOS EN RIESGO:	29.59% (203 INSTANCIAS)

Tabla 14. Resumen del dataset del experimento 3.

Resultados (medias de las 10 ejecuciones)

SUBCICLO 0			
Algoritmos	Precisión	Especificidad	Sensibilidad

SUBCICLO 1			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9187	NA	1,0000
C50	0,9187	NA	1,0000
J48	0,9187	NA	1,0000
JRip	0,9187	NA	1,0000
NaiveBayes	0,4571	NA	NA

SUBCICLO 2			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,8433	0,4357	0,9147
C50	0,7900	0,7112	0,8049
J48	0,8533	0,0000	1,0000
JRip	0,8533	0,0000	1,0000
NaiveBayes	0,7269	0,1436	1,0000

SUBCICLO 3			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9300	0,6519	0,9793
C50	0,8300	0,6519	0,8622
J48	0,8533	0,0000	1,0000
JRip	0,8533	0,0000	1,0000
NaiveBayes	0,7395	0,1560	0,9900

SUBCICLO 4			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,8800	0,5226	0,9411
C50	0,8467	0,4940	0,9102
J48	0,8533	0,0000	1,0000
JRip	0,8533	0,0000	1,0000
NaiveBayes	0,7237	0,1810	0,9569

SUBCICLO 5			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,8800	0,4690	0,9529
C50	0,8433	0,3324	0,9291
J48	0,8533	0,0000	1,0000
JRip	0,8533	0,0000	1,0000
NaiveBayes	0,7302	0,3810	0,8906

SUBCICLO 6			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9233	0,7705	0,9487
C50	0,9233	0,7705	0,9487
J48	0,8533	0,0000	1,0000
JRip	0,8533	0,0000	1,0000
NaiveBayes	0,7318	1,0000	0,6707

Tabla 15. Resultados del experimento 3.

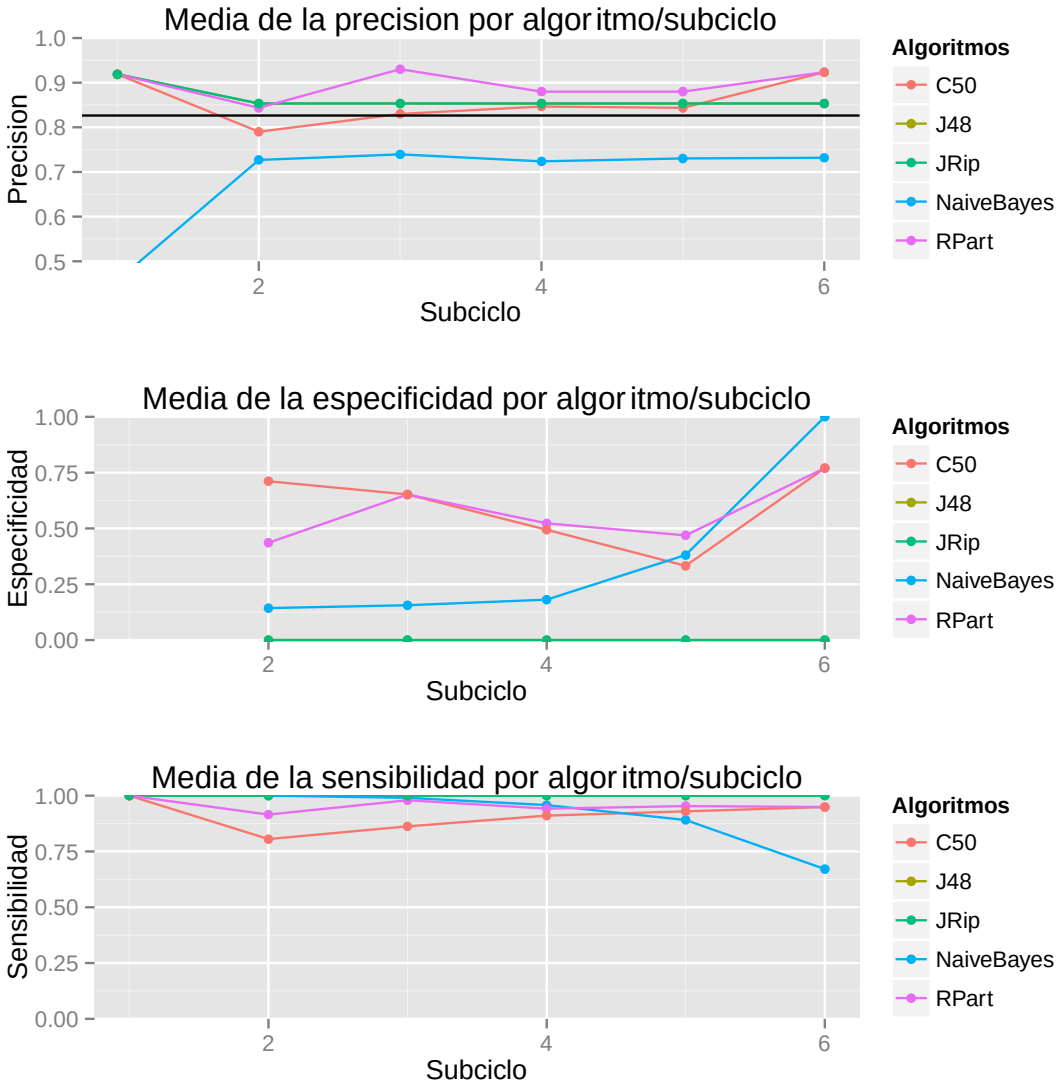

Figura 24. Comparativa de algoritmos del experimento 3.

4.3.4 Experimento 4

RESUMEN DEL DATASET	
PAIS:	ARGENTINA
COLEGIOID:	2202
NIVEL:	SECUNDARIA (36)
GRADO:	SEGUNDO SECUNDARIA (103)
AREA:	GEOGRAFIA
Nº DE INSTANCIAS:	840
% ALUMNOS EN RIESGO:	42.5% (357 INSTANCIAS)

Tabla 16. Resumen del dataset del experimento 4.

Resultados (medias de las 10 ejecuciones)

SUBCICLO 0			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,7087	0,0000	1,0000
C50	0,7087	0,0000	1,0000
J48	0,7087	0,0000	1,0000
JRip	0,7087	0,0000	1,0000
NaiveBayes	0,7087	0,0000	1,0000

SUBCICLO 1			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,6102	0,1353	0,9263
C50	0,5989	0,0000	1,0000
J48	0,5989	0,0000	1,0000
JRip	0,5989	0,0000	1,0000
NaiveBayes	0,5989	0,0000	1,0000

SUBCICLO 2			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,8260	0,7248	0,8951
C50	0,7894	0,5779	0,9215
J48	0,6105	0,0000	1,0000
JRip	0,6105	0,0000	1,0000
NaiveBayes	0,5976	0,7774	0,4758

SUBCICLO 3			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,8444	0,7518	0,9083
C50	0,8167	0,6767	0,9038
J48	0,6056	0,0000	1,0000
JRip	0,6056	0,0000	1,0000
NaiveBayes	0,6727	0,7652	0,6169

SUBCICLO 4			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9278	0,8895	0,9540
C50	0,7556	0,5635	0,8816
J48	0,6056	0,0000	1,0000
JRip	0,6056	0,0000	1,0000
NaiveBayes	0,7122	0,8096	0,6494

SUBCICLO 5			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9389	0,9152	0,9536
C50	0,7556	0,5635	0,8816
J48	0,6056	0,0000	1,0000
JRip	0,6056	0,0000	1,0000
NaiveBayes	0,7122	0,8096	0,6494

SUBCICLO 6			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9694	0,9740	0,9627
C50	0,7556	0,5635	0,8816
J48	0,6056	0,0000	1,0000
JRip	0,6056	0,0000	1,0000
NaiveBayes	0,7123	0,8217	0,6386

Tabla 17. Resultados del experimento 4.

