



**VNiVERSiDAD
D SALAMANCA**

DEPARTAMENTO DE INFORMÁTICA Y AUTOMÁTICA

Programa de innovación docente

CONVOCATORIA DE AYUDAS 2013

MEMORIA FINAL DEL PROYECTO

**Aplicación de técnicas de minería de datos en
el seguimiento y análisis de resultados de
evaluación en las asignaturas de Ingeniería del
Software del Grado en Ingeniería Informática**

Coordinadora del proyecto:

María N. Moreno García

Tabla de contenido

DATOS DEL PROYECTO.....	1
MEMORIA DEL PROYECTO	2
1. Introducción	2
2. Descripción del estudio	2
3. Datos analizados	4
Ingeniería del Software II	5
Gestión de Proyectos	6
4. Resultados.....	7
4.1 Ingeniería del Software II.....	7
4.2 Gestión de Proyectos.....	18
5. Conclusiones	27

DATOS DEL PROYECTO

Título

Aplicación de técnicas de minería de datos en el seguimiento y análisis de resultados de evaluación en las asignaturas de Ingeniería del Software del Grado en Ingeniería Informática

Código: ID2013/228

Coordinadora

María N. Moreno García

Profesores participantes

Francisco José García Peñalvo

María Dolores Muñoz Vicente

Vivian López Batista

Emilio Corchado Rodríguez

Financiación

200 €

SALAMANCA, 30 de junio de 2014

María N. Moreno García

MEMORIA DEL PROYECTO

1. Introducción

En la Universidad de Salamanca las asignaturas relacionadas con las materias de Ingeniería del Software que se impartían en las titulaciones de Ingeniería Informática han sufrido un cambio sustancial con la implantación de los nuevos grados, en los que el número de horas de dicha materia han aumentado significativamente. El incremento en un curso de los nuevos grados respecto a las Ingenierías Técnicas extinguidas ha hecho que parte de los contenidos de las asignaturas de Ingeniería del Software del plan de estudios de la antigua titulación de Ingeniería Informática (2º Ciclo) hayan sido incorporados a nuevas asignaturas de dichos grados. Además de los cambios relacionados con la actualización y reorganización de contenidos, ha sido necesario realizar una modificación de mayor trascendencia como es la relacionada con la implantación de una nueva metodología docente basada en la adquisición de competencias y un sistema de evaluación diferente que permita valorar de forma progresiva dichas competencias acorde a la filosofía del EEES.

Teniendo en cuenta dichos aspectos, en un proyecto anterior se realizó un estudio de la repercusión de la implantación de la nueva metodología docente y los nuevos sistemas de evaluación en relación con los resultados obtenidos en las asignaturas de los planes antiguos. Se estudiaron las diferencias en la organización de los contenidos, cambios en la impartición de los contenidos teóricos, cambios en la forma de adquirir las competencias prácticas y en la realización de trabajos así como la Implantación de nuevas actividades formativas. El análisis de los aspectos anteriores permitió realizar un estudio comparativo de los resultados obtenidos en las asignaturas de los grados respecto a los de las antiguas asignaturas.

En este proyecto hemos profundizado en las causas del éxito o fracaso de los resultados en las diferentes asignaturas mediante la aplicación de técnicas de minería de datos a los datos pormenorizados de todos los estudiantes para encontrar los factores más influyentes en la obtención de buenos resultados tanto desde el punto de vista de las características y forma de realizar las actividades por los propios estudiantes como de aspectos relacionados con la metodología docente seguida. Asimismo, mediante estas técnicas se podrán construir modelos predictivos que nos permitirán modificar convenientemente la metodología docente o intentar modificar el comportamiento de los estudiantes si la predicción de los modelos en base a esos factores es de fracaso.

2. Descripción del estudio

Como paso previo a la descripción de las actividades realizadas, y con el fin de lograr una mejor comprensión de las mismas, se realiza una breve introducción de las técnicas aplicadas en este estudio.

La minería de datos comenzó a surgir a finales de los años 80 en el entorno de los negocios como parte de un proceso genérico denominado KDD (*Knowledge Discovery in Databases*). KDD engloba todas las fases y procedimientos destinados a encontrar patrones y similitudes en los datos. La minería de datos ha recuperado algunas técnicas matemáticas y estadísticas clásicas que han tenido que ser actualizadas para adaptarse al tratamiento de grandes muestras de datos y a los requisitos del procesamiento automático de la información. La minería de datos ha dado lugar a una paulatina sustitución del análisis de datos dirigido a

la verificación por un enfoque de análisis de datos dirigido al descubrimiento del conocimiento. La aplicación automatizada de algoritmos de minería de datos permite detectar fácilmente patrones en los datos, razón por la cual esta técnica es mucho más eficiente que el análisis dirigido a la verificación cuando se intenta explorar datos procedentes de almacenes de gran tamaño y complejidad elevada.

Los algoritmos de minería de datos se clasifican en dos grandes categorías: supervisados o predictivos y no supervisados o de descubrimiento del conocimiento.

- **Algoritmos supervisados o predictivos:** permiten inducir modelos que serán utilizados para predecir el valor de un atributo de un conjunto de datos, conocidos otros atributos.
- **Algoritmos no supervisados o de descubrimiento del conocimiento:** descubren patrones y tendencias en los datos actuales. El descubrimiento de esa información sirve para la toma de decisiones y la obtención de un beneficio (científico o de negocio) de ellas.

En este proyecto se han aplicado ambos tipos de algoritmos para obtener la máxima información posible que pueda ser útil en la toma de decisiones sobre la metodología aplicada, la realización de actividades y la forma de conducir a los estudiantes de cara a la obtención de buenos resultados.

Seguidamente se detallan los algoritmos utilizados, los modelos que se obtienen a partir de ellos y las mejoras docentes que se pueden conseguir:

- **Árboles de decisión:** Una de las técnicas supervisadas más conocidas y utilizadas son los algoritmos de inducción de árboles de decisión, mediante los cuales se construyen clasificadores en forma de árbol. En cada nodo del árbol se evalúan condiciones relativas a los valores de un atributo, el cumplimiento o no de las mismas permite separar los registros entre las ramas del árbol. Las hojas del árbol contendrán los registros clasificados, es decir, contendrán registros de una sola clase.

En el caso de nuestro estudio, los atributos que se evalúan son factores relacionados con las características y comportamiento académico de los estudiantes así como con la metodología docente y las clases pueden ser éxito o fracaso (aprueba o suspende) o las distintas calificaciones (MH, sobresaliente, notable, etc.).

Con los modelos inducidos se podrá predecir el éxito o fracaso (o las calificaciones) de un estudiante en función de los factores comentados anteriormente.

Por otra parte, analizando las reglas del árbol se podrá conocer qué factores individuales o combinación de los mismos pueden ser determinantes en el éxito o fracaso de los estudiantes.

- **Redes bayesianas:** Son modelos probabilísticos en los que se modela de forma cualitativa y cuantitativa las relaciones entre variables. El aprendizaje de redes bayesianas para un conjunto de datos D consiste en encontrar, entre todos los grafos posibles, el grafo G que mejor represente el conjunto de dependencias/independencias entre los datos.

Al igual que en el árbol de decisión, en este proyecto las variables que constituyen los nodos del grafo son los factores comentados anteriormente y el resultado de aplicar el modelo inducido a un registro correspondiente a un estudiante permitirá predecir sus resultados. Igualmente podremos conocer cuáles son los factores más influyentes en la obtención de un resultado u otro.

- **Reglas de asociación:** El análisis de asociación persigue el establecimiento de relaciones entre registros individuales o grupos de registros de la base de datos. Una regla de asociación tiene la forma: si X entonces Y donde X e Y representan los valores de uno o varios atributos.

El análisis del modelo de reglas puede resultar de gran interés a la hora de tomar decisiones que permitan mejorar las metodologías docentes puesto que nos proporcionan información sobre las circunstancias (valores de los factores) que deben darse conjuntamente para obtener buenos o malos resultados.

- **Algoritmos de *clustering*:** Dado un gran conjunto multidimensional de datos, en el que cada atributo constituye una dimensión del espacio, los métodos de agrupación (*clustering*) descubren patrones de distribución del conjunto de datos en forma de grupos de registros con características similares.

En nuestro proyecto se han creado grupos que representen los diferentes resultados, bien en la forma de éxito o fracaso o en forma de las diferentes calificaciones. De esta manera se tendrá información sobre las características de cada uno de esos grupos que lógicamente se corresponden con los factores de metodología docente y con los relacionados con los estudiantes.

Los **beneficios docentes esperados** con la aplicación de los métodos mencionados dependen del tipo de algoritmo que se aplique:

- En el caso de los algoritmos supervisados como los árboles de decisión o las redes bayesianas, la inducción de modelos con datos de cursos anteriores y del curso actual nos permitirán realizar predicciones sobre estudiantes de cursos futuros, con lo que se podrá guiar a ese estudiante o adaptar la metodología docente en caso de que se predigan malos resultados.
- Los algoritmos no supervisados como las reglas de asociación y los algoritmos de *clustering* nos proporcionan información muy valiosa sobre aquellas prácticas docentes que conducen a buenos resultados por lo que el análisis de esa información nos permitirá en cursos sucesivos aplicar de forma conveniente los métodos o actividades para conseguir mayor tasa de éxito y guiar de forma adecuada a los estudiantes para que consigan mejores resultados.

