Predicción del Riesgo de Repetición para Alumnos de Primaria

Recuperación de Información y Recomendaciones en la Web
Facultad de Ingeniería, Universidad de la República

Noviembre de 2017

Grupo 16

Stephanie Casas Oscar Montañés Sebastián Herrera Danilo Amaral

Introducción

Contexto

GURI, es un sistema de información web, que se encuentra enmarcado dentro de las políticas del gobierno electrónico y las políticas educativas del Consejo de Educación Inicial y Primaria(CEIP, 2010-2017) órgano de la Administración Nacional de Educación Pública a cargo de impartir la Educación Inicial y Primaria del País.

Permite tener una base de datos actualizada de docentes, no docentes y alumnos y unificar las gestiones a nivel nacional. La sistematización de datos realizada por GURI propicia la mejora de la calidad de las respuestas al instante y en solicitudes específicas del propio organismo y de otros, al tener información en tiempo real, favorece la toma de decisiones de manera oportuna.

Este sistema otorga beneficios notorios para el país, para el organismo y para los funcionarios docentes, no docentes, alumnos y padres.

A partir de estos datos se plantea poder clasificar si un alumno se encuentra en riesgo de repetición, en función de variables tales como la escuela, el nivel socio cultural, el barrio, la edad, la nota, el juicio y las faltas de los primeros carné.

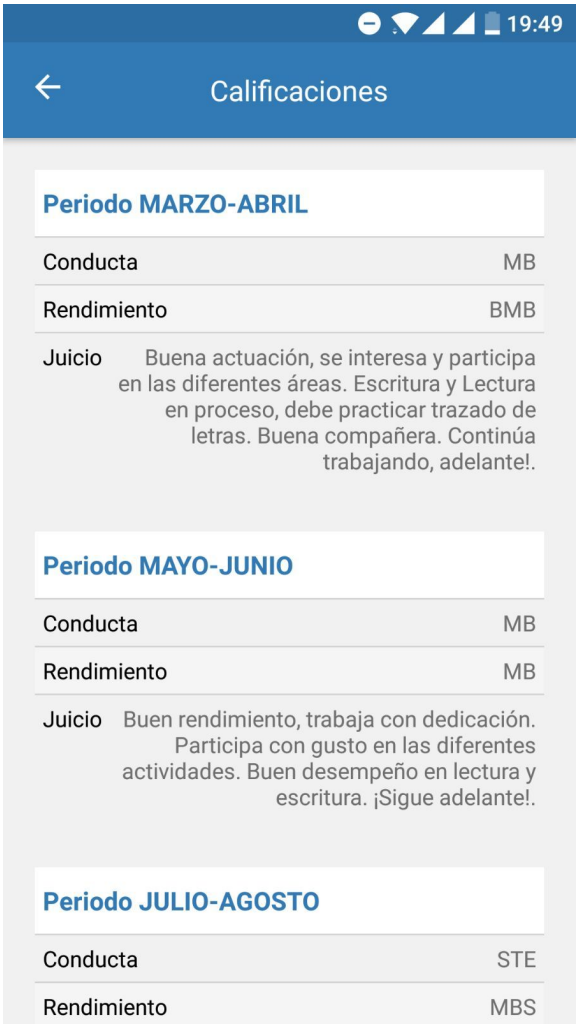
Objetivo

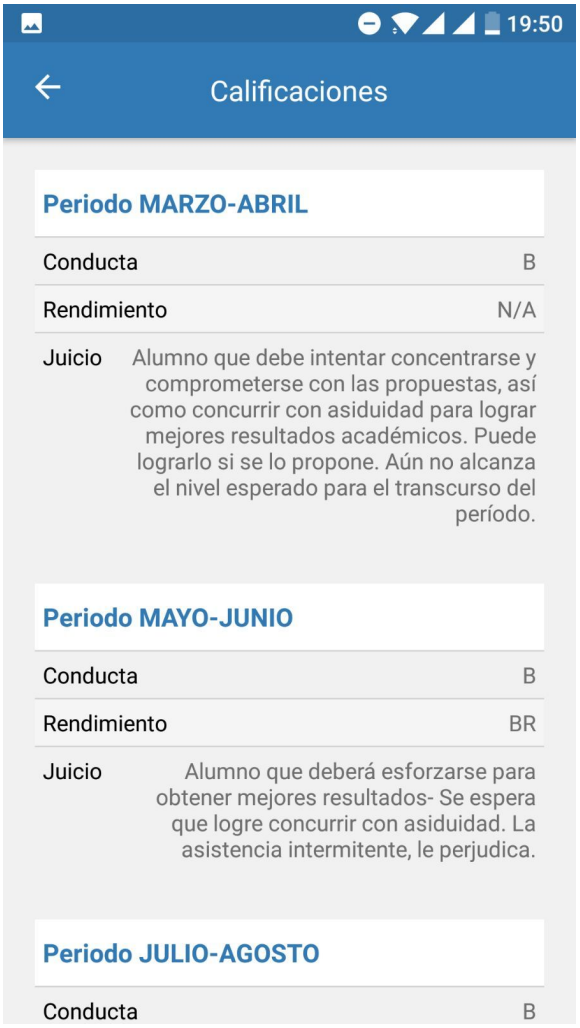
Estudios estadísticos realizados sobre los datos de los alumnos de primaria demuestran que existe correlación entre el nivel sociocultural, la cantidad de faltas, la maduración del niño (edad) y el rendimiento. Este último es difícil de medir dado que la nota carece de distintas dimensiones cognitivas y del aprendizaje.

