



Trabajo de Fin de Grado

Predicción de abandonos en cursos online masivos y abiertos

Predicting dropouts in MOOC(Massive Open Online Courses) .

Manuel Alejandro Bacallado López

La Laguna, 5 de septiembre de 2016

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 <i>La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion</i>	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA <i>En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ</i>	Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA <i>En nombre de JESUS MANUEL JORGE SANTISO</i>	2016/09/05 14:39:27

D. **Jesús Manuel Jorge Santiso**, con N.I.F. 42.097.398-S profesor Titular de Universidad adscrito al Departamento de Ingeniería Informática y de Sistemas de la Universidad de La Laguna, como tutor

C E R T I F I C A

Que la presente memoria titulada:

“Predicción de abandonos en cursos online masivos y abiertos.”

ha sido realizada bajo su dirección por D. **Manuel Alejandro Bacallado López**, con N.I.F. 42.221.012-G.

Y para que así conste, en cumplimiento de la legislación vigente y a los efectos oportunos firman la presente en La Laguna a 5 de septiembre de 2016

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 <i>La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion</i>	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: <i>UNIVERSIDAD DE LA LAGUNA</i> <i>En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ</i>	Fecha 2016/09/05 14:36:16
<i>UNIVERSIDAD DE LA LAGUNA</i> <i>En nombre de JESUS MANUEL JORGE SANTISO</i>	2016/09/05 14:39:27

Agradecimientos

A mi familia, sobre todo a mi madre Adela, mis tios Esperanza, Begoña, Aurora, Antonio y Jose por haber sido un apoyo constante en los momentos más difíciles de mi vida.

A mis primos Julio y Yesica, por ser ejemplos y fuentes de inspiración a seguir y haber confiado en que finalmente podría lograrlo.

A mi amigo Orlandy, por el perfecto equipo que formamos desde que nos conocemos y por hacer que no me rinda cada vez que me lo planteo.

A mi amiga Luciana, porque a pesar de la distancia siempre ha estado disponible y ha seguido manteniendo su interés hacia mi persona.

A mis compañeros de facultad, en especial a Eliana, Joaquin y Erik por compartir estos tres años y todo lo que eso conlleva.

A mi tutor Jesús, por ser un ejemplo a seguir y haber explicado con tanta pasión la asignatura "Modelado de Sistemas Software", en especial lo concerniente al K- nn y que gracias a eso me haya embarcado en el fantástico mundo de la Minería de datos.

Y finalmente, a todas las personas que de una forma u otra han formado parte de mi vida estos tres últimos años.

Gracias a todos.

Sólo con constancia y paciencia se obtiene la recompensa.

Mamu

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 <i>La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion</i>	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: <i>UNIVERSIDAD DE LA LAGUNA</i> <i>En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ</i>	Fecha 2016/09/05 14:36:16
<i>UNIVERSIDAD DE LA LAGUNA</i> <i>En nombre de JESUS MANUEL JORGE SANTISO</i>	2016/09/05 14:39:27

Licencia

* Si quiere permitir que se compartan las adaptaciones de tu obra mientras se comparta de la misma manera y quieres permitir usos comerciales de tu obra (licencia de Cultura Libre) indica:



© Esta obra está bajo una licencia de Creative Commons Reconocimiento-CompartirIgual 4.0 Internacional.

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 <i>La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion</i>		
Identificador del documento: 757360		Código de verificación: gHnJNBEL
Firmado por: <i>UNIVERSIDAD DE LA LAGUNA</i> <i>En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ</i>		Fecha 2016/09/05 14:36:16
<i>UNIVERSIDAD DE LA LAGUNA</i> <i>En nombre de JESUS MANUEL JORGE SANTISO</i>		2016/09/05 14:39:27

Resumen

El objetivo de este trabajo ha sido profundizar y utilizar la Minería de datos para realizar una clasificación sobre los alumnos matriculados en los cursos online masivos y abiertos. El conjunto de datos ha sido el utilizado en la competición KDD Cup 2015.

Esta clasificación se ha obtenido mediante la herramienta RapidMiner Studio 7.0 y se ha desarrollado una aplicación escrita en el lenguaje Java, haciendo uso de los operadores internos de RadpiMiner Studio 7.0. También se han utilizado herramientas y técnicas de la ingeniería del software para un desarrollo profesional de la aplicación (gestión de proyectos, modelado de software, control de versiones, documentación, prueba unitarias y patrones de diseño).

Palabras clave: Minería de datos, Clasificación de datos, KDD Cup 2015, RapidMiner Studio 7.0, Java

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ	Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO	2016/09/05 14:39:27

Abstract

The aim of this work has been to study and apply data mining techniques to do a classification of students enrolled in the massive and open online courses. The data set used has been downloaded from KDD Cup 2015.

This classification has been achieved through RapidMiner Studio 7.0. We have developed a software application in Java, using internal operators of RadpiMiner Studio 7.0. We have applied tools and techniques of software engineering for professional software development (project management , software modeling , version control , documentation , unit testing and design patterns).

Keywords: *Data Mining, Data Classification, KDD Cup 2015, RapidMiner Studio 7.0, Java*

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 <i>La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion</i>	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: <i>UNIVERSIDAD DE LA LAGUNA</i> <i>En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ</i>	Fecha 2016/09/05 14:36:16
<i>UNIVERSIDAD DE LA LAGUNA</i> <i>En nombre de JESUS MANUEL JORGE SANTISO</i>	2016/09/05 14:39:27

Índice general

1. Introducción	1
1.1. Antecedentes y estado actual del tema	2
1.2. Estado del arte	3
1.2.1. Weka	4
1.2.2. KNIME	4
1.2.3. IBM SPSS Modeler	5
1.2.4. RapidMiner Studio 7.0	6
1.3. Objetivos	7
1.4. Planificación del proyecto	8
1.5. Estructura de la memoria	8
2. Fundamentos de la minería de datos	10
2.1. Introducción	10
2.2. Fases del KDD	11
2.3. Tareas de minería de datos	12
2.4. Técnicas de minería de datos	13
3. Tecnologías	16
3.1. SGBD Maria DB	16
3.2. Apache Maven	17
3.3. Bibliotecas de clases de RapidMiner Studio 7.0 e integración en el lenguaje Java	18
3.4. Java y Eclipse	18
3.4.1. Java	18
3.4.2. Eclipse	19
3.5. Git y Github	19
3.5.1. Git	19
3.5.2. Github	19
3.6. Javadoc y Doxygen	20
3.6.1. Javadoc	20
3.6.2. Doxygen	20
3.7. JUnit	21
4. Aplicación software y resultados	22

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ	Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO	2016/09/05 14:39:27

4.1. Introducción	22
4.2. Estructura de la interfaz gráfica de usuario	24
4.3. Configuración de un operador	25
4.4. Flujo de proceso	27
4.5. Patrones de diseño	32
4.5.1. Patrón observador	32
4.5.2. Patrón estrategia	34
4.5.3. Patrón MVC	35
5. Caso de estudio	38
5.1. Introducción	38
5.2. Generación de nuevas características y creación de la vista minable	38
5.3. Configuración del proceso utilizado y resultados usando Rapid-	
Miner Studio 7.0	48
5.4. Resultados utilizando la aplicación Java	54
6. Conclusiones y líneas futuras	58
6.1. Líneas futuras	58
7. Summary and Conclusions	59
7.1. Future lines	59
8. Presupuesto	60
A. Relación Operador RapidMiner Studio 7.0 - Clase Java	61
A.1. Data Access	61
A.2. Blending	62
A.3. Cleansing	65
A.4. Modeling	67
A.5. Scoring	68
A.6. Validation	69
Bibliografía	70

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA

Fecha 2016/09/05 14:36:16

En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

UNIVERSIDAD DE LA LAGUNA

2016/09/05 14:39:27

En nombre de JESUS MANUEL JORGE SANTISO

Índice de figuras

1.1. Diagrama de integridad referencial	3
1.2. Herramienta Weka	4
1.3. Herramienta Knime	5
1.4. Herramienta IBM SPSS Modeler	6
1.5. Herramienta RapidMiner Studio 7.0	7
1.6. Planificación del proyecto mediante la herramienta Gantt Project.	8
2.1. Proceso de KDD	11
2.2. Ejemplo de Árbol de decisión	14
2.3. Ejemplo de K-vecinos más cercanos	15
3.1. Herramienta Heidi SQL.	17
3.2. Fichero pom.xml característico de Apache Maven.	17
3.3. Añadidas dependencias para la configuración de las librerías de clase de RapidMiner en Eclipse.	18
3.4. Repositorio del proyecto almacenado en Github.	20
3.5. Documentación de la aplicación desarrollada.	21
3.6. Ejemplos de tests básicos realizados en la aplicación.	21
4.1. Paquetes de clases	22
4.2. Ejemplo de operadores dentro de un paquete	23
4.3. Ejemplo de atributos en la clase NumericToBinomial	23
4.4. Interfaz gráfica de usuario	24
4.5. Configuración Aplicación vs RapidMiner Studio 7.0	25
4.6. Ejemplo de configuración de un operador	26
4.7. Configuración Aplicación vs RapidMiner Studio 7.0	26
4.8. Configuración del operador Normalize	27
4.9. Configuración del parámetro method del operador Normalize	27
4.10. Diagrama de flujo de la aplicación	28
4.11. Contenido de la clase Apply	29
4.12. Diagrama de clases en las que interviene la variable de tipo Process	30
4.13. Añadiendo al proceso los operadores del conjunto de entrenamiento	30
4.14. Añadiendo al proceso los operadores del conjunto de test	30
4.15. Añadiendo al proceso los operadores del conjunto de apply	31
4.16. Conectando los operadores del conjunto de entrenamiento	31

III

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion		
Identificador del documento: 757360		Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ	Fecha 2016/09/05 14:36:16	
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO	2016/09/05 14:39:27	

4.17. Conectando al proceso los operadores del conjunto de test . . .	31
4.18. Conectando al proceso los operadores del conjunto de apply . .	31
4.19. Clase: ProcessKDDCup	32
4.20. Clase: Training con el método configurado	33
4.21. Diagrama de clases con las clases observadoras y observadas . .	33
4.22. Clase: TrainingTableModel	34
4.23. Cambio de comportamiento dependiendo de la elección	34
4.24. Cuadro de dialogo en donde se puede seleccionar el algortimo . .	35
4.25. Cambio de comportamiento dependiendo de la elección	35
4.26. Diagrama de clases que representan el uso del patrón MVC . . .	36
4.27. Método que dependiendo del evento invoca a una función u otra del controlador	36
4.28. Constructor de la clase ModelingController	37
4.29. Constructor de la clase ModelingDialog	37
5.1. Registros totales	39
5.2. Registros que abandonan	39
5.3. Registros que no abandonan	39
5.4. Consulta: Inscritosporcurso_train	40
5.5. Resultados: inscritosporcurso_train	40
5.6. Consulta: cursosporalumno_train	40
5.7. Resultados: cursosporalumno_train	40
5.8. Consulta: Curso con alumnos aprobados	41
5.9. Resultados: Curso con alumnos aprobados	41
5.10. Consulta: Alumnos con cursos aprobados	41
5.11. Resultados: Alumnos con cursos aprobados	41
5.12. Consulta: Días accedidos a un curso	42
5.13. Resultados: Días accedidos a un curso	42
5.14. Consulta: Accesos totales a los módulos	42
5.15. Resultados: Accesos totales a los módulos	43
5.16. Consulta: Evento más accedido	43
5.17. Resultados: Evento más accedido	43
5.18. Consulta: Número de categorías por curso	44
5.19. Resultados: Número de categorías por curso	44
5.20. Consulta: Número de módulos por curso	44
5.21. Resultados: Número de módulos por curso	44
5.22. Consulta: Número de categorías	45
5.23. Resultados: Número de categorías	45
5.24. Consulta: Número de accesos	45
5.25. Resultados: Número de accesos	45
5.26. Consulta: Número de días por curso	46
5.27. Resultados: Número de días por curso	46
5.28. Consulta: Número de módulos	46

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA

Fecha 2016/09/05 14:36:16

En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

UNIVERSIDAD DE LA LAGUNA

2016/09/05 14:39:27

En nombre de JESUS MANUEL JORGE SANTISO

5.29. Resultados: Número de módulos	46
5.30. Consulta: Cursos simultáneos	47
5.31. Resultados: Cursos simultáneos	47
5.32. Consulta: Rango	47
5.33. Resultados: Rango	47
5.34. Resultados: Vista minable	48
5.35. Proceso resultante en RapidMiner Studio 7.0	49
5.36. Configuración del operador READ CSV	49
5.37. Configuración del operador SELECT ATTRIBUTES	50
5.38. Configuración del operador SELECT ATTRIBUTES	50
5.39. Configuración del operador NORMALIZE	51
5.40. Configuración del operador NORMALIZE	51
5.41. Configuración del operador NUMERICAL TO BINOMIAL	52
5.42. Configuración del operador SET ROLE	52
5.43. Configuración del operador COMPARE ROCS	53
5.44. Algoritmos dentro del proceso COMPARE ROCS	53
5.45. Curva Roc resultante	54
5.46. Configuración de la técnica: Árbol de decisión	55
5.47. Resultado del Árbol de decisión	55
5.48. Configuración de la técnica: KNN	56
5.49. Resultado de KNN	56
5.50. Configuración de la técnica: Naive Bayes	57
5.51. Resultado de Naive Bayes	57

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA

Fecha 2016/09/05 14:36:16

En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

UNIVERSIDAD DE LA LAGUNA

2016/09/05 14:39:27

En nombre de JESUS MANUEL JORGE SANTISO

Índice de tablas

4.1. Operadores de RapidMiner y su paquete de ubicación	24
5.1. Atributos seleccionados para la vista minable	48

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 <i>La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion</i>	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA <i>En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ</i>	Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA <i>En nombre de JESUS MANUEL JORGE SANTISO</i>	2016/09/05 14:39:27

Capítulo 1

Introducción

El aumento del volumen y variedad de información que se encuentra informatizada en bases de datos digitales y otras fuentes ha crecido espectacularmente en las últimas décadas. Gran parte de esa información es histórica, representando situaciones que se han producido. Esta información es útil para explicar situaciones pasadas, comprender el presente y poder hacer predicciones sobre el futuro.

La mayoría de las decisiones de empresas, organizaciones e instituciones se basan también en información sobre experiencias pasadas extraídas de diversas fuentes. Además, ya que los datos pueden proceder de fuentes muy diversas y pertenecer a diferentes dominios, parece inminente la necesidad de analizar los mismos para obtener información útil para la organización.

En muchas situaciones, el método tradicional de convertir los datos en conocimiento consiste en un análisis e interpretación realizada de forma manual. El especialista en la materia, analiza los datos y elabora un informe o hipótesis que refleja las tendencias o pautas de los mismos.

Esta forma de actuar es lenta, cara y altamente subjetiva. El análisis de datos de forma manual viene a ser impracticable en dominios donde el volumen de los datos crece exponencialmente. El increíble conjunto de datos sobrepasa la capacidad humana de comprenderlos manualmente sin el apoyo de herramientas adecuadas. Se añade a esto, que parte de las decisiones que se realizan se hacen en base a la intuición del usuario que realiza el proceso.

Existen muchos contextos, como en la medicina, ciencia, educación, en los que los datos por si solos tienen un valor relativo. Lo realmente importante es el conocimiento que se puede extraer a partir de esos datos y más importante aún, la capacidad de poder usar ese conocimiento.

Para solucionar este problema, se integran numerosas técnicas de bases de datos, análisis de datos y extracción de modelos, dando lugar a un campo de conocimiento denominado Minería de Datos.

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ	Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO	2016/09/05 14:39:27

Se trata de extraer patrones, describir tendencias y regularidades, predecir comportamientos y sacar el máximo potencial a la información almacenada ya sea en ficheros, sistemas gestores de bases de datos o almacenes de datos mediante técnicas que resuelvan la problemática. De esta forma, las personas y empresas tienen a su disposición una forma eficiente de actuar y de tomar decisiones.

En este proyecto, vamos a utilizar la Minería de Datos para realizar una clasificación de los alumnos matriculados en los cursos online masivos y abiertos, prestando atención a su posible abandono o finalización con éxito. Para ello se ha implementado una aplicación en Java que utilice algunos algoritmos incluidos en RapidMiner Studio 7.0

1.1. Antecedentes y estado actual del tema

El proyecto que se va a describir en este documento trata de analizar los datos del desafío planteado en el **KDD Cup 2015 (Predicting dropouts in MOOC)**.

KDD Cup es una competición anual de minería de datos y descubrimiento de conocimiento organizado por la entidad **ACM (Association of Computing Machinery)**. En esta competición se plantea un problema, se proporcionan los datos, los objetivos para superar la prueba, las restricciones de la misma y la recompensa para los primeros ganadores.