En definitiva, el estudio permite identificar los puntos fuertes y débiles de las metodologías aplicadas y del comportamiento de los estudiantes, por lo que se podrá conocer en qué aspectos están funcionando bien y en qué otros es necesario mejorar.

3. Datos analizados

En el estudio se han tenido en cuenta las calificaciones pormenorizadas de los alumnos obtenidas durante el curso 2013-2014 correspondientes a las asignaturas de Ingeniería del Software II y Gestión de Proyectos. A medida que estas asignaturas se vayan impartiendo en cursos sucesivos se podrá ir aumentando la cantidad de información disponible y obtener resultados más fiables.

A continuación se ofrecen representaciones gráficas de las distribuciones de valores de los atributos analizados en función de los valores de las clases. Se han realizado dos estudios, en el primero se analiza la repercusión de los valores de dichas variables con respecto a las clases “éxito” y “fracaso” correspondientes respectivamente a suspenso y aprobado con cualquier calificación respectivamente. En el segundo estudio se estudia dicha influencia con respecto a todas las posibles calificaciones cualitativas que serán las diferentes clases para este último caso.

Ingeniería del Software II

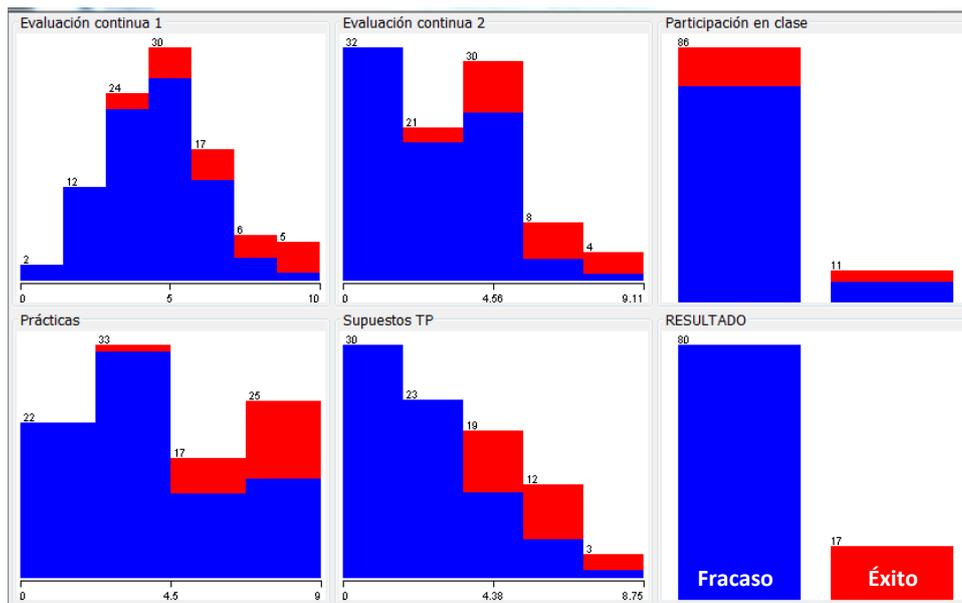


Figura 1. Distribución de valores de los atributos en función de las clases *fracaso* y *éxito*

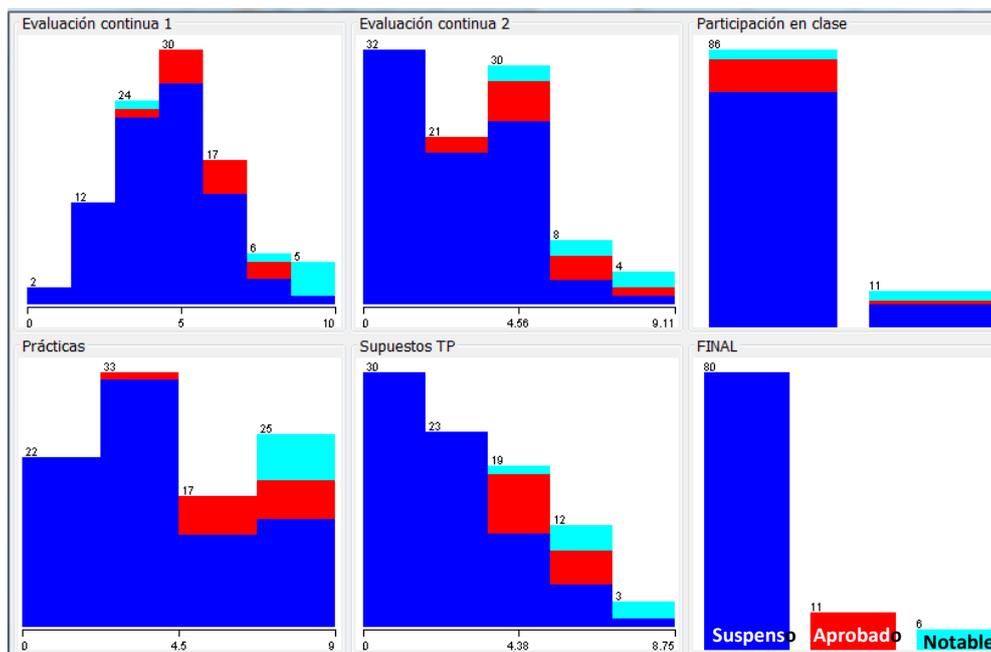


Figura 2. Distribución de valores de los atributos en función de las clases *suspensado*, *aprobado* y *notable*

Gestión de Proyectos

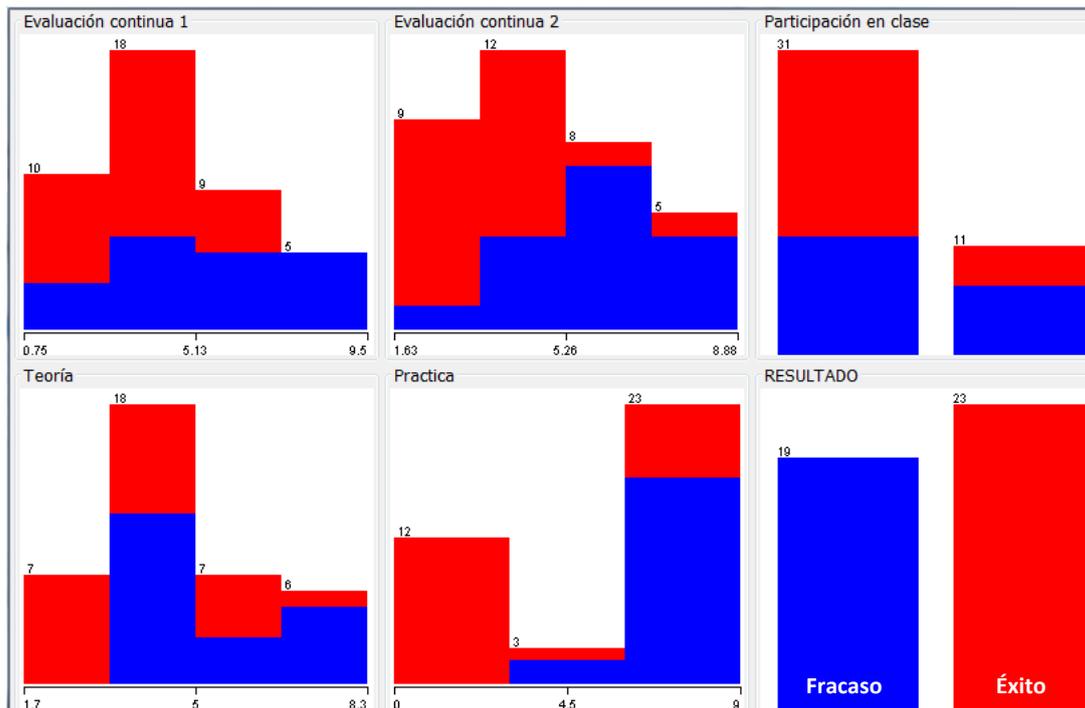


Figura 3. Distribución de valores de los atributos en función de las clases fracaso y éxito

