MEMORIA DE LAS ACCIONES DESARROLLADAS.
PROYECTOS DE MEJORA DE LA CALIDAD DOCENTE.
VICERRECTORADO DE PLANIFICACIÓN Y CALIDAD.
XII CONVOCATORIA (2010-2011)

UNIDAD DE CALIDAD
UNIVERSIDAD DE CÓRDOBA

DATOS IDENTIFICATIVOS:

1. Título del Proyecto

Modelos para la predicción del abandono del alumnado en las nuevas titulaciones de grado de la Escuela Politécnica Superior mediante técnicas de minería de datos

2. Código del Proyecto

106025

3. Resumen del Proyecto

El objetivo principal de este proyecto ha consistido en la obtención, mediante técnicas de minería de datos de modelos de predicción del abandono de alumnos en la Universidad de Córdoba. Para ello, se ha recabado información relacionada con este fenómeno, se ha procesado para tenerla disponible en formato electrónico, se han aplicado distintas técnicas de minería de datos para la obtención de modelos, y se ha comprobado la calidad de los mismos. Posteriormente, habrá que analizar la utilidad de dichos modelos, comprobando si es posible utilizarlos para intervenir sobre el alumnado en situación de riesgo, y si dichas intervenciones ayudan a reducir la tasa de fracasos de nuestros estudiantes. Nuestra investigación se ha centrado en las titulaciones de la Escuela Politécnica Superior de la Universidad de Córdoba, el centro al que estamos adscritos todos los miembros del equipo de trabajo.

4. Coordinador del Proyecto

Nombre y Apellidos	Departamento	Código del Grupo Docente	Categoría Profesional
Sebastián Ventura Soto	Informática y Análisis Numérico	4	PDI

5. Otros Participantes

Nombre y Apellidos	Departamento	Código del Grupo Docente	Categoría Profesional
Cristóbal Romero Morales	Informática y Análisis Numérico	4	PDI
Amelia Zafra Gómez	Informática y Análisis Numérico	4	PDI
María Luque Rodríguez	Informática y Análisis Numérico	4	PDI
Eva L. Gibaja Galindo	Informática y Análisis Numérico	4	PDI
Jose L. Ávila Jimenez	Informática y Análisis Numérico	4	PDI
Juan Luis Olmo Ortiz	Informática y Análisis Numérico	4	Becario
José María Luna Ariza	Informática y Análisis Numérico	4	Becario
Alberto Cano Rojas	Informática y Análisis Numérico	4	Alumno
Aurora Ramírez Quesada	Informática y Análisis Numérico	4	Alumna

6. Asignaturas afectadas

Nombre de la asignatura	Área de conocimiento	Titulación/es
Introducción a la Programación	Ciencias de la Computación	Grado en Ingeniería Informática
Metodología de la Programación	Ciencias de la Computación	Grado en Ingeniería Informática
Fundamentos de Informática	Ciencias de la Computación	Grado en Ingeniería Electrónica Industrial
Fundamentos de Informática	Ciencias de la Computación	Grado en Ingeniería Electricidad
Fundamentos de Informática	Ciencias de la Computación	Grado en Ingeniería Mecánica

MEMORIA DE LA ACCIÓN
Especificaciones

Utilice estas páginas para la redacción de la Memoria de la acción desarrollada. La Memoria debe contener un mínimo de cinco y un máximo de diez páginas, incluidas tablas y figuras, en el formato indicado (tipo y tamaño de fuente: Times New Roman, 12; interlineado: sencillo) e incorporar todos los apartados señalados (excepcionalmente podrá excluirse alguno). En el caso de que durante el desarrollo de la acción se hubieran producido documentos o material gráfico dignos de reseñar (CD, páginas Web, revistas, vídeos, etc.) se incluirá como anexo una copia de buena calidad.

Apartados

1. Introducción (justificación del trabajo, contexto, experiencias previas etc.)

Los nuevos títulos de grado llevan asociado un gran cambio en la forma tanto de impartir como de evaluar la enseñanza universitaria. Uno de los problemas que intenta resolver el cambio metodológico que se avecina es el del fracaso o abandono universitario. En los últimos años, el número de alumnos que abandonan los estudios universitarios es muy importante, y desde la administración se insiste en la necesidad de desarrollar mecanismos para combatir este fenómeno. De hecho, en el apartado 8 de la memoria de los títulos de grados de nuestra Universidad, denominado "Resultados Previstos" hay que presentar una estimación realista de las tasas de graduación, abandono y eficiencia y el progreso de los alumnos de esa titulación. Estos valores deben de ser lo más realistas posibles (normalmente obtenidos de años anteriores), ya que uno de los aspectos a considerar en la evaluación de las titulaciones por la ANECA será, precisamente, el cumplimiento de estos objetivos.