En la edición del año 2015 trataban de realizar una predicción sobre el abandono en los cursos online y masivos abiertos. Los conjuntos de datos iniciales están almacenados en la página web de la competición, los cuales se detallarán brevemente a continuación:

- **date.csv**: Cada línea de este conjunto de datos representa el identificador del curso, su fecha de inicio y su fecha de fin, siendo la clave principal `course_id`. La representación de los atributos es la siguiente:
 - `<course id, from, to >`
- **object.csv**: Este conjunto de datos describe un módulo perteneciente a un curso con su correspondiente categoría, objetos hijo y fecha de comienzo. Los módulos representan diferentes materiales online utilizados en los cursos, como por ejemplo, videos, hojas de problemas, etc... Los módulos se estructuran como árboles, esto quiere decir, que un curso contiene varios capítulos, cada capítulo varias secciones y cada sección varios objetos, siendo la clave principal `course_id`. La representación de los atributos es la siguiente:

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: **UNIVERSIDAD DE LA LAGUNA**

Fecha 2016/09/05 14:36:16

En nombre de **MANUEL ALEJANDRO BACALLADO LOPEZ**

UNIVERSIDAD DE LA LAGUNA

2016/09/05 14:39:27

En nombre de **JESUS MANUEL JORGE SANTISO**

- <**course_id**, module_id, category, children, start >
- enrollment_train.csv: Cada línea perteneciente a este conjunto de datos es un registro de inscripción(matrícula) de un curso con su identificador, el identificador de inscripción y nombre de usuario. Este archivo pertenece al conjunto de entrenamiento, siendo la clave principal enrollment_id. La representación de los atributos es la siguiente:
 - <**enrollment_id**, username, course_id >
- log_train.csv: Cada registro de este conjunto de datos es un evento. La información almacenada es el id de inscripción de un alumno, la hora en que se produce ese evento, la fuente donde se origina, el tipo de evento que se produce y su objeto de acceso de los estudiantes. Este archivo pertenece al conjunto de entrenamiento, siendo la clave principal enrollment_id. La representación de los atributos es la siguiente:
 - <**enrollment_id**, time, source, event, object >
- Truth_train.csv: Cada línea contiene información sobre el grado de veracidad de abandono de las inscripciones en el conjunto de entrenamiento. Si se trata de abandono le corresponde el valor 1 y 0 si terminó el curso con éxito. La representación de los datos es la siguiente:
 - <enrollment_id, success >

Los archivos del conjunto de prueba tienen exactamente la misma estructura que los del conjunto de entrenamiento. Existirá, por lo tanto, dos archivos con extensión csv cuyo nombre corresponde a enrollment_test.csv y log_test.csv

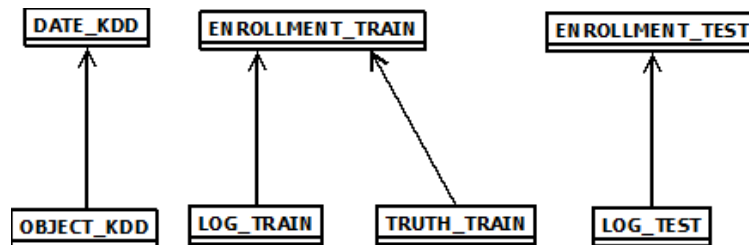


Figura 1.1: Diagrama de integridad referencial

1.2. Estado del arte

Para el análisis de la información, existen multitud de herramientas que permiten automatizar este proceso sin necesidad de ser experto en el tema ni poseer

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ	Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO	2016/09/05 14:39:27

grandes conocimientos de las técnicas de Minería de datos que pueden ser utilizadas. Algunas de estas herramientas son:

1.2.1. Weka

Weka[1] es un software de código abierto bajo la licencia GPL. Contienen un conjunto de algoritmos de aprendizaje automático para tareas de minerías de datos. Los algoritmos o bien se pueden aplicar directamente a un conjunto de datos a través de su interfaz gráfica o bien utilizando la librería de clases Java. Weka contiene herramientas para el procesamiento previo de datos, técnicas de clasificación, regresión, reglas de asociación y visualización.

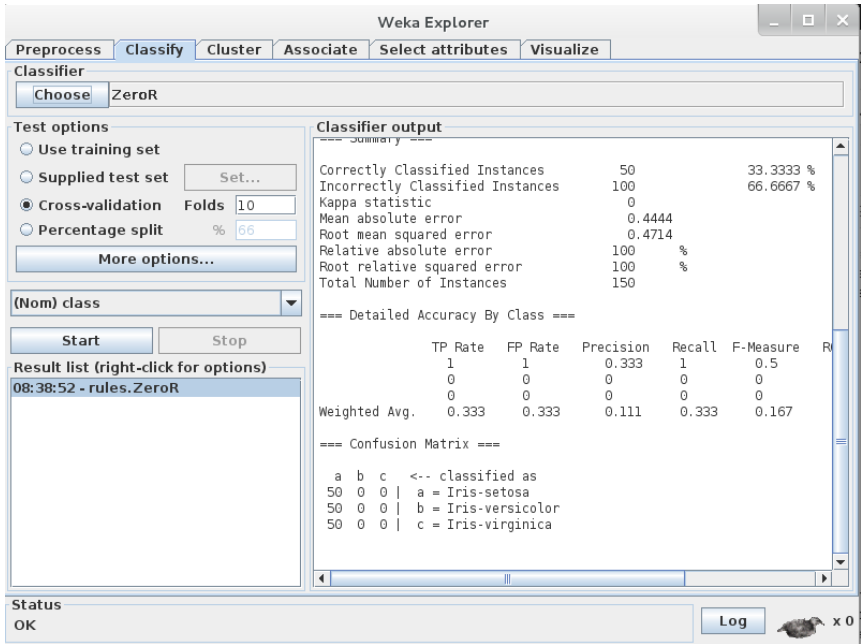


Figura 1.2: Herramienta Weka

1.2.2. KNIME

KNIME Analytics Platform[2] es una solución abierta que permite la innovación mediante el análisis de datos. Contiene 1000 módulos, herramientas integradas y una multitud de algoritmos avanzados.

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion		
Identificador del documento: 757360		Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ		Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO		2016/09/05 14:39:27

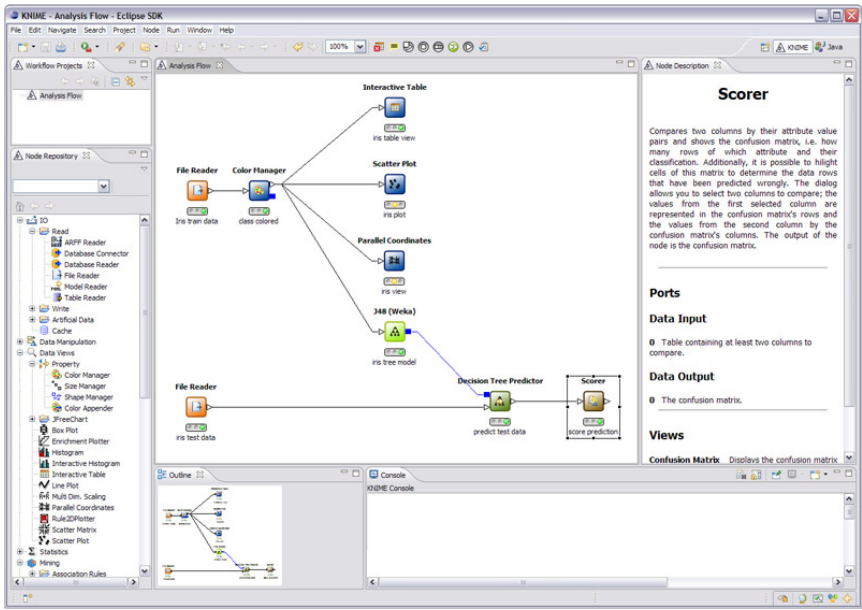


Figura 1.3: Herramienta Knime

1.2.3. IBM SPSS Modeler

IBM SPSS Modeler[3], antigua herramienta clementine, es una plataforma de pago de análisis predictiva diseñada para aportar inteligencia predictiva a decisiones llevadas a cabo por personas, grupos, sistemas y la empresa. Proporciona un rango de algoritmos y técnicas avanzadas, análisis de texto, la gestión y optimización de decisiones, para ayudar a seleccionar las acciones que proporcionan un mejor resultado.

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion		
Identificador del documento: 757360		Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ	Fecha 2016/09/05 14:36:16	
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO	2016/09/05 14:39:27	

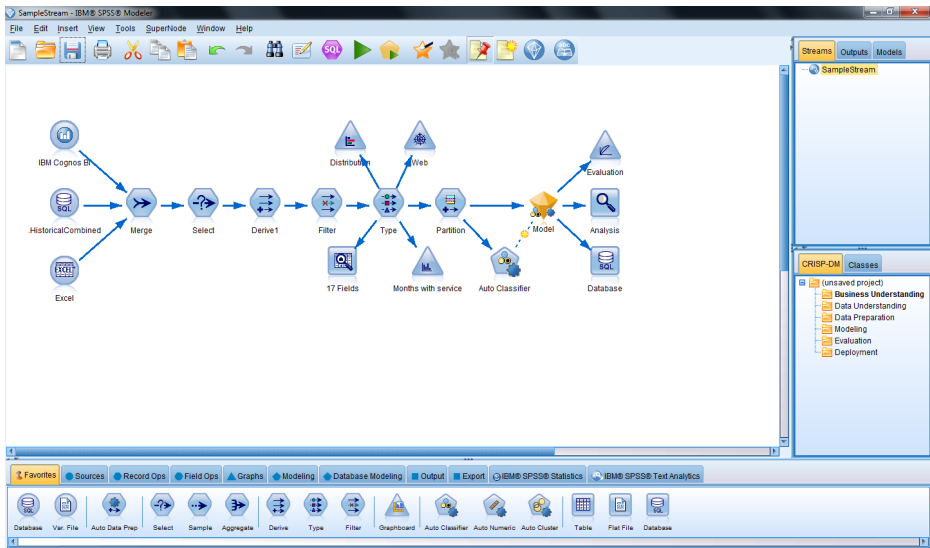


Figura 1.4: Herramienta IBM SPSS Modeler

1.2.4. RapidMiner Studio 7.0

RapidMiner Studio 7.0[4] es una herramienta software para el análisis de datos, empleando técnicas de minería de datos.

La forma de trabajar de RapidMiner es mediante el uso de operadores que representan tanto a los pasos previos de un proceso de KDD, como la preparación de los datos, como a los algoritmos de minería de datos a utilizar.

Una de las características más importantes de RapidMiner es su sistema “Drag and drop”, el cual permite arrastrar los operadores a su panel central, conectarlos y configurarlos sin problema alguno.

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion		
Identificador del documento: 757360		Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ		Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO		2016/09/05 14:39:27

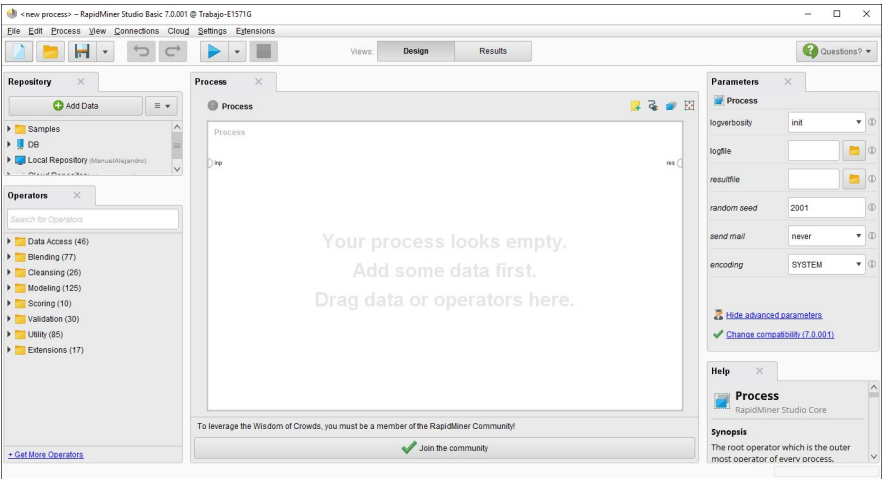


Figura 1.5: Herramienta RapidMiner Studio 7.0

1.3. Objetivos

Acorde con los antecedentes de este proyecto, los objetivos de éste constituyen los siguientes puntos:

- 1. Clasificar, mediante el uso de las técnicas de minería de datos, los alumnos matriculados en los cursos online masivos y abiertos, prestando atención a su posible abandono o finalización con éxito.
- 2. Utilizar herramientas de software libre para:
 - o La creación de una aplicación de escritorio usando el lenguaje de programación Java que utilice los operadores de la aplicación de minería de datos RapidMiner Studio 7.0.
 - o El almacenamiento de los conjuntos de datos en tablas de base de datos y creación de nuevas tablas con conocimiento nuevo generado.
 - o El desarrollo de la aplicación, desde el inicio hasta su finalización, mediante un sistema de control de versiones, pruebas unitarias y documentación de la misma.

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion		
Identificador del documento: 757360		Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ		Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO		2016/09/05 14:39:27

1.4. Planificación del proyecto

La duración del proyecto se ha establecido con una lectura sobre minería de datos utilizando el libro “Introducción a la minería de datos”, el cual ha llevado unas tres semanas. La comprensión del caso de estudio y la generación de conocimiento para poder construir la vista minable han costado unos tres meses. El aprendizaje de la herramienta RapidMiner Studio, tanto de su biblioteca de clases, para conocer el funcionamiento de los operadores y su configuración, como de su interfaz gráfica y el uso de todas las herramientas anteriormente explicadas en el capítulo 3 han costado alrededor de cuatro meses. Las tareas concernientes al caso de estudio y la biblioteca de clases se han desarrollado en paralelo.

El tiempo inicial establecido de entrega era en Junio del 2016, pero los problemas que se han generado durante el desarrollo del proyecto han llevado a su retraso hasta septiembre de ese mismo año.

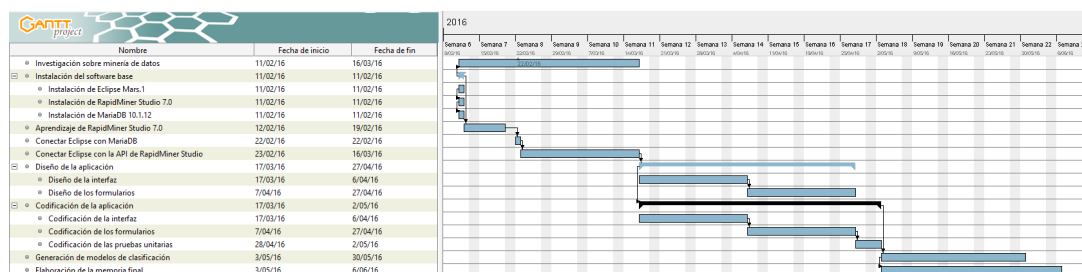


Figura 1.6: Planificación del proyecto mediante la herramienta Gantt Project.

1.5. Estructura de la memoria

Para describir las tareas que conllevan al alcance de los objetivos del proyecto, esta memoria se estructura en 5 capítulos, además de esta breve introducción:

- En el capítulo 2 se hace una revisión de los fundamentos de la minería de datos, las fases del descubrimiento de conocimiento en bases de datos, las tareas de minería de datos y las técnicas para resolver estas tareas, los métodos de evaluación y las técnicas específicas que se van a utilizar para poder realizar la clasificación.
- En el capítulo 3, se hace una revisión de las tecnologías utilizadas para llevar a cabo el proyecto.

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA
En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

Fecha 2016/09/05 14:36:16

UNIVERSIDAD DE LA LAGUNA
En nombre de JESUS MANUEL JORGE SANTISO

2016/09/05 14:39:27

- En el capítulo 4 se explica detalladamente la aplicación software basada en la configuración de los operadores y la creación de la interfaz gráfica.
- En el capítulo 5 se explica de manera concisa el nuevo conocimiento generado en tablas de bases de datos, la configuración de un proceso de minería de datos usando RapidMiner Studio 7.0 y los resultados obtenidos, tanto en RapidMiner Studio como en la interfaz gráfica desarrollada.
- En el capítulo 6 se establecen las conclusiones y el planteamiento de las líneas futuras para mejorar la aplicación.

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ	Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO	2016/09/05 14:39:27

Capítulo 2

Fundamentos de la minería de datos

2.1. Introducción

La minería de datos se define como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos. Fundamentalmente, la labor de la minería de datos es encontrar modelos a partir de los datos disponibles, por lo que este proceso debería ser automático o asistido para que fuera realmente efectivo y aporte beneficios a la organización.

Este conocimiento puede ser en forma de relaciones, patrones, reglas inferidas de los datos o simplemente, en forma de resumen de los datos. Existen muchas maneras diferentes de representar los modelos, también denominada tarea de minería de datos y cada una determina el tipo de técnica a utilizar.

Comentando esto, es razonable decir que la minería de datos se enfrenta a dos problemas principales: Por un lado, trabajar con grandes volúmenes de datos procedentes de diversos sistemas de información. Esto conlleva diversos problemas: ruido, datos ausentes, información no tratada. Por otro lado, la utilización de algoritmos ó técnicas adecuadas para analizar los mismos y extraer ese conocimiento novedoso y útil.

Otra definición que recibe la minería de datos es la extracción o descubrimiento de conocimiento en bases de datos (Knowledge Discovery in Databases, KDD). Estos términos, utilizados como sinónimos, presentan diferencias entre ambos. El KDD es un proceso complejo que incluye no solo la obtención de los modelos o patrones, sino la evaluación, interpretación y visualización del mismo, además de la generación de nuevas características, por lo que finalmente la minería de datos es una fase dentro de este proceso.

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ	Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO	2016/09/05 14:39:27

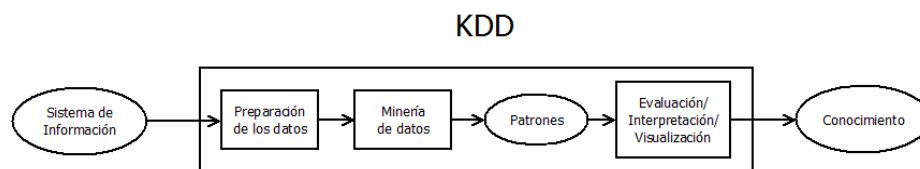


Figura 2.1: Proceso de KDD

La minería de datos guarda relación con otras tecnologías, desarrollándose en paralelo o como una prolongación de las mismas. A continuación, se destacan las más influyentes:

- Bases de datos
- Recuperación de la información
- Estadística
- Aprendizaje automático (Machine Learning)
- Sistemas de toma de decisiones
- Visualización de datos
- Computación paralela y distribuida

2.2. Fases del KDD

Como se explicó en el apartado anterior, la fase de la minería de datos es la más característica del KDD. Anterior y posterior a esta, existen otras fases de igual ó mayor importancia, que sin su correcta ejecución, impedirían obtener finalmente el modelo deseado o representativo de esos datos. Estas fases son:

- Integración y recopilación de datos: Se determinan las fuentes de información que pueden ser útiles y como obtenerlas. Se transforman los datos mediante el almacenamiento en bases de datos.
- Selección, limpieza y transformación: En esta fase se tratan los valores anómalos, errores y/o faltantes en los datos, eliminándolos o corrigiéndolos. Posteriormente a eso, se realiza una proyección, también denominada vista minable, para considerar únicamente aquellos atributos que realmente van a ser relevantes, con el fin de hacer más fácil la tarea de minería de datos (disminuyendo la dimensionalidad de los datos).

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA

Fecha 2016/09/05 14:36:16

En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

UNIVERSIDAD DE LA LAGUNA

2016/09/05 14:39:27

En nombre de JESUS MANUEL JORGE SANTISO

- Evaluación e interpretación: Fase posterior a la de minería de datos, se evalúa el modelo resultante y se analiza. Si es necesario, se vuelve a una fase anterior para una nueva iteración y comprobar si existen conflictos con el conocimiento previamente generado.
- Fase de difusión: Se utiliza el nuevo conocimiento y se hace disponible para los usuarios.

El objetivo de la fase de minería de datos es producir conocimiento nuevo. Esto se realiza construyendo un modelo a partir de los datos recopilados en fases anteriores. El modelo es una descripción de los patrones y relaciones entre los datos que se pueden usar para realizar explicaciones sobre situaciones pasadas, entender mejor el presente y predicciones a un futuro cercano. Antes de empezar este proceso, es necesario tomar una serie de decisiones:

- Establecer el tipo de tarea que es necesario resolver.
- Seleccionar el tipo de modelo (técnica de minería de datos).
- Elegir el algoritmo de minería que resuelva la tarea dentro de la técnica seleccionada y proporcione el tipo de modelo deseado. Esta elección es relativa, ya que para resolver una tarea existen muchos tipos de técnicas disponibles a utilizar y dentro de cada una muchos algoritmos.