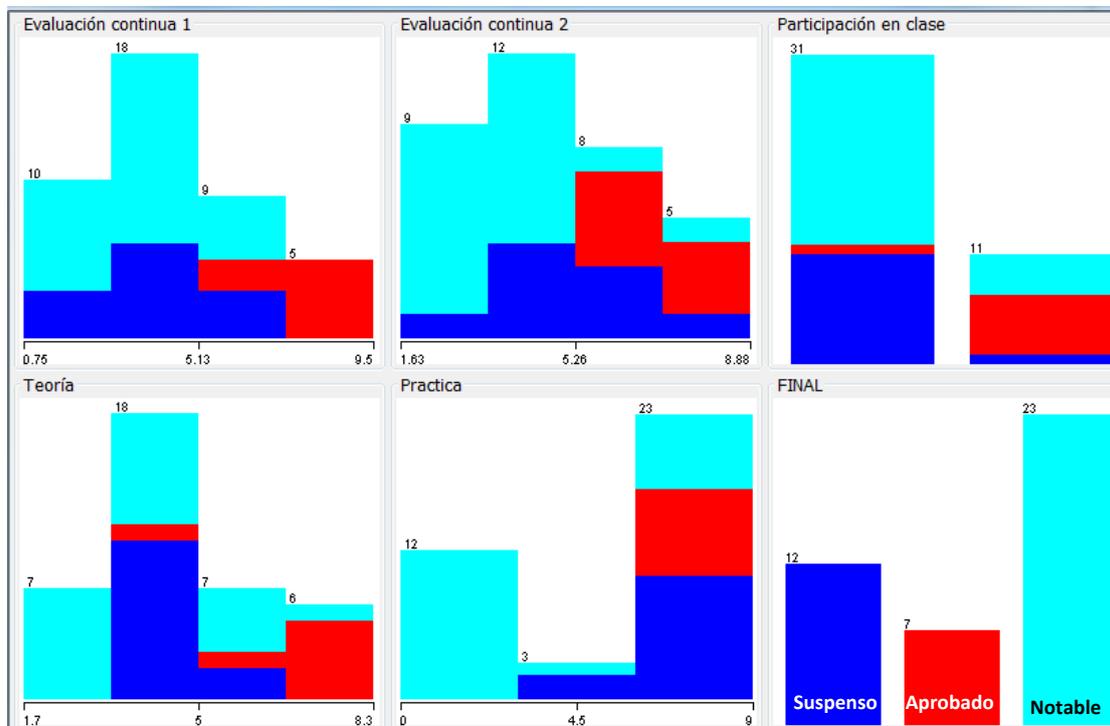


Figura 4. Distribución de valores de los atributos en función de las clases suspense, aprobado y notable

4. Resultados

4.1 Ingeniería del Software II

4.1.1 Estudio del éxito o fracaso

Árbol de decisión J48

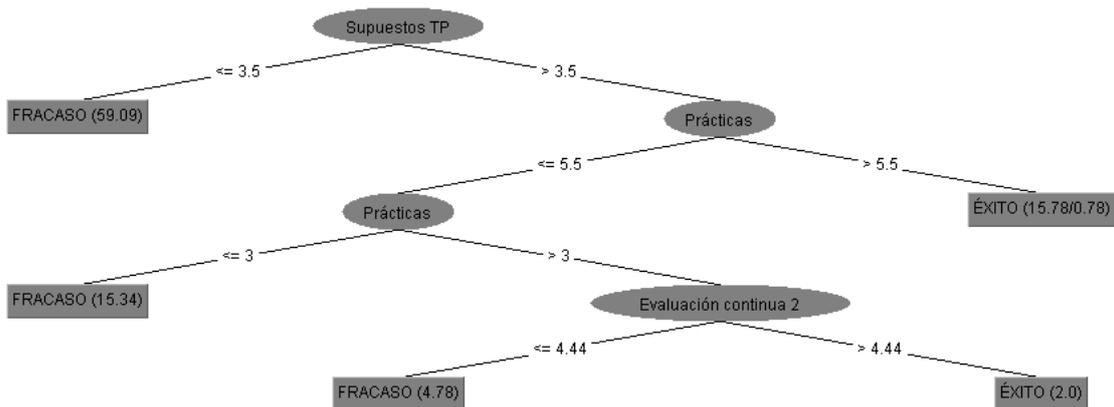


Figura 5. Modelo de reglas del árbol J48

Exactitud	88.66 %
Kappa	0.6166
Error absoluto medio	0.1403

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Clase
0.925	0.294	0.937	0.925	0.931	0.804	FRACASO
0.706	0.075	0.667	0.706	0.686	0.804	ÉXITO

En la figura 5 se puede observar la mayor influencia de las calificaciones obtenidas en el examen teórico práctico en la determinación del fracaso seguido de las calificaciones prácticas y de una parte de la evaluación continua.

Respecto a la calidad del modelo de predicción se puede señalar que éste tiene una exactitud relativamente alta aunque la precisión en la predicción de la clase fracaso es bastante mayor que en la de éxito.

Random tree

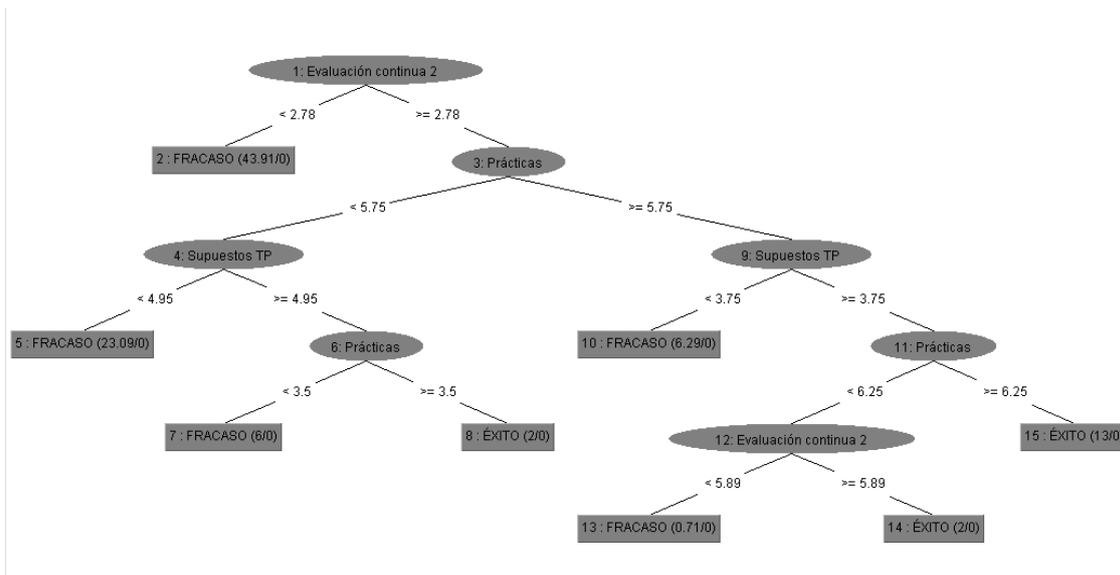


Figura 6. Modelo de reglas del árbol *Random*

Exactitud	87.6289 %
Kappa	0.5721
Error absoluto medio	0.1321

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.925	0.353	0.925	0.925	0.925	0.773	FRACASO
0.647	0.075	0.647	0.647	0.647	0.773	ÉXITO

Este modelo presenta una complejidad mayor que el anterior y en este caso la segunda parte de la evaluación continua aparece como un factor determinante del éxito o fracaso. Tanto en este modelo como en el anterior no aparecen como factores determinantes ni la participación en clase ni la primera parte de la evaluación continua. Los valores de exactitud y precisión de ambas clase son similares a los obtenidos con el algoritmo J48.

Bayes net

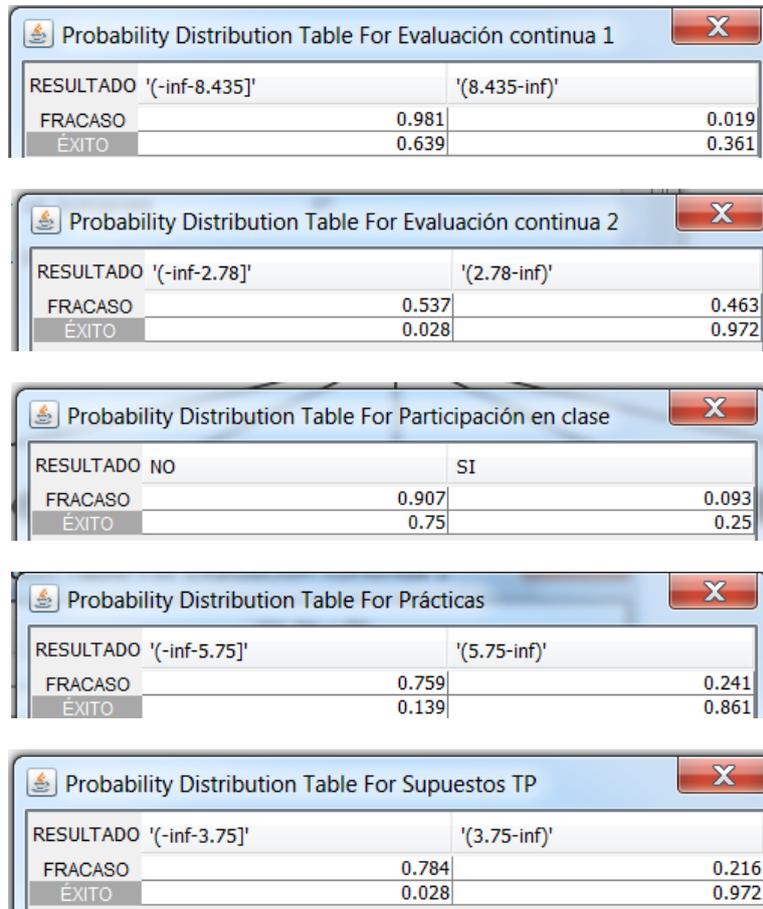


Figura 7. Red de Bayes y distribuciones de probabilidad

Exactitud	91.7526 %
Kappa	0.7008
Error absoluto medio	0.1152

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.963	0.294	0.939	0.963	0.951	0.935	FRACASO
0.706	0.038	0.800	0.706	0.750	0.935	ÉXITO

En el caso de las redes bayesianas los resultados de precisión son bastante mejores que con los árboles de decisión y además presentan otro aspecto positivo ya que las precisiones para ambas clases son más parecidos que en los dos modelos anteriores.