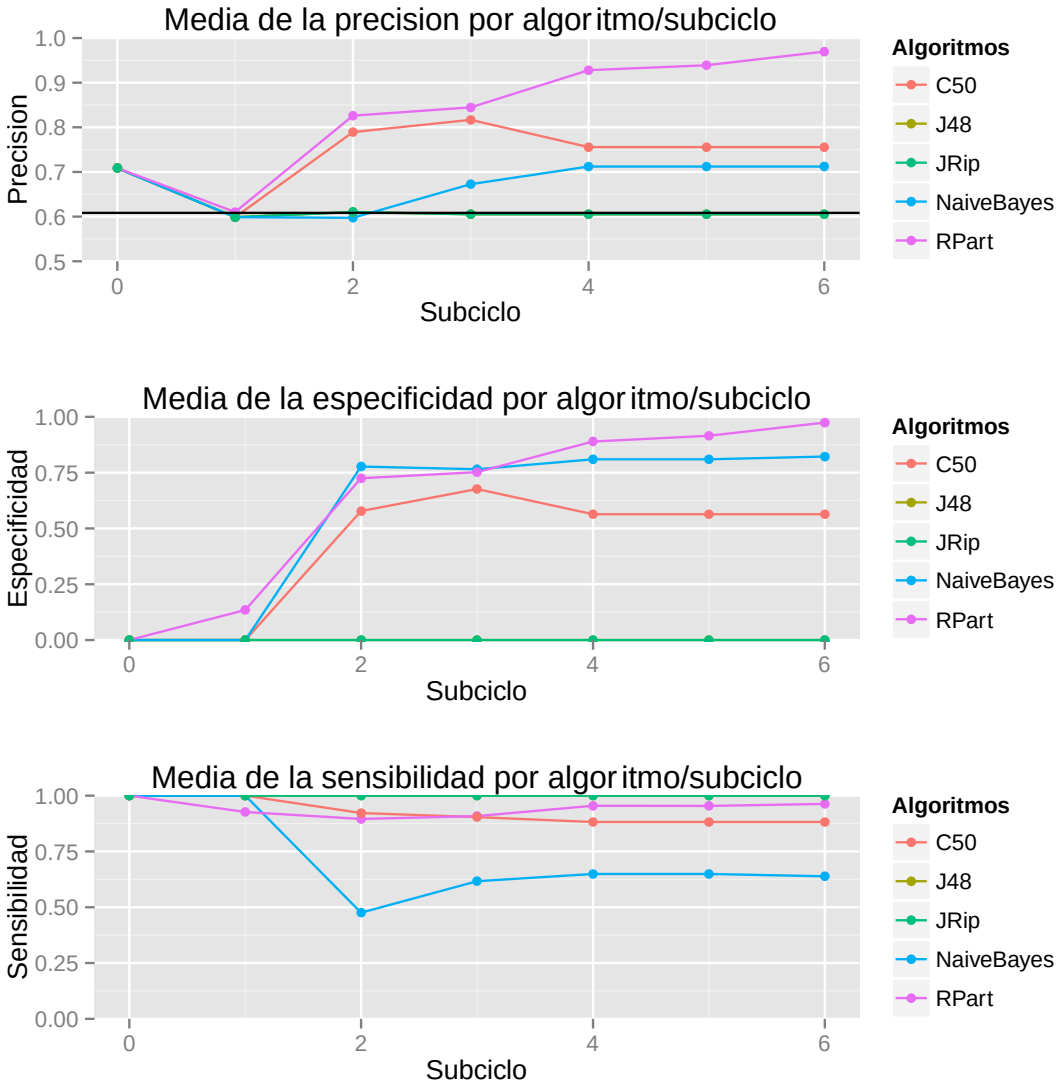

Figura 25. Comparativa de algoritmos del experimento 4.

4.3.5 Experimento 5

RESUMEN DEL DATASET	
PAIS:	COLOMBIA
COLEGIOID:	2253
NIVEL:	MEDIA (24)
GRADO:	DÉCIMO MEDIA (103)
AREA:	CIENCIAS SOCIALES
Nº DE INSTANCIAS:	5215
% ALUMNOS EN RIESGO:	7.11% (371 INSTANCIAS)

Tabla 18. Resumen del dataset del experimento 5.

Resultados (medias de las 10 ejecuciones)

SUBCICLO 0			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9212	0	1
C50	0,9212	0	1
J48	0,9212	0	1
JRip	0,9212	0	1
NaiveBayes	0,9212	0	1

SUBCICLO 1			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9082	0,243	0,9879
C50	0,8939	0	1
J48	0,8939	0	1
JRip	0,8939	0	1
NaiveBayes	0,9032	0,2653	0,9795

SUBCICLO 2			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,94	0,4633	0,9936
C50	0,9117	0,1069	1
J48	0,901	0	1
JRip	0,901	0	1
NaiveBayes	0,7554	0,652	0,7683

SUBCICLO 3			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9462	0,6376	0,9788
C50	0,9359	0,4259	0,9888
J48	0,9073	0	1
JRip	0,9073	0	1
NaiveBayes	0,7414	0,5904	0,7594

SUBCICLO 4			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,961	0,6265	0,9898
C50	0,9502	0,614	0,9791
J48	0,9235	0	1
JRip	0,9235	0	1
NaiveBayes	0,8246	0,6437	0,8417

SUBCICLO 5			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9701	0,6441	0,9966
C50	0,967	0,6563	0,9923
J48	0,9272	0	1
JRip	0,9272	0	1
NaiveBayes	0,821	0,6716	0,836

SUBCICLO 6			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,979	0,8303	0,9909
C50	0,9737	0,8467	0,9841
J48	0,9272	0	1
JRip	0,9272	0	1
NaiveBayes	0,8268	0,6694	0,8429

Tabla 19. Resultados del experimento 5.

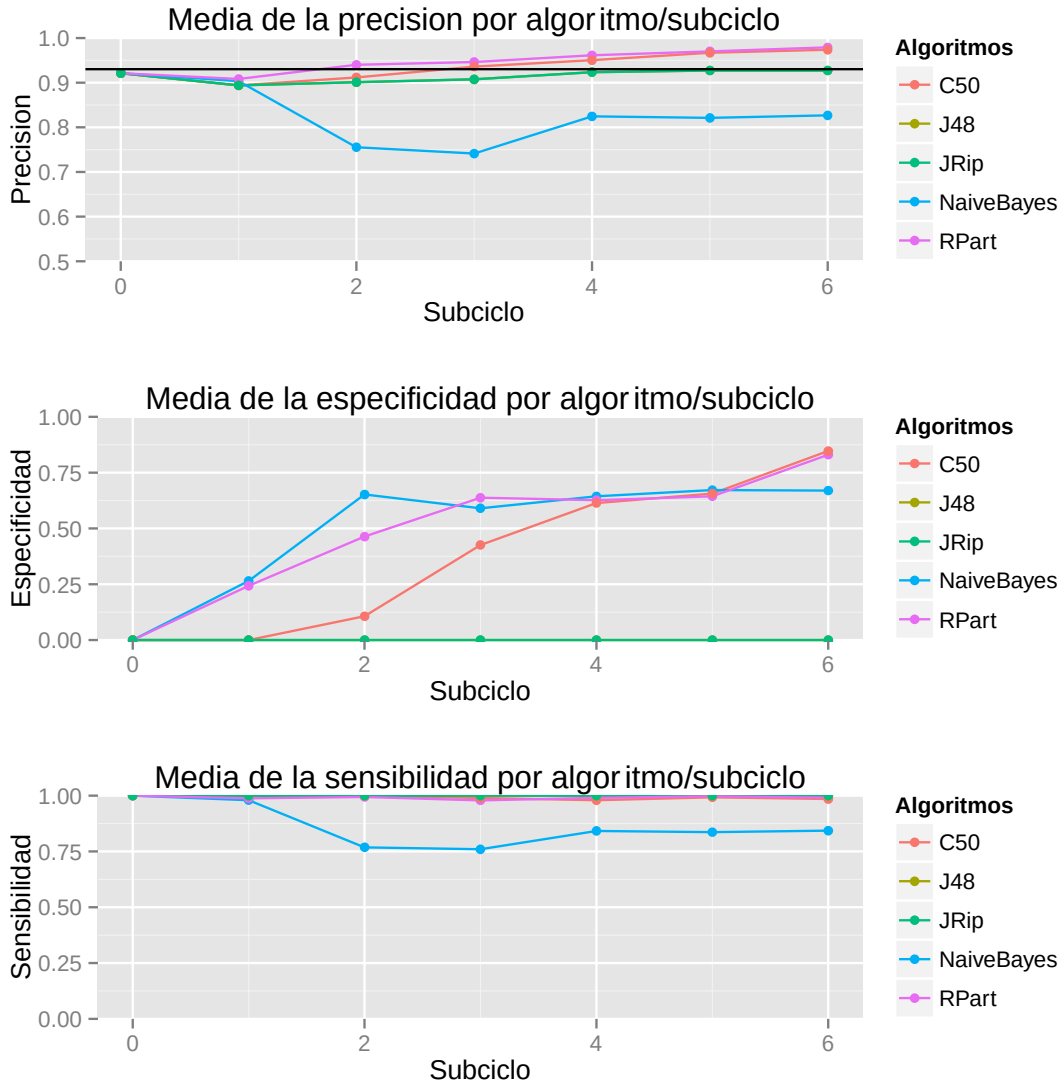

Figura 26. Comparativa de algoritmos del experimento 5.

4.3.6 Experimento 6

RESUMEN DEL DATASET	
PAIS:	COLOMBIA
COLEGIOID:	2030
NIVEL:	PRIMARIA (22)
GRADO:	CUARTA PRIMARIA(44)
AREA:	MATEMATICAS
Nº DE INSTANCIAS:	1685
% ALUMNOS EN RIESGO:	11.87% (200 INSTANCIAS)

Tabla 20. Resumen del dataset del experimento 6.

