Spark for Learning Analytics

Dropout prediction in MOOCs

by

David Torres Pascual

13 July 2016

Content

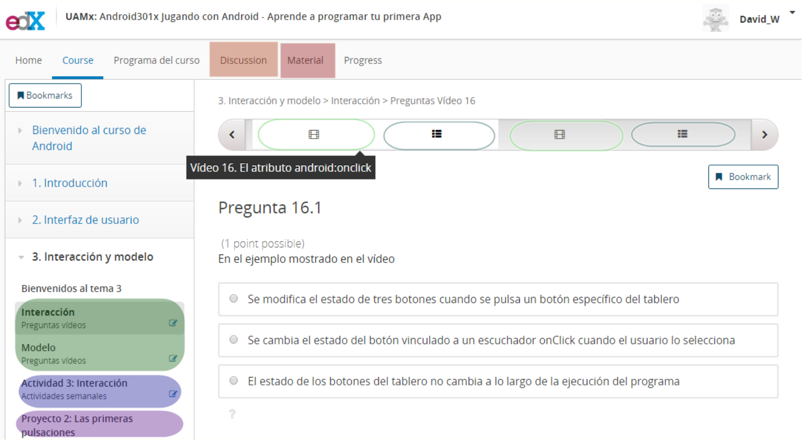
1. Introduction
   - Motivation
   - Goals
2. Methodology
   - Data
   - Tools
   - Model
3. Results
4. Conclusions & future work

Introduction

Motivation (I)

- What is a MOOC?
- What are their main ploblems?

Motivation (II)



Pregunta 16.1

(1 point possible)

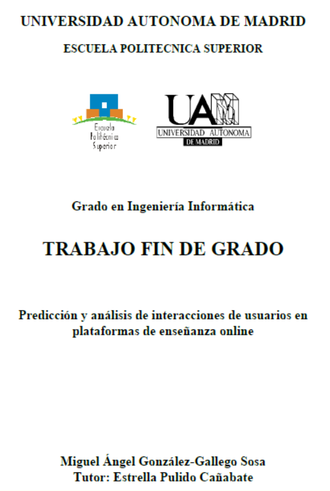
En el ejemplo mostrado en el vídeo

○ Se modifica el estado de tres botones cuando se pulsa un botón específico del tablero

○ Se cambia el estado del botón vinculado a un escuchador onClick cuando el usuario lo selecciona

○ El estado de los botones del tablero no cambia a lo largo de la ejecución del programa

Goals (I)



UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

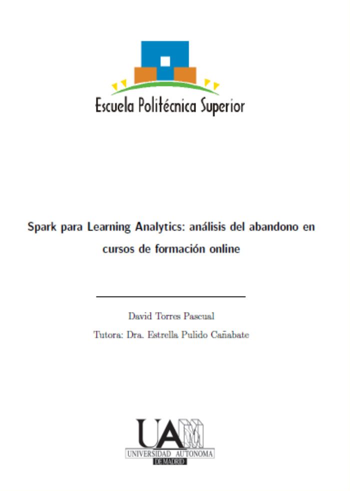
Grado en Ingeniería Informática

TRABAJO FIN DE GRADO

Predicción y análisis de interacciones de usuarios en plataformas de enseñanza online

Miguel Ángel González-Gallego Sosa
Tutor: Estrella Pulido Cañabate

(a) Fase 1



Spark para Learning Analytics: análisis del abandono en cursos de formación online

David Torres Pascual

Tutora: Dra. Estrella Pulido Cañabate

(b) Fase 2

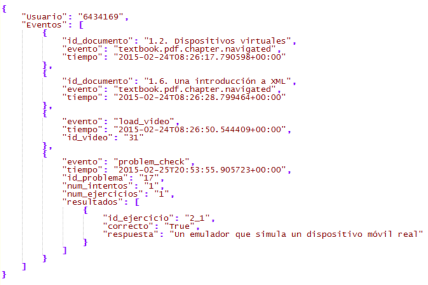
Goals (II)

- Apply Big Data tools and machine learning algorithms to *Learning Analytics* problems.
- Understand MOOC's user behavior.
- Identify dropout in order to help those students with problems with individual educational plans.

Methodology

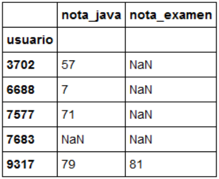
Data (I)

- JSON file with the events in the platform (a).
- Structured tables (b).



```
{
    "Usuario": "6434169",
    "Eventos": [
        {
            "id_documento": "1.2. Dispositivos virtuales",
            "evento": "textbook.pdf.chapter.navigated",
            "tiempo": "2015-02-24T08:26:17.790598+00:00"
        },
        {
            "id_documento": "1.6. Una introducción a XML",
            "evento": "textbook.pdf.chapter.navigated",
            "tiempo": "2015-02-24T08:26:28.799464+00:00"
        },
        {
            "evento": "load_video",
            "tiempo": "2015-02-24T08:26:50.544409+00:00",
            "id_video": "31"
        },
        {
            "evento": "problem_check",
            "tiempo": "2015-02-25T20:53:55.905723+00:00",
            "id_problema": "17",
            "num_intentos": "1",
            "num_ejercicios": "1",
            "resultados": [
                {
                    "id_ejercicio": "2_1",
                    "correcto": "True",
                    "respuesta": "Un emulador que simula un dispositivo móvil real"
                }
            ]
        }
    ]
}
```

(a) *JSON* file



	nota_java	nota_examen
usuario		
3702	57	NaN
6688	7	NaN
7577	71	NaN
7683	NaN	NaN
9317	79	81

(b) *pandas* dataframe

- N° users = 7172 → 2906 (users with ≤ 50 events were removed).