Las distribuciones de probabilidad nos indican el grado en que los valores de los atributos permiten discriminar entre las diferentes clases. Es de destacar el hecho de que la no participación en clase incide significativamente en el fracaso en la asignatura.

Clustering

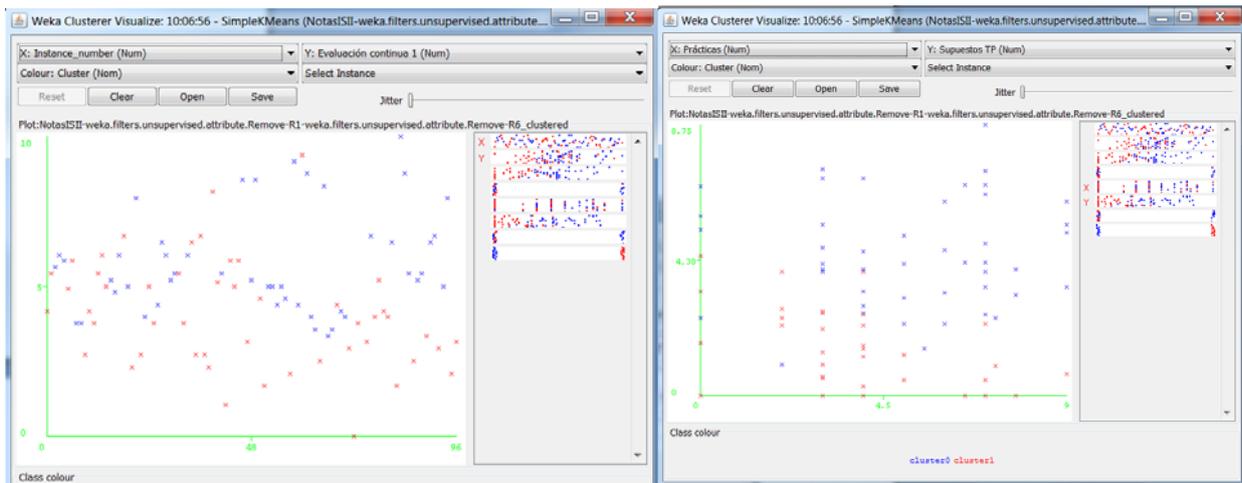


Figura 8. Representación de los *clusters* en función de diferentes variables

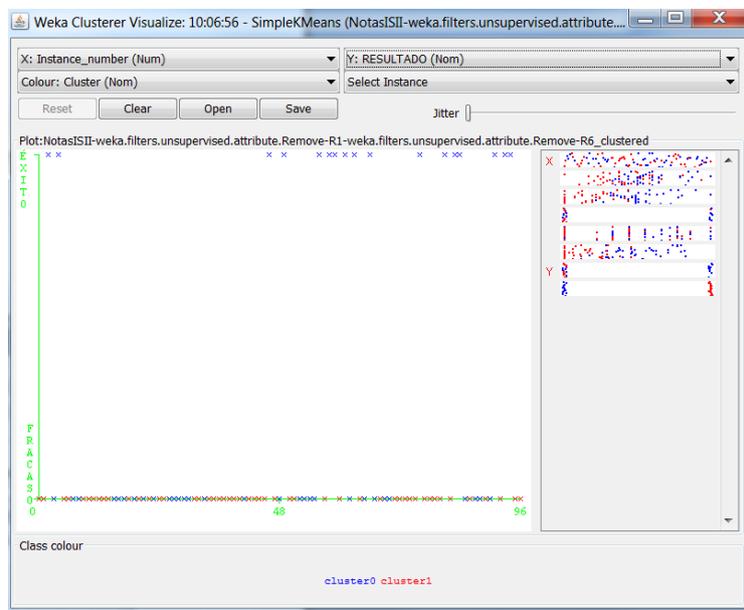


Figura 9. Representación de los *clusters* en función de la clase

La aplicación del método K-medias se aplicó para obtener dos clusters que se esperaba que coincidieran con las diferentes clases sin embargo, tal como puede observarse en la figura 9 los registros de la clase fracaso se encuentran distribuidos entre los dos clusters. Otras representaciones de otras variables en relación con los clusters como las de la figura 8 tampoco permitieron derivar mucha información debido a que en ninguna puede observarse una clara separación de las agrupaciones.

Reglas de asociación de clases

Se aplicó el algoritmo Apriori pero con la restricción de que la parte consecuente de las reglas fuese el atributo de clase. Seguidamente se presentan las diez mejores reglas inducidas, la mayoría de las cuales presentan el factor de confianza máximo (1 o 100%):

1. Evaluación continua 2= $[-\infty, -1.822]$ 32 \Rightarrow RESULTADO=FRACASO conf:(1)
2. Evaluación continua 2= $[-\infty, -1.822]$ Participación en clase=NO \Rightarrow RESULTADO=FRACASO 31 conf:(1)
3. Supuestos TP= $[-\infty, -1.75]$ 30 \Rightarrow RESULTADO=FRACASO conf:(1)
4. Participación en clase=NO Supuestos TP= $[-\infty, -1.75]$ 29 \Rightarrow RESULTADO=FRACASO conf:(1)
5. Supuestos TP= $[1.75, 3.5]$ 23 \Rightarrow RESULTADO=FRACASO conf:(1)
6. Prácticas= $[1.8, 3.6]$ 22 \Rightarrow RESULTADO=FRACASO conf:(1)
7. Participación en clase=NO Prácticas= $[1.8, 3.6]$ 22 \Rightarrow RESULTADO=FRACASO conf:(1)
8. Evaluación continua 1= $[2, 4]$ Participación en clase=NO 28 \Rightarrow RESULTADO=FRACASO conf:(0.96)
9. Evaluación continua 1= $[2, 4]$ 30 \Rightarrow RESULTADO=FRACASO conf:(0.93)
10. Prácticas= $[3.6, 5.4]$ 23 \Rightarrow RESULTADO=FRACASO conf:(0.91)

En las reglas nuevamente vuelve a aparecer la segunda parte de la evaluación continua como factor de influencia significativa en los resultados.

4.1.2 Estudio en función de las calificaciones

Árbol de decisión J48

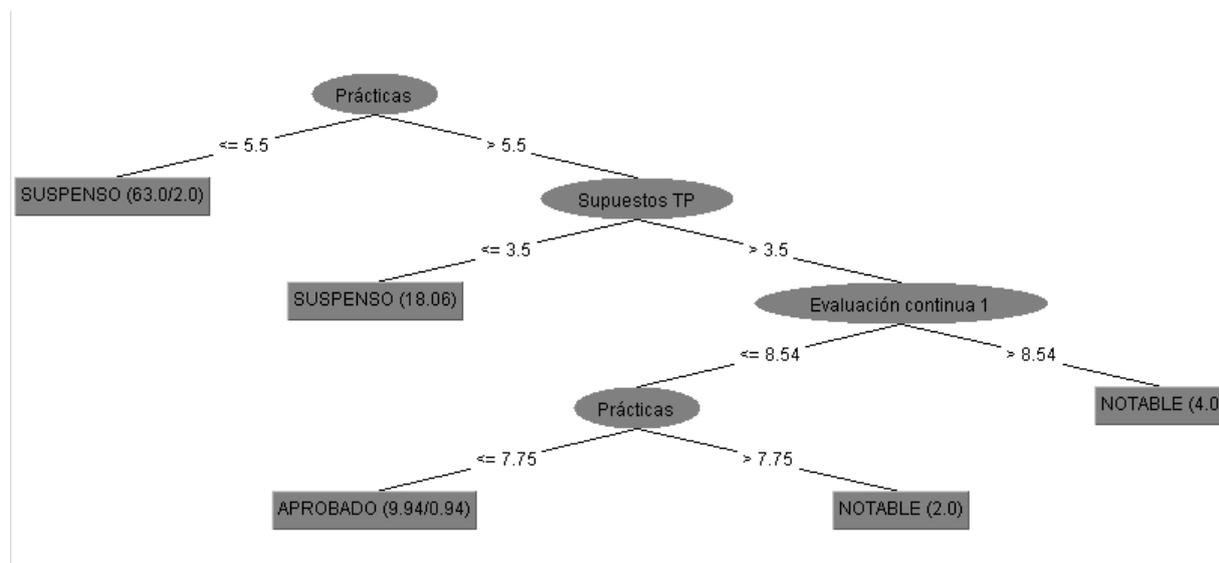


Figura 10. Modelo de reglas del árbol J48

Exactitud	87.6289 %
Kappa	0.6114
Error absoluto medio	0.1018

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.938	0.176	0.962	0.938	0.949	0.868	SUSPENSO
0.545	0.070	0.500	0.545	0.522	0.689	APROBADO
0.667	0.033	0.571	0.667	0.615	0.896	NOTABLE

El árbol de decisión J48 aporta como factor predominante del suspenso la calificación de la parte práctica. Un aspecto a señalar es la aparición de la calificación de la primera parte de la evaluación continua como factor discriminador entre las calificaciones de notable y aprobado.