Resultados (medias de las 10 ejecuciones)

SUBCICLO 0			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,8395	0	1
C50	0,8395	0	1
J48	0,8395	0	1
JRip	0,8395	0	1
NaiveBayes	0,8307	0	1

SUBCICLO 1			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9255	0,5696	0,9779
C50	0,8687	0	1
J48	0,8687	0	1
JRip	0,8687	0	1
NaiveBayes	0,901	0,5097	0,9751

SUBCICLO 2			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9507	0,8115	0,9676
C50	0,8761	0	1
J48	0,8761	0	1
JRip	0,8761	0	1
NaiveBayes	0,9123	0,8703	0,9186

SUBCICLO 3			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9479	0,757	0,974
C50	0,8761	0	1
J48	0,8761	0	1
JRip	0,8761	0	1
NaiveBayes	0,925	0,7265	0,9546

SUBCICLO 4			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9592	0,8418	0,9744
C50	0,931	0,5459	0,9872
J48	0,8761	0	1
JRip	0,8761	0	1
NaiveBayes	0,9084	0,7017	0,9377

SUBCICLO 5			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9481	0,7686	0,9695
C50	0,9442	0,8412	0,9561
J48	0,8883	0	1
JRip	0,8883	0	1
NaiveBayes	0,8747	0,8679	0,8752

SUBCICLO 6			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9649	0,8979	0,9723
C50	0,9416	0,747	0,9649
J48	0,8883	0	1
JRip	0,8883	0	1
NaiveBayes	0,889	0,8679	0,8912

Tabla 21. Resultados del experimento 6.

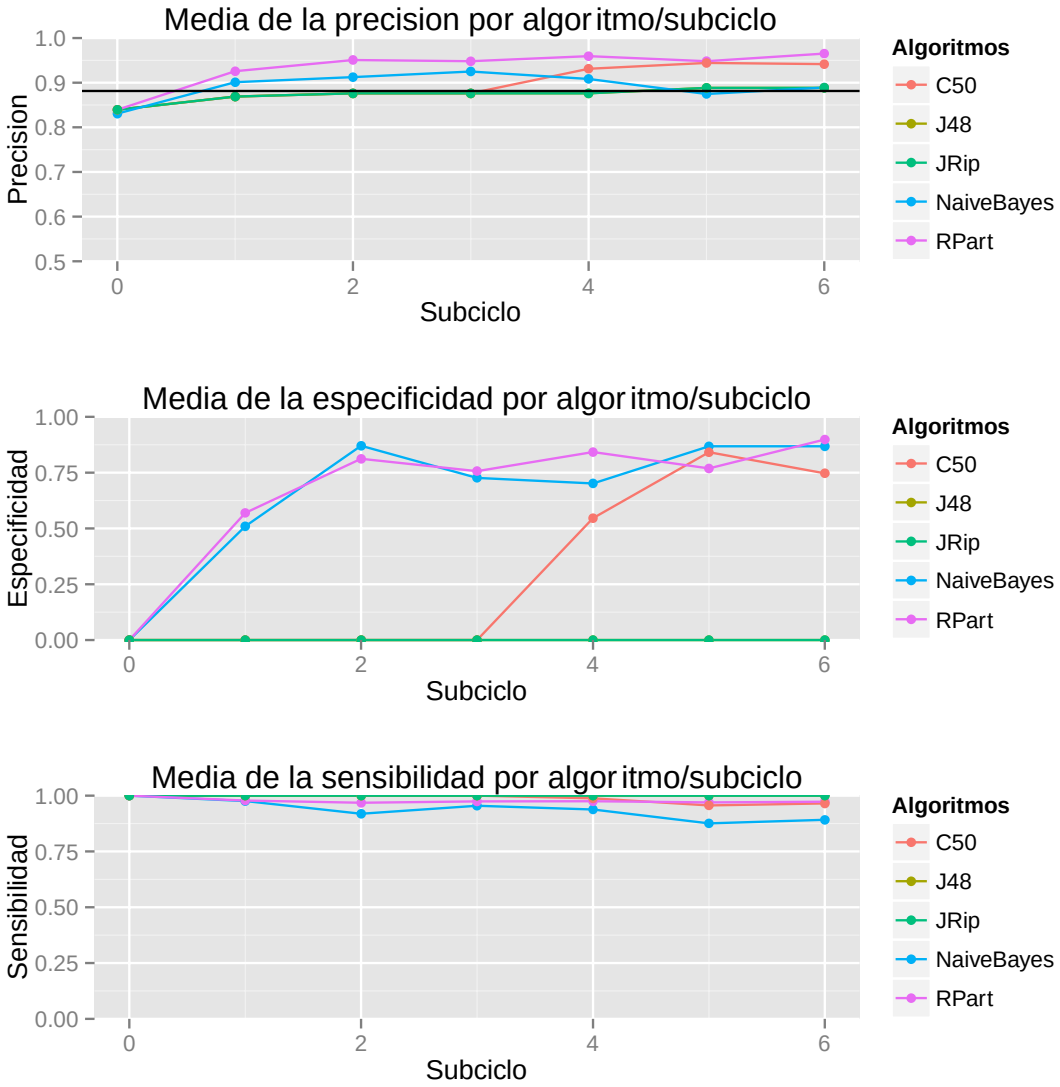

Figura 27. Comparativa de algoritmos del experimento 6.

4.3.7 Experimento 7

RESUMEN DEL DATASET	
PAIS:	COLOMBIA
COLEGIOID:	2030
NIVEL:	PRIMARIA (22)
GRADO:	TERCERO PRIMARIA (43)
AREA:	INGLES
Nº DE INSTANCIAS:	1505
% ALUMNOS EN RIESGO:	7.44% (112 INSTANCIAS)

Tabla 22. Resumen del dataset del experimento 7.

Resultados (medias de las 10 ejecuciones)

SUBCICLO 0			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,8979	0	1
C50	0,8979	0	1
J48	0,8979	0	1
JRip	0,8979	0	1
NaiveBayes	0,9008	0,02	1

SUBCICLO 1			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9428	0,4783	0,9883
C50	0,915	0	1
J48	0,915	0	1
JRip	0,915	0	1
NaiveBayes	0,8944	0,2967	0,9541

SUBCICLO 2			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9706	0,6758	0,9934
C50	0,9567	0,4175	1
J48	0,9258	0	1
JRip	0,9258	0	1
NaiveBayes	0,9243	0,6592	0,9507

SUBCICLO 3			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9814	0,8117	0,9968
C50	0,9258	0	1
J48	0,9258	0	1
JRip	0,9258	0	1
NaiveBayes	0,9253	0,6192	0,9562

SUBCICLO 4			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9631	0,6317	0,9883
C50	0,9262	0	1
J48	0,9262	0	1
JRip	0,9262	0	1
NaiveBayes	0,9116	0,4867	0,9505

SUBCICLO 5			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9631	0,5058	1
C50	0,9554	0,6592	0,9786
J48	0,9262	0	1
JRip	0,9262	0	1
NaiveBayes	0,9133	0,435	0,9586

SUBCICLO 6			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9615	0,55	0,995
C50	0,9615	0,6592	0,9852
J48	0,9262	0	1
JRip	0,9262	0	1
NaiveBayes	0,9113	0,435	0,959

Tabla 23. Resultados del experimento 7.