En este proyecto solo se van a manejar algoritmos pertenecientes a una tarea de clasificación, por lo que se comentará en detalle dicha tarea y únicamente se nombrarán las restantes.

2.3. Tareas de minería de datos

Una tarea de minería de datos es un tipo de problema de minería de datos.

Las tareas en minería de datos se pueden clasificar en predictivas y descriptivas:

- Predictivas: Son tareas en las que hay que predecir uno o más valores para uno o más casos. Los casos van acompañados de una etiqueta (clase, categoría o valor numérico) seguidos de un orden entre ellos. Entre las tareas predictivas más importantes tenemos:
 - Clasificación: Cada registro o instancia en la base de datos pertenece a una clase, la cual se establece mediante el valor de un atributo

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA

Fecha 2016/09/05 14:36:16

En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

UNIVERSIDAD DE LA LAGUNA

2016/09/05 14:39:27

En nombre de JESUS MANUEL JORGE SANTISO

que se llama clase de la instancia. Este atributo en particular, puede tomar diferentes valores, cada uno de ellos pertenece a una clase distinta. Los demás atributos pertenecientes a dicha instancia se utilizan para predecir la clase. El objetivo de esta tarea es maximizar la precisión predictiva de la clasificación de las nuevas instancias, la cual se calcula como en cociente entre las predicciones correctas y el número total de predicciones (correctas e incorrectas).

- Regresión
- Descriptivas: Los casos constituyen un conjunto sin etiquetas ni orden. El objetivo de este tipo de tareas es describir los datos existentes. Existen varias tareas descriptivas:
 - Agrupamiento (Clustering)
 - Correlaciones
 - Reglas de asociación
 - Reglas de asociación secuenciales

Comúnmente, las tareas predictivas reciben el nombre de “aprendizaje supervisado” y las tareas descriptivas “aprendizaje no supervisado”.

2.4. Técnicas de minería de datos

Las tareas nombradas en la sección anterior, se agrupan en diversas categorías o técnicas para llegar a su resolución. Estos algoritmos pueden recibir también el nombre de métodos. Para una tarea, existe una gran variedad de algoritmos para su resolución y el mismo método puede utilizarse para resolver multitud de tareas. A continuación, se mostrarán los tipos de técnicas existentes y se comentarán las utilizadas en este proyecto:

- Técnicas de modelización estadística
- Técnicas bayesianas: Se basan en estimar la probabilidad de pertenencia, mediante la estimación de probabilidades condicionadas o usando el teorema de Bayes. La dificultad de estas técnicas es baja, pueden tratar multitud de atributos y son muy robustos frente a los problemas de ruido en los datos.

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA

Fecha 2016/09/05 14:36:16

En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

UNIVERSIDAD DE LA LAGUNA

2016/09/05 14:39:27

En nombre de JESUS MANUEL JORGE SANTISO

- Técnicas basadas en arboles de decisión y sistemas de aprendizaje de reglas: Son técnicas que su forma de representación es en forma de reglas. Principalmente se utiliza la estrategia de divide y vencerás:
 - “Divide y vencerás”: ID3, C4.5, Cart...

Es una de las técnicas fundamentales en minería de datos. Son fáciles de utilizar, admiten atributos continuos y discretos y mediante la poda resuelve bien los problemas de valores faltantes o ruido en los atributos. Son muy eficientes y obtienen excelentes resultados para la clasificación. La mayor ventaja de este tipo de técnicas es que los modelos resultantes pueden ser expresados como un conjunto de reglas, resultando comprensibles.

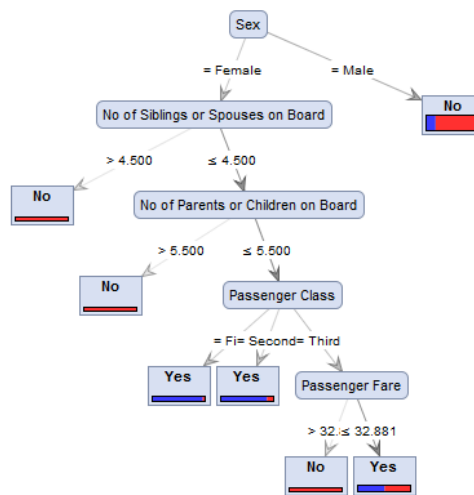


Figura 2.2: Ejemplo de Árbol de decisión

- Técnicas relacionales y declarativas
- Técnicas basadas en redes neuronales artificiales
- Técnicas basadas en núcleo y máquinas de soporte vectorial
- Técnicas estocásticas y difusas
- Técnicas basadas en casos, en densidad o distancia: Técnicas que se basan en distancias respecto a los otros elementos, de forma directa, como el k-vecinos más cercano o mediante la estimación de funciones de densidad. Estas técnicas son fáciles de utilizar y tienen buenos resultados si el

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA

Fecha 2016/09/05 14:36:16

En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

UNIVERSIDAD DE LA LAGUNA

2016/09/05 14:39:27

En nombre de JESUS MANUEL JORGE SANTISO

número de ejemplos no es excesivamente grande. Al basarse en distancias, se aumenta la dimensionalidad al no saber expresar adecuadamente los atributos no significativos y se construyen distancias ficticias. No construyen un modelo, por lo que no producen un conocimiento comprensible.

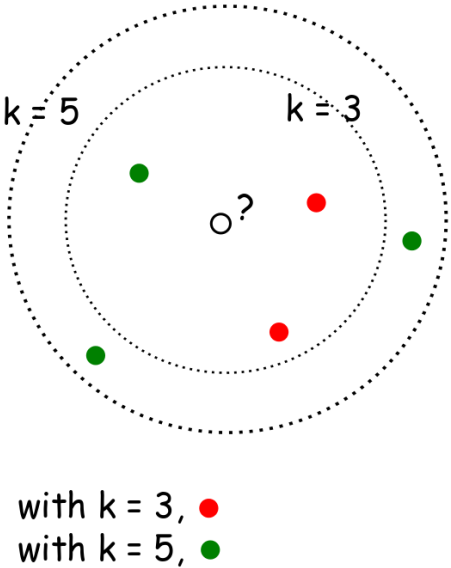


Figura 2.3: Ejemplo de K-vecinos más cercanos

En esta sección se han expuesto técnicas de minería de datos en las que se basan construyendo un modelo y otras en las que no. Las técnicas sin modelo reciben el nombre de “retardadas o perezosas” (lazy) y las que sí tienen un modelo son “anticipadas o impacientes” (eager):

- Técnicas retardadas o perezosas: No se construye modelo. Utilizan optimización local. Los ejemplos deben conservarse porque son necesarios para realizar la siguiente predicción. El tiempo de respuesta decae en medida que se aumenta el número de instancias. Por el contrario, la ventaja que tienen este tipo de técnicas es que no es necesario entrenar el modelo.
- Técnicas anticipativas o impacientes: Las instancias de entrenamiento pueden desecharse, ya que se obtiene un modelo a partir de ellas. Utilizan optimización global. La desventaja es que se requiere mucho tiempo de entrenamiento, pero una vez realizado este proceso, su aplicación es instantánea.

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003	
La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ	Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO	2016/09/05 14:39:27

Capítulo 3

Tecnologías

Las tecnologías a utilizar durante el desarrollo del proyecto son las siguientes:

3.1. SGBD Maria DB

Resulta muy útil para el almacenamiento y manipulación de los datos contar con un sistema gestor de base de datos.

MariaDB[5] está desarrollado por los creadores de MySQL. En su sitio web se indica que todos los comandos, interfaces, librerías y Apis que están disponibles en MySQL, también están disponibles en MariaDB. El software que se instala tiene por nombre “heidiSQL”, el cual nos permitirá las operaciones básicas de cualquier sistema gestor de bases de datos: crear una base de datos, consultas, etc.

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 <i>La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion</i>	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA <i>En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ</i>	Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA <i>En nombre de JESUS MANUEL JORGE SANTISO</i>	2016/09/05 14:39:27

The screenshot shows the HeidiSQL interface with a table named 'kbbcup2015' selected. The table has columns: Nombre, Fecha, Turno, Creado, Actualizado, Motor, Comentario, and Tipo. The data is organized into a grid with rows for various course-related entries.

Figura 3.1: Herramienta Heidi SQL.

3.2. Apache Maven

Apache Maven[6] es una herramienta de gestión de proyectos software basados en el lenguaje de programación Java. Permite la creación de los proyectos a través de su modelo de objeto de proyecto (Project Object Model en inglés) y un conjunto de módulos (plugins) que son compartidos por todos los proyectos, proporcionando un sistema de construcción firme. También permite gestionar las dependencias de los módulos y los componentes, el orden de construcción de los mismos y tareas como la compilación y empaquetado del código fuente.

The screenshot shows a snippet of a pom.xml file. It defines a project with the following structure:

```

1 <?xml encoding="UTF-8" ?>
2 <project xmlns="http://maven.apache.org/POM/4.0.0" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://maven.apache.org/POM/4.0.0 http://maven.apache.org/xsd/maven-4.0.0.xsd">
3   <groupId>pruebaMaven</groupId>
4   <artifactId>pruebaMaven</artifactId>
5   <version>0.0.1-SNAPSHOT</version>
6   <build>
7     <sourceDirectory>src</sourceDirectory>
8     <plugins>
9       <plugin>
10        <groupId>maven-compiler-plugin</groupId>
11        <version>3.3</version>
12        <configuration>
13          <source>1.8</source>
14          <target>1.8</target>
15        </configuration>
16      </plugin>
17    </plugins>
18  </build>
19 </project>

```

Figura 3.2: Fichero pom.xml característico de Apache Maven.

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA

Fecha 2016/09/05 14:36:16

En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

UNIVERSIDAD DE LA LAGUNA

2016/09/05 14:39:27

En nombre de JESUS MANUEL JORGE SANTISO

3.3. Bibliotecas de clases de RapidMiner Studio 7.0 e integración en el lenguaje Java

Los operadores disponibles en RapidMiner Studio 7.0 para realizar las operaciones, internamente son clases en Java. Para su uso, debido a su licencia GPL, este conjunto de clases está disponible para que los desarrolladores ajenos a la organización puedan crear sus propios operadores y, si lo desean, publicarlos en “Rapidminer Marketplace”. Para su publicación, es necesario registrarse en la comunidad, llamada “RapidMiner Community”.

El acceso a los operadores puede ser a través del github de la organización [7], o como se ha realizado en este proyecto, a través de una dependencia alojada en el archivo de configuración “pom.xml”. A continuación, se muestra una imagen de la dependencia:

Finalmente, una vez incluida la dependencia, la utilización de las clases es idéntica al de otro paquete de bibliotecas alojado en java.

```
<dependency>
  <groupId>com.rapidminer.studio</groupId>
  <artifactId>rapidminer-studio-core</artifactId>
  <version>7.0.1</version>
</dependency>
</dependencies>

<repositories>
  <repository>
    <id>rapidminer-repository</id>
    <url>https://maven.rapidminer.com/content/groups/public</url>
  </repository>
</repositories>
```

Figura 3.3: Añadidas dependencias para la configuración de las librerías de clase de RapidMiner en Eclipse.

3.4. Java y Eclipse

3.4.1. Java

Java[8] es un lenguaje de programación y una plataforma informática comercializada por primera vez en 1995 por Sun Microsystems. La sintaxis propia del lenguaje deriva de otros lenguajes de programación tales como C/C++, aunque tiene menos utilidades a bajo nivel que los mencionados.

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ	Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO	2016/09/05 14:39:27

Uno de sus principales objetivos está en permitir que los desarrolladores codifiquen el programa una vez y no necesite ninguna otra compilación para ejecutarlo en otro dispositivo. Está basado en la licencia GNU/ GPL.

Java está presente multitud de aparatos electrónicos: portátiles, centros de datos, consolas, súper computadoras, teléfonos móviles e Internet, por esta razón y la facilidad de su aprendizaje, es uno de los lenguajes de programación más usados en el mundo.

Soporta el paradigma de programación imperativa, orientada a objetos, programación concurrente y en la última versión (Java SE 8) añade funcionalidad para la programación funcional mediante expresiones lambda.

3.4.2. Eclipse

Eclipse[9] es un entorno de desarrollo integrado (IDE) para codificar aplicaciones. Soporta varios lenguajes de programación como Java, C++, Php, etc. Se puede instalar el específico de cada lenguaje, a través de su web, o a través de módulos (plugins) disponibles en el “marketplace” de eclipse. Para hacerlo de esta última forma, se accede a través de la opción help -¿Eclipse Marketplace. Aparecerá una ventana en donde se podrá buscar el modulo deseado y posteriormente instalarlo.

3.5. Git y Github

3.5.1. Git

Git[10] es un sistema libre y de código abierto de control de versiones para la eficiencia y seguimiento de las versiones de aplicaciones que alberguen un número considerable de archivos fuente. Las operaciones pueden realizarse mediante líneas de comandos, ya que tiene incorporado una consola que emula a un sistema Linux y otra como la de un sistema Windows, o mediante una interfaz gráfica también disponible.

3.5.2. Github

Github[11] es la plataforma de desarrollo colaborativo de software para alojar proyectos utilizando el sistema de control de versiones Git. Por defecto, la cuenta de usuario es pública, esto quiere decir que el código fuente se almacenará de forma pública y el resto de usuarios de la plataforma o ajenos a ella, podrán visualizar y descargar el proyecto en cuestión. Para contrarrestar esa problemática, existen cuentas privadas en donde el código fuente se aloja de

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ	Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO	2016/09/05 14:39:27

manera privada. Con estas herramientas se fomenta el trabajo en equipo dentro de los proyectos.

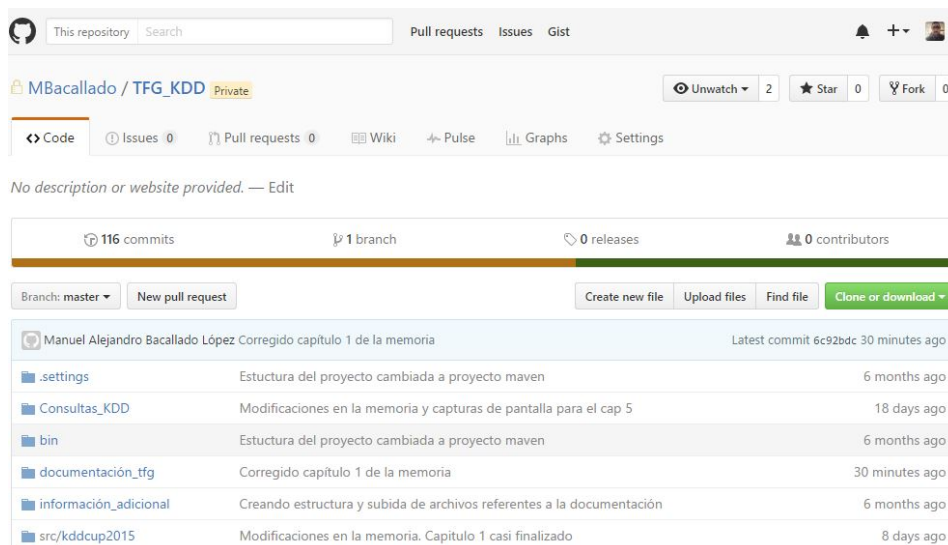


Figura 3.4: Repositorio del proyecto almacenado en Github.

3.6. Javadoc y Doxygen

3.6.1. Javadoc

Javadoc[12] es la utilidad de Oracle para la generación de documentación de las clases de Java en formato HTML. La mayoría de los IDEs utilizan javadoc para generar la documentación, como, por ejemplo, Eclipse, Blue J, etc.

3.6.2. Doxygen

Doxygen[13] es la herramienta para la generación de documentación que soporta los siguientes lenguajes de programación: C++, C, Java, Objective-C, C#, etc. Permite generar la documentación a través de un navegador en formato HTML a partir de un conjunto de ficheros fuentes previamente documentados, soporte para salida en diversos formatos como RTF, PDF. La documentación se extrae directamente de las fuentes, lo que reduce la dificultad de mantener la documentación en relación al código fuente.

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA

Fecha 2016/09/05 14:36:16

En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

UNIVERSIDAD DE LA LAGUNA

2016/09/05 14:39:27

En nombre de JESUS MANUEL JORGE SANTISO

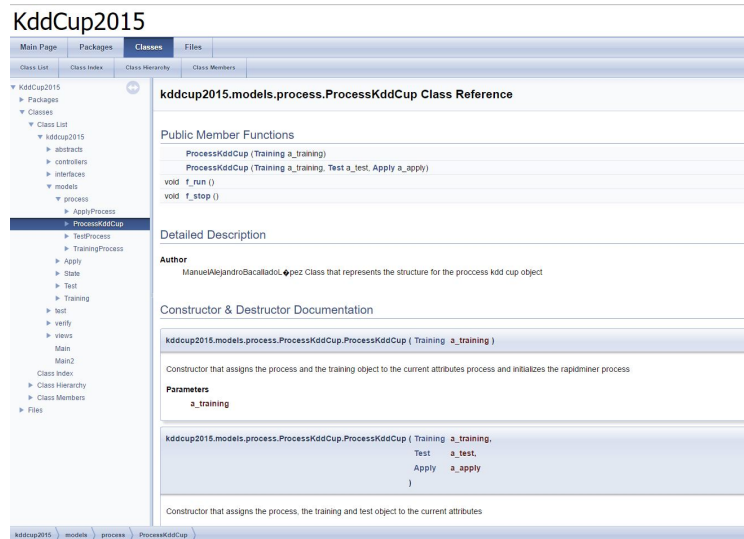


Figura 3.5: Documentación de la aplicación desarrollada.

3.7. JUnit

JUnit[14] es un framework para realizar pruebas unitarias que utiliza anotaciones para identificar los métodos en donde se incluyen dichas pruebas. Todos estos métodos se especifican en una clase que solo se utiliza para dicho cometido. Las anotaciones pueden hacerse mediante importaciones de las clases o encima del método incluir “@org.junit.X” donde X representa la clase.

```
@org.junit.Test
public void test1() {
    assertEquals("Test1: Checks that training list is not empty", 0, this.m_training.f_getTraining().size());
}

@org.junit.Test
public void test2() {
    assertEquals("Test2: Checks that test list is not empty", 0, this.m_test.f_getTest().size());
}

@org.junit.Test
public void test3() {
    assertEquals("Test3: Checks that apply list is not empty", 0, this.m_apply.f_getApply().size());
}

@org.junit.Test
public void test4() {
    assertTrue("Test4: Checks that the items in the training list are operator type", this.m_training.f_getTraining().get(0) instanceof Operator);
}

@org.junit.Test
public void test5() {
    assertTrue("Test5: Checks that the items in the test list are operator type", this.m_test.f_getTest().get(0) instanceof Operator);
}

@org.junit.Test
public void test6() {
    assertTrue("Test6: Checks that the items in the apply list are operator type", this.m_apply.f_getApply().get(0) instanceof Operator);
}
```

Figura 3.6: Ejemplos de tests básicos realizados en la aplicación.