Los resultados relativos a la calidad de la predicción son muy similares a los obtenidos en el estudio del éxito o fracaso, como puede observarse en las tablas anteriores.

Random tree

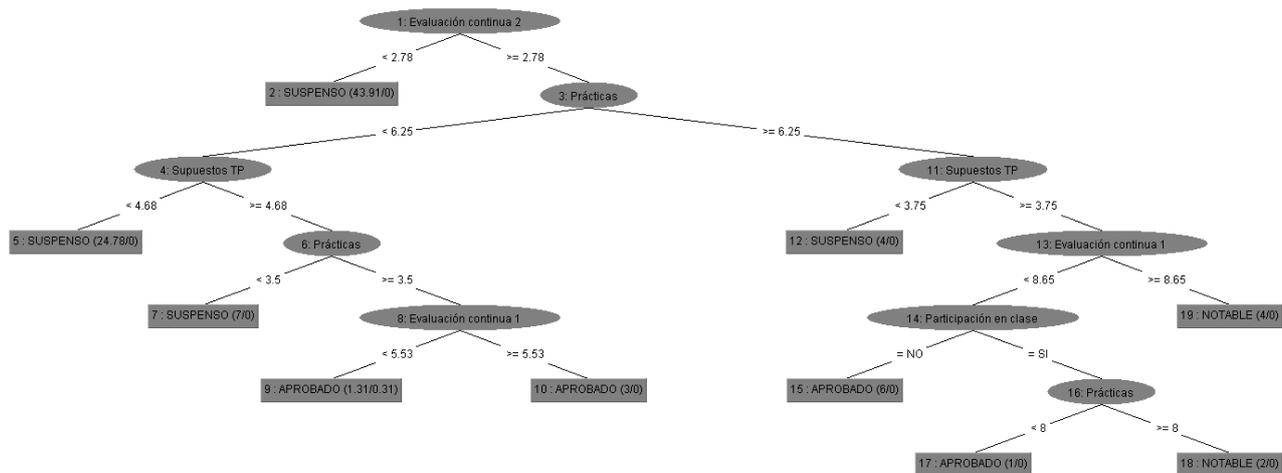


Figura 11. Modelo de reglas del árbol *Random*

Exactitud	88.6598 %
Kappa	0.6061
Error absoluto medio	0.0816

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.963	0.294	0.939	0.963	0.951	0.832	SUSPENSO
0.455	0.058	0.5	0.455	0.476	0.688	APROBADO
0.667	0.011	0.8	0.667	0.727	0.829	NOTABLE

Este modelo corrobora la influencia de la calificación práctica y la segunda parte de la evaluación continua en la calificación de suspenso y el papel discriminador de la primera parte de la evaluación continua entre las notas de aprobado y suspenso.

Los resultados de exactitud, precisión y otras métricas de calidad de la predicción son similares a las del algoritmo J48.

Bayes net

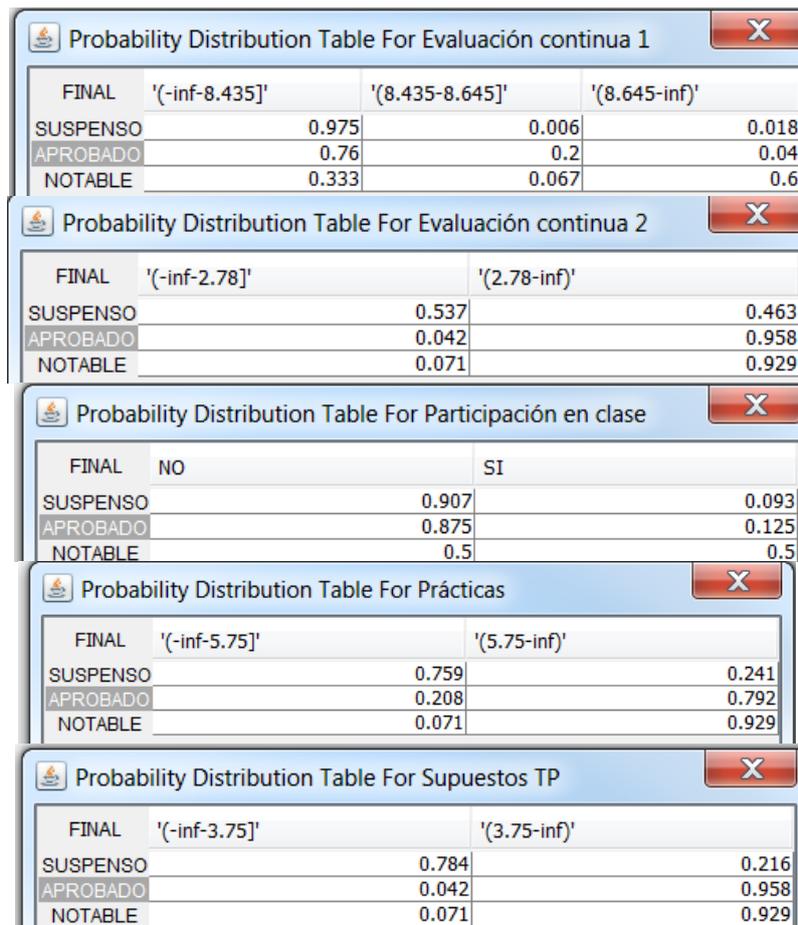
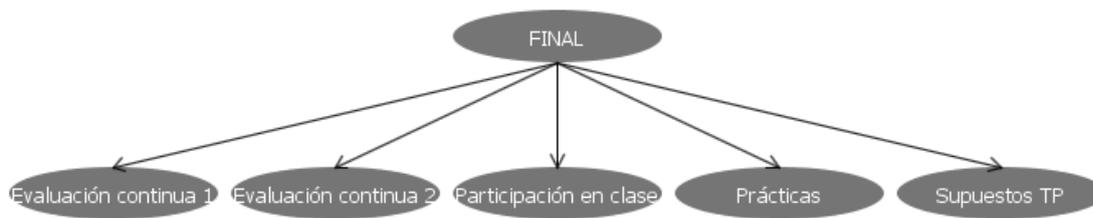


Figura 12. Red de Bayes y distribuciones de probabilidad

Exactitud	86.5979 %
Kappa	0.4785
Error absoluto medio	0.1196

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.988	0.412	0.919	0.988	0.952	0.933	SUSPENSO
0.364	0.058	0.444	0.364	0.400	0.851	APROBADO
0.167	0.011	0.500	0.167	0.25	0.94	NOTABLE

Las distribuciones de probabilidad de la red confirman los resultados de los modelos anteriores y al igual que en el estudio del éxito o fracaso se observa que la no participación en clase es un factor determinante en la obtención de suspenso.

A diferencia de lo ocurrido en el estudio del éxito o fracaso, en este estudio la exactitud de la red de Bayes no mejora respecto a la de los árboles de decisión.

Clustering

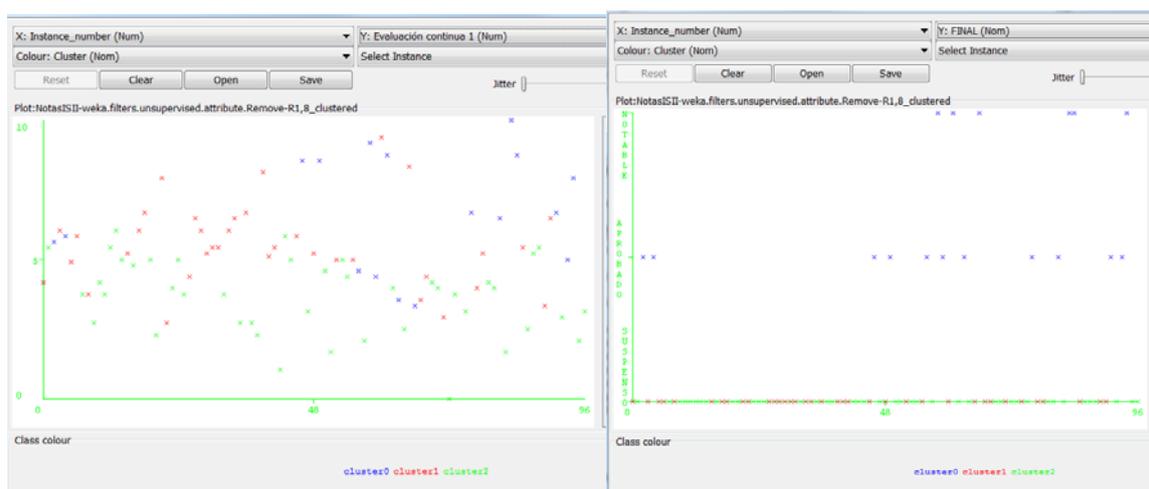


Figura 13. Representación de los *clusters* en función de diferentes variables incluida la clase

Al igual que en el estudio anterior la generación de cluster no proporciona información relevante.