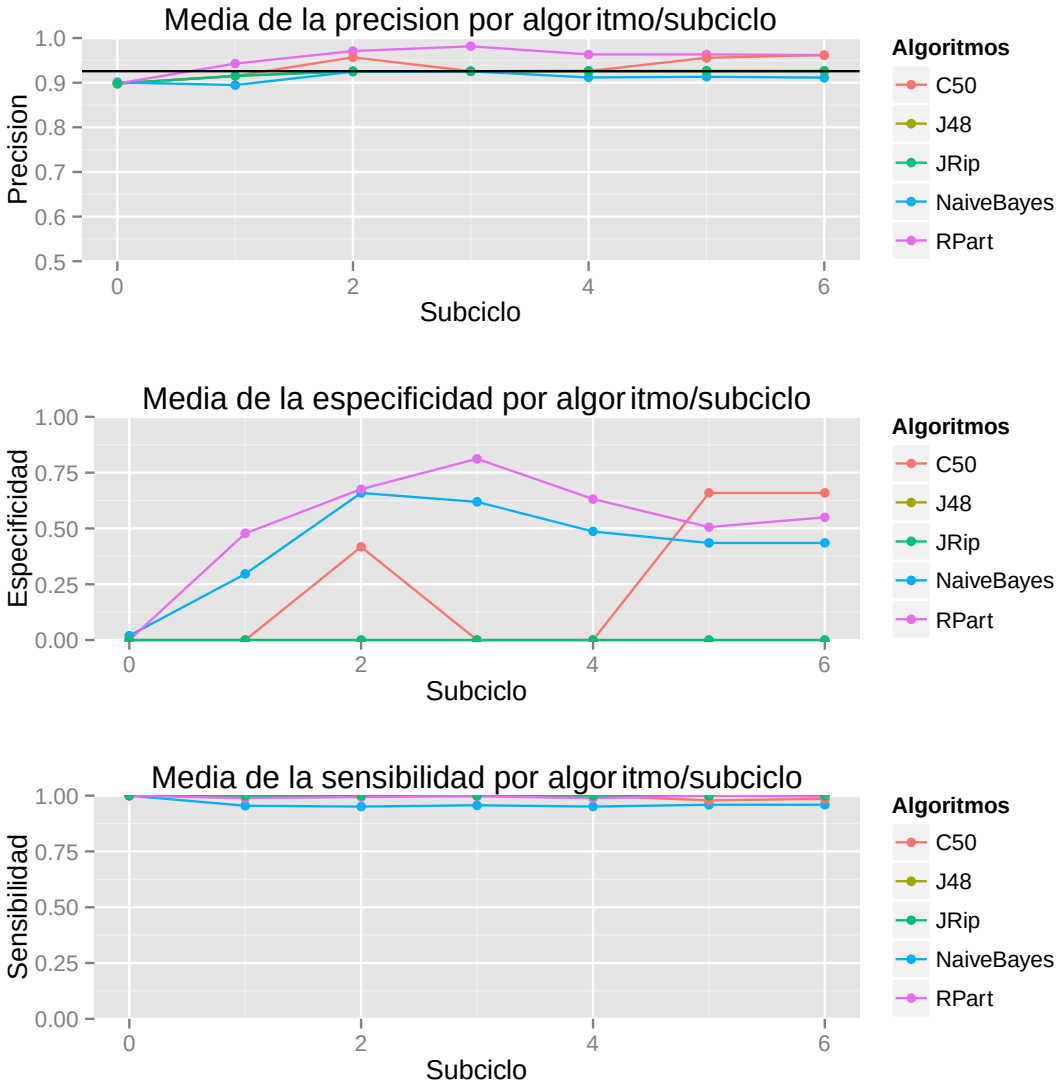

Figura 28. Comparativa de algoritmos del experimento 7.

4.3.8 Experimento 8

RESUMEN DEL DATASET	
PAIS:	COLOMBIA
COLEGIOID:	3170
NIVEL:	BACHILLERATO (23)
GRADO:	OCTAVO SECUNDARIA (48)
AREA:	CIENCIAS NATURALES
Nº DE INSTANCIAS:	1211
% ALUMNOS EN RIESGO:	25.43% (308 INSTANCIAS)

Tabla 24. Resumen del dataset del experimento 8.

Resultados (medias de las 10 ejecuciones)

SUBCICLO 0			
Algoritmos	Precisión	Especificidad	Sensibilidad

SUBCICLO 1			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9439	0,8157	1
C50	0,906	0,8417	0,9345
J48	0,7071	0	1
JRip	0,7071	0	1
NaiveBayes	0,7211	0,0659	1

SUBCICLO 2			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,8914	0,7911	0,9301
C50	0,8683	0,5204	1
J48	0,7248	0	1
JRip	0,7248	0	1
NaiveBayes	0,8658	0,5252	0,9611

SUBCICLO 3			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9481	0,8756	0,9765
C50	0,8712	0,5262	1
J48	0,2731	1	0
JRip	0,7269	0	1
NaiveBayes	0,8542	0,454	0,9636

SUBCICLO 4			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9519	0,9224	0,9636
C50	0,9231	0,7145	1
J48	0,2731	1	0
JRip	0,7269	0	1
NaiveBayes	0,8543	0,5649	0,9333

SUBCICLO 5			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9423	0,8884	0,9636
C50	0,9365	0,9088	0,9471
J48	0,2731	1	0
JRip	0,7269	0	1
NaiveBayes	0,8543	0,5649	0,9333

SUBCICLO 6			
Algoritmos	Precisión	Especificidad	Sensibilidad
RPart	0,9423	0,8884	0,9636
C50	0,9365	0,9088	0,9471
J48	0,2731	1	0
JRip	0,7269	0	1
NaiveBayes	0,8543	0,5649	0,9333

Tabla 25. Resultados del experimento 8.

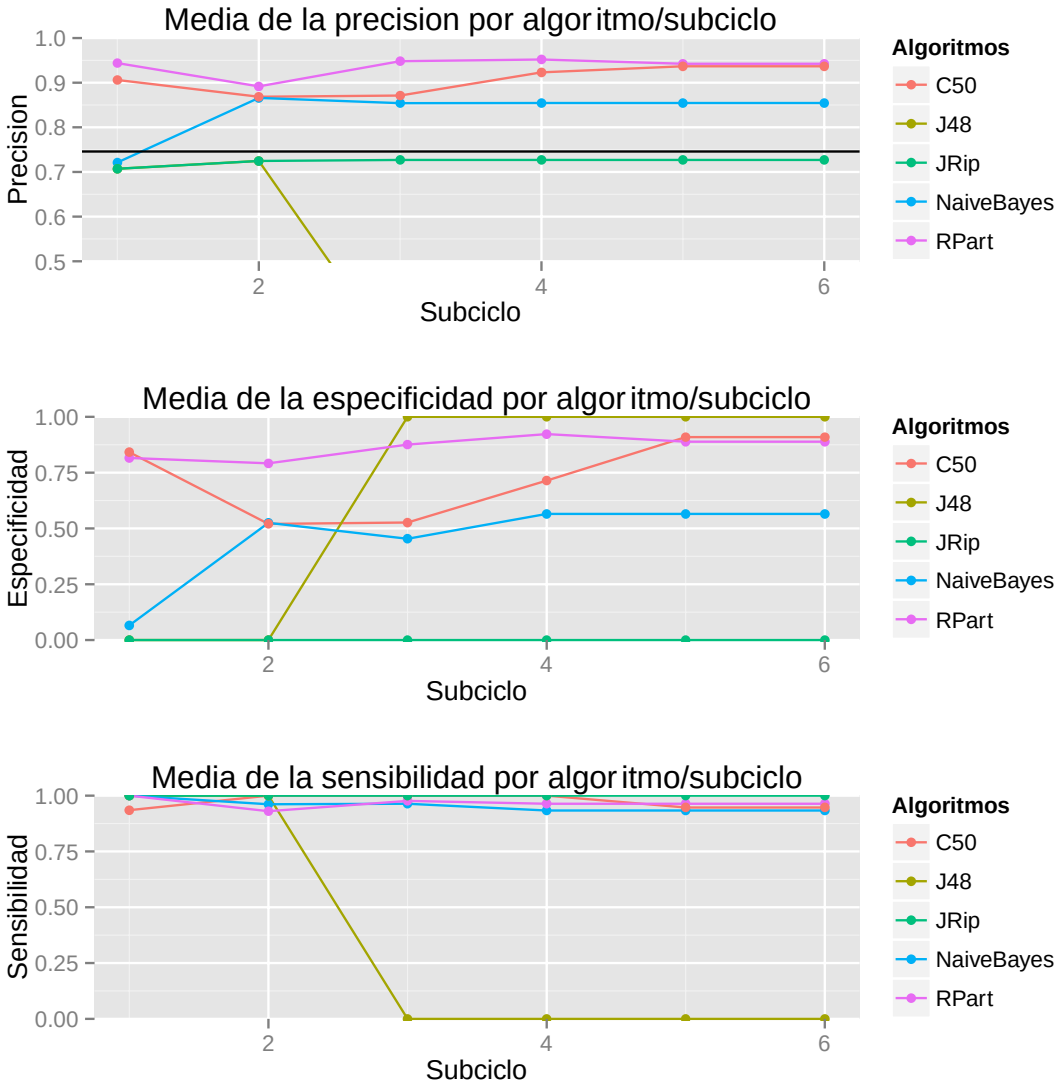

Figura 29. Comparativa de algoritmos del experimento 8.

4.3.9 Resumen de la experimentación

En este apartado vamos a mostrar los valores medios de los resultados obtenidos en los experimentos. De esta forma podremos ver de forma más clara qué algoritmo está teniendo mejor comportamiento.

Primero mostraremos tablas de las tres medidas (precisión, sensibilidad y especificidad) para cada subciclo. Esto nos permitirá ver qué algoritmo funciona mejor en cada etapa.

Nos interesa principalmente las etapas más tempranas. Descartaremos 2 subciclos:

- Subciclo 0: No tenemos suficientes datos en esta etapa para realizar una predicción. Es posible que lo tengamos en los próximos años, pero no en el momento de realizar este estudio.
- Subciclo 6: En este subciclo ya se ha terminado el curso y tenemos las notas finales por lo que la predicción no es necesaria.

Por otro lado, como ya indicamos anteriormente, nos centraremos en la especificidad (tasa de aciertos de la clase con menor representación que normalmente corresponde al número de alumnos en riesgo) y en la precisión.

Finalmente mostraremos todas las medias en una sola tabla para tener una visión global de los resultados.