2. Objetivos (concretar qué se pretendió con la experiencia)

Como ya se ha comentado, el objetivo fundamental de este proyecto ha sido la obtención de modelos de predicción de abandono de los estudiantes en los primeros cursos de las titulaciones de la Escuela Politécnica Superior. Dicho objetivo principal puede desglosarse en los siguientes objetivos secundarios:

- Identificación de las posibles variables que afectan al abandono. Se trata de analizar cuál es la información que vamos a recopilar para obtener nuestros modelos. Para ello, consultaremos la bibliografía disponible.
- Captura y recogida de la mayor cantidad posible de datos referentes a las anteriores variables.
  - Encuestas.
  - Información proporcionada por el profesor.
  - Información proporcionada por sistemas de enseñanza on-line.
- Aplicación de algoritmos de minería de datos sobre los datos anteriores.
  - Preprocesado de los datos
  - Clasificación/Predicción
- Analizar los resultados y modelos obtenidos.

3. Descripción de la experiencia (exponer con suficiente detalle lo realizado en la experiencia)

Las principales actividades realizadas a lo largo de la experiencia han sido las siguientes:

1. Identificación de posibles variables que puedan afectar al abandono. En esta fase del proyecto realizamos una búsqueda bibliográfica sobre el problema de la predicción del abandono de estudiantes universitarios, para recopilar cuáles son las variables que se utilizan con más frecuencia en la construcción de este tipo de modelos. Una vez recopilada, diseñamos una serie de encuestas que han sido utilizadas en la siguiente fase de recogida de datos.
2. Recogida y captura de datos. La fase de recogida de datos comprendió varias etapas distintas. En primer lugar, se elaboró una encuesta inicial, que se pasó a los alumnos de primer curso el día comienzo del curso, aprovechando la jornada de recepción de alumnos de nuevo ingreso. Además de esta encuesta inicial, se realizaron algunas encuestas durante el desarrollo del curso. Para el desarrollo de estas encuestas, utilizamos la plataforma MOODLE de la Universidad de Córdoba. También recopilamos información de uso de la plataforma MOODLE en distintas asignaturas. Esta información se recabó periódicamente, así como la de otras actividades que nos proporcionen los profesores de las titulaciones. Una vez recopilada la información, hay que pasarla a formato electrónico. Para ello, instalamos un sistema de gestión de bases de datos en un servidor suficientemente potente, que también alberga el software de minería de datos que utilizamos en la siguiente fase. Para finalizar, realizamos una serie de operaciones para preparar los datos para el proceso de minería.
3. Aplicar técnicas de minería de datos. Obtenidos todos los datos, procedimos a aplicar distintas técnicas de extracción de conocimiento para obtener modelos predictivos de la tasa de abandono de los estudiantes. Para ello utilizamos algoritmos de clasificación estándar como, por ejemplo, el J48 (mejora del C4.5) disponible en la herramienta WEKA.
4. Analizar los resultados y modelos obtenidos. En esta última fase se analizaron los modelos obtenidos. En primer lugar, se analizó la calidad predictiva de los mismos, utilizando las métricas que se emplean convencionalmente en minería de datos (exactitud, precisión, especificidad…). En segundo lugar, estudiamos la utilidad de los modelos obtenidos, es decir, si es posible detectar e intervenir sobre los alumnos detectados, para reducir la tasa de fracaso.

4. Materiales y métodos (describir la metodología seguida y, en su caso, el material utilizado)

Para la recogida de información de los alumnos se han utilizado las siguientes fuentes y métodos:

- Se han obtenido información de entrada de los alumnos a la Universidad de Córdoba. Para ello se han utilizado los datos de las distintas fases de adjudicación de plazas tanto en Junio como en Septiembre de 2010: número, dni, nombre, via/opción/ciclo formación, nota y fecha de matrícula.
- Se han contabilizado la asistencia de los alumnos a las horas de prácticas de las asignaturas. Para ello se ha utilizado la hoja de firmas que pasa el profesor al comienzo de cada sesión de

prácticas. Se ha contabilizado el número total de horas de prácticas y el número de asistencia por cada alumno y se ha calculado