Reglas de asociación

Las diez mejores reglas inducidas confirman los resultados de los modelos predictivos.

1. Evaluación continua 2= $(-\infty-1.822]$ 32 \implies FINAL=SUSPENSO 32 conf:(1)
2. Evaluación continua 2= $(-\infty-1.822]$ Participación en clase=NO 31 \implies FINAL=SUSPENSO 31 conf:(1)
3. Supuestos TP= $(-\infty-1.75]$ 30 \implies FINAL=SUSPENSO 30 conf:(1)
4. Participación en clase=NO Supuestos TP= $(-\infty-1.75]$ 29 \implies FINAL=SUSPENSO 29 conf:(1)
5. Supuestos TP= $(1.75-3.5]$ 23 \implies FINAL=SUSPENSO 23 conf:(1)
6. Prácticas= $(1.8-3.6]$ 22 \implies FINAL=SUSPENSO 22 conf:(1)
7. Participación en clase=NO Prácticas= $(1.8-3.6]$ 22 \implies FINAL=SUSPENSO 22 conf:(1)
8. Evaluación continua 1= $(2-4]$ Participación en clase=NO 28 \implies FINAL=SUSPENSO 27 conf:(0.96)
9. Evaluación continua 1= $(2-4]$ 30 \implies FINAL=SUSPENSO 28 conf:(0.93)
10. Prácticas= $(3.6-5.4]$ 23 \implies FINAL=SUSPENSO 21 conf:(0.91)

4.2 Gestión de Proyectos

4.2.1 Estudio del éxito o fracaso

Árbol de decisión J48

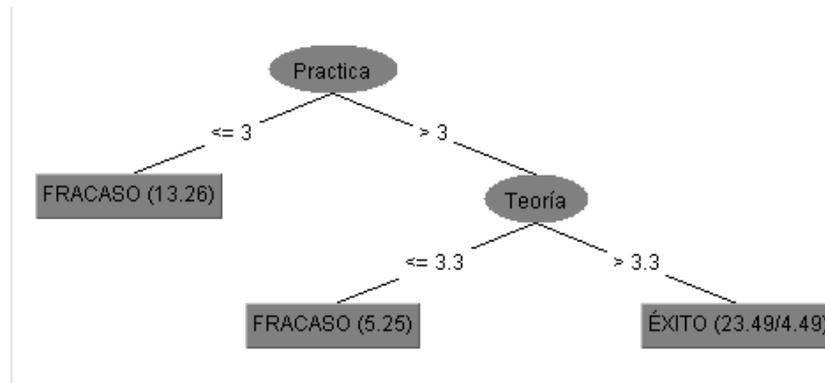


Figura 14. Modelo de reglas del árbol J48

Exactitud	85.7143 %
Kappa	0.7143
Error absoluto medio	0.2025

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Clase
0.895	0.174	0.810	0.895	0.85	0.907	FRACASO
0.826	0.105	0.905	0.826	0.864	0.907	ÉXITO

En la asignatura de Gestión de Proyectos el árbol obtenido es muy simple y únicamente considera los factores teoría y prácticas como determinantes del éxito o fracaso. La exactitud y precisión pueden considerarse aceptables y además, a diferencia de lo que ocurría en la asignatura anterior, no existe mucha diferencia en la precisión para ambas clases.

Random tree

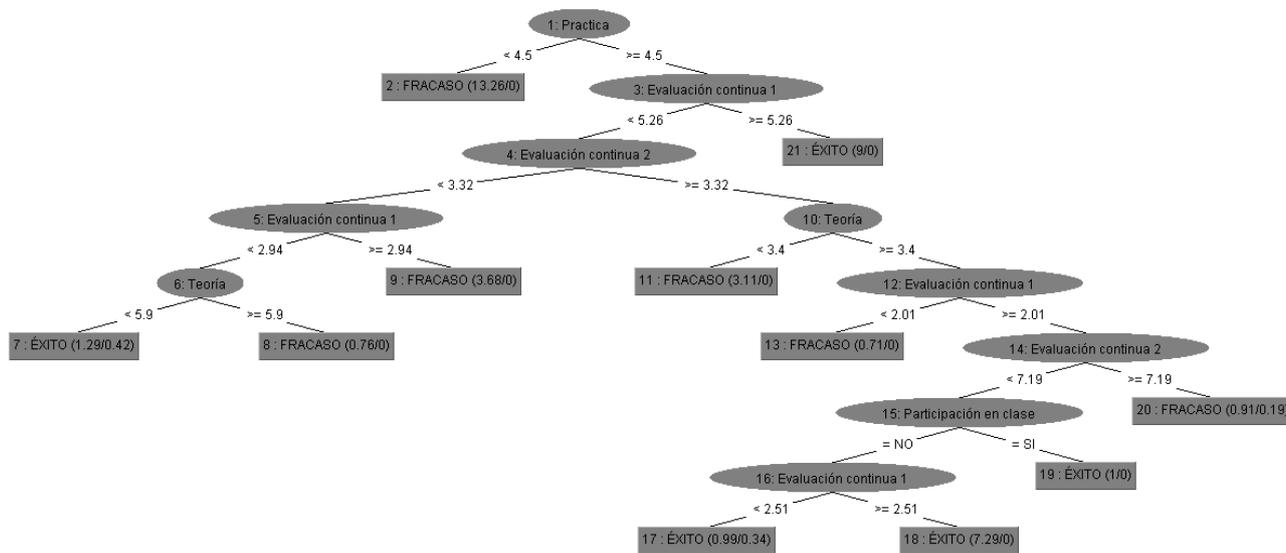


Figura 15. Modelo de reglas del árbol *Random*

Exactitud	85.7143 %
Kappa	0.7117
Error absoluto medio	0.3601

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.842	0.13	0.842	0.842	0.842	0.874	FRACASO
0.87	0.158	0.870	0.870	0.870	0.862	ÉXITO

Con el algoritmo random tree se obtiene un modelo más complejo en el que nuevamente aparece la calificación de prácticas como el factor más influyente. La diferencia principal respecto al árbol J48 se encuentra en que las calificaciones de las dos partes de la evaluación continua aparecen en nodos superiores del árbol por lo que pueden considerarse factores influyentes en el éxito o fracaso.

Las métricas de calidad del modelo predictivo son similares para ambos algoritmos.

Bayes net

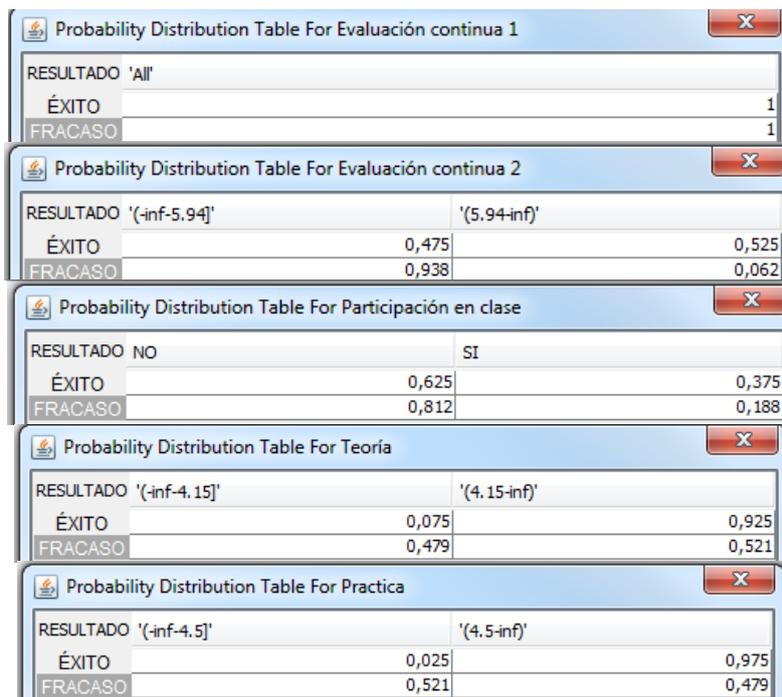


Figura 16. Red de Bayes y distribuciones de probabilidad

Exactitud	71.4286 %
Kappa	0.4233
Error absoluto medio	0.3298

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.684	0.261	0.684	0.684	0.684	0.771	FRACASO
0.739	0.316	0.739	0.739	0.739	0.770	ÉXITO

Las distribuciones de probabilidad nos indican que la calificación de la parte práctica es el factor más relevante. En esta asignatura los resultados de precisión y exactitud son peores que en los árboles de decisión a diferencia de lo que ocurría en la asignatura de Ingeniería del Software.