SUBCICLO 1:

ESPECIFICIDAD SUBCICLO 1								
ColegioID	**NivelId**	**GradoId**	**Area**	**RPart**	**C50**	**J48**	**JRip**	**NaiveBayes**
2100	24	50	Matematicas	0,6044	0,0000	0,0000	0,0000	0,1682
1047	18	25	Lenguaje	0,6895	0,3271	0,0000	0,0000	0,5571
2336	18	27	Tecnologia e Informatica	NA	NA	NA	NA	NA
2202	36	103	Geografia	0,1353	0,0000	0,0000	0,0000	0,0000
2053	24	50	Ciencias Sociales	0,2430	0,0000	0,0000	0,0000	0,2653
2030	22	44	Matematicas	0,5696	0,0000	0,0000	0,0000	0,5097
2030	22	43	Ingles	0,4783	0,0000	0,0000	0,0000	0,2967
3170	23	48	Ciencias Naturales	0,8157	0,8417	0,0000	0,0000	0,0659
PROMEDIO:				0,5051	0,1670	0,0000	0,0000	0,2661

Figura 30. Resumen de los valores de la especificidad obtenidos en el subciclo 1.

PRECISIÓN SUBCICLO 1								
ColegioID	**NivelId**	**GradoId**	**Area**	**RPart**	**C50**	**J48**	**JRip**	**NaiveBayes**
2100	24	50	Matematicas	0,7503	0,6005	0,6005	0,6005	0,5636
1047	18	25	Lenguaje	0,9120	0,8896	0,8379	0,8379	0,8217
2336	18	27	Tecnologia e Informatica	0,9187	0,9187	0,9187	0,9187	0,4571
2202	36	103	Geografia	0,6102	0,5989	0,5989	0,5989	0,5989
2053	24	50	Ciencias Sociales	0,9082	0,8939	0,8939	0,8939	0,9032
2030	22	44	Matematicas	0,9255	0,8687	0,8687	0,8687	0,9010
2030	22	43	Ingles	0,9428	0,9150	0,9150	0,9150	0,8944
3170	23	48	Ciencias Naturales	0,9439	0,9060	0,7071	0,7071	0,7211
PROMEDIO:				0,8640	0,8239	0,7926	0,7926	0,7326

Figura 31. Resumen de los valores de la precisión obtenidos en el subciclo 1.

SENSIBILIDAD SUBCICLO 1								
ColegioID	**NivelId**	**GradoId**	**Area**	**RPart**	**C50**	**J48**	**JRip**	**NaiveBayes**
2100	24	50	Matematicas	0,8467	1,0000	1,0000	1,0000	0,8277
1047	18	25	Lenguaje	0,9629	1,0000	1,0000	1,0000	0,8760
2336	18	27	Tecnologia e Informatica	1,0000	1,0000	1,0000	1,0000	NA
2202	36	103	Geografia	0,9263	1,0000	1,0000	1,0000	1,0000
2053	24	50	Ciencias Sociales	0,9879	1,0000	1,0000	1,0000	0,9795
2030	22	44	Matematicas	0,9779	1,0000	1,0000	1,0000	0,9751
2030	22	43	Ingles	0,9883	1,0000	1,0000	1,0000	0,9541
3170	23	48	Ciencias Naturales	1,0000	0,9345	1,0000	1,0000	1,0000
PROMEDIO:				0,9613	0,9918	1,0000	1,0000	0,9446

Figura 32. Resumen de los valores de la sensibilidad obtenidos en el subciclo 1.

SUBCICLO 2:

ESPECIFICIDAD SUBCICLO 2								
ColegioID	NivelId	GradoId	Area	RPart	C50	J48	JRip	NaiveBayes
2100	24	50	Matematicas	0,3486	0,0000	0,0000	0,0000	0,3266
1047	18	25	Lenguaje	0,4837	0,4417	0,0000	0,0000	0,6709
2336	18	27	Tecnologia e Informatica	0,4357	0,7112	0,0000	0,0000	0,1436
2202	36	103	Geografia	0,7248	0,5779	0,0000	0,0000	0,7774
2053	24	50	Ciencias Sociales	0,4633	0,1069	0,0000	0,0000	0,6520
2030	22	44	Matematicas	0,8115	0,0000	0,0000	0,0000	0,8703
2030	22	43	Ingles	0,6758	0,4175	0,0000	0,0000	0,6592
3170	23	48	Ciencias Naturales	0,7911	0,5204	0,0000	0,0000	0,5252
PROMEDIO:				0,5918	0,3469	0,0000	0,0000	0,5781

Figura 33. Resumen de los valores de la especificidad obtenidos en el subciclo 2.

PRECISIÓN SUBCICLO 2								
ColegioID	NivelId	GradoId	Area	RPart	C50	J48	JRip	NaiveBayes
2100	24	50	Matematicas	0,6944	0,6109	0,6109	0,6109	0,6100
1047	18	25	Lenguaje	0,9152	0,9109	0,8413	0,8413	0,8882
2336	18	27	Tecnologia e Informatica	0,8433	0,7900	0,8533	0,8533	0,7269
2202	36	103	Geografia	0,8260	0,7894	0,6105	0,6105	0,5976
2053	24	50	Ciencias Sociales	0,9400	0,9117	0,9010	0,9010	0,7554
2030	22	44	Matematicas	0,9507	0,8761	0,8761	0,8761	0,9123
2030	22	43	Ingles	0,9706	0,9567	0,9258	0,9258	0,9243
3170	23	48	Ciencias Naturales	0,8914	0,8683	0,7248	0,7248	0,8658
PROMEDIO:				0,8790	0,8392	0,7929	0,7929	0,7851

Figura 34. Resumen de los valores de la precisión obtenidos en el subciclo 2.

SENSIBILIDAD SUBCICLO 2								
ColegioID	NivelId	GradoId	Area	RPart	C50	J48	JRip	NaiveBayes
2100	24	50	Matematicas	0,9130	1,0000	1,0000	1,0000	0,7899
1047	18	25	Lenguaje	1,0000	1,0000	1,0000	1,0000	0,9338
2336	18	27	Tecnologia e Informatica	0,9147	0,8049	1,0000	1,0000	1,0000
2202	36	103	Geografia	0,8951	0,9215	1,0000	1,0000	0,4758
2053	24	50	Ciencias Sociales	0,9936	1,0000	1,0000	1,0000	0,7683
2030	22	44	Matematicas	0,9676	1,0000	1,0000	1,0000	0,9186
2030	22	43	Ingles	0,9934	1,0000	1,0000	1,0000	0,9507
3170	23	48	Ciencias Naturales	0,9301	1,0000	1,0000	1,0000	0,9611
PROMEDIO:				0,9509	0,9658	1,0000	1,0000	0,8498

Figura 35. Resumen de los valores de la sensibilidad obtenidos en el subciclo 2.

SUBCICLO 3:

ESPECIFICIDAD SUBCICLO 3								
ColegioID	**NivelId**	**GradoId**	**Area**	**RPart**	**C50**	**J48**	**JRip**	**NaiveBayes**
2100	24	50	Matematicas	0,4884	0,5489	0,0000	0,0000	0,3467
1047	18	25	Lenguaje	0,7395	0,4716	0,0000	0,0000	0,7086
2336	18	27	Tecnologia e Informatica	0,6519	0,6519	0,0000	0,0000	0,1560
2202	36	103	Geografia	0,7518	0,6767	0,0000	0,0000	0,7652
2053	24	50	Ciencias Sociales	0,6376	0,4259	0,0000	0,0000	0,5904
2030	22	44	Matematicas	0,7570	0,0000	0,0000	0,0000	0,7265
2030	22	43	Ingles	0,8117	0,0000	0,0000	0,0000	0,6192
3170	23	48	Ciencias Naturales	0,8756	0,5262	1,0000	0,0000	0,4540
PROMEDIO:				0,7142	0,4127	0,1250	0,0000	0,5458

Figura 36. Resumen de los valores de la especificidad obtenidos en el subciclo 3.

PRECISIÓN SUBCICLO 3								
ColegioID	**NivelId**	**GradoId**	**Area**	**RPart**	**C50**	**J48**	**JRip**	**NaiveBayes**
2100	24	50	Matematicas	0,7159	0,6455	0,5992	0,5992	0,5841
1047	18	25	Lenguaje	0,9391	0,9152	0,8413	0,8413	0,8935
2336	18	27	Tecnologia e Informatica	0,9300	0,8300	0,8533	0,8533	0,7395
2202	36	103	Geografia	0,8444	0,8167	0,6056	0,6056	0,6727
2053	24	50	Ciencias Sociales	0,9462	0,9359	0,9073	0,9073	0,7414
2030	22	44	Matematicas	0,9479	0,8761	0,8761	0,8761	0,9250
2030	22	43	Ingles	0,9814	0,9258	0,9258	0,9258	0,9253
3170	23	48	Ciencias Naturales	0,9481	0,8712	0,2731	0,7269	0,8542
PROMEDIO:				0,9066	0,8520	0,7352	0,7919	0,7920

Figura 37. Resumen de los valores de la precisión obtenidos en el subciclo 3.