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion		
Identificador del documento: 757360		Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ		Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO		2016/09/05 14:39:27

Capítulo 4

Aplicación software y resultados

4.1. Introducción

Utilizando los operadores de RapidMiner Studio 7.0, se ha decidido desarrollar una interfaz gráfica de usuario (GUI) para el uso de las técnicas de minería de datos, necesarias para resolver el caso de estudio, explicado en el capítulo siguiente. La aplicación está desarrollada en el lenguaje Java, utilizando la biblioteca gráfica Swing y cabe destacar que en su interior no estarán todos los operadores existentes en RapidMiner configurados, únicamente los necesarios para realizar un proceso de clasificación.

Estos operadores se encuentran en el paquete de clases “com.rapidminer.operator.*” y está distribuido en diferentes secciones, las cuales ayudan en la búsqueda de las clases necesarias.

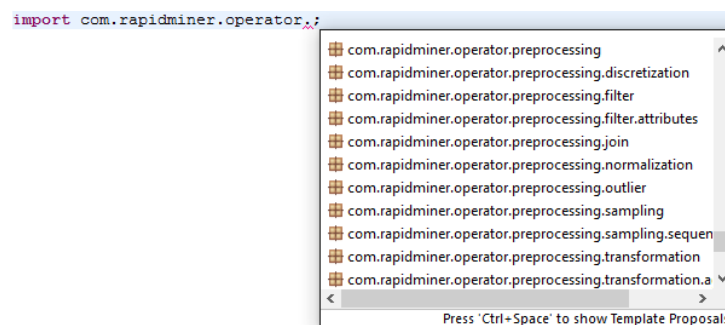


Figura 4.1: Paquetes de clases

Por ejemplo, hay clases cuyo nombre es idéntico o prácticamente igual que el operador en RapidMiner. En la figura 4.2 se pueden apreciar las clases “Nume-

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ	Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO	2016/09/05 14:39:27

ricToBinomial” y “NominalToBinomial”, las cuales posteriormente serán utilizadas para crear los operadores “NumericalToBinomial” y “NominalToBinomial”. Accediendo a “NumericToBinomial”, se puede observar que incluyen los parámetros necesarios para posteriormente ser configurados.

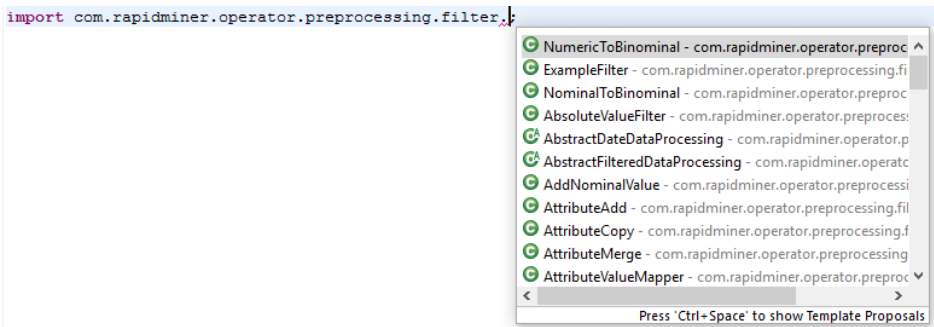


Figura 4.2: Ejemplo de operadores dentro de un paquete

```
public class NumericToBinominal extends NumericToNominal {  
  
    /** The parameter name for "The minimal value which is mapped to false (included)." */  
    public static final String PARAMETER_MIN = "min";  
  
    /** The parameter name for "The maximal value which is mapped to false (included)." */  
    public static final String PARAMETER_MAX = "max";  
  
}
```

Figura 4.3: Ejemplo de atributos en la clase NumericToBinomial

En otros casos, el nombre no tiene similitud alguna y no se ha encontrado información en donde establezca que un operador X se corresponda con una clase X[15], por lo que se ha optado por la comparación de parámetros de las clases seleccionadas con el operador en RapidMiner hasta encontrar el exacto. Esta relación se ha recogido y está presente en el anexo de esta memoria, ya que en esta sección, solo se va a mostrar una pequeña lista con los operadores que aparecen en esta aplicación y el paquete Java al que corresponde.

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion		
Identificador del documento: 757360		Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ		Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO		2016/09/05 14:39:27

Nombre	Paquete Java + Clase
Read CSV	nio.CSVExampleSource
Result Writer	io.ResultWriter
Write CSV	io.CSVExampleSetWriter
Set Role	preprocessing.filter.ChangeAttributeRole
Numerical to Binomial	preprocessing.filter.NumericToBinominal
Nominal to Binomial	preprocessing.filter.NominalToBinominal
Select Attributes	preprocessing.filter.attributes.AttributeFilter
Normalize	preprocessing.normalization.Normalization
K-nn	learner.lazy.KNNLearner
Naive Bayes	learner.bayes.NaiveBayes
Decision tree	learner.tree.DecisionTreeLearner (Depercated)
Apply Model	ModelApplier
Performance	performance.PerformanceEvaluator
Performance Binomial	performance.BinominalClassificationPerformanceEvaluator
Performance to Data	performance.PerformanceVectorToExampleSet

Tabla 4.1: Operadores de RapidMiner y su paquete de ubicación

4.2. Estructura de la interfaz gráfica de usuario

La aplicación está estructurada en diversas secciones reconocibles en la vista principal:

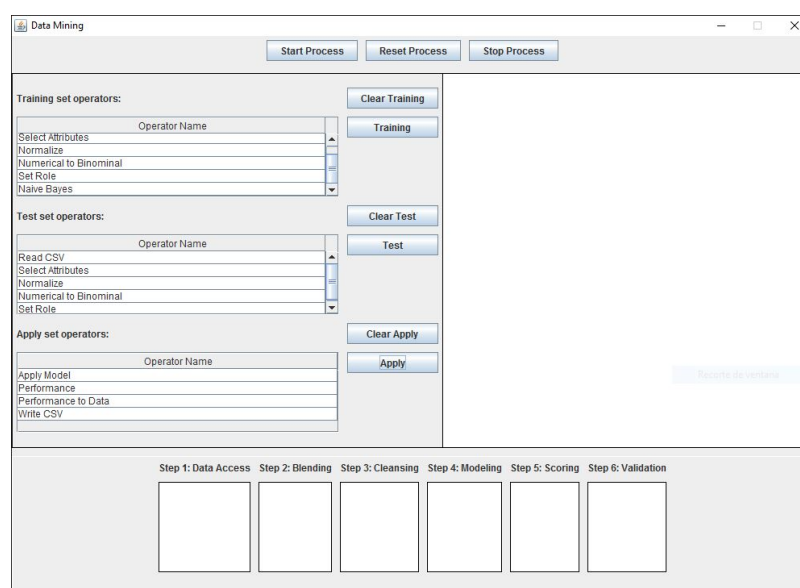


Figura 4.4: Interfaz gráfica de usuario

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA

Fecha 2016/09/05 14:36:16

En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

UNIVERSIDAD DE LA LAGUNA

2016/09/05 14:39:27

En nombre de JESUS MANUEL JORGE SANTISO

- Sección superior: En esta sección se encuentran los botones que inicia el proceso una vez configurado (Start Process), el de resetear el proceso (Reset Process) y el de parar el proceso (Stop Process).
- Sección lateral izquierda: En esta sección están los operadores del conjunto de entrenamiento, del conjunto de prueba y los operadores necesarios para aplicar el modelo. Para seleccionar el conjunto al que añadir operadores, bastará únicamente con pulsar los botones Training, Test y Apply. Los botones Clear Training y Clear Test borran los operadores almacenados en los conjuntos. Por defecto, los operadores configurados se almacenarán en Training.
- Sección lateral derecha: En esta sección se mostrará la predicción obtenida una vez ejecutado el proceso.
- Sección inferior: En esta sección se encuentra organizado por “pasos” las distintas secciones que componen un proceso de minería de datos. Su interior está compuesto por un cuadro de dialogo en donde se podrá seleccionar la categoría y configurar el operador que mejor satisfaga la necesidad del usuario.

4.3. Configuración de un operador

Una vez seleccionado el operador, se mostrará un panel con sus parámetros. La configuración del mismo es exactamente igual que en la aplicación de RapidMiner Studio.

Figura 4.5: Configuración Aplicación vs RapidMiner Studio 7.0

Internamente, cada panel pertenece a una clase específica en donde está ubicado el operador a configurar, mediante una variable de tipo “Operator”. Todas estas clases incorporan una función denominada `f.createOperator()`. Este método, asignará a la variable “Operator” el tipo deseado y también será el encargado de asignar los parámetros recogidos por la interfaz. Esto se consigue con las dos primeras líneas del método, si no se encuentran, ejecutará un mensaje de error diciendo que no se encuentra la descripción de dicho operador.

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ	Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO	2016/09/05 14:39:27

```

/**
 * function that creates and configures the naive bayes operator
 */
@Override
public void f_createOperator() {
    RapidMiner.setExecutionMode(ExecutionMode.COMMAND_LINE);
    RapidMiner.init();
    try {
        this.m_naiveBayesOperator = OperatorService.createOperator(NaiveBayes.class);
        if(this.m_laplaceCorrectionCheck.isSelected()) {
            this.m_naiveBayesOperator.setParameter(NaiveBayes.PARAMETER_LAPLACE_CORRECTION, "true");
        } else {
            this.m_naiveBayesOperator.setParameter(NaiveBayes.PARAMETER_LAPLACE_CORRECTION, "false");
        }
    } catch (OperatorCreationException e) {
        e.printStackTrace();
    }
}
}

```

Figura 4.6: Ejemplo de configuración de un operador

La diferencia entre operadores radica en su propia configuración. Para facilitar su comprensión se expone un ejemplo simple:

El operador Read CSV (CSVExampleSource.class) utiliza únicamente su clase para el uso de sus parámetros.

```

/**
 * function that creates and configures the read csv operator
 */
@Override
public void f_createOperator() {
    RapidMiner.setExecutionMode(ExecutionMode.COMMAND_LINE);
    RapidMiner.init();
    try {
        this.m_readCsvOperator = OperatorService.createOperator(com.rapidminer.operator.nio.CSVExampleSource.class);
        this.m_readCsvOperator.setParameter(com.rapidminer.operator.nio.CSVExampleSource.PARAMETER_CSV_FILE, this.m_csvRouteText.getText());
        this.m_readCsvOperator.setParameter(com.rapidminer.operator.nio.CSVExampleSource.PARAMETER_COLUMN_SEPARATORS, this.m_columnText.getText());
        this.m_readCsvOperator.setParameter(com.rapidminer.operator.nio.CSVExampleSource.PARAMETER_QUOTES_CHARACTER, this.m_quotesCharacterText.getText());
        if(this.m_firstRowsBox.isSelected()) {
            this.m_readCsvOperator.setParameter(com.rapidminer.operator.nio.CSVExampleSource.PARAMETER_FIRST_ROW_AS_NAMES, "True");
        } else {
            this.m_readCsvOperator.setParameter(com.rapidminer.operator.nio.CSVExampleSource.PARAMETER_FIRST_ROW_AS_NAMES, "False");
        }
    } catch (OperatorCreationException e) {
        e.printStackTrace();
    }
}
}

```

Figura 4.7: Configuración Aplicación vs RapidMiner Studio 7.0

En cambio otros, parte de su funcionalidad depende de otras clases. El operador Normalize (Normalization.class) utiliza la clase “AttributeSubsetSelector.class” y los valores “AttributeSubsetSelector.CONDITION_SINGLE” si se quiere un único valor o “AttributeSubsetSelector.CONDITION_SUBSET” si se quiere un subconjunto de valores.

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ	Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO	2016/09/05 14:39:27

```

/**
 * function that creates and configures the normalization operator
 */
@Override
public void f_createOperator(){
    RapidMiner.setExecutionMode(ExecutionMode.COMMAND_LINE);
    RapidMiner.init();
    try {
        this.m_normalizationOperator = OperatorService.createOperator(Normalization.class);
        f_selectAttributes(this.m_attributeFilterTypeCombo);
        if(this.m_createViewCheck.isSelected()){
            this.m_normalizationOperator.setParameter(Normalization.PARAMETER_CREATE_VIEW, "true");
        }else{
            this.m_normalizationOperator.setParameter(Normalization.PARAMETER_CREATE_VIEW, "false");
        }
        f_selectMethod(this.m_methodCombo);
    } catch (OperatorCreationException e) {
        e.printStackTrace();
    }
}

/**
 * function that depends on the selected value, configures the filter parameters
 * @param a_attributeFilterTypeCombo
 */
@SuppressWarnings("rawtypes")
private void f_selectAttributes(JComboBox a_attributeFilterTypeCombo) {
    if(((String)a_attributeFilterTypeCombo.getSelectedItem()).equals("single")){
        this.m_normalizationOperator.setParameter(AttributeSubsetSelector.PARAMETER_FILTER_TYPE, ""+(AttributeSubsetSelector.CONDITION_SINGLE));
        this.m_normalizationOperator.setParameter(SingleAttributeFilter.PARAMETER_ATTRIBUTE, this.m_attributeText.getText());
    }else if(((String)a_attributeFilterTypeCombo.getSelectedItem()).equals("subset")){
        this.m_normalizationOperator.setParameter(AttributeSubsetSelector.PARAMETER_FILTER_TYPE, ""+(AttributeSubsetSelector.CONDITION_SUBSET));
        this.m_normalizationOperator.setParameter(SubsetAttributeFilter.PARAMETER_ATTRIBUTES, this.m_attributesText.getText());
    }
}

```

Figura 4.8: Configuración del operador Normalize

También a la hora de seleccionar un método de normalización, si se utiliza el método “METHOD_RANGE_TRANSFORMATION” se utiliza la clase “RangeNormalizationMethod.class” para asignar el parámetro mínimo (RangeNormalizationMethod.PARAMETER_MIN) y máximo (RangeNormalizationMethod.PARAMETER_MAX) por el cual se quiere normalizar.

```

/**
 * function that depends on the selected value, configures the normalization method parameters
 * @param a_methodCombo
 */
@SuppressWarnings("rawtypes")
private void f_selectMethod(JComboBox a_methodCombo) {
    if(((String)a_methodCombo.getSelectedItem()).equals("Z-transformation")){
        this.m_normalizationOperator.setParameter(Normalization.PARAMETER_NORMALIZATION_METHOD, ""+Normalization.METHOD_Z_TRANSFORMATION);
    }else if(((String)a_methodCombo.getSelectedItem()).equals("range transformation")){
        this.m_normalizationOperator.setParameter(Normalization.PARAMETER_NORMALIZATION_METHOD, ""+Normalization.METHOD_RANGE_TRANSFORMATION);
        this.m_normalizationOperator.setParameter(RangeNormalizationMethod.PARAMETER_MIN, this.m_minValueText.getText());
        this.m_normalizationOperator.setParameter(RangeNormalizationMethod.PARAMETER_MAX, this.m_maxValueText.getText());
    }else if(((String)a_methodCombo.getSelectedItem()).equals("proportion transformation")){
        this.m_normalizationOperator.setParameter(Normalization.PARAMETER_NORMALIZATION_METHOD, ""+Normalization.METHOD_PROPORTION_TRANSFORMATION);
    }
}

```

Figura 4.9: Configuración del parámetro method del operador Normalize

Finalmente, con este ejemplo se ha querido mostrar que la dificultad se basa en el número de clases dependientes que utilice un operador para que esté perfectamente configurado.

4.4. Flujo de proceso

Después de que todos los operadores estén configurados, tanto si se utiliza un conjunto de entrenamiento únicamente o entrenamiento y prueba, al hacer

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ	Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO	2016/09/05 14:39:27

clic sobre el botón “Start Process”, comenzará la conexión entre los operadores para posteriormente ejecutar el proceso. Mediante un diagrama de flujo se expresará el flujo de trabajo paso a paso desde el inicio y un diagrama de clases con las clases implicadas en esta sección.

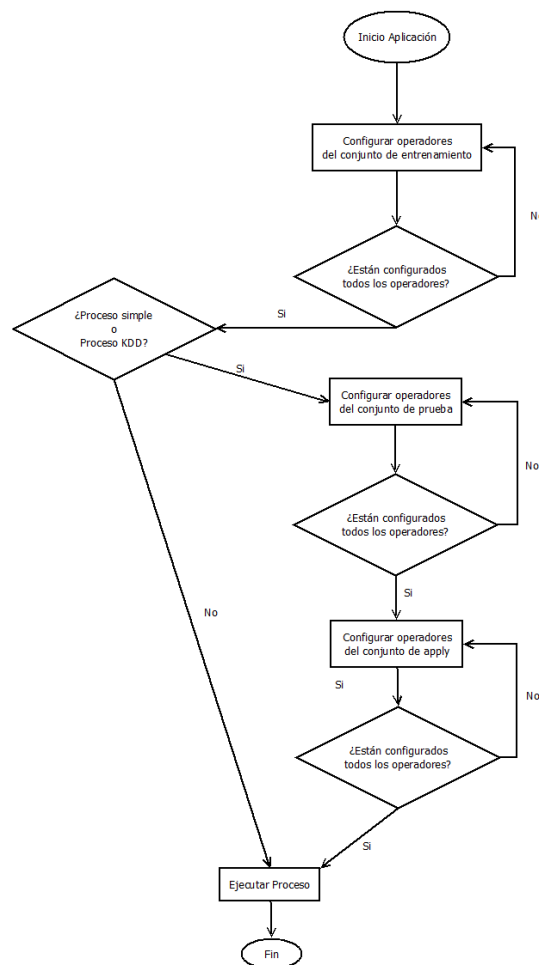


Figura 4.10: Diagrama de flujo de la aplicación

Las clases Training, Test y Apply representan los conjuntos de entrenamiento, prueba y el conjunto donde estarán los operadores para aplicar el modelo, teniendo en su interior, una lista de operadores correspondientes a cada conjunto.