Clustering

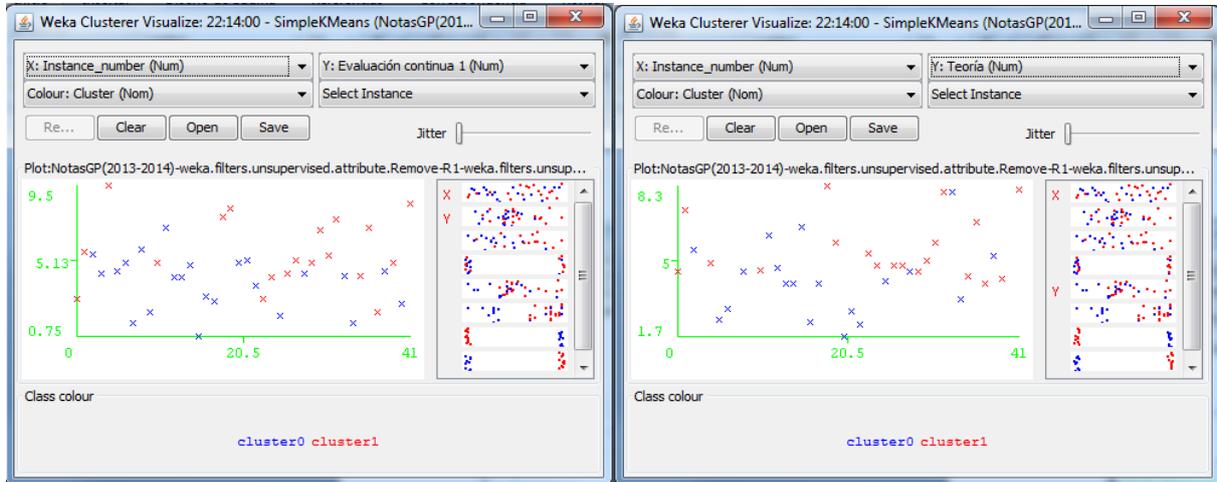


Figura 17. Representación de los *clusters* en función de diferentes variables

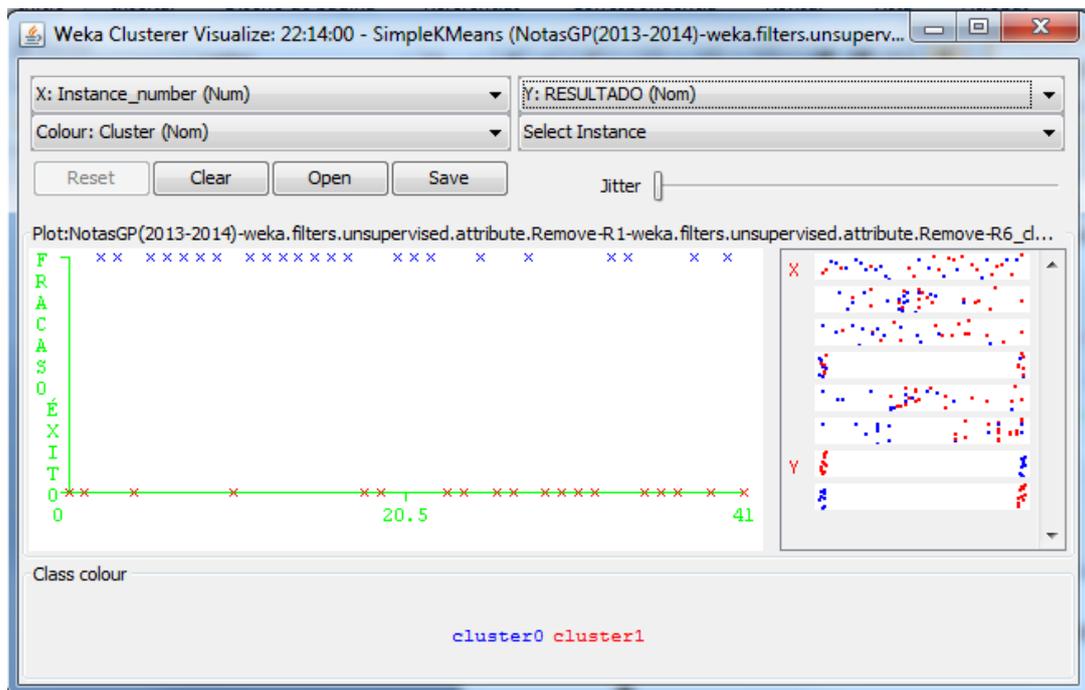


Figura 18. Representación de los *clusters* en función de la clase

Una diferencia significativa respecto a la asignatura anterior es el hecho de que en esta asignatura los clusters se corresponden exactamente con las clases éxito y fracaso. Asimismo la diferenciación entre clusters en las representaciones de las diferentes variables es más clara que en el caso anterior.

Reglas de asociación de clases

Las diez mejores reglas obtenidas tienen todas la confianza máxima del 100%. : En ellas podemos observar la combinación de valores de atributos que conducen al éxito o fracaso. La calificación práctica vuelve a ser un factor determinante.

1. Practica='(1.8-3.6]' 10 ==> RESULTADO=FRACASO conf:(1)
2. Participación en clase=NO Practica='(1.8-3.6]' 9 ==> RESULTADO=FRACASO conf:(1)
3. Evaluación continua 2='(5.98-7.43]' 7 ==> RESULTADO=ÉXITO conf:(1)
4. Teoría='(-inf-3.02]' 6 ==> RESULTADO=FRACASO conf:(1)
5. Evaluación continua 2='(5.98-7.43]' Practica='(7.2-inf)' 6 ==> RESULTADO=ÉXITO conf:(1)
6. Evaluación continua 1='(4.25-6]' Practica='(1.8-3.6]' 5 ==> RESULTADO=FRACASO conf:(1)
7. Participación en clase=NO Teoría='(-inf-3.02]' 5 ==> RESULTADO=FRACASO conf:(1)
8. Teoría='(-inf-3.02]' Practica='(7.2-inf)' 5 ==> RESULTADO=FRACASO conf:(1)

4.2.2 Estudio en función de las calificaciones

Árbol de decisión J48

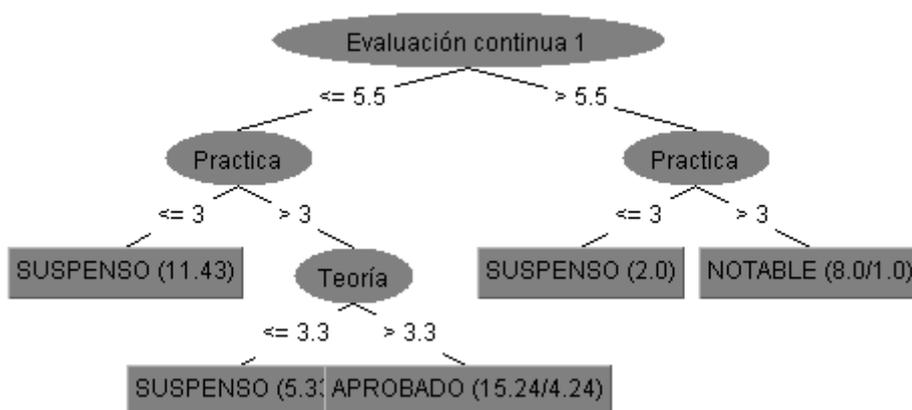


Figura 19. Modelo de reglas del árbol J48

Exactitud	78.5714 %
Kappa	0.6516
Error absoluto medio	0.1968

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.750	0.133	0.692	0.750	0.720	0.767	SUSPENSO
0.714	0.114	0.556	0.714	0.625	0.831	APROBADO
0.826	0.053	0.95	0.826	0.884	0.878	NOTABLE

Al igual que en la asignatura de Ingeniería del Software, la primera parte de la evaluación continua aparece como factor influyente en las calificaciones a pesar de que en el estudio del éxito o fracaso no aparecía. La calificación práctica se confirma como factor determinante en esta asignatura.

Las métricas de calidad no son excesivamente buenas ya que la exactitud no es muy alta y existen diferencias significativas en las precisiones para las diferentes clases.