Data (II)

Features:

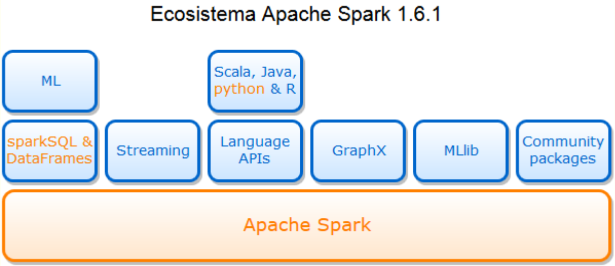
Video	Start a video, finish a video, click play, click pause, seek along video total nº of interactions with each video (37)
Exercise	nº of problems asnwered, nº of attempts and scores; nº of proyect problems and scores; Java exercise scores and exam scores
Forum	nº of threads, nº of comments, nº of replies, nº of threads with problems, nº of words and nº of searches in the forum

Target:

Continue / Dropout

Tools

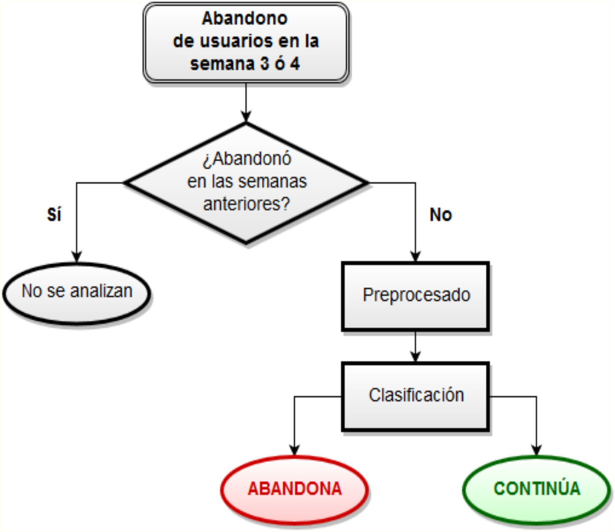
Apache Spark...



Ecosistema Apache Spark 1.6.1

... ¡and Python! (*pandas, numpy, sklearn...*)

Model (I)


Abandono de usuarios en la semana 3 ó 4
¿Abandonó en las semanas anteriores?
Sí
No
No se analizan
Preprocesado
Clasificación
ABANDONA
CONTINÚA

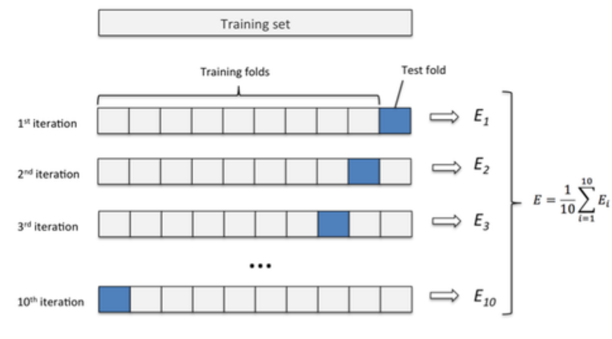
Model (II)

Algorithms for classification:

- Support Vector Classifier
- Random Forest
- Gradient Boosting

Model (III)

Preprocess, train-test split, and feature selection with k-Fold cross-validation.



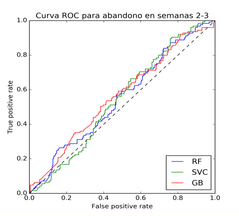
Results

Results I

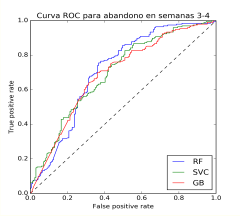
Accuracy (Acc.), Precision (Pr.) y Recall (Rec.)

	Weeks 2-3			Weeks 3-4			Weeks 4-5			Weeks 5-6		
Alg.	Acc.	Pr.	Rec.	Acc.	Pr.	Rec.	Acc.	Pr.	Rec.	Acc.	Pr.	Rec.
SVM	0.55	0.59	0.53	0.65	0.66	0.66	0.75	0.75	0.75	0.74	0.79	0.72
RF	0.55	0.57	0.53	0.70	0.70	0.70	0.79	0.79	0.79	0.76	0.78	0.74
GB	0.33	0.65	0.40	0.56	0.67	0.61	0.66	0.74	0.68	0.62	0.75	0.57

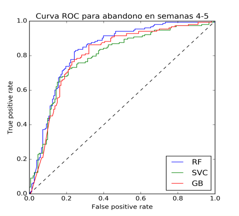
Results II



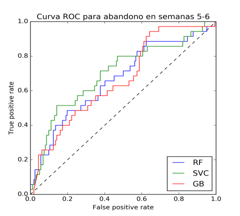
(a) Dropout in week 2 or 3



(b) Dropout in week 3 or 4



(c) Dropout in week 4 or 5



(d) Dropout in week 5 or 6

Conclusions and future work

Conclusions

- We obtain good results with our classification models after the third week of the course.
- We get great results with weighted-classes in the algorithms if we want to focus on dropouts.
- There are some good features for our classification model:
  - score of the problems.
  - involvement in the course project.

Future work

“Perhaps the most important principle for the good algorithm designer is to refuse to be content.”(Aho, Hopcroft & Ullman)

- Improve our results: better classifiers (tuning hyperparameters).
- Look for new goals related to MOOCs:
  - Cluster analysis.
  - Recommender systems.
  - Predict results of the final exam.

Thanks!

Comments, questions?