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA
En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

Fecha 2016/09/05 14:36:16

UNIVERSIDAD DE LA LAGUNA
En nombre de JESUS MANUEL JORGE SANTISO

2016/09/05 14:39:27


```
public class Apply extends java.util.Observable{
    /**
     * Operator list
     */
    private List<Operator> m_apply;

    /**
     * Constructor that initializes the operator list
     */
    public Apply(){
        this.m_apply = new ArrayList<Operator>();
    }

    /**
     * function that assigns the new operator list to the current list
     * @param a_list
     */
    public void f_setApply(List<Operator> a_list) {
        this.m_apply = a_list;
    }

    /**
     * function that returns the operator list
     * @return List<Operator>
     */
    public List<Operator> f_getApply() {
        return this.m_apply;
    }

    /**
     * function that adds to list a rapidminer operator
     * @param a_operator
     */
    public void f_addApply(Operator a_operator) {
        this.m_apply.add(a_operator);
        setChanged();
        notifyObservers();
    }

    /**
     * function that deletes all operators stored in the list
     */
    public void f_clear(){
        this.m_apply.clear();
        setChanged();
        notifyObservers();
    }
}
```

Figura 4.11: Contenido de la clase Apply

En las clases TrainingProcess, TestProcess y ApplyProcess, se introducen los operadores de cada conjunto en una variable de tipo “Process”, común en ambas clases. Este objeto provisto de la librería de clases de RapidMiner, almacenará los operadores como si de una lista se tratara.

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ	Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO	2016/09/05 14:39:27

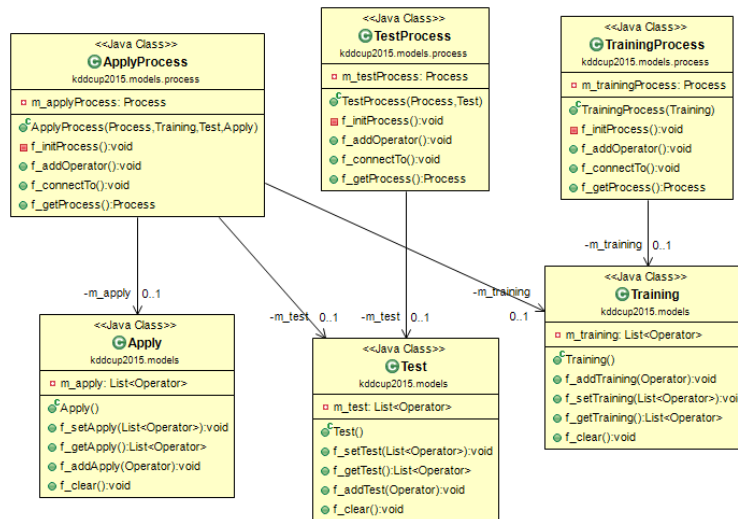


Figura 4.12: Diagrama de clases en las que interviene la variable de tipo Process

```
/**
 * overloaded function that adds the operators to the rapidminer process
 */
@Override
public void f_addOperator() {
    for(int i=0;i<this.m_training.f_getTraining().size();i++){
        this.m_trainingProcess.getRootOperator().getSubprocess(0).addOperator(this.m_training.f_getTraining().get(i));
    }
}
```

Figura 4.13: Añadiendo al proceso los operadores del conjunto de entrenamiento

```
/**
 * overloaded function that adds the operators to the rapidminer process
 */
@Override
public void f_addOperator() {
    for(int i=0;i<this.m_test.f_getTest().size();i++){
        this.m_testProcess.getRootOperator().getSubprocess(0).addOperator(this.m_test.f_getTest().get(i));
    }
}
```

Figura 4.14: Añadiendo al proceso los operadores del conjunto de test

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA
En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

Fecha 2016/09/05 14:36:16

UNIVERSIDAD DE LA LAGUNA
En nombre de JESUS MANUEL JORGE SANTISO

2016/09/05 14:39:27

```

/**
 * overloaded function that adds the operators to the rapidminer process
 */
@Override
public void f_addOperator() {
    for(int i=0;i<this.m_apply.f_getApply().size();i++){
        this.m_applyProcess.getRootOperator().getSubprocess(0).addOperator(this.m_apply.f_getApply().get(i));
    }
}

```

Figura 4.15: Añadiendo al proceso los operadores del conjunto de apply

Una vez almacenados en la lista, los operadores tienen puertos de entrada y salida. Por regla general, todos se conectarán mediante el primer puerto ya sea de entrada o salida, salvo en la clase “ApplyProcess”, que deberá incluir el operador “Apply Model”. Para conectarse a él, el último operador del conjunto de entrenamiento conectará su primer puerto de salida con el primer puerto de entrada de este operador y el conjunto de prueba conectará su primer puerto de salida con el segundo puerto de entrada.

```

/**
 * overloaded function that connects the operators by his ports
 */
@Override
public void f_connectTo() {
    for(int i=0;i<this.m_training.f_getTraining().size()-1;i++){
        this.m_training.f_getTraining().get(i).getOutputPorts().getPortByIndex(0).connectTo(this.m_training.f_getTraining().get(i+1).getInputPorts().getPortByIndex(0));
    }
}

```

Figura 4.16: Conectando los operadores del conjunto de entrenamiento

```

/**
 * overloaded function that connects the operators by his ports
 */
@Override
public void f_connectTo() {
    int t_it;
    for(t_it=0;t_it<this.m_test.f_getTest().size()-1;t_it++){
        this.m_test.f_getTest().get(t_it).getOutputPorts().getPortByIndex(0).connectTo(this.m_test.f_getTest().get(t_it+1).getInputPorts().getPortByIndex(0));
    }
}

```

Figura 4.17: Conectando al proceso los operadores del conjunto de test

```

/**
 * overloaded function that connects the operators by his ports
 * if exists in the apply list the "Apply Model" operator,
 * the last element of the training list and the test list establishes a connection with the previous operator...
 */
@Override
public void f_connectTo() {
    int t_it;
    for(t_it=0;t_it<this.m_apply.f_getApply().size()-1;t_it++){
        this.m_apply.f_getApply().get(t_it).getOutputPorts().getPortByIndex(0).connectTo(this.m_apply.f_getApply().get(t_it+1).getInputPorts().getPortByIndex(0));
    }

    this.m_training.f_getTraining().get(this.m_training.f_getTraining().size()-1).getOutputPorts().getPortByIndex(0).connectTo(this.m_apply.f_getApply().get(0).getInputPorts().getPortByIndex(0));
    this.m_test.f_getTest().get(this.m_test.f_getTest().size()-1).getOutputPorts().getPortByIndex(0).connectTo(this.m_apply.f_getApply().get(0).getInputPorts().getPortByIndex(1));
}

```

Figura 4.18: Conectando al proceso los operadores del conjunto de apply

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA

Fecha 2016/09/05 14:36:16

En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

UNIVERSIDAD DE LA LAGUNA

2016/09/05 14:39:27

En nombre de JESUS MANUEL JORGE SANTISO

La clase ProcessKddCup es la que se encargará de instanciar los objetos de las clases anteriormente mencionadas (TrainingProcess, TestProcess y ApplyProcess). En ella, también incorpora los métodos f_run() y f_stop(), que invocarán a la función run() y stop() del atributo Process y serán las que ejecuten y paren el proceso.

```
/**
 * Constructor that assigns the process and the training object to the current attributes
 * process and initializes the rapidminer process
 * @param a_training
 */
public ProcessKddCup(Training a_training){
    this.m_training = new TrainingProcess(a_training);
    this.m_process = this.m_training.f_getProcess();
}

/**
 * Constructor that assigns the process, the training and test object to the current attributes
 * @param a_training
 * @param a_test
 */
public ProcessKddCup(Training a_training,Test a_test,Apply a_apply){
    this(a_training);
    this.m_test = new TestProcess(this.m_process,a_test);
    this.m_process = this.m_test.f_getProcess();
    this.m_apply = new ApplyProcess(this.m_process,a_training, a_test,a_apply);
    this.m_process = this.m_apply.f_getProcess();
}

/**
 * function that executes the rapidminer process
 */
public void f_run(){
    try {
        this.m_process.run();
    } catch (OperatorException e) {
        e.printStackTrace();
    }
}

/**
 * function that stops the rapidminer process
 */
public void f_stop(){
    this.m_process.stop();
}
```

Figura 4.19: Clase: ProcessKDDCup

4.5. Patrones de diseño

En la realización de esta aplicación se han utilizado diversos patrones de diseño, con el fin de resolver problemas que se puedan producir a la hora de codificar. Los patrones utilizados son:

4.5.1. Patrón observador

La aplicabilidad de este patrón se encuentra en las clases Training y Test. Estas clases son las observadas y notificaran a las clases observadoras cuando se añada un operador o se elimine en la lista, esto se produce mediante los métodos setChanged() y notifyObservers().

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ	Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO	2016/09/05 14:39:27

```

public class Training extends java.util.Observable{
    /**
     * Operator list
     */
    private List<Operator> m_training;

    /**
     * Constructor that initializes the operator list
     */
    public Training() {
        this.m_training = new ArrayList<Operator>();
    }

    /**
     * function that adds to the list a rapidminer operator
     * @param a_operator
     */
    public void f_addTraining(Operator a_operator) {
        this.m_training.add(a_operator);
        setChanged();
        notifyObservers();
    }
}

```

Figura 4.20: Clase: Training con el método configurado

Las clases observadoras son TrainingTableModel, TestTableModel y ApplyTableModel, las cuales representan las tablas que muestran el nombre de los operadores configurados. Estas clases, actualizarán los objetos observados automáticamente cuando estos aumenten en número los operadores o se eliminen, esto se consigue con el método update() realizando una conversión del valor que recibe por parámetro a la variable de tipo Training y actualizando la tabla.

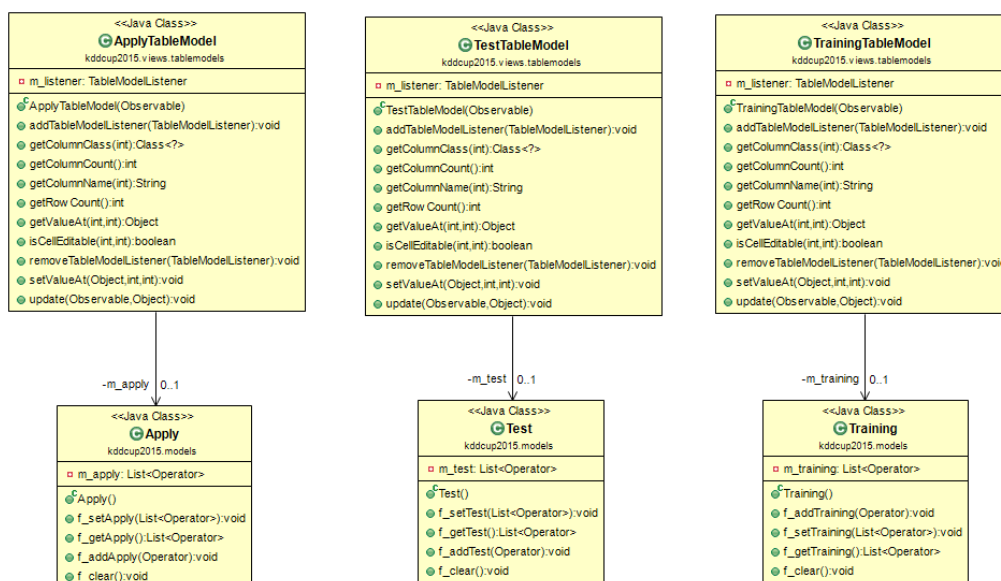


Figura 4.21: Diagrama de clases con las clases observadoras y observadas

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA

Fecha 2016/09/05 14:36:16

En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

UNIVERSIDAD DE LA LAGUNA

2016/09/05 14:39:27

En nombre de JESUS MANUEL JORGE SANTISO

```
public class TrainingTableModel implements TableModel, java.util.Observer{

    private Training m_training;
    private TableModelListener m_listener;

    public TrainingTableModel(Observable a_training){
        this.m_training = (Training) a_training;
        this.m_training.addObserver(this);
    }
    @Override
    public void update(Observable a_training, Object arg) {
        this.m_training = (Training) a_training;
        this.m_listener.tableChanged(new TableModelEvent(this));
    }
}
```

Figura 4.22: Clase: TrainingTableModel

4.5.2. Patrón estrategia

La aplicabilidad de este patrón se encuentra en todos los controladores, concretamente en el método `f_selectAccess()`, donde en tiempo de ejecución cambian el comportamiento de los objetos dependiendo del valor del parámetro del método.

```
/**
 * overloaded function that depending on the string, changes the current comboBox
 * @param a_access
 */
@Override
public void f_selectAccess(String a_access) {
    this.m_modelingDialog.getContentPane().remove(this.m_actualPanel);
    if(a_access.equals("K-nn")){
        this.m_actualPanel = new KNNPanel(this.m_modelingDialog.f_getWidth(),this.m_modelingDialog.f_getHeight()/2);
    }else if(a_access.equals("Naive Bayes")){
        this.m_actualPanel = new NaiveBayesPanel(this.m_modelingDialog.f_getWidth(),this.m_modelingDialog.f_getHeight()/2);
    }else{
        this.m_actualPanel = new DecisionTreePanel(this.m_modelingDialog.f_getWidth(), this.m_modelingDialog.f_getHeight());
    }
    this.f_addContentPane();
}
```

Figura 4.23: Cambio de comportamiento dependiendo de la elección

Uno de los inconvenientes que tiene el patrón estrategia es que tiene que conocer todas las clases que están involucradas, esto conlleva que finalmente haya varias sentencias `if...else` asignando al panel la instancia correspondiente. En este caso, se refiere a los cuadros de dialogo donde depende de la opción seleccionada, se instancia un panel con los parámetros para configurar un operador.

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ	Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO	2016/09/05 14:39:27

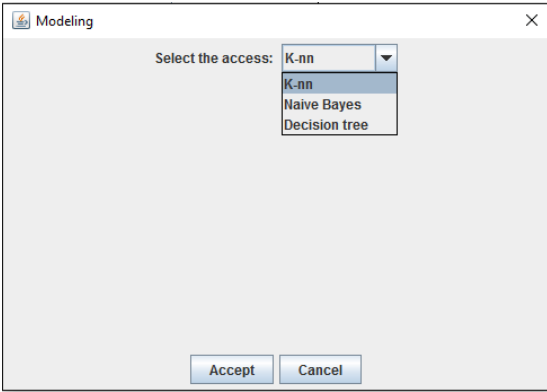


Figura 4.24: Cuadro de dialogo en donde se puede seleccionar el algoritmo

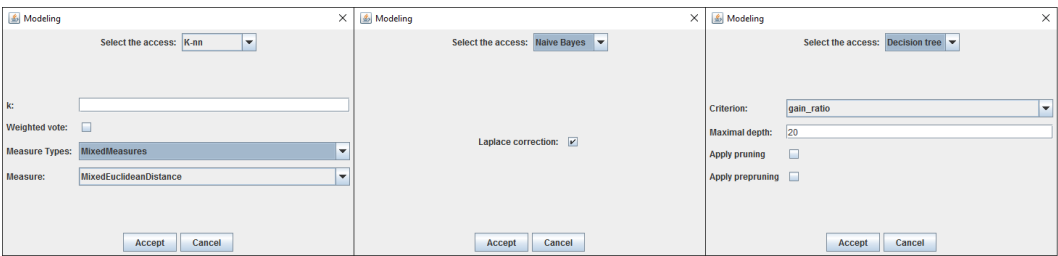


Figura 4.25: Cambio de comportamiento dependiendo de la elección

4.5.3. Patrón MVC

La aplicabilidad de este patrón reside en separar el código fuente de los eventos producidos por la interfaz gráfica, también denominada vista e incluirlo en clases que supervisen esas interfaces, de ahí el nombre de controlador.

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion		
Identificador del documento: 757360		Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ		Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO		2016/09/05 14:39:27

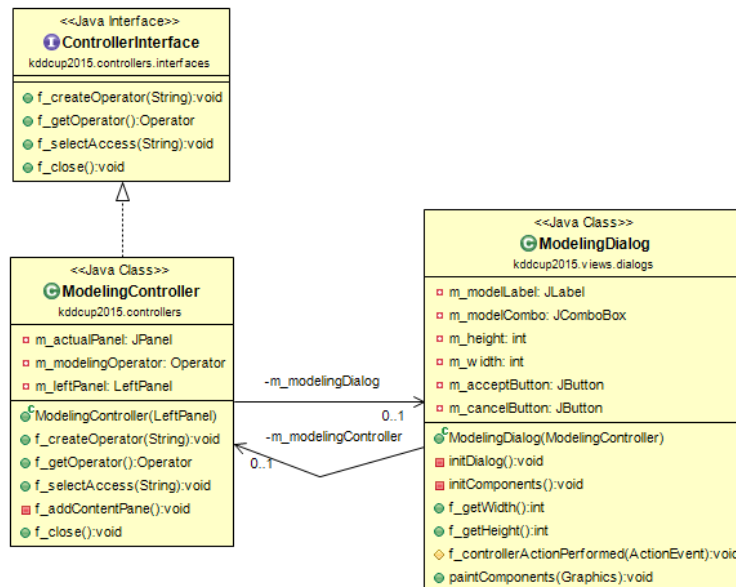


Figura 4.26: Diagrama de clases que representan el uso del patrón MVC

```

/**
 * function that depends on the selected event and selects the controller's function
 * @param a_event
 */
protected void f_controllerActionPerformed(ActionEvent a_event) {
    if(a_event.getSource() == this.m_modelCombo){
        this.m_modelingController.f_selectAccess((String)this.m_modelCombo.getSelectedItem());
    }else if(a_event.getSource() == this.m_acceptButton){
        this.m_modelingController.f_createOperator((String)this.m_modelCombo.getSelectedItem());
    }else{
        this.m_modelingController.f_close();
    }
}

```

Figura 4.27: Método que dependiendo del evento invoca a una función u otra del controlador

El controlador también es el que crea la vista y en la vista se utiliza el controlador, como se pudo observar en la imagen anterior, por lo que el controlador se pasa como parámetro al constructor de la vista, para así poder utilizar sus métodos dependiendo del evento seleccionado.

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ	Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO	2016/09/05 14:39:27


```
/**
 * Constructor that initializes the attributes
 * @param a_leftPanel
 */
public ModelingController(LeftPanel a_leftPanel){
    this.m_modelingDialog = new ModelingDialog(this);
    this.m_modelingDialog.setVisible(true);
    this.m_actualPanel = new JPanel();
    this.m_leftPanel = a_leftPanel;
}
```

Figura 4.28: Constructor de la clase ModelingController

```
/**
 * Constructor that initializes the attributes
 * @param a_modelingController
 */
public ModelingDialog(ModelingController a_modelingController){
    this.m_height = 350;
    this.m_width = 500;
    this.m_modelingController = a_modelingController;
    initDialog();
    initComponents();
}
```

Figura 4.29: Constructor de la clase ModelingDialog

El modelo es una variable de tipo operador que se obtiene del panel configurado, correspondiendo con el operador. Este, se incluye en la variable “m_leftPanel”, en alguno de los conjuntos disponibles (training, test o apply) dependiendo de cuál ha sido seleccionado.