Random tree

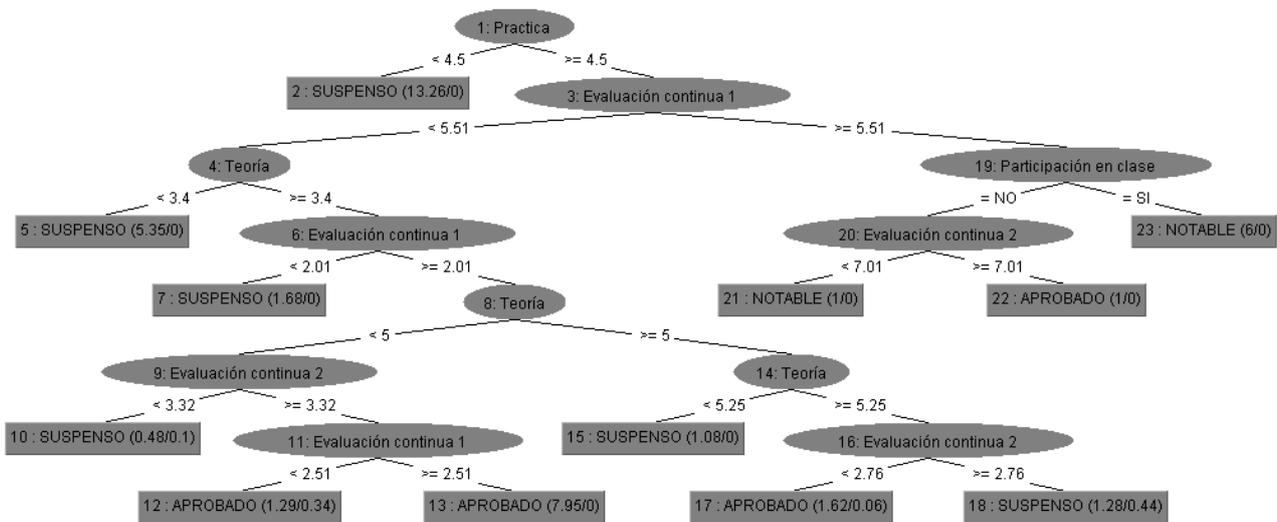


Figura 20. Modelo de reglas del árbol *Random*

Exactitud	71.4286 %
Kappa	0.5218
Error absoluto medio	0.2171

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.75	0.233	0.563	0.75	0.643	0.756	SUSPENSO
0.571	0.029	0.800	0.571	0.667	0.780	APROBADO
0.739	0.211	0.81	0.739	0.773	0.780	NOTABLE

A pesar de que el árbol es bastante más complejo que el J48 los factores predictores son más o menos los mismos, al igual que las métricas de calidad de la predicción.

Bayes net



Probability Distribution Table For Evaluación continua 1			
FINAL	'(-inf-5.565]'	'(5.565-inf)'	
APROBADO	0,885	0,115	0,115
NOTABLE	0,062	0,938	0,938
SUSPENSO	0,896	0,104	0,104

Probability Distribution Table For Evaluación continua 2			
FINAL	'(-inf-6.19]'	'(6.19-inf)'	
APROBADO	0,808	0,192	0,192
NOTABLE	0,062	0,938	0,938
SUSPENSO	0,938	0,062	0,062

Probability Distribution Table For Participación en clase			
FINAL	NO	SI	
APROBADO	0,885	0,115	0,115
NOTABLE	0,188	0,812	0,812
SUSPENSO	0,812	0,188	0,188

Probability Distribution Table For Teoría			
FINAL	'(-inf-6.6]'	'(6.6-inf)'	
APROBADO	0,962	0,038	0,038
NOTABLE	0,312	0,688	0,688
SUSPENSO	0,938	0,062	0,062

Probability Distribution Table For Practica			
FINAL	'(-inf-4.5]'	'(4.5-inf)'	
APROBADO	0,038	0,962	0,962
NOTABLE	0,062	0,938	0,938
SUSPENSO	0,521	0,479	0,479

Figura 21. Red de Bayes y distribuciones de probabilidad

Exactitud	64.2857 %
Kappa	0.3719
Error absoluto medio	0.2541

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.25	0.133	0.429	0.25	0.316	0.743	SUSPENSO
0.857	0.057	0.75	0.857	0.800	0.959	APROBADO
0.783	0.474	0.667	0.783	0.720	0.807	NOTABLE

Las distribuciones de probabilidad nos indican que las dos partes de la evaluación continua tienen una influencia significativa en la calificación de notable, al igual que la nota de prácticas y la participación en clase.

Los resultados de exactitud y precisión son peores que en los modelos de árboles.

Clustering

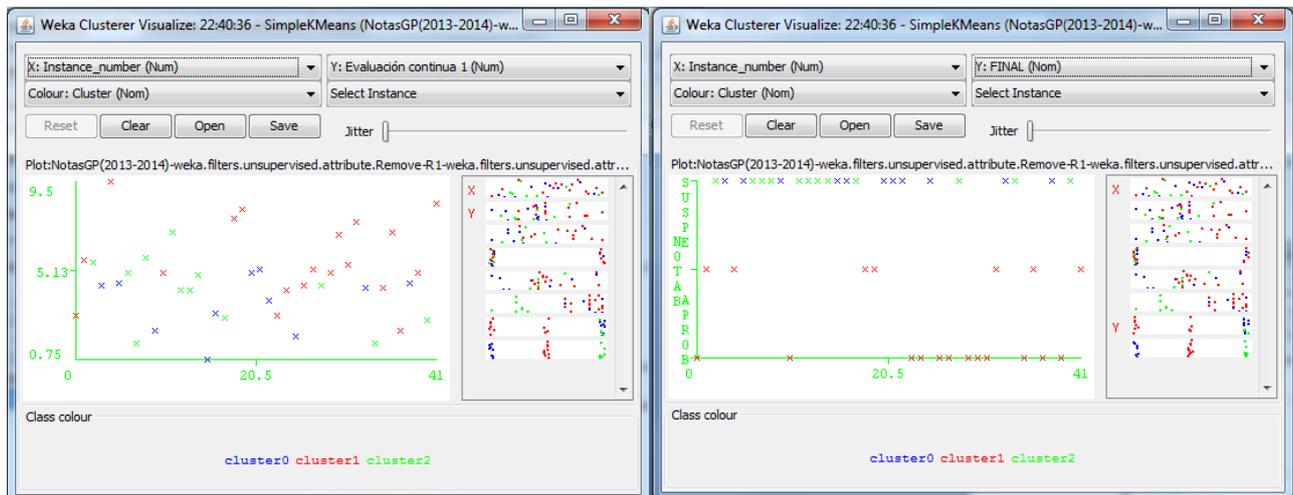


Figura 22. Representación de los *clusters* en función de diferentes variables incluida la clase

No se observa una diferenciación clara de los clusters con respecto a los valores de ninguna de las variables, por lo que no se pueden extraer conclusiones claras de los mismos.

Reglas de asociación de clases

Sólo se han obtenido 6 reglas con un factor de confianza superior al umbral del 90% establecido. En las mismas vuelve a aparecer la nota de prácticas como atributo determinante de la calificación final. Las reglas son las siguientes

1. Practica='(1.8-3.6]' 10 ==> FINAL=SUSPENSO conf:(1)
2. Participación en clase=NO Practica='(1.8-3.6]' ==> FINAL=SUSPENSO 9 conf:(1)
3. Teoría='(-inf-3.02]' 6 ==> FINAL=SUSPENSO conf:(1)
4. Evaluación continua 1='(4.25-6]' Practica='(1.8-3.6]' 5 ==> FINAL=SUSPENSO conf:(1)
5. Participación en clase=NO Teoría='(-inf-3.02]' 5 ==> FINAL=SUSPENSO conf:(1)
6. Teoría='(-inf-3.02]' Practica='(7.2-inf)' 5 ==> FINAL=SUSPENSO conf:(1) 9. Evaluación continua 1='(2-4]' 30 ==> RESULTADO=FRACASO conf:(0.93)

5. Conclusiones

En este proyecto se han aplicado diferentes técnicas de minería de datos a los resultados obtenidos en la evaluación de asignaturas relacionadas con la materia de Ingeniería del Software del Grado en Ingeniería Informática de la Universidad de Salamanca. El objetivo fundamental ha sido la identificación de los factores que conducen a buenos o malos resultados en la evaluación de los estudiantes y el inducir modelos predictivos que nos permitan estimar resultados en futuros cursos académicos.

Tal como se describe en el apartado de resultados, se han identificado algunos factores de significativa influencia en los resultados que sin embargo no son los que mayor peso tienen en el cálculo de la nota final por lo que esta información puede considerarse muy valiosa para incidir especialmente en estos factores en sucesivos cursos. Por ejemplo, en la asignatura de Ingeniería del Software II, la segunda parte de la evaluación continua, la cual solo tiene un peso de un 10% en la nota final, se ha revelado como uno de los factores determinantes para superar la asignatura. En Gestión de Proyectos también aparece ese factor como significativo aunque superado por la parte práctica de la asignatura, la cual tiene un peso de un 35% en la nota final. En ambas asignaturas se ha observado que la primera parte de la evaluación continua incide directamente en la obtención de la calificación de notable. Otro aspecto a destacar es el hecho de que la no participación en clase, la cual tiene un peso despreciable en el cálculo de la nota final, repercute significativamente en la obtención de suspenso.

En cuanto a la precisión de los modelos predictivos se puede considerar que ésta es aceptable dada la pequeña cantidad de datos que se maneja. A medida que se vayan obteniendo más resultados en sucesivos cursos académicos se pueden ir obteniendo nuevos modelos con mayor precisión.