El objetivo entonces será poder anticipar el resultado de la promoción del alumno en base a los diferentes datos que se tengan, incluyendo la información que se pueda extraer de los juicios de los carné de la vida inicial de un niño y su primer carné en primer año.

Motivación

Para visualizar el trabajo realizado, consideremos las siguientes dos imágenes tomadas desde la aplicación mobile de GURI lanzada en Noviembre de este año

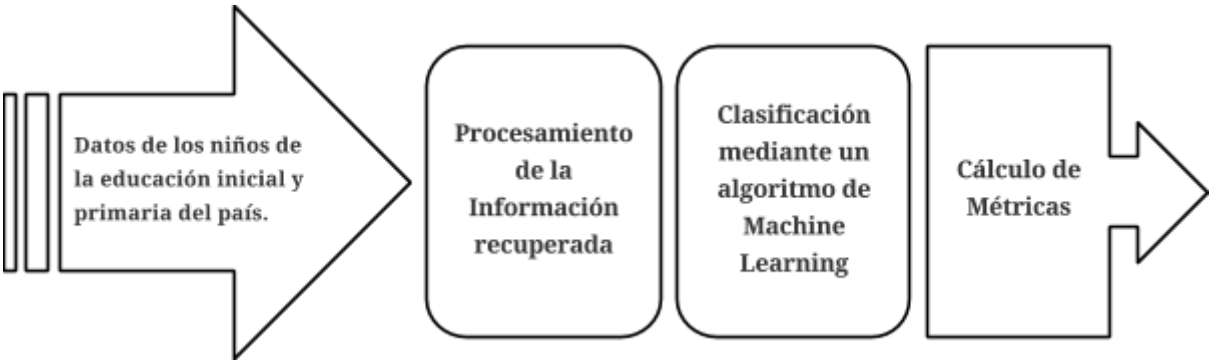




El centro de nuestro análisis se basa en si existe alguna correlación entre la repetición en primero y los juicios de las calificaciones..

Como se puede observar en la figura, tendríamos juicios con tendencia negativa o juicios con tendencia positiva, cada uno de los cuales utiliza un vocabulario que corresponde (como es de esperar) la actividad docente.

La idea es tomar ese vocabulario y representarlo de manera que un algoritmo de aprendizaje automático pueda clasificar si un alumno se encuentra en riesgo de repetición como ilustra la siguiente figura.

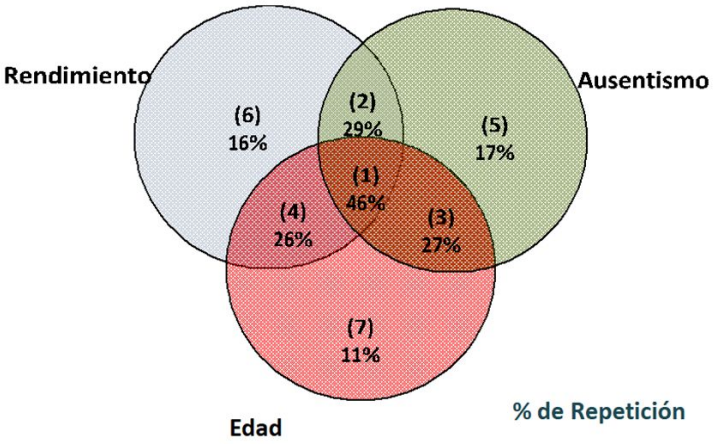


Análisis

En función de la estadística que el CEIP cuenta, dado que el mayor porcentaje de repetición de los alumnos de la enseñanza primaria se da en el primer año, que a su vez existen estudios que la extraedad es un problema importante en la vida estudantil de un niño se realizará el estudio solamente sobre alumnos de dicho año.

En base a los datos de GURI el Monitor Educativo de Primaria del 2015 arrojo posibles resultados de alerta temprana o riesgos respecto al niño, los mismos son:

