Predicción de abandonos en cursos online masivos y abiertos

Manuel Alejandro Bacallado López

Trabajo Final de Grado

II Congreso de Estudiantes de Informática de la Universidad de La Laguna

Escuela Superior de Ingeniería y Tecnología

30 de Noviembre de 2016

Índice

Índice

Introducción

- El proyecto planteado está basado en un problema real:
  - Los cursos online masivos y abiertos se realizan por todo el mundo.
  - Muchos alumnos se inscribirán a estos cursos debido a su bajo coste y gran cantidad de materiales disponibles, pero no llegarán a terminarlos.
- Se analizarán los datos del problema para realizar una clasificación de los alumnos matriculados en los cursos, prestando atención a su abandono o finalización con éxito.
- La predicción se llevará a cabo mediante técnicas de minería de datos, ya que este proceso manual es lento y subjetivo.

Índice

Antecedentes y estado actual del tema

- Los datos provienen del desafío KDD Cup 2015 (Predicting dropouts in MOOC)
- KDD Cup es una competición a nivel mundial de minería de datos y descubrimiento de conocimiento.
- Organizado por la ACM (Association of Computing Machinery).
- Vigencia desde 1997-Actualidad.
- Cada año se realiza el mismo proceso:
  - Habilitan los datasets con la información a tratar.
  - Instrucciones a seguir (Plazos de entrega, objetivos a cumplir).
  - Premio en $ para los primeros puestos.

Antecedentes y estado actual del tema

- Los datos se estructuran en las siguientes tablas:
  - date.csv
  - object.csv
  - enrollment_train.csv - enrollment_test.csv
  - log_train.csv - log_test.csv
  - truth_train.csv

Índice

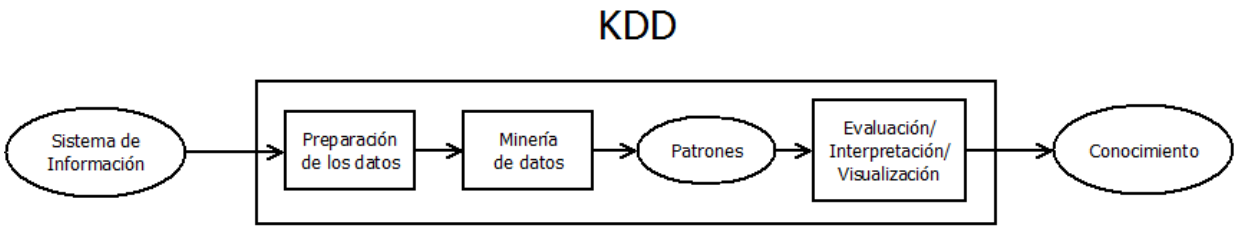
Objetivos

- Los objetivos de este proyecto se constituyen en:
  - Realizar una clasificación sobre los alumnos matriculados en los cursos online masivos y abiertos.
  - Utilizar herramientas de software libre para:
    - Crear una aplicación en Java que use los operadores internos de la aplicación RapidMiner Studio 7.0.
    - Almacenamiento de los datos iniciales en un base de datos y creación de tablas con conocimiento nuevo generado(Ingeniería de características).
    - Técnicas de ingeniería del software para un desarrollo profesional de la aplicación.

Índice

Fundamentos de minería de datos

- Se define como el proceso de extraer conocimiento útil y comprensible desde grandes cantidades de datos almacenados en distintos formatos.



Fundamentos de minería de datos

- Tareas
  - Predictivas: Predecir uno o más valores para uno o más casos. Los casos van acompañados de una etiqueta(clase, categoría o valor numérico).
    - Clasificación: Cada registro en la base de datos pertenece a una clase, la cual se establece mediante el valor de un atributo denominado clase de la instancia.
  - Descriptivas: Los casos constituyen un conjunto sin etiquetas. El objetivo es describir los datos existentes.
- Técnicas
  - Árboles de decisión
  - Casos y vecindad(K-nn)
  - Técnicas probabilistas(Naive Bayes)

Índice

Herramientas utilizadas

- SGBD Maria DB
- Apache Maven
- Biblioteca de clases de RapidMiner Studio 7.0
- Java
- Eclipse
- Git
- Github
- Doxygen
- JUnit