- Se ha realizado un seguimiento de la realización de actividades dentro del sistema de enseñanza virtual Moodle. Para ello se han utilizado la información de logs o histórico de acciones realizadas por cada alumno dentro de la asignatura, ya que cada asignatura presencial tiene una asignatura virtual con material y actividades. La información disponible ha sido: total actividades, número de actividades realizadas, número de accesos, número de menos de 30 segundos, número de recursos visitados. A partir de toda esta información se ha obtenido un único valor que mide el nivel de participación del alumno en la asignatura virtual que puede tener 3 valores (alto, medio y bajo).
- Se ha realizado de forma electrónica (como una actividad tipo test dentro de MOODLE) un encuesta de 20 preguntas, normalmente con 5 posibles respuestas/opciones (Muy alto, Alto, Medio, Bajo y Muy bajo):

    1. Sobre la calidad de los recursos y materiales utilizados para impartir tanto las clases (aulas, laboratorios, ordenadores, etc.), ¿Qué nivel de calidad tienen para ti?
    2. ¿Qué nivel de dificultad y exigencia tiene para ti en general esta carrera/titulación?
    3. ¿En qué nivel crees que los conocimientos que estas adquiriendo te serán luego de utilidad en tu vida profesional?
    4. ¿Con qué grado has utilizado las tutorías y a los asesores académicos?
    5. ¿Cúal es tu grado de asistencia a todas las clases?
    6. ¿Qué edad tienes?
    7. Cuándo te matriculaste en esta carrera/titulación era o no la que realmente querías estudiar. ¿Que ganas/interés tenías por estudiar esta carrera en concreto y no otra?
    8. Tus expectativas de aprobar este curso son:
    9. Con respecto a cuántos amigos tienes en clase. ¿Cúal consideras que es tu nivel de integración en clase?
    10. Tu nivel de motivación por aprender lo que se enseña en la titulación en general es:
    11. Con respecto a tu persistencia para acabar la titulación a pesar de los obstáculos que te puedas encontrar, tu nivel de persistencia es:
    12. Las notas que has obtenido en tus anteriores estudios (bachillerato o ciclo formativo) antes de entrar a la Universidad, han sido en general:
    13. Una vez que ya conoces mejor en que consiste tu titulación y que es lo que se enseña realmente: ¿Cuánto te gusta? ¿Cuál es tu grado de satisfación con la titulación?
    14. Con respecto a las técnicas de estudio como por ejemplo: subrayar los apuntes, realizar resumenes y esquemas generales,lectura previa al comenzar a estudiar y repasar temas estudiados, adoptar una actitud crítica realizandote preguntas sobre lo que lees, realizacion de sintesis, cuadros, diagramas, etc. ¿En qué grado utilizas estas y otras técnicas de estudio?
    15. Con respecto al tiempo que dedicas a las asignaturas, es decir, al trabajo personal diario adicional a la asistencia a clase. Consideras que el tiempo diario que le dedicas es:
    16. ¿Qué nivel de apoyo, preocupación y motivación tienen tus padres ante tus estudios?
    17. ¿Qué nivel económico tiene tu familia?
    18. Con respecto a si los profesores dialogan con los alumno sobre la marcha de las clases, si tienen en cuenta las opiniones de los alumnos y si motivan a los alumnos en la asignatura. En general, ¿Cual crees que es el nivel de implicación de los profesores?

19. Sobre la forma de enseñar de los profesores, los métodos docentes utilizados y la calidad educativa en las clases. Tu grado de satisfaccion es:
20. Con respecto a las técnicas de evaluación que utilizan los profesores en las asignaturas: exámenes finales, correción de trabajos y prácticas, exposiciones orales, etc. Tu nivel de satisfacción sobre las técnicas utilzadas es:

- Se han recogido las notas finales obtenidas por los alumnos en las asignaturas. Para ello se han utilizado las notas de las actas publicadas por el profesor tras finalizar el cuatrimestre. Se ha utilizado los valores categóricos: Aprobado, Suspenso, y No presentado. Finalmente, se han integrado toda la información anterior sobre los alumnos (26 variables) en una única tabla resumen donde cada fila o registro contiene toda la información disponible de cada alumno.

Información de Entrada					Información a Predecir/Salida
ID Alumno (1)	Info Entrada UCO (2)	Asistencia Prácticas (3)	Utilización Moodle (4)	Encuesta (5- 25)	Examen (26)

Indicar tambien, que se han eliminado del fichero de datos (es decir, no se han tenido en cuenta) a todos los alumnos que no realizaron la encuesta. Esto es debido a que la encuesta es la fuente de información más amplia de la que disponemos (20 variables de las 26 totales).