SENSIBILIDAD SUBCICLO 3								
ColegioID	**NivelId**	**GradoId**	**Area**	**RPart**	**C50**	**J48**	**JRip**	**NaiveBayes**
2100	24	50	Matematicas	0,8680	0,7101	1,0000	1,0000	0,7426
1047	18	25	Lenguaje	0,9791	1,0000	1,0000	1,0000	0,9317
2336	18	27	Tecnologia e Informatica	0,9793	0,8622	1,0000	1,0000	0,9900
2202	36	103	Geografia	0,9083	0,9038	1,0000	1,0000	0,6169
2053	24	50	Ciencias Sociales	0,9788	0,9888	1,0000	1,0000	0,7594
2030	22	44	Matematicas	0,9740	1,0000	1,0000	1,0000	0,9546
2030	22	43	Ingles	0,9968	1,0000	1,0000	1,0000	0,9562
3170	23	48	Ciencias Naturales	0,9765	1,0000	0,0000	1,0000	0,9636
PROMEDIO:				0,9576	0,9331	0,8750	1,0000	0,8644

Figura 38. Resumen de los valores de la sensibilidad obtenidos en el subciclo 3.

SUBCICLO 4:

ESPECIFICIDAD SUBCICLO 4								
ColegioID	**NivelId**	**GradoId**	**Area**	**RPart**	**C50**	**J48**	**JRip**	**NaiveBayes**
2100	24	50	Matematicas	0,5778	0,5567	0,0000	0,0000	0,3362
1047	18	25	Lenguaje	0,4716	0,6786	0,0000	0,0000	0,7131
2336	18	27	Tecnologia e Informatica	0,5226	0,4940	0,0000	0,0000	0,1810
2202	36	103	Geografia	0,8895	0,5635	0,0000	0,0000	0,8096
2053	24	50	Ciencias Sociales	0,6265	0,6140	0,0000	0,0000	0,6437
2030	22	44	Matematicas	0,8418	0,5459	0,0000	0,0000	0,7017
2030	22	43	Ingles	0,6317	0,0000	0,0000	0,0000	0,4867
3170	23	48	Ciencias Naturales	0,9224	0,7145	1,0000	0,0000	0,5649
PROMEDIO:				0,6855	0,5209	0,1250	0,0000	0,5546

Figura 39. Resumen de los valores de la especificidad obtenidos en el subciclo 4.

PRECISION SUBCICLO 4								
ColegioID	**NivelId**	**GradoId**	**Area**	**RPart**	**C50**	**J48**	**JRip**	**NaiveBayes**
2100	24	50	Matematicas	0,7053	0,6439	0,5992	0,5992	0,5015
1047	18	25	Lenguaje	0,9152	0,9457	0,8413	0,8413	0,8819
2336	18	27	Tecnologia e Informatica	0,8800	0,8467	0,8533	0,8533	0,7237
2202	36	103	Geografia	0,9278	0,7556	0,6056	0,6056	0,7122
2053	24	50	Ciencias Sociales	0,9610	0,9502	0,9235	0,9235	0,8246
2030	22	44	Matematicas	0,9592	0,9310	0,8761	0,8761	0,9084
2030	22	43	Ingles	0,9631	0,9262	0,9262	0,9262	0,9116
3170	23	48	Ciencias Naturales	0,9519	0,9231	0,2731	0,7269	0,8543
PROMEDIO:				0,9079	0,8653	0,7373	0,7940	0,7898

Figura 40. Resumen de los valores de la precisión obtenidos en el subciclo 4.

SENSIBILIDAD SUBCICLO 4								
ColegioID	**NivelId**	**GradoId**	**Area**	**RPart**	**C50**	**J48**	**JRip**	**NaiveBayes**
2100	24	50	Matematicas	0,7914	0,7027	1,0000	1,0000	0,6120
1047	18	25	Lenguaje	1,0000	0,9949	1,0000	1,0000	0,9147
2336	18	27	Tecnologia e Informatica	0,9411	0,9102	1,0000	1,0000	0,9569
2202	36	103	Geografia	0,9540	0,8816	1,0000	1,0000	0,6494
2053	24	50	Ciencias Sociales	0,9898	0,9791	1,0000	1,0000	0,8417
2030	22	44	Matematicas	0,9744	0,9872	1,0000	1,0000	0,9377
2030	22	43	Ingles	0,9883	1,0000	1,0000	1,0000	0,9505
3170	23	48	Ciencias Naturales	0,9636	1,0000	0,0000	1,0000	0,9333
PROMEDIO:				0,9503	0,9320	0,8750	1,0000	0,8495

Figura 41. Resumen de los valores de la sensibilidad obtenidos en el subciclo 4.

SUBCICLO 5:

ESPECIFICIDAD SUBCICLO 5								
ColegioID	NivelId	GradoId	Area	RPart	C50	J48	JRip	NaiveBayes
2100	24	50	Matematicas	0,7425	0,5921	0,0000	0,0000	0,7358
1047	18	25	Lenguaje	0,8051	0,8051	0,0000	0,0000	0,7129
2336	18	27	Tecnologia e Informatica	0,4690	0,3324	0,0000	0,0000	0,3810
2202	36	103	Geografia	0,9152	0,5635	0,0000	0,0000	0,8096
2053	24	50	Ciencias Sociales	0,6441	0,6563	0,0000	0,0000	0,6716
2030	22	44	Matematicas	0,7686	0,8412	0,0000	0,0000	0,8679
2030	22	43	Ingles	0,5058	0,6592	0,0000	0,0000	0,4350
3170	23	48	Ciencias Naturales	0,8884	0,9088	1,0000	0,0000	0,5649
PROMEDIO:				0,7173	0,6698	0,1250	0,0000	0,6473

Figura 42. Resumen de los valores de la especificidad obtenidos en el subciclo 5.

PRECISIÓN SUBCICLO 5								
ColegioID	NivelId	GradoId	Area	RPart	C50	J48	JRip	NaiveBayes
2100	24	50	Matematicas	0,8732	0,8444	0,6288	0,6288	0,8190
1047	18	25	Lenguaje	0,9696	0,9696	0,8413	0,8413	0,9035
2336	18	27	Tecnologia e Informatica	0,8800	0,8433	0,8533	0,8533	0,7302
2202	36	103	Geografia	0,9389	0,7556	0,6056	0,6056	0,7122
2053	24	50	Ciencias Sociales	0,9701	0,9670	0,9272	0,9272	0,8210
2030	22	44	Matematicas	0,9481	0,9442	0,8883	0,8883	0,8747
2030	22	43	Ingles	0,9631	0,9554	0,9262	0,9262	0,9133
3170	23	48	Ciencias Naturales	0,9423	0,9365	0,2731	0,7269	0,8543
PROMEDIO:				0,9356	0,9020	0,7430	0,7997	0,8285

Figura 43. Resumen de los valores de la precisión obtenidos en el subciclo 5.

SENSIBILIDAD SUBCICLO 5								
ColegioID	NivelId	GradoId	Area	RPart	C50	J48	JRip	NaiveBayes
2100	24	50	Matematicas	0,9512	0,9938	1,0000	1,0000	0,8690
1047	18	25	Lenguaje	1,0000	1,0000	1,0000	1,0000	0,9437
2336	18	27	Tecnologia e Informatica	0,9529	0,9291	1,0000	1,0000	0,8906
2202	36	103	Geografia	0,9536	0,8816	1,0000	1,0000	0,6494
2053	24	50	Ciencias Sociales	0,9966	0,9923	1,0000	1,0000	0,8360
2030	22	44	Matematicas	0,9695	0,9561	1,0000	1,0000	0,8752
2030	22	43	Ingles	1,0000	0,9786	1,0000	1,0000	0,9586
3170	23	48	Ciencias Naturales	0,9636	0,9471	0,0000	1,0000	0,9333
PROMEDIO:				0,9734	0,9598	0,8750	1,0000	0,8695

Figura 44. Resumen de los valores de la sensibilidad obtenidos en el subciclo 1.

RESUMEN:

A continuación mostraremos el resumen de los resultados anteriores en una sola tabla por medida.

ESPECIFICIDAD					
Subciclo	**RPart**	**C50**	**J48**	**JRip**	**NaiveBayes**
1	0,5051	0,1670	0,0000	0,0000	0,2661
2	0,5918	0,3469	0,0000	0,0000	0,5781
3	0,7142	0,4127	0,1250	0,0000	0,5458
4	0,6855	0,5209	0,1250	0,0000	0,5546
5	0,7173	0,6698	0,1250	0,0000	0,6473

Tabla 26. Resumen especificidad por subciclo y algoritmo.