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ	Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO	2016/09/05 14:39:27

Capítulo 5

Caso de estudio

5.1. Introducción

Los cursos masivos online y abiertos (en inglés, Massive Open Online Course (MOOC)) revolucionan la educación, ya que proveen fácil acceso a los materiales de un curso mediante Internet. Este tipo de cursos tienen una gran aceptación respecto a los cursos en los centros tradicionales, gracias a la disponibilidad de los materiales y su bajo coste. Sin embargo, aparte de estas ventajas, también tienen una enorme desventaja: la tasa de abandono de los cursos suele ser muy alta.

Como se ha comentado a lo largo de esta memoria, en este proyecto se usarán las técnicas de minería de datos para finalmente, realizar una predicción sobre los estudiantes que abandonan los cursos. Para ello, previamente se han analizado los datos ofrecidos por la plataforma, para extraer nuevas características y generar nuevo conocimiento, creando otras tablas en la base de datos.

En los siguientes apartados, se describirán las nuevas características y tablas generadas, así como la configuración del proceso en RapidMiner Studio 7.0 y la curva resultante.

5.2. Generación de nuevas características y creación de la vista minable

El primer paso que se ha realizado al introducir las tablas iniciales en la base de datos, ha sido renombrarlas al idioma español para una mejor comprensión de los datos de problema.

Antes de analizar los datos, se han hecho unas consultas previas sobre la tabla “resultados_train”, ya que el caso práctico se va a resolver unicamente sobre el conjunto de entrenamiento del conjunto de datos global proporcionado

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ	Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO	2016/09/05 14:39:27

por la entidad, para saber el número de alumnos que abandonan o continúan estudiando:

- El número total de registros de dicha tabla es de 120.542.

```
1 use kddcup2015;
2 |
3 /*Número total de inscripciones por curso en el conjunto de entrenamiento*/
4 SELECT *
5 FROM resultados_train
```

resultados_train (2x120.542)

Figura 5.1: Registros totales

- El número de registros de los alumnos que abandonan es de 95.581.

```
1 use kddcup2015;
2 |
3 /*Número total de inscripciones que abandonan por curso en el conjunto de entrenamiento*/
4 SELECT *
5 FROM resultados_train
6 WHERE Resultado = 1
```

resultados_train (2x95.581)

Figura 5.2: Registros que abandonan

- El número de registros de los alumnos que no abandonan es de 24.961

```
1 use kddcup2015;
2 |
3 /*Número total de inscripciones que no abandonan por curso en el conjunto de entrenamiento*/
4 SELECT *
5 FROM resultados_train
6 WHERE Resultado = 0
```

resultados_train (2x24.961)

Figura 5.3: Registros que no abandonan

Por lo tanto, se puede demostrar que, sin aplicar ningún conocimiento nuevo o técnicas de minería de datos, el abandono de los cursos es de un 79 % y los que no abandonan es de un 21 %. El objetivo de este proyecto es obtener una precisión predictiva superior al 79 %.

Para extraer conocimiento nuevo, se ha generado una batería de cuestiones sobre los datos iniciales. Estas preguntas se han resuelto mediante consultas en la base de datos, creando nuevas tablas. En esta sección, se expondrán las preguntas más relevantes, su consulta en lenguaje SQL y finalmente, unos ejemplos de los datos resultantes:

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ	Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO	2016/09/05 14:39:27

- ¿Cuántos alumnos hay por curso?

```
/*Alumnos inscritos por curso en el conjunto de entrenamiento*/
CREATE TABLE inscritosporcurso_train AS
SELECT Curso_Id, COUNT(Usuario) AS NE
FROM inscritos_train
GROUP BY Curso_Id;
```

Figura 5.4: Consulta: Inscritosporcurso_train

Curso_Id	NE
1pvLqtotBskv7QSOsLicJDQMhX3lui6d	2392
3cnZpv6ReApmCaZyaQwi2izDZxVRdC01	2207
3VktHkmOtom3jM2wCu94xgzZu1d6Dn7or	2008
5Gyp41oLV07Gg7vF4vpmggWP5MU70QO6	2992

Figura 5.5: Resultados: inscritosporcurso_train

- ¿Cuántos cursos tiene un alumno?

```
/*Cursos en los que está inscrito un alumno en el conjunto de entrenamiento*/
CREATE TABLE cursosporalumno_train AS
SELECT Usuario, Count(DISTINCT Curso_Id) AS NC
FROM inscritos_train
GROUP BY Usuario
ORDER BY Count(DISTINCT Curso_Id) DESC;
```

Figura 5.6: Consulta: cursosporalumno_train

Usuario	NC
okqN4nBVKRYrkYonKidKGOMpqVGyJZh	27
viuKt4GgrX0rYL5EpHoxJbsiC9TbISuk	25
s6iksgptFmyh2vfiarKdJkiKnp6RLi3f	25
nYSFDE8RCbcZhD46fRSvPHUvA1ahfupi	21

Figura 5.7: Resultados: cursosporalumno_train

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA

Fecha 2016/09/05 14:36:16

En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

UNIVERSIDAD DE LA LAGUNA

2016/09/05 14:39:27

En nombre de JESUS MANUEL JORGE SANTISO

- ¿Cuántos alumnos de un curso aprobaron?

```
/*Estadísticas del número de éxito y fracaso de alumnos por curso
en el conjunto de entrenamiento*/
CREATE TABLE estadisticasdealumnosporcurso_train AS
SELECT Curso_Id, NE, NAS, (NE-NAS) AS FRACASO, (NAS/NE) AS PS, ((NE-NAS)/NE) AS PF
FROM inscritosporcurso_train NATURAL JOIN numscritosfinalizancurso_train;
```

Figura 5.8: Consulta: Curso con alumnos aprobados

Curso_Id	NE	NAS	FRACASO	PS	PF
1pvLqtoBskv7QSOsLicJDQMhX3lui6d	2392	546	1846	0.2283	0.7717
3cnZpv6ReApmCaZyaQwi2izDZxVRdC01	2207	520	1687	0.2356	0.7644
3VkhKmOtom3jM2wCu94xgzdu1d6Dn7or	2008	376	1632	0.1873	0.8127
5Gyp41oLV07Gg7vF4vpmggWp5MU70Q06	2992	389	2603	0.1300	0.8700

Figura 5.9: Resultados: Curso con alumnos aprobados

- ¿Cuántos cursos aprobó un alumno?

```
/*Estadísticas del número de éxito y fracaso de los cursos por alumnos
en el conjunto de entrenamiento*/
CREATE TABLE estadisticasdecursosporalumnos_train AS
SELECT Usuario, NC, NCS, (NC-NCS) AS NCF, (NCS/NC) AS PS, ((NC-NCS)/NC) AS PF
FROM cursosporalumno_train NATURAL JOIN numcursosfinalizanolosalumnos_train
```

Figura 5.10: Consulta: Alumnos con cursos aprobados

Usuario	NC	NCS	NCF	PS	PF
okqN4nBVKRYrYonKicKGomPqVGyJZh	27	25	2	0.9259	0.0741
viuKt4GgrX0rYL5EpHoxJbsiC9TbISuk	25	23	2	0.9200	0.0800
s6iksgptFmyh2vfiarKdJkiKnp6RLi3f	25	23	2	0.9200	0.0800
nYSFDE8RCbcZhD46fRSvPHUvA1ahfupi	21	3	18	0.1429	0.8571

Figura 5.11: Resultados: Alumnos con cursos aprobados

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA

Fecha 2016/09/05 14:36:16

En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

UNIVERSIDAD DE LA LAGUNA

2016/09/05 14:39:27

En nombre de JESUS MANUEL JORGE SANTISO

- ¿Número de días en los que un alumno accede a un curso?

```
/*Número de días accedidos al curso por inscripcion en el conjunto de entrenamiento*/
CREATE TABLE numdiasaccedidosalcurso_eventologtrainunique AS
SELECT Curso_Id,Inscripcion_Id,COUNT(DISTINCT DATE_FORMAT(Momento_Evento, '%Y-%m-%d')) AS NEDV
FROM evento_log_train_unique NATURAL JOIN inscritos_train
GROUP BY Curso_Id,Inscripcion_Id;
```

Figura 5.12: Consulta: Días accedidos a un curso

Curso_Id	Inscripcion_Id	NEDV
1pvLqtotBsKv7QSOsLicJDQMhX3lui6d	13797	9
1pvLqtotBsKv7QSOsLicJDQMhX3lui6d	13798	4
1pvLqtotBsKv7QSOsLicJDQMhX3lui6d	13799	10
1pvLqtotBsKv7QSOsLicJDQMhX3lui6d	13800	7

Figura 5.13: Resultados: Días accedidos a un curso

- ¿Número de accesos totales a los módulos de un curso?

```
/*Número de accesos a eventos pertenecientes de un modulo del curso en el conjunto de entrenamiento*/
CREATE TABLE numaccesosaeventosporcurso_eventologtrainunique AS
SELECT Curso_Id,Inscripcion_Id,COUNT(DATE_FORMAT(Momento_Evento, '%Y-%m-%d')) AS NEV
FROM evento_log_train_unique NATURAL JOIN inscritos_train
GROUP BY Curso_Id,Inscripcion_Id;
```

Figura 5.14: Consulta: Accesos totales a los módulos

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA
En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

Fecha 2016/09/05 14:36:16

UNIVERSIDAD DE LA LAGUNA
En nombre de JESUS MANUEL JORGE SANTISO

2016/09/05 14:39:27

Curso_Id	Inscripcion_Id	NEV
1pvLqtotBskv7QSOsLicJDQMhX3lui6d	13797	270
1pvLqtotBskv7QSOsLicJDQMhX3lui6d	13798	90
1pvLqtotBskv7QSOsLicJDQMhX3lui6d	13799	318
1pvLqtotBskv7QSOsLicJDQMhX3lui6d	13800	189

Figura 5.15: Resultados: Accesos totales a los módulos

- ¿Qué evento perteneciente a un curso es el más accedido por un alumno?

```
/*Número de accesos por tipo de evento en el conjunto de entrenamiento*/
CREATE TABLE numaccesosportipoevento_eventologtrainunique AS
SELECT Inscripcion_Id,Tipo_Evento,COUNT(Tipo_Evento) AS NEL
FROM evento_log_train_unique
GROUP BY Inscripcion_Id,Tipo_Evento;
```

Figura 5.16: Consulta: Evento más accedido

Inscripcion_Id	Tipo_Evento	NEL
1	access	107
1	navigate	25
1	page_close	66
1	problem	80
1	video	29
3	access	79
3	discussion	26
3	navigate	14
3	page_close	22
3	problem	81
3	video	9
4	access	61
4	navigate	15
4	page_close	8
4	problem	6
4	video	4

Figura 5.17: Resultados: Evento más accedido

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA

Fecha 2016/09/05 14:36:16

En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

UNIVERSIDAD DE LA LAGUNA

2016/09/05 14:39:27

En nombre de JESUS MANUEL JORGE SANTISO

- ¿Número de categorías por curso?

```
/*Número de categorías que tiene el curso*/
CREATE TABLE numcategoriasporcurso_modulo AS
SELECT Curso_Id,COUNT(DISTINCT Categoria) AS NCD
FROM estadisticasdemodulosporcatdecorso_modulo
GROUP BY Curso_Id;
```

Figura 5.18: Consulta: Número de categorías por curso

Curso_Id	NCD
1pvLqtoTBskv7QSOsLicJDQMhx3lui6d	10
3cnZpv6ReApmCaZyaQwi2izDZxVRdC01	11
3VkhkmOtom3jM2wCu94xgzuz1d6Dn7or	12
5Gyp41oLV07Gg7vF4vpmggWP5MU70QO6	12

Figura 5.19: Resultados: Número de categorías por curso

- ¿Número de módulos por curso?

```
/*Número de módulos pertenecientes a un curso*/
CREATE TABLE modulosporcurso_modulo AS
SELECT Curso_Id,COUNT(Distinct Modulo_Id) AS NM
FROM modulo
GROUP BY Curso_Id;
```

Figura 5.20: Consulta: Número de módulos por curso

Curso_Id	NM
1pvLqtoTBskv7QSOsLicJDQMhx3lui6d	333
3cnZpv6ReApmCaZyaQwi2izDZxVRdC01	713
3VkhkmOtom3jM2wCu94xgzuz1d6Dn7or	402
5Gyp41oLV07Gg7vF4vpmggWP5MU70QO6	488

Figura 5.21: Resultados: Número de módulos por curso

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA

Fecha 2016/09/05 14:36:16

En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

UNIVERSIDAD DE LA LAGUNA

2016/09/05 14:39:27

En nombre de JESUS MANUEL JORGE SANTISO

- ¿Número de categorías que tiene un módulo perteneciente a un curso?

```
/*Número de módulos pertenecientes a un curso por categoría*/
CREATE TABLE modulosporcatcurso_modulo AS
SELECT Curso_Id, Categoria,COUNT(Distinct Modulo_Id) AS NMPC
FROM modulo
GROUP BY Curso_Id,Categoria
ORDER BY Curso_Id,Categoria;
```

Figura 5.22: Consulta: Número de categorías

Curso_Id	Categoria	NMPC
1pvLqtotBskv7QSOsLicJDQMh3lui6d	about	3
1pvLqtotBskv7QSOsLicJDQMh3lui6d	chapter	11
1pvLqtotBskv7QSOsLicJDQMh3lui6d	course	1
1pvLqtotBskv7QSOsLicJDQMh3lui6d	course_info	2
1pvLqtotBskv7QSOsLicJDQMh3lui6d	discussion	17
1pvLqtotBskv7QSOsLicJDQMh3lui6d	problem	33
1pvLqtotBskv7QSOsLicJDQMh3lui6d	sequential	28
1pvLqtotBskv7QSOsLicJDQMh3lui6d	static_tab	2
1pvLqtotBskv7QSOsLicJDQMh3lui6d	vertical	132
1pvLqtotBskv7QSOsLicJDQMh3lui6d	video	104

Figura 5.23: Resultados: Número de categorías

- ¿Número de accesos a los tipos de eventos por inscripción?

```
/*Número de accesos a los tipos de eventos de un curso por inscripción*/
CREATE TABLE numtipoeventosaccedidosporinscripcion_eventologtrainunique AS SELECT Inscripcion_Id,COUNT(DISTINCT Tipo_Evento) AS NTA
FROM evento_log_train_unique
GROUP BY Inscripcion_Id;
```

Figura 5.24: Consulta: Número de accesos

Inscripcion_Id	NTA
1	5
3	6
4	5
5	6
6	5

Figura 5.25: Resultados: Número de accesos

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA

Fecha 2016/09/05 14:36:16

En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

UNIVERSIDAD DE LA LAGUNA

2016/09/05 14:39:27

En nombre de JESUS MANUEL JORGE SANTISO

- ¿Cuántos días tienen de duración los cursos?

```
/*Número de días de duración de los cursos*/
CREATE TABLE rangodiasporcurso AS SELECT *, DATEDIFF(Fin,Inicio) as RangoCurso
FROM curso
```

Figura 5.26: Consulta: Número de días por curso

Curso_Id	Inicio	Fin	RangoCurso
1pvLqtotBskV7Q5OsLicJDQMhX3lui6d	2013-11-26	2013-12-25	29
3cnZpv6ReApmCaZyaQwi2zDZxVRdC01	2014-05-25	2014-06-23	29
3VkhkmOtom3jM2wCu94xgzuz1d6Dn7or	2013-11-01	2013-11-30	29
5Gyp41oLV07Gg7vF4vpmggWP5MU70QO6	2013-12-11	2014-01-09	29

Figura 5.27: Resultados: Número de días por curso

- ¿Número de módulos únicos visitados por inscripción?

```
/*Número de módulos únicos visitados de un curso por inscripción*/
CREATE TABLE nummodulosdistintosvisitadosporinscripcion_eventlogtrainunique AS
SELECT Incripcion_Id,COUNT(DISTINCT Modulo) AS NMV
FROM evento_log_train_unique
GROUP BY Incripcion_Id
```

Figura 5.28: Consulta: Número de módulos

Inscripcion_Id	NMV
1	95
3	99
4	29
5	111
6	16

Figura 5.29: Resultados: Número de módulos

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA

Fecha 2016/09/05 14:36:16

En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

UNIVERSIDAD DE LA LAGUNA

2016/09/05 14:39:27

En nombre de JESUS MANUEL JORGE SANTISO

- ¿Número de cursos simultáneos que tiene un alumno en un determinado mes?

```
/*Número de cursos simultáneos que tiene un alumno en un determinado mes*/
CREATE TABLE numcursossimultaneosporusuario_train AS SELECT Usuario,DATE_FORMAT(Inicio,'%m-%Y') as Fecha,COUNT(Inscripcion_Id) as CS
FROM Curso NATURAL JOIN inscritos_train
GROUP BY Usuario,DATE_FORMAT(Inicio,'%m-%Y')
ORDER BY Usuario,DATE_FORMAT(Inicio,'%m-%Y')
```

Figura 5.30: Consulta: Cursos simultáneos

Usuario	Fecha	CS
00038q9lTDdhWUJPTqFNEuMEA6pdK5n	11-2013	1
0011RjfJQIzbeqatOj8hS4pIQzsOFbeu	11-2013	1
001Wosm650x4ktE3NPV76K6QhVwIKDP6	10-2013	1
001Wosm650x4ktE3NPV76K6QhVwIKDP6	12-2013	1
0089b3aJIRi14gwpkv8EJVR4pjkVvye8	11-2013	1
0089b3aJIRi14gwpkv8EJVR4pjkVvye8	12-2013	1
008XUUt5rc6hUrg7SHj9oW0rG0FE0dkF	05-2014	1
00DCGVn7t4aRvR.2CsgdD8ZhYzNhFGq86	05-2014	3

Figura 5.31: Resultados: Cursos simultáneos

- ¿Número de días de acceso entre el primero y último acceso por inscripción?

```
/*Rango entre la primera fecha de acceso al curso y el ultimo por alumno al curso
en el conjunto de entrenamiento*/
CREATE TABLE fechaminymaxdeaccesoporinscripcion_eventologtrainunique AS
SELECT DISTINCT Inscripcion_Id,Curso_Id,MIN DATE_FORMAT(Momento_Evento,'%Y-%m-%d')) AS MIN,MAX DATE_FORMAT(Momento_Evento,'%Y-%m-%d')) AS MAX,
DATEDIFF(MAX DATE_FORMAT(Momento_Evento,'%Y-%m-%d')),MIN DATE_FORMAT(Momento_Evento,'%Y-%m-%d')) AS Rango
FROM evento_log_train_unique NATURAL JOIN inscritos_train
GROUP BY Inscripcion_Id,Curso_Id
ORDER BY Inscripcion_Id ASC;
```