5. Resultados obtenidos y disponibilidad de uso (concretar y discutir los resultados obtenidos y aquéllos no logrados, incluyendo el material elaborado y su grado de disponibilidad)

En la siguiente Tabla se muestra el número total de alumnos utilizados en la experiencia en cada una de las asignaturas de grado, junto con las diferentes tasas o porcentajes de aprobación, reprobación y abandono, obtenidas en la convocatoria de Junio 2011.

Asignatura	**Número Total de Alumnos**	**Tasa de Aprobación**	**Tasa de Reprobación**	**Tasa de Abandono**
Grado en Informática	88	41,37%	27,58%	31’03%
Grado en Electrónica	80	61,25%	26,25%	12,50%
Grado en Electricidad	25	32%	64%	4%
Grado en Mecánica	72	47,22%	26,38%	26,38%

En un primer experimento se han intentado predecir los tres anteriores valores (aprobado, suspenso y no presentado) en función de toda la información previa de los alumnos (notas de entrada a la Universidad, asistencia a prácticas, realización de actividades en Moodle, y encuesta). Para ello se ha utilizado la herramienta de minería de datos Weka y el ampliamente conocido algoritmo de clasificación J48 (versión del C45) que es además muy utilizado en sistemas de toma de decisiones debido a la alta comprensibilidad de los modelos obtenidos. Tras ejecutar un 10 cross-validation (los datos se dividen en 10 ficheros para entrenamiento y 10 para pruebas) con los parámetros por defecto del algoritmo J48 los resultados obtenidos son los siguientes:

Alumnos correctamente clasificados: 38 (43.1818 %)
Alumnos clasificados incorrectamente: 50 (56.8182 %)

Por lo que se puede ver, el Error Medio Absoluto: 0.396 es muy alto, es decir, el modelo es poco preciso en la clasificación. Podemos ver estas clasificaciones en cada una de las clases en la matriz de confusión:

a b c <-- classified as
20 8 8 | a = Aprobado
10 **16** 2 | b = No-presentado
13 9 **2** | c = Suspenso

Se puede apreciar como más de la mitad de los Aprobados (20) y de los No-presentados (16) son clasificados correctamente, pero muy pocos Suspensos (2) son correctamente clasificados. Finalmente, con respecto al modelo de predición finalmente obtenido es el siguiente árbol de clasificación, donde se obtienen reglas de clasificación IF-THEN directamente con sólo recorrer el árbol desde el nodo raíz a cada una los nodos terminales:

```
ASISTENCIAPRACTICAS = ALTA
|   InterésPorEstudiarTitulacion = Ato
|   |   NivelApoyoPadres = Muy-alto
|   |   |   NOTAENTRADA = MEDIA: Suspenso (6.0/2.0)
|   |   |   NOTAENTRADA = ALTA: Aprobado (3.0)
|   |   |   NOTAENTRADA = BAJA: Aprobado (0.0)
|   |   NivelApoyoPadres = Medio: Aprobado (3.0/1.0)
|   |   NivelApoyoPadres = Alto
|   |   |   GradoUtilizadoTutoríasyAsesorias = Muy-bajo: Aprobado (1.0)
|   |   |   GradoUtilizadoTutoríasyAsesorias = Medio: Aprobado (2.0)
|   |   |   GradoUtilizadoTutoríasyAsesorias = Bajo: No-presentado (3.0)
|   |   |   GradoUtilizadoTutoríasyAsesorias = Alto: Aprobado (0.0)
|   |   NivelApoyoPadres = Bajo: Aprobado (0.0)
|   |   NivelApoyoPadres = Muy-bajo: Aprobado (0.0)
|   InterésPorEstudiarTitulacion = Muy-alto
|   |   GradoAsistenciaClases = Muy-alto: Aprobado (26.0/8.0)
|   |   GradoAsistenciaClases = Alto
|   |   |   NivelCalidadRecursos = Alto
|   |   |   |   NivelApoyoPadres = Muy-alto: No-presentado (5.0/1.0)
|   |   |   |   NivelApoyoPadres = Medio: No-presentado (0.0)
```