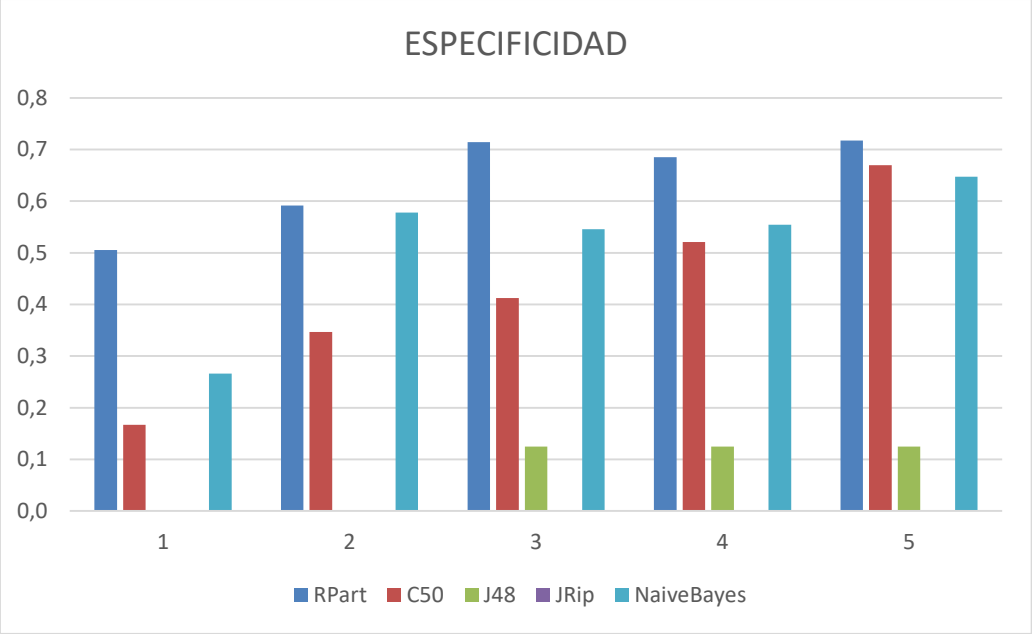

Figura 45. Comparativa especificidad por subciclo y algoritmo.

PRECISIÓN					
Subciclo	**RPart**	**C50**	**J48**	**JRip**	**NaiveBayes**
1	0,8640	0,8239	0,7926	0,7926	0,7326
2	0,8790	0,8392	0,7929	0,7929	0,7851
3	0,9066	0,8520	0,7352	0,7919	0,7920
4	0,9079	0,8653	0,7373	0,7940	0,7898
5	0,9356	0,9020	0,7430	0,7997	0,8285

Tabla 27. Resumen precisión por subciclo y algoritmo.

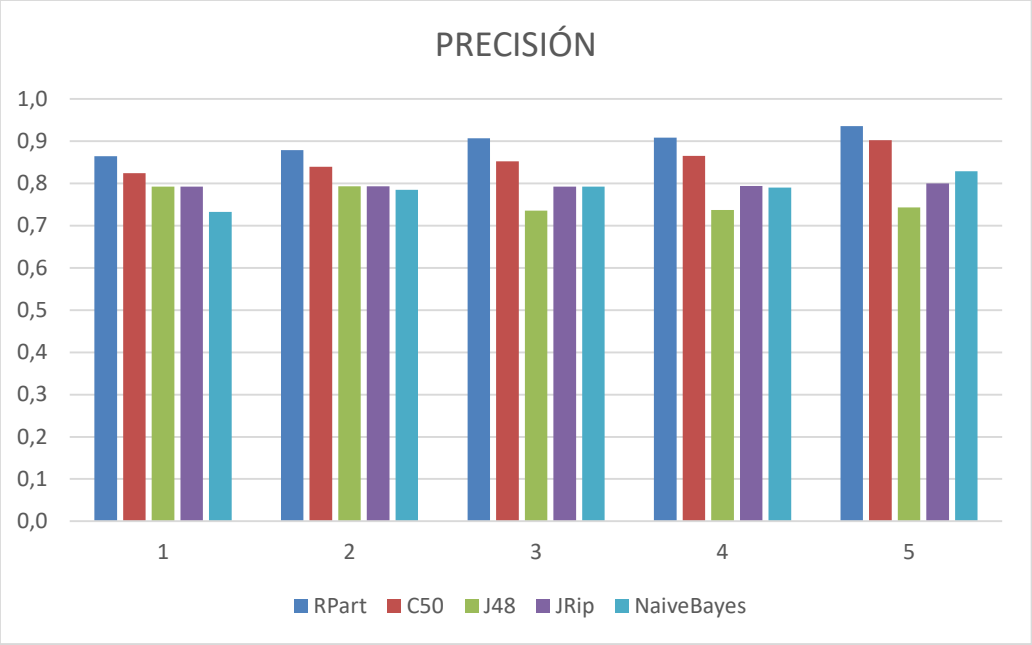

Figura 46. Comparativa precisión por subciclo y algoritmo.

Por lo tanto, basándonos en los resultados de la especificidad y la precisión, observamos que el algoritmo RPart es el más adecuado.

SENSIBILIDAD					
Subciclo	**RPart**	**C50**	**J48**	**JRip**	**NaiveBayes**
1	0,5051	0,1670	0,0000	0,0000	0,2661
2	0,9509	0,9658	1,0000	1,0000	0,8498
3	0,9576	0,9331	0,8750	1,0000	0,8644
4	0,9503	0,9320	0,8750	1,0000	0,8495
5	0,9734	0,9598	0,8750	1,0000	0,8695

Tabla 28.Resumen sensibilidad por subciclo y algoritmo.

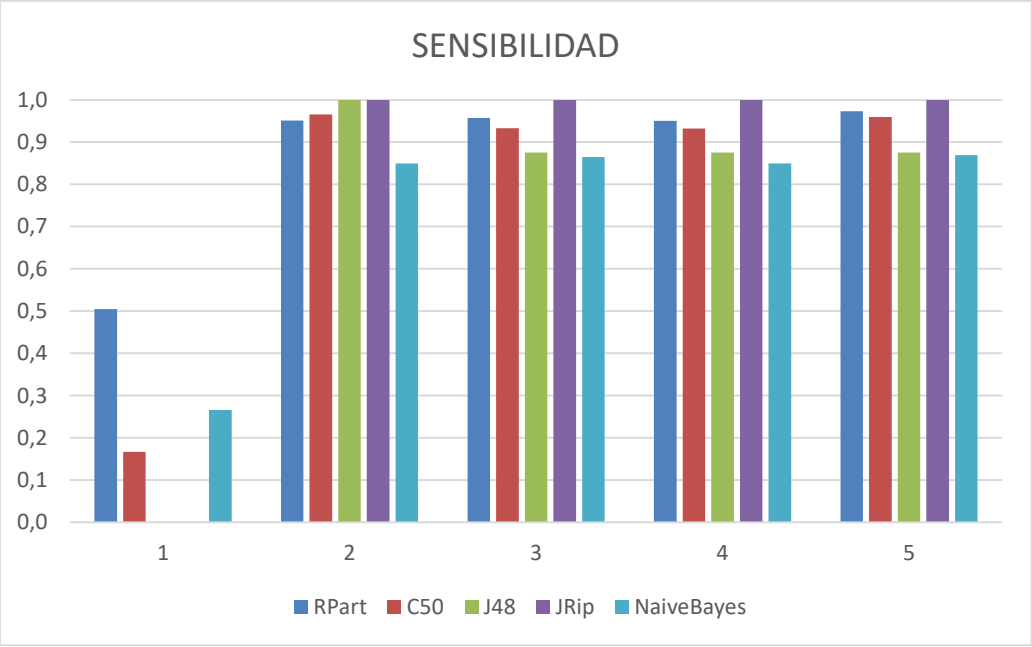

Figura 47. Comparativa sensibilidad por subciclo y algoritmo.

5 Conclusiones

Los objetivos marcados al comienzo de este proyecto fueron los siguientes:

Objetivo: Poder predecir si un alumno va a suspender una determinada asignatura.

Subobjetivos:

1. Creación de un dataset.
2. Selección de varios colegios con datos suficientes para hacer un estudio.
3. Selección del mejor algoritmo para predecir con este dataset.
4. Facilitar la generación automática de modelos predictivos por parte del personal de desarrollo no experto en data mining.

Subobjetivo 1. Creación de un Dataset:

Hemos creado un dataset con más de 6 millones de instancias y 47 variables. El dataset permite relacionar información de un mismo alumno en dos direcciones:

- Entre las dos bases de datos. Los datos de Pleno y Datamart se encuentran unidos en cada instancia.
- Entre dos ciclos. Al introducir el concepto de área hemos podido relacionar la información de un alumno de un ciclo con la información que tenía en el ciclo anterior, lo que nos ha permitido la generación de variables con valores en el subciclo 0 (como la nota del año anterior o saber si ha cambiado de nivel) y la predicción de la nota de un alumno en base a los resultados de otros alumnos en el ciclo anterior (esto no se podía hacer usando clase puede cambiar de un año para otro). Por ejemplo:
  Tenemos una clase que se “Matemáticas de primero” y corresponden al ciclo uno. En el ciclo dos el profesor le cambia el nombre y pasa a llamarse “Clase de matemáticas de Juan”. Ambas pertenecen al mismo grado, pero, al no llamarse igual, no podemos saber que imparten los mismos contenidos. Al crear el concepto de área, ahora ambas pertenecen al área “Matemáticas” por lo que se considera que ambas imparten la misma materia y podemos utilizar los datos de la clase del ciclo uno para predecir las notas del ciclo 2.

  Consideramos que este subobjetivo se ha cumplido.

Subobjetivo 2. Selección de varios colegios con datos suficientes para hacer un estudio.