Figura 5.32: Consulta: Rango

Inscripcion_Id	Curso_Id	MIN	MAX	Rango
1	DPnLzkJJqOOPRJfBxiHbQEERiYHu5ila	2014-06-14	2014-07-11	27
3	7GRhBDsirIGkRZBTSMZNTyDr2JQm4xx	2014-06-19	2014-07-17	28
4	DPnLzkJJqOOPRJfBxiHbQEERiYHu5ila	2014-06-15	2014-07-02	17
5	7GRhBDsirIGkRZBTSMZNTyDr2JQm4xx	2014-06-19	2014-07-18	29
6	AXUJZGmZ0xaYSWazu8RQ1G5c76ECT1Kd	2014-07-01	2014-07-02	1
7	7GRhBDsirIGkRZBTSMZNTyDr2JQm4xx	2014-06-19	2014-07-15	26

Figura 5.33: Resultados: Rango

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA
En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

Fecha 2016/09/05 14:36:16

UNIVERSIDAD DE LA LAGUNA
En nombre de JESUS MANUEL JORGE SANTISO

2016/09/05 14:39:27

Nombre	Descripción	Tipo
Usuario	Nombre de usuario	Nominal
Curso_Id	Nombre del curso	Nominal
NEDV	Número de días accedidos al curso	Numérico
Rango	Diferencia entre el primer y último acceso al curso	Numérico
RangoCurso	Duración del curso	Numérico
NMV	Número de módulos visitados en un curso	Numérico
NM	Número de módulos	Numérico
NTA	Número de tipo de evento accedido	Numérico
CS	Número de cursos simultáneos	Numérico
Resultado (Clase)	0 si continúa y 1 si abandona el curso	Numérico

Tabla 5.1: Atributos seleccionados para la vista minable

A partir de todos estos atributos generados, se han seleccionado los más relevantes y se han insertado en una única tabla, la cual se denomina vista minable. Esta tabla es sobre la que los algoritmos de minería de datos trabajaran, sin prestar atención al atributo que forma la clase, para posteriormente realizar la predicción predictiva. Los atributos seleccionados en esta nueva tabla son:

Usuario	Curso_Id	NEDV	Rango	RangoCurso	NMV	NM	NTA	CS	Resultado
1qXC7Fjbwp66GPQc6pHlfeU08WKozxG4	7GRhBDsrIGkRZBISMEzNTyDr2JQm4kcx	9	28	29	99	699	6	1	0
1ELM1tXjpjncZU4WkxvrVri8AjR2gf	AXUJZGmZ0xaYSWazu8RQ1G5c76ECT1Kd	5	15	29	49	264	6	1	0
0K6JPqivQzicxY4EV4nqMLCL3a08A97	DPnLzkJ3qOOPRJfBxIHbQEERIYHu5ila	11	25	29	77	398	7	1	0
088SASUPOVYUGhoEly8virkGvRB3oNwp	7GRhBDsrIGkRZBISMEzNTyDr2JQm4kcx	18	29	29	101	699	7	2	0
0XSM8IXWzvML1r2AhrvzvWxbczqfFP	DPnLzkJ3qOOPRJfBxIHbQEERIYHu5ila	13	22	29	59	398	7	1	1
1h4cVFontLTW8vs6Jg6kDELHITwYJukub	TAyXodh39I2LZnfbpL0LFF2NxrCKplox	1	0	29	2	333	1	1	1
0U6Ls9kSlxf9NGu5KLvjyV4KIOonn33	AXUJZGmZ0xaYSWazu8RQ1G5c76ECT1Kd	8	17	29	66	264	7	1	0
1T8tUcZn49U0rZp09wSozsY7mAk229	DPnLzkJ3qOOPRJfBxIHbQEERIYHu5ila	8	24	29	39	398	7	1	0
0fEolC14nqwTVtr21LrdvEqXhccEQPz	TAyXodh39I2LZnfbpL0LFF2NxrCKplox	10	19	29	45	333	7	2	1
0SiWaSfGyO3je3WqDu3MOT4Rr4grcql3	DPnLzkJ3qOOPRJfBxIHbQEERIYHu5ila	10	16	29	55	398	7	1	0

Figura 5.34: Resultados: Vista minable

5.3. Configuración del proceso utilizado y resultados usando RapidMiner Studio 7.0

La vista minable generada en la sección anterior se utilizará como conjunto de datos. Los objetivos de la competición establecían el área bajo la curva ROC, por lo que se usará el operador que cumpla con ese criterio. En esta sección se mostrarán los operadores seleccionados y su configuración en RapidMiner Studio, dando como resultado el siguiente proceso:

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003	
La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ	Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO	2016/09/05 14:39:27

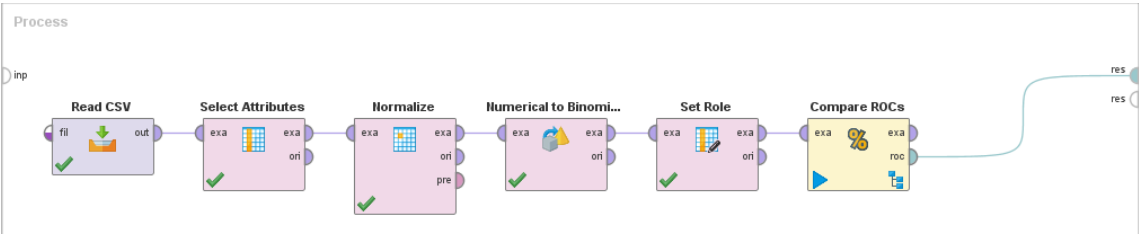


Figura 5.35: Proceso resultante en RapidMiner Studio 7.0

1. Read CSV: Operador para leer archivos con formato CSV. Para que los atributos formen parte del operador, hay que configurar el archivo mediante el botón “Import Configuration Wizard” y no seleccionándolo en el parámetro “csv file”. Los demás parámetros vienen así por defecto.

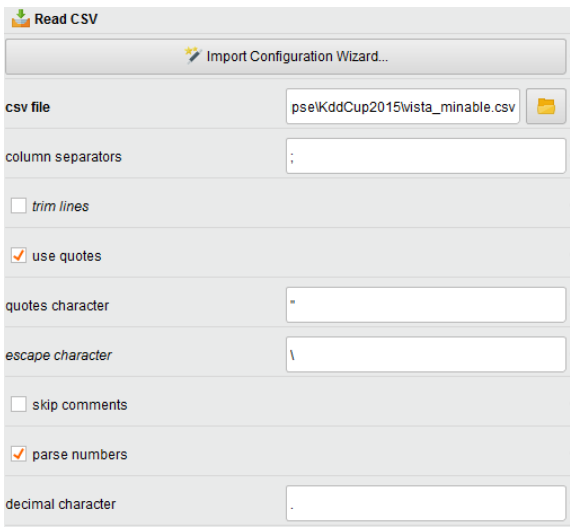


Figura 5.36: Configuración del operador READ CSV

2. Select Attributes: Operador que selecciona un atributo o conjunto de atributos. Esto se realiza mediante el parámetro “attribute filter type” y “attribute” si se selecciona single o “attributes” si se selecciona subset. Se han seleccionado todos los atributos salvo el “Usuario” y “Curso_Id”, ya que no tienen relevancia en el proceso.

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion		
Identificador del documento: 757360		Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ		Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO		2016/09/05 14:39:27

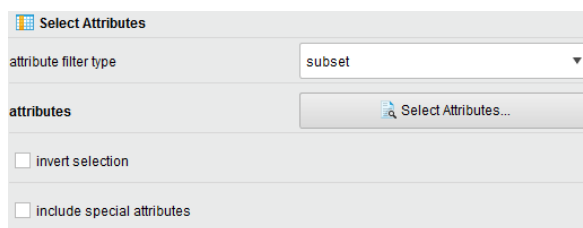


Figura 5.37: Configuración del operador SELECT ATTRIBUTES

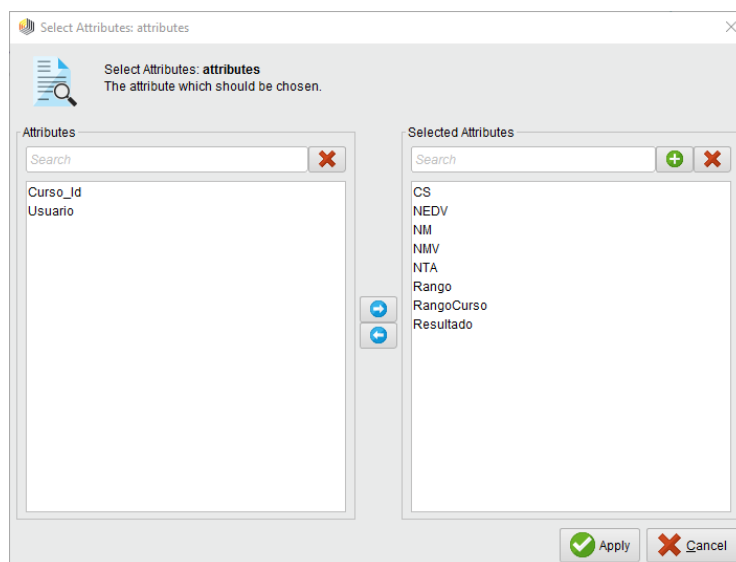


Figura 5.38: Configuración del operador SELECT ATTRIBUTES

3. Normalize: Operador que transforma los datos entre un rango de valores. Se ha seleccionado un subconjunto de datos excluyendo el atributo que representa a la clase y el método de transformación por rango 0-1.

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA
En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

Fecha 2016/09/05 14:36:16

UNIVERSIDAD DE LA LAGUNA
En nombre de JESUS MANUEL JORGE SANTISO

2016/09/05 14:39:27

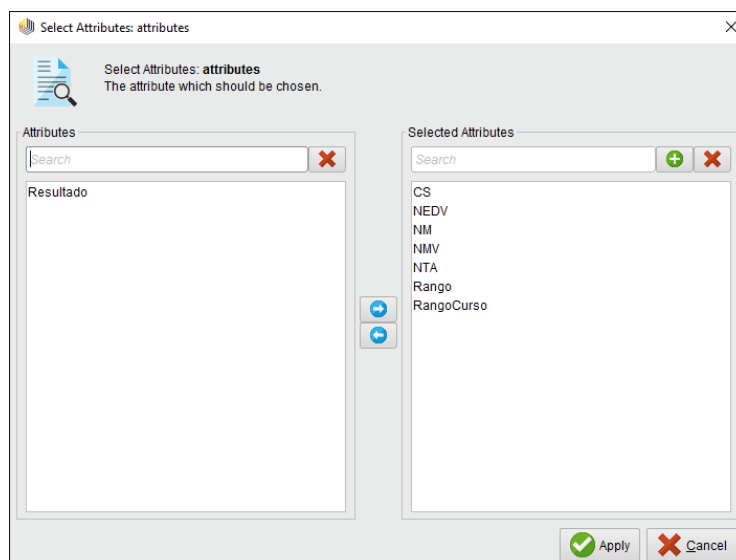


Figura 5.39: Configuración del operador NORMALIZE

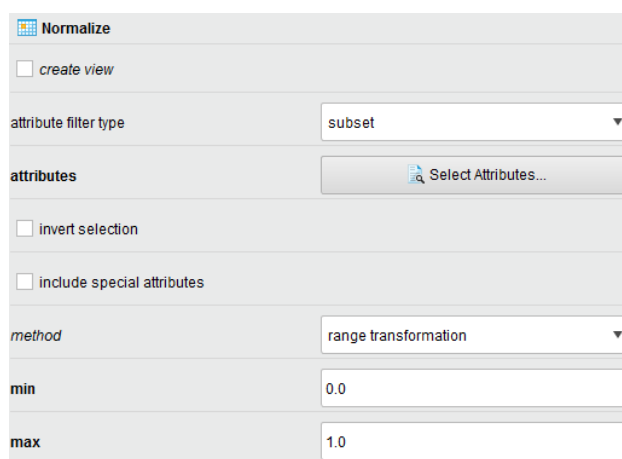


Figura 5.40: Configuración del operador NORMALIZE

4. Numerical to Binomial: Operador que transforma el valor de un atributo numérico en valor true-false. Al tener el valor 1 tanto en los parámetros min y max, significa que 1 será tratado como false y 0 como true, tal y como se especificaba en los objetivos de la competición.

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ	Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO	2016/09/05 14:39:27

Figura 5.41: Configuración del operador NUMERICAL TO BINOMIAL

5. Set Role: Operador que asigna un rol al atributo seleccionado. El rol que se asigna es de tipo clase “label”.

Figura 5.42: Configuración del operador SET ROLE

6. Compare ROCs: Para utilizar este operador no se realiza una división del conjunto de datos en otros dos subconjuntos (entrenamiento y prueba), ya que solo tiene un puerto de entrada, por lo que el proceso tendrá menos operadores que el mostrado en el capítulo 4 de esta memoria. Sin embargo, utiliza un número de pliegues y un tipo de muestreo aleatorio para poder llevar a cabo su tarea. Por defecto, el número de pliegues es 10 y el tipo de muestreo es estratificado, esto quiere decir que se realizaran subconjuntos para llevar a cabo la tarea.

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA
En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

Fecha 2016/09/05 14:36:16

UNIVERSIDAD DE LA LAGUNA
En nombre de JESUS MANUEL JORGE SANTISO

2016/09/05 14:39:27

Compare ROCs

number of folds

10

split ratio

0.7

sampling type

stratified sampling

☐ use local random seed

☒ use example weights

roc bias

optimistic

Figura 5.43: Configuración del operador COMPARE ROCS

Los algoritmos de minería de datos se incluyen dentro de este operador. Independientemente de su configuración, únicamente han de conectarse los puertos de entrada y salida con los que vienen dentro del panel.

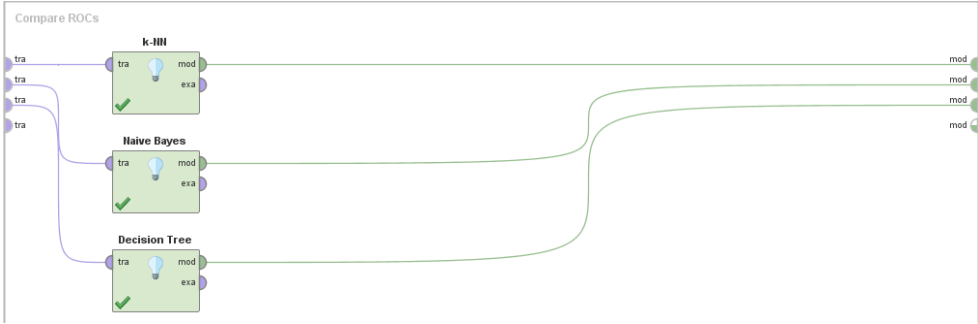


Figura 5.44: Algoritmos dentro del proceso COMPARE ROCS

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion		
Identificador del documento: 757360		Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ		Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO		2016/09/05 14:39:27

Si se desea obtener la curva roc, se conecta el segundo puerto de salida del operador al del panel, tal y como se mostró en la imagen al inicio de este capítulo.

Para finalizar, se ejecuta el proceso, que tardará varios minutos por la ejecución simultánea de los tres algoritmos y al completarse, mostrará la curva ROC.

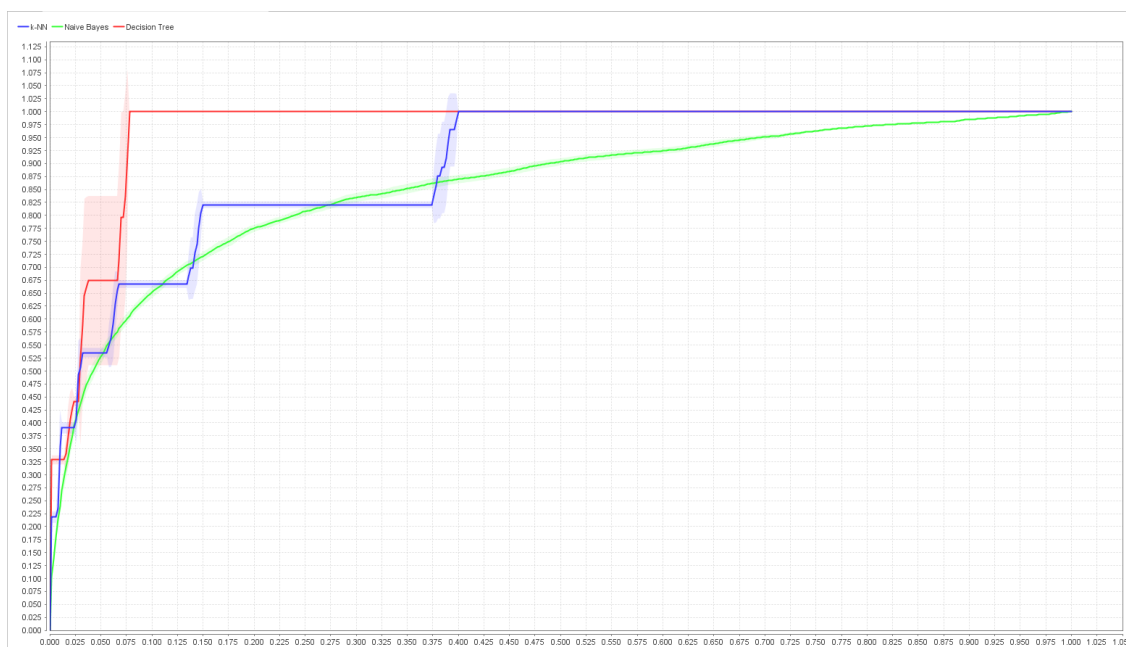


Figura 5.45: Curva Roc resultante

Se puede observar que el algoritmo que ofrece el mejor resultado es el árbol de decision (Decision Tree), seguido del K-nn y por último el Naive Bayes. Esto se debe a la altura de las curvas que genera el operador.

5.4. Resultados utilizando la aplicación Java

En la versión actual de la aplicación, los resultados se almacenarán un archivo csv.

En el capítulo siguiente se usará la aplicación de RapidMiner Studio para resolver en detalle el caso de estudio. En este capítulo también se resolverá, realizando una partición de la vista minable en dos conjuntos: uno de entrenamiento y otro de prueba. La partición será de un 70 % para el conjunto de entrenamiento y un 30 % para el conjunto de prueba.