| | | | NivelApoyoPadres = Alto: Aprobado (2.0)
| | | | NivelApoyoPadres = Bajo: No-presentado (0.0)
| | | | NivelApoyoPadres = Muy-bajo: No-presentado (0.0)
| | | NivelCalidadRecursos = Medio: Suspenso (2.0)
| | | NivelCalidadRecursos = Muy-bajo: Aprobado (0.0)
| | | NivelCalidadRecursos = Muy-alto: Aprobado (1.0)
| | | NivelCalidadRecursos = Bajo: Aprobado (0.0)
| | GradoAsistenciaClases = Medio: No-presentado (2.0)
| | GradoAsistenciaClases = Bajo: Aprobado (1.0)
| InterésPorEstudiarTitulacion = Bajo: No-presentado (1.0)
| InterésPorEstudiarTitulacion = Medio: Suspenso (4.0)
| InterésPorEstudiarTitulacion = Muy-bajo: Suspenso (1.0)
ASISTENCIAPRACTICAS = MEDIA
| VIA/OPCION/CIC.FORM. = PruebaAccesoUniversidad
| | NivelDificultadExigenciaTitulación = Muy-alta: No-presentado (8.0/1.0)
| | NivelDificultadExigenciaTitulación = Alta: Aprobado (2.0/1.0)
| | NivelDificultadExigenciaTitulación = Media: No-presentado (0.0)
| VIA/OPCION/CIC.FORM. = CicloFormativo-FormacionProfesional: Aprobado (3.0)
| VIA/OPCION/CIC.FORM. = Deportista Alto Nivel PruebaAccesoUniversidad: No-presentado (0.0)
ASISTENCIAPRACTICAS = BAJA
| NivelIntegraciónEnClase = Bajo: No-presentado (2.0)
| NivelIntegraciónEnClase = Alto: No-presentado (5.0)
| NivelIntegraciónEnClase = Medio: Suspenso (4.0/1.0)
| NivelIntegraciónEnClase = Muy-bajo: No-presentado (0.0)
| NivelIntegraciónEnClase = Muy-alto: Suspenso (1.0)

6. **Utilidad** (comentar para qué ha servido la experiencia y a quiénes o en qué contextos podría ser útil)

   Esta experiencia nos ha servido para darnos cuenta de la dificultad de la predicción tanto del abandono de los alumnos como de su aprobación/reprobación. Y aunque se dispone de una gran cantidad de información sobre el alumno, procedente de diferentes fuentes, no se han obtenido unos porcentajes de clasificación/predición aceptables, de por lo menos de más del 50%.

7. **Observaciones y comentarios** (comentar aspectos no incluidos en los demás apartados)

Es importante indicar que casi la totalidad del tiempo y esfuerzo realizado durante el proyecto se ha dedicado a la captura y preprocesado de la información. Esto es debido a la gran cantidad de información que se quería recoger de muy distintas fuentes. De este modo, este proceso ha requerido las siguientes fases:

1. Primera: obtener/capturar toda la información disponible en sus diferentes formatos.
2. Segunda, integrar toda la información en una única fuente de datos.
3. Tercera, transformar la información a un formato válido para la aplicación de las técnicas de minería de datos.

8. Autoevaluación de la experiencia (señalar la metodología utilizada y los resultados de la evaluación de la experiencia)

Creemos que el bajo valor en las prediciones (% de clasificaciones obtenidos) podría ser debido a los siguientes motivos:

1. El algoritmo de clasificación utilizado y/o sus parámetros no son los correctos. Pero se han utilizado otros algoritmos de clasificación, y tambien se han realizado pruebas utilizando otros parámetros y se han obtenidos siempre unos resultados muy similares de baja clasificación. Por tanto este parece no ser un motivo.
2. Los datos utilizados no son fiables, es decir, no reflejan la realidad o son ruidosos, etc. Se han comprobado los datos y son correctos, sólo nos queda la duda de que los alumnos hayan sido sinceros al rellenar las encuestas. Por otro lado, creemos que el motivo de estos resultados haya sido debido a que las notas de Junio no son totalmente fiables y definitivas. Un valor más real y definitivo podría ser utilizar las notas de Septiembre, dado que esperamos que algunos de los alumnos mal clasificados (suspensos en junio) superen la asignatura en esta convocatoria.

9. Bibliografía

Areque, F., Roldan, C., Salguero, A. Factors influencing university drop out rates. Computers & Education 53, 563-574, 2009.

Kotsiantis, S. Educational Data Mining: a case study for predicting drop-out prone students. Knowledge Engineering and Soft Data Paradigms. 1 (2), 101-111, 2009.

Romero C., Ventura. S. Educational Data Mining: A Survey from 1995 to 2005. Expert Systems with Applications, 33(1), 135-146, 2007.

Witten, I.H., Frank, E. Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann. 2005.

Lugar y fecha de la redacción de esta memoria

Córdoba, 6 de Septiembre de 2011.