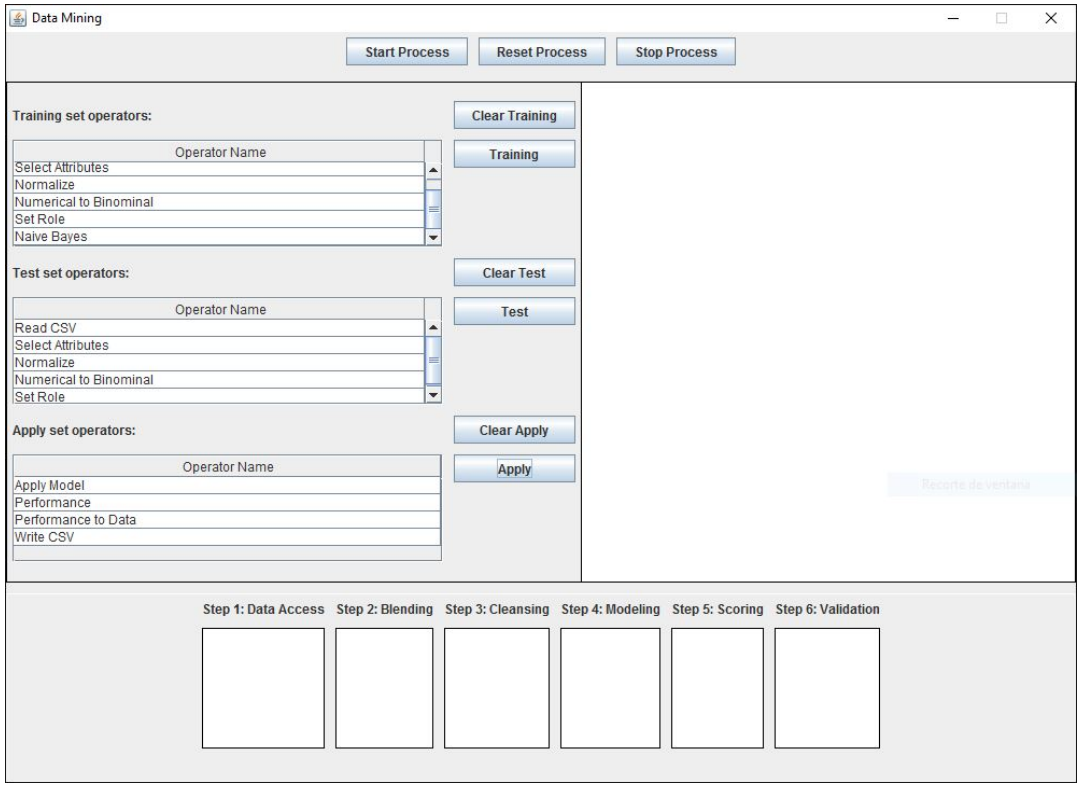
Índice

Aplicación software

- La aplicación software ha sido realizada utilizando:
  - Java y Swing.
  - Los operadores internos de RapidMiner Studio 7.0.
  - Patrones de diseño:
    - Patrón Observador(Observer Pattern).
    - Patrón Estrategia(Strategy Pattern).
    - Patrón Modelo-Vista-Controlador(MVC Pattern).

Aplicación software

- La aplicación software presenta la siguiente estructura:



Índice

Caso de estudio

- Previamente a la generación de nuevas características se han realizado varias tareas:
  - Se han renombrado las tablas iniciales al castellano para una mejor comprensión.
  - Se eliminó la duplicidad de los registros en las tablas.
  - Se eliminaron atributos en las tablas que no aportaban información relevante.
  - Se analizó la tabla resultados_train.csv(truth_train.csv) para comprobar el número de abandonos sin realizar minería de datos, teniendo el valor 1 si abandona y valor 0 si continúa:
    - El número total de registros es de: 120.542.
    - El número de alumnos que abandonan es de: 95.581.
    - El número de alumnos que no abandonan es de: 24.691.

Caso de estudio

- La generación de nuevo conocimiento se ha producido mediante una batería de cuestiones. A continuación se expondrán las más relevantes:
  - ¿Cuántos alumnos hay por curso?
  - ¿Cuántos cursos tiene un alumno?
  - ¿Cuántos alumnos de un curso aprobaron?
  - ¿Cuántos cursos aprobó un alumno?
  - ¿Número de días en los que un alumno accede a un curso?
  - ¿Número de cursos simultáneos que tiene un alumno en un determinado mes?
  - ¿Número de días entre el primero y último acceso por inscripción?