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA

Fecha 2016/09/05 14:36:16

En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

UNIVERSIDAD DE LA LAGUNA

2016/09/05 14:39:27

En nombre de JESUS MANUEL JORGE SANTISO

Utilizando las técnicas comentadas en el capítulo 2, la única variación en el proceso resultante es la configuración de las mismas técnicas. A continuación, se muestra una ejecución de cada técnica con su configuración y el resultado final:

1. Árbol de decisión

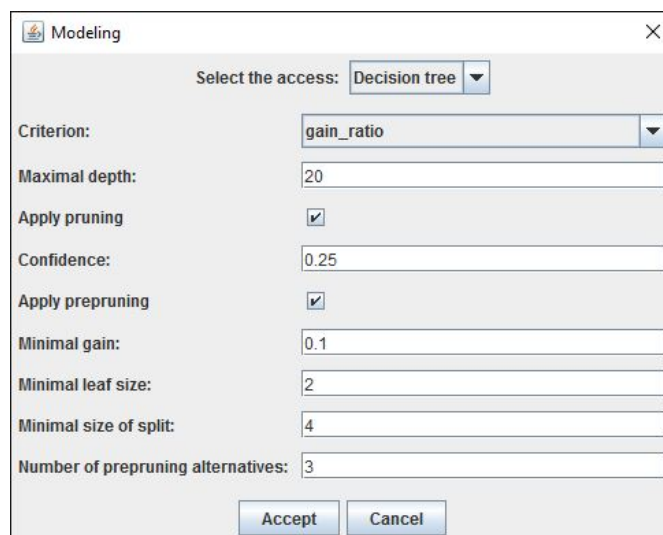


Figura 5.46: Configuración de la técnica: Árbol de decisión

Criterion	Value	Standard Deviation	Variance
accuracy	0.854968448789959		

Figura 5.47: Resultado del Árbol de decisión

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA
En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

Fecha 2016/09/05 14:36:16

UNIVERSIDAD DE LA LAGUNA
En nombre de JESUS MANUEL JORGE SANTISO

2016/09/05 14:39:27

2. KNN

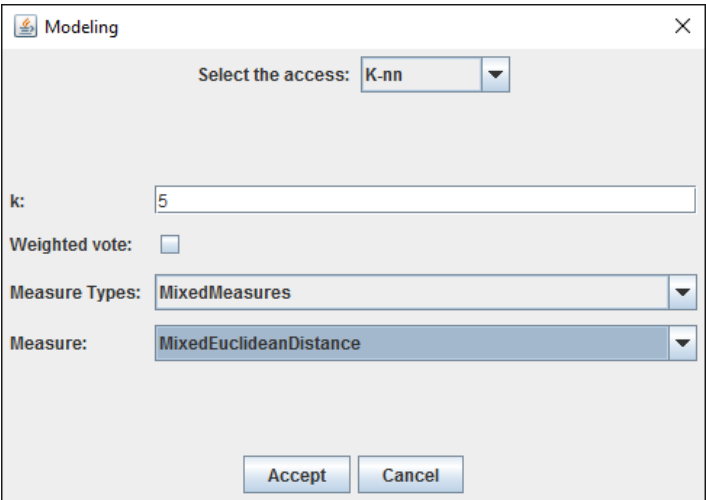


Figura 5.48: Configuración de la técnica: KNN

Criterion	Value	Standard Deviation	Variance
accuracy	0.8950488870397337		

Figura 5.49: Resultado de KNN

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ	Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO	2016/09/05 14:39:27

3. Naive Bayes

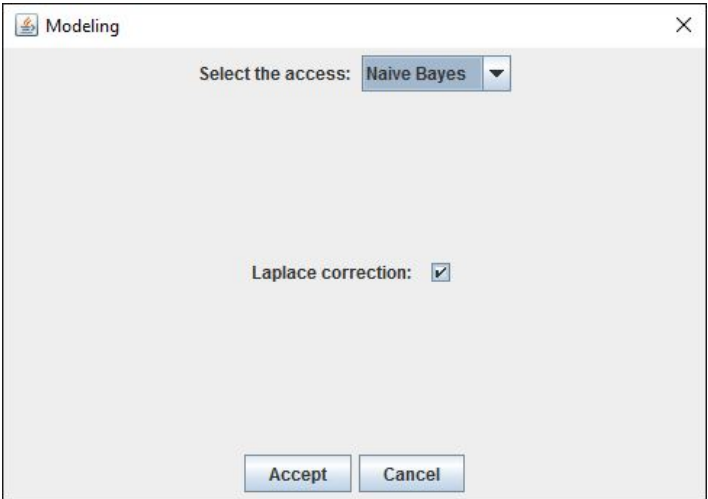


Figura 5.50: Configuración de la técnica: Naive Bayes

Criterion	Value	Standard Deviation	Variance
accuracy	0.8474100270438943		

Figura 5.51: Resultado de Naive Bayes

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ	Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO	2016/09/05 14:39:27

Capítulo 6

Conclusiones y líneas futuras

Con este trabajo se ha profundizado en el fantástico mundo de la minería de datos, realizando un caso práctico para los cursos online masivos y abiertos y desarrollando una aplicación en Java utilizando los operadores internos de RapidMiner Studio 7.0. Ambos son temas de actualidad, los cursos online por la facilidad de inscripción y su bajo coste y la minería de datos por su continuo crecimiento y resolución de problemas enfocados a cualquier ámbito.

6.1. Líneas futuras

En esta sección se indicarán las mejoras planteadas para la aplicación software realizada:

- La inclusión en la aplicación de más operadores de RapidMiner Studio 7.0 o versiones más actualizadas, ya sea tanto de selección, limpieza, transformación de atributos o de técnicas de minería de datos.
- La visualización de resultados gráficamente, ya sea utilizando las facilidades proporcionadas por las librerías de clases de RapidMiner Studio 7.0 o creándola desde cero o con ayuda de otras herramientas de visualización.
- Plantear la creación de una aplicación web que utilice los operadores de RapidMiner Studio 7.0 o versiones más actualizadas y permita realizar los mismos procesos de minería de datos. Así se evita la limitación de recursos de una máquina de sobremesa o portátil.
- Modificar la interfaz gráfica de usuario para que sea del estilo “Drag and Drop” y que el usuario pueda arrastrar los componentes a un panel y realizar su configuración.

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ	Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO	2016/09/05 14:39:27

Capítulo 7

Summary and Conclusions

This work has provided a deeper understanding of the fantastic world of data mining, performing a study case for massive and open online courses and developing an application in Java using internal operators of RapidMiner Studio 7.0. Both of them are hot topics, online courses for ease of entry and low cost and data mining for their continued growth and problem-solving focused on any field.

7.1. Future lines

In this section we describe the improvements proposed for the software application:

- The inclusion on the application of more operators of RapidMiner Studio 7.0 or latest versions, such that selection, cleaning, transformation of attributes or data mining techniques.
- Graphical displaying of the results using the RapidMiner Studio 7.0 class libraries or from scratch or using other visualization tools.
- Developing a web application using RapidMiner operators that allow to perform data mining processes. So the limited resources of a machine or laptop is avoided.
- Modifying the application to be style "Drag and Drop." and that the user can drag components to a panel and perform configuration.

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ	Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO	2016/09/05 14:39:27

Capítulo 8

Presupuesto

En este capítulo se mostrarán el coste de una estación de trabajo[16] para poder realizar los procesos de minería de datos y el del personal para llevar a cabo un proyecto de estas características:

Estación de trabajo	Coste
HP Z640 Tower Workstation	2.479,29 €

Personal	Horas trabajadas	Coste/Hora	Total Personal
1	300	30 €	9000 €

Estación de trabajo + Personal	Total Presupuesto
2.479,29 €+ 9000 €	11.479,29 €

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ	Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO	2016/09/05 14:39:27

Apéndice A

Relación Operador RapidMiner Studio 7.0 - Clase Java

En este apéndice se incluirán en forma de tablas distribuidos por secciones, al igual que ocurre al seleccionar un operador en RapidMiner Studio 7.0, la relación de un operador con su correspondiente clase en Java. No se ha podido establecer una relación de todos, por lo que únicamente estarán los descubiertos.

A.1. Data Access

- Files

- Read

Operador	Clase Java
Read CSV	CSVExampleSource
Read Excel	ExcelExampleSource
Read Excel With Format	ExcelFormatExampleSource

- Write

Operador	Clase Java
Write CSV	CSVExampleSetWriter
Write Excel	ExcelExampleSetWriter
Write Excel With Format	SpecialFormatExampleSetWriter

A.2. Blending

- Attributes

- Names & Roles

Operador	Clase Java
Rename	ChangeAttributeName
Rename by Replacing	ChangeAttributesNamesReplace
Rename By Generic Names	ChangeAttributesNames2Generic
Rename By Constructions	Construction2Names
Set Role	ChangeAttributeRole
Exchange Roles	ExchangeAttributeRoles

- Types

Operador	Clase Java
Numerical to Binomial	NumericToBinomial
Numerical to Polynomial	NumericToPolynomial
Numerical to Real	Numerical2Real
Numerical to Date	Numerical2Date
Real to Integer	Real2Integer
Nominal to Binomial	NominalToBinomial
Nominal to Text	Nominal2String
Nominal to Numerical	NominalToNumeric
Nominal to Date	Nominal2Date
Text To Nominal	String2Nominal
Date To Numerical	Date2Numerical
Date to Nominal	Date2Nominal
Parse Numbers	NominalNumbers2Numerical
Format Numbers	NumericToFormattedNominal
Guess Types	GuessValueTypes

- Selection

Operador	Clase Java
Select Attributes	AttributeFilter
Select By Weights	AttributeWeightSelection
Select By Random	RandomSelection
Remove Attribute Range	ExampleRangeFilter
Remove Useless Attributes	RemoveUselessFeatures
Remove Correlated Attributes	RemoveCorrelatedFeatures
Work On Subset	AttributeSubsetPreprocessing

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA

Fecha 2016/09/05 14:36:16

En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

UNIVERSIDAD DE LA LAGUNA

2016/09/05 14:39:27

En nombre de JESUS MANUEL JORGE SANTISO

- Generation

Operador	Clase Java
Generate ID	IDTagging
Generate EmptyAttributes	AttributeAdd
Generate Copy	AttributeCopy
Generate Concatenation	AttributeMerge
Generate Aggregation	AggregationOperator
Generate Absolutes	AbsoluteValueFilter
Generate Products	ProductGenerationOperator
Generate Gaussians	GaussFeatureConstructionOperator
Generate TFIDF	TFIDFFilter
Generate Item Set Indicators	FrequentItemSetAttributeCreate
Generate Weight (LPR)	LocalPolynomialRegressionOperator
Reorder Attributes	AttributeOrderingOperator

- Examples

Operador	Clase Java
Filter Examples	ExampleFilter
Filter Examples Range	ExampleRangeFilter
Sample	AbsoluteSampling
Sample(Stratified)	StratifiedSamplingOperator
Sample(BootStrapping)	BootStrappingOperator
Sample(Kennard-Stone)	KennardStoneSampling
Sample(Model-Based)	ModelBasedSampling
Sort	Sorting

- Table

- Rotation

Operador	Clase Java
Transpose	ExampleSetTranspose

- Joins

Operador	Clase Java
Append	ExampleSetMerge
Join	ExampleSetJoin
Set Minus	ExampleSetMinus
Intersect	ExampleSetIntersect
Union	ExampleSetUnion
Superset	ExampleSetSuperset
Cartesian Product	ExampleSetCartesian

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA

Fecha 2016/09/05 14:36:16

En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

UNIVERSIDAD DE LA LAGUNA

2016/09/05 14:39:27

En nombre de JESUS MANUEL JORGE SANTISO

- Values

Operador	Clase Java
Map	AttributeValueMapper
Replace	AttributeValueReplace
Replace Directory	ExampleSetToDictionary
Split	AttributeValueSplit
Trim	AttributeValueTrim
Merge	MergeNomialValues
Add	AddNominalValue
Remap Binomials	InternalBinomialRemapping

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA

Fecha 2016/09/05 14:36:16

En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

UNIVERSIDAD DE LA LAGUNA

2016/09/05 14:39:27

En nombre de JESUS MANUEL JORGE SANTISO

A.3. Cleansing

- Normalization

Operador	Clase Java
Normalize	Normalization
De-Normalize	DenormalizationOperator

- Binning

Operador	Clase Java
Size	AbsoluteDiscretization
Binning	BinDiscretization
Frequency	FrequencyDiscretization
User Specification	UserBasedDiscretization
Entropy	MinimalEntropyDiscretization

- Missing

Operador	Clase Java
Impute Missing Values	MissingValueImputation
Declare Missing Values	DeclareMissingValueOperator
Replace Infinite Values	InfiniteValueReplenishment
Remove Unusued Values	RemoveUnusuedNominalValuesOperator
Fill Data Gaps	FillDataGaps

- Duplicates

Operador	Clase Java
Remove Duplicates	RemoveDuplicates

- Outliers

Operador	Clase Java
Detect Outlier (Distances)	DKNOutlierOperator
Detect Outlier (Densities)	DBOOutlierOperator
Detect Outlier (LOF)	LOFOutlierOperator
Detect Outlier (COF)	EcoDbOperator

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA

Fecha 2016/09/05 14:36:16

En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

UNIVERSIDAD DE LA LAGUNA

2016/09/05 14:39:27

En nombre de JESUS MANUEL JORGE SANTISO

■ Dimensionality Reduction

Operador	Clase Java
Principal Component Analysis	PCA
Principal Component Analysis (Kernel)	KernelPCA
Independent Component	FastICA
Generalized Hebbian Algorithm	GHA
Singular Value Decomposition	SVDReduction
Self-Organizing Map	SOMDimensionalityReduction
Fourier Transformation	FourierTransformation

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA

Fecha 2016/09/05 14:36:16

En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

UNIVERSIDAD DE LA LAGUNA

2016/09/05 14:39:27

En nombre de JESUS MANUEL JORGE SANTISO

A.4. Modeling

- Predictive

- Lazy

- K-nn

Operador	Clase Java
K-nn	KNNLearner

- Bayesian

- Naive Bayes

Operador	Clase Java
Naive Bayes	NaiveBayes

- Trees

- Decision Tree

Operador	Clase Java
Decision Tree	DecisionTreeLearner

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA

Fecha 2016/09/05 14:36:16

En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

UNIVERSIDAD DE LA LAGUNA

2016/09/05 14:39:27

En nombre de JESUS MANUEL JORGE SANTISO

A.5. Scoring

- Confidences

Operador	Clase Java
Rescale Confidences	PlattScaling
Drop Uncertain Predictions	SimpleUncertainPredictionsTransformation
Generate Prediction	GeneratePredictionOperator
Generate Prediction Ranking	DenormalizationOperator
Find Threshold	ThresholdFinderOperator
Create Threshold	ThresholdCreator
Apply Threshold	ThresholdApplier
Select Recall	RecallChooser

Operador	Clase Java
Apply Model	ModelApplier

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA

Fecha 2016/09/05 14:36:16

En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

UNIVERSIDAD DE LA LAGUNA

2016/09/05 14:39:27

En nombre de JESUS MANUEL JORGE SANTISO

A.6. Validation

■ Performance

● Predictive

Operador	Clase Java
Performance (Classification)	PerformanceEvaluator
Performance (Binomial Classification)	BinomialClassificationPerformance
Performance (Regression)	RegressionPerformanceEvaluator
Performance (Costs)	CostEvaluator
Performance (Ranking)	RankingEvaluator
Performance (Support Vector Count)	SupportVectorCounter
Performance (Attribute Count)	AttributeCounter

● Segmentation

Operador	Clase Java
Cluster Count Performance	ClusterNumberEvaluator
Cluster Distance Performance	CentroidBasedEvaluator
Cluster Density Performance	ClusterDensityEvaluator
Item Distribution Performance	DistributionEvaluator
Map Clustering On Labels	ClusterToPrediction

● Significance Tests

Operador	Clase Java
T-Test	TtestSignificanceTestOperator
ANOVA	AnovaSignificanceTestOperator

Operador	Clase Java
Combine Performance	WeightPerformanceCreator
Performance (User-Based)	UserBasedPerformanceEvaluator
Performance (Min-Max)	MinMaxWrapper
Performance to Data	AnovaSignificanceTestOperator

■ Visual

Operador	Clase Java
Create Lift Chart	LiftParetoChartGenerator
Compare ROCs	ROCBasedComparisionOperator
Create Learning Curve	LearningCurveOperator
Visualize Model by SOM	SOMModelVisualization

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA

Fecha 2016/09/05 14:36:16

En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

UNIVERSIDAD DE LA LAGUNA

2016/09/05 14:39:27

En nombre de JESUS MANUEL JORGE SANTISO

Operador	Clase Java
Split Validation	SplitValidationOperator
Compare ROCs	XValidation
BootStrapping Validation	BootStrappingValidation
Batch-X-Validation	BatchXValidation
Wrapper Split Validation	RandomSplitWrapperValidationChain
Wrapper-X-Validation	WrapperXValidation

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003

La autenticidad de este documento puede ser comprobada en la dirección: <https://sede.ull.es/validacion>

Identificador del documento: 757360

Código de verificación: gHnJNBEL

Firmado por: UNIVERSIDAD DE LA LAGUNA

Fecha 2016/09/05 14:36:16

En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ

UNIVERSIDAD DE LA LAGUNA

2016/09/05 14:39:27

En nombre de JESUS MANUEL JORGE SANTISO

Bibliografía

- [1] <http://www.cs.waikato.ac.nz/ml/weka/>
- [2] <https://www.knime.org/>
- [3] <http://www-03.ibm.com/software/products/es/spss-modeler>
- [4] <https://rapidminer.com/products/studio/>
- [5] <https://mariadb.org/>
- [6] <https://maven.apache.org/>
- [7] <https://github.com/rapidminer/rapidminer-studio>
- [8] <https://www.java.com/es/>
- [9] <https://eclipse.org/>
- [10] <https://git-scm.com/>
- [11] <https://github.com/>
- [12] <http://www.oracle.com/technetwork/articles/java/index-jsp-135444.html>
- [13] <http://www.stack.nl/~dimitri/doxygen/>
- [14] <http://junit.org/junit4/>
- [15] <http://docs.rapidminer.com/studio/operators/rapidminer-studio-operator-reference.pdf>
- [16] <http://store.hp.com/SpainStore/Merch/Product.aspx?id=T4K60EA&opt=ABE&sel=DTP>

Este recibo incorpora firma electrónica de acuerdo a la Ley 59/2003 La autenticidad de este documento puede ser comprobada en la dirección: https://sede.ull.es/validacion	
Identificador del documento: 757360	Código de verificación: gHnJNBEL
Firmado por: UNIVERSIDAD DE LA LAGUNA En nombre de MANUEL ALEJANDRO BACALLADO LOPEZ	Fecha 2016/09/05 14:36:16
UNIVERSIDAD DE LA LAGUNA En nombre de JESUS MANUEL JORGE SANTISO	2016/09/05 14:39:27