Hemos realizado una exploración de los datos y hemos seleccionado los colegios que tenían suficientes datos, tanto en número de instancias como en número de valores en cada columna (ya que tenemos muchos valores perdidos o vacíos). En la sección de exploración de datos se detalla los criterios seguidos para esta selección.

Dado que hemos podido realizar los experimentos con los datos seleccionados y los resultados han sido aceptables consideramos el subobjetivos cumplido.

Subobjetivo 3. Selección del mejor algoritmo para predecir con este dataset.

Hemos realizado algunas pruebas y los resultados parecen indicar que el algoritmo RPart es el más adecuado en la mayoría de los casos. Como no podemos hacer una selección en cada caso ya que hay muchísimas combinaciones, se debe escoger el que mejores resultados dé generalmente. El algoritmo C5.0 también parece dar buenos resultados, pero nos quedaremos con Rpart.

Se ha seleccionado un algoritmo que da buenos resultados por lo que se considera cumplido este subobjetivo.

Subobjetivo 4. Facilitar la generación automática de modelos predictivos por parte del personal de desarrollo no experto en data mining.

Se han automatizado los procesos para la generación de modelos predictivos. Se han definido y se han especificado sus entradas y salidas por lo que son fácilmente reutilizables por expertos programadores, aunque no tengan conocimientos de data mining.

Por otro lado, se guardan los modelos generados de manera que, para predecir si un alumno va a aprobar o suspender, tan solo debe cargar el modelo (con la función load) y realizar la predicción con una llamada a la función predict con dos parámetros (el modelo cargado y el data.frame con los datos del alumno a predecir). Por lo tanto, se puede realizar una predicción con dos líneas de código y sin necesidad de tener conocimientos de minería de datos.

Consideramos que este subobjetivo se ha cumplido.

Al haberse cumplido todos los subobjetivos consideramos el objetivo del proyecto cumplido.

6 Trabajos futuros

Son muchos los trabajos y experimentos que se pueden realizar sobre estos datos.

A continuación, se expondrán algunas ideas:

- No contamos con suficientes datos de algunas materias o áreas. Podríamos intentar determinar si se puede predecir los resultados de una asignatura con los resultados de otra.
- De igual manera, algunos colegios llevan poco tiempo y no tenemos suficientes datos de los mismos. Lo mismo ocurrirá con los nuevos colegios que empiecen a utilizar estas herramientas. Para poder predecir sobre estos alumnos podríamos aplicar técnicas de clustering para agrupar colegios con características similares (similar sistema de estudios, condiciones económicas o sociales de los alumnos, etc) para intentar predecir un colegio con los resultados de otros con similares características.
- Se puede cambiar el objetivo (target) redefiniendo fracaso escolar. Por ejemplo, podríamos modificar el dataset para predecir la probabilidad de que un alumno apruebe un curso o termine sus estudios.
- Iremos añadiendo fuentes de datos conforme vayan apareciendo (en principio ya existen otras dos fuentes, una de datos económicos y de gestión y otra aún por determinar) que aumentarán las posibilidades del dataset. De igual manera podemos realizar divisiones de los datos y realizar los mismos experimentos sobre estos subconjuntos (por ejemplo, separar datos de primaria y secundaria, separar por países, por fuente de datos, etc).
- El dataset está preparado para intentar detectar test de conocimiento (de Pleno) que sean relevantes o que contengan fallos (por ejemplo, si detectamos que una pregunta casi siempre se falla puede deberse a que se haya seleccionado mal la respuesta correcta).
- Podemos utilizar lógica fuzzy para determinar el grado de riesgo del alumno y no liminarnos a indicar si está o no en riesgo.

7 Definiciones, siglas y abreviaturas

- **Clase**: Una clase podría entenderse como una asignatura. Sin embargo, el concepto de clase es más complejo. En realidad, se refiere a un conjunto de alumnos que cursan una o más materias en común. Por lo tanto, podríamos tener una clase denominada "matemáticas 3° B" que correspondería a un conjunto de alumnos que cursan una materia (seguramente matemáticas). Sin embargo, podemos encontrarnos la clase "Clase de Nicolás" que tenga una o más materias (por ejemplo matemáticas, lenguaje y Ciencias Sociales) y cuya interpretación sería más compleja.
- **Datamart**: Vista de la base de datos que contiene la información obtenida del LMS.
- **Subciclo**: Distintas particiones de un ciclo. En principio un ciclo se dividirá en 6 subciclos del mismo tamaño (con el mismo número de días con una posible variación mínima por el redondeo) más un subciclo inicial (subciclo 0) con cero días de duración para indicar la situación en la fecha de inicio.
- **LMS**: Learning Management System (Sistema de gestión de aprendizaje). También se utiliza para referirse a la base de datos que contiene la información obtenida del sistema de gestión de aprendizaje o a la vista sobre dicha base de datos (Datamart).
- **Pleno(PL)**: Base de datos que contiene la información de los tests de conocimientos que los alumnos pueden realizar de forma voluntaria. También se utiliza para nombrar a la vista sobre dicha base de datos.
- **CicloId:** Periodo durante el que se desarrolla el curso escolar (normalmente cercano al año).
- **Dataset:** Conjunto de datos. En general está formado por una tabla o matriz donde cada columna representa una variable y cada fila representa una instancia del conjunto de datos en cuestión.
- **EDM**: Es una disciplina emergente dedicada al desarrollo de métodos para la exploración de datos provenientes de entornos educativos con el fin de entender mejor a los estudiantes y sus métodos de aprendizaje.
- **SGBD**: Un Sistema de Gestión de Bases de Datos es un conjunto de programas que permiten el almacenamiento, modificación y extracción de la información en una base de datos, además de proporcionar herramientas para añadir, borrar, modificar y analizar los datos.

8 Bibliografía

- Definición funcional de los paquetes SSIS implementados en el proyecto SantillanaNeo de la plataforma Santillana. Fichero pdf.

[1] C. Romero and S. Ventura. Educational data mining: A review of the state of the art. *IEEE Transactions on Systems Man and Cybernetics Part C-Applications and Reviews*, 40(6):601-618, 2010.

[2] C. Romero, S. Ventura, M. Pechenizkiy, and R. S. J. Baker. *Handbook of Educational Data Mining*. Chapman & Hall/CRC Data Mining and Knoyledge Discovery Series. Taylor & Francis, 2011. ISBN 9781439804582.

[3] George H John and Pat Langley. Estimating continuous distributions in Bayesian classifiers. In Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, 338-345. Morgan Kaufmann Publishers Inc., 1995.

[4] C. Romero, S. Ventura and E. García. Data Mining in Course Management Systems: MOODLE Case Study and Tutorial. *Computers and Education*, 51(**1**), 368-384, 2008.

[5] Romero, C., & Ventura, S. (2007). Educational Data Mining: A Survey from 1995 to 2005. *Expert System with Applications*(33), 135 - 146.

[6] Romero, C., & Ventura, S. (2013). Data mining education. *WIREs Data Mining Knowledge Discovery., 3*, 12 - 27.

[7] Magaña Hernández, M. (2002). Causas del Fracaso Escolar. *XIII Congreso de la Sociedad Española de Medicina Adolescente*. España.

[8] S.A.H.a Basha, A.b Govardhan, S.c Viswanadha Raju, and N.d Sultana. A comparative analysis of prediction techniques for predicting graduate rate of university. *European Journal of Scientific Research*, 46(2):186-193, 2010.

[9] E.P.I Garcia and P.M. Mora. Model prediction of academic performance for first year students. *Proceedings – 2011 10th Mexican International Conference on Artificial Intelligence: Advances in Artificial Intelligente and Aplications, MICAI 2011 – Proceedings of Special Session*, pages 169-174, 2011.

[10] K.a Bunkar, U.K.b Singh, B.a Pandya, and R.a Bunkar. Data mining: Prediction for performance improvement of graduate students using classification. *IFIP International Conference on Wireless and Optical Communications Networks, WOCN*, 2012.

[11] S.A. Kumar and M.N. Vijayalakshmi. Mining of student academic evaluation records in higher education. *Proceedings of the 2012 International Conference on Recent Advances in Computing and Software Systems, RACSS 2012*, pages 67-70, 2012.

[12] T. Sevindik. Prediction of student academic performance by using an adaptive neuro-fuzzy inference system. *Energy Education Science and Technology Part B: Social and Educational Studies*, 3(4):635-646, 2011.

[13] P. Gu and Q. Zhou. Student performances prediction based on improved c4.5 decision tree algorithm. *Advances in Intelligent and Soft Computing*, 146 AISC: 1-8, 2012.

[14] L.S. Affendey, I.H.M. Paris, N. Mustapha, M.N. Sulaiman, and Z. Muda. Ranking of influencing factors in predicting student's academic performance. *Information Technology Journal*, 9(4): 832-837, 2010.

[15] Jiménez, M., Luna, J. M., & Ventura, S. (2013). EDM para la detección precoz del fracaso escolar en secundaria., (págs. 1353 - 1362). Madrid.