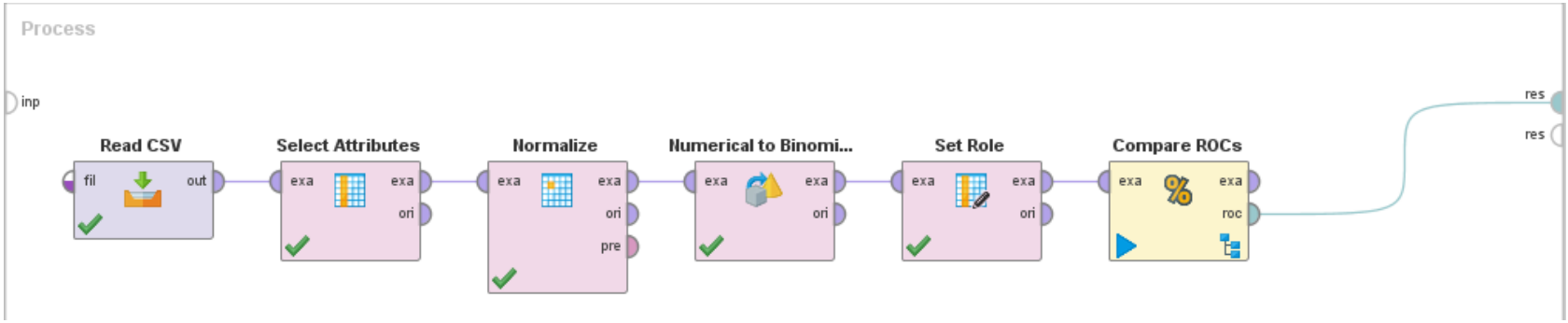
Caso de estudio

- Los atributos seleccionados para la creación de la vista minable son:

Usuario	Curso_Id	NEDV	Rango	RangoCurso	NMV	NM	NTA	CS	Resultado
1qXC7Fjbwp66GPQc6pHLfEuO8WKozxG4	7GRhBDsirIGkRZBtSMEzNTyDr2JQm4xx	9	28	29	99	699	6	1	0
1ELMItXjpijnCZU4WKxvrVri8AjkR2gf	AXUJZGmZ0xaYSWazu8RQ1G5c76ECT1Kd	5	15	29	49	264	6	1	0
0K6JPqiivQzicxY4EV4nqMLCL3a08A97	DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila	11	25	29	77	398	7	1	0
088SASUPOVYUGhoEly8vlrkGvRBJoNwp	7GRhBDsirIGkRZBtSMEzNTyDr2JQm4xx	18	29	29	101	699	7	2	0
0XSMtliXiWzvML1r2AHrvvzWxbczqfFP	DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila	13	22	29	59	398	7	1	1
1h4cVFonLTW8vs6Jg6kDELHITwYJukUb	TAYxxh39I2LZnftBpL0LfF2NxzrCKpkx	1	0	29	2	333	1	1	1
0U6Ls9kSlXfs9NGu5KLvjyV4KJIOnn33	AXUJZGmZ0xaYSWazu8RQ1G5c76ECT1Kd	8	17	29	66	264	7	1	0
1T8ttUcZNn49U0rZpO9wSozsY7mAk229	DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila	8	24	29	39	398	7	1	0
0fEolC14nqwTVtr21LrdveqXhccEiQPz	TAYxxh39I2LZnftBpL0LfF2NxzrCKpkx	10	19	29	45	333	7	2	1
0SiWaSfGyOJje3Wq0u3MOT4Rr4grcqL3	DPnLzkJJqOOPRJfBxIHbQEERiYHu5ila	10	16	29	55	398	7	1	0

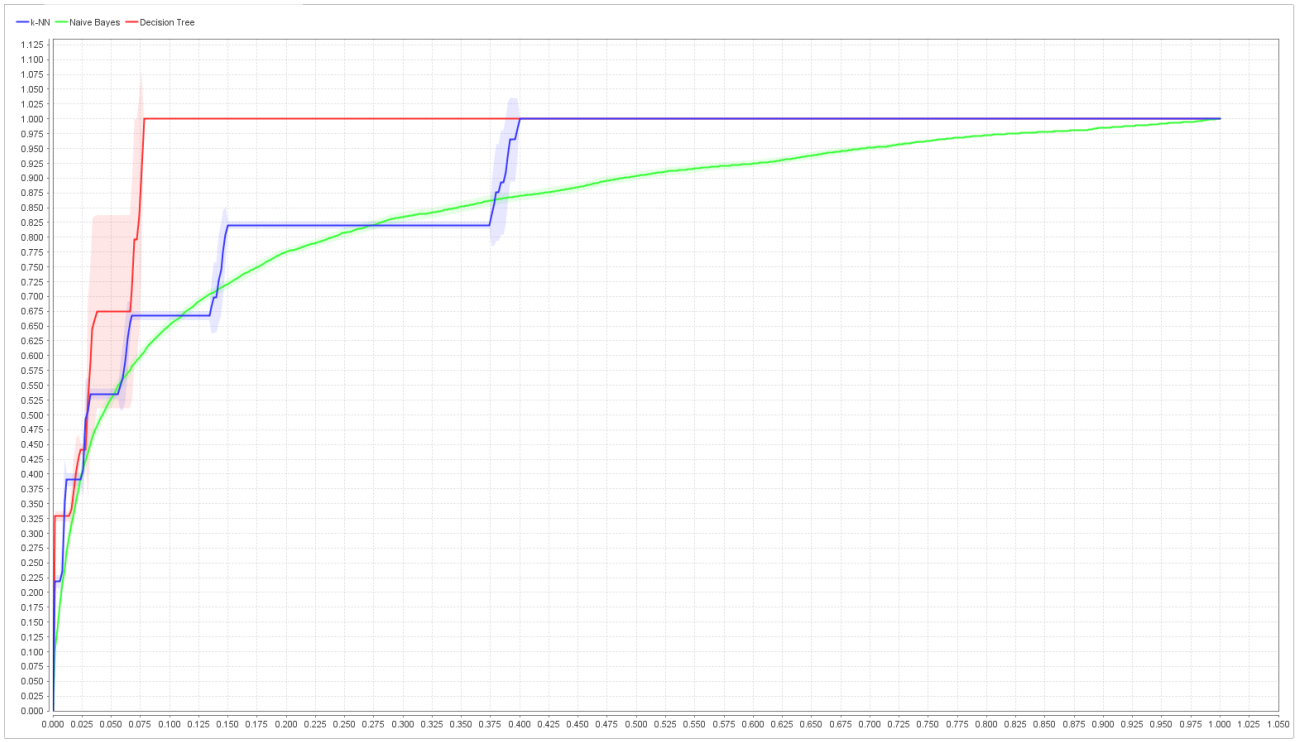
Caso de estudio

- El proceso resultante en RapidMiner Studio 7.0 para resolver el caso de estudio es el siguiente:



Caso de estudio

- El resultado utilizando RapidMiner Studio 7.0 es el siguiente:



Caso de estudio

- Los resultados utilizando los algoritmos en la aplicación software son los siguientes:
  - Árbol de decisión:

Criterion	Value	Standard Deviation	Variance
accuracy	0.854968448789959		

  - K-nn:

Criterion	Value	Standard Deviation	Variance
accuracy	0.8950488870397337		

  - Naive Bayes:

Criterion	Value	Standard Deviation	Variance
accuracy	0.8474100270438943		

Índice

Conclusiones

- 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.

Índice

Líneas futuras

- La inclusión en la aplicación de más operadores de RapidMiner Studio 7.0 o versiones más actualizadas.
- 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.
- Modificar la interfaz gráfica de usuario para que sea del estilo "Drag and Drop".

Índice

Bibiliografía

- SGBD Maria DB. https://mariadb.org/
- Apache Maven. https://maven.apache.org/
- Biblioteca de clases de RapidMiner Studio 7.0. https://github.com/rapidminer/rapidminer-studio
- Java. https://www.java.com/es/
- Eclipse. https://eclipse.org/
- Git. https://git-scm.com/
- Github. https://github.com/

Bibiliografía

- Javadoc. http://www.oracle.com/technetwork/articles/java/index-jsp-135444.html
- Doxygen. http://www.stack.nl/ dimitri/doxygen/
- JUnit. http://junit.org/junit4/
- Observer Pattern. E. Freeman, E. Freeman. Head First Design Pattern, O´Reilly, 2004.
- Strategy Pattern. E. Freeman, E. Freeman. Head First Design Pattern, O´Reilly, 2004.
- MVC Pattern. E. Freeman, E. Freeman. Head First Design Pattern, O´Reilly, 2004.

Fin

Gracias por su atención.

E-mail: manuelbacallado89@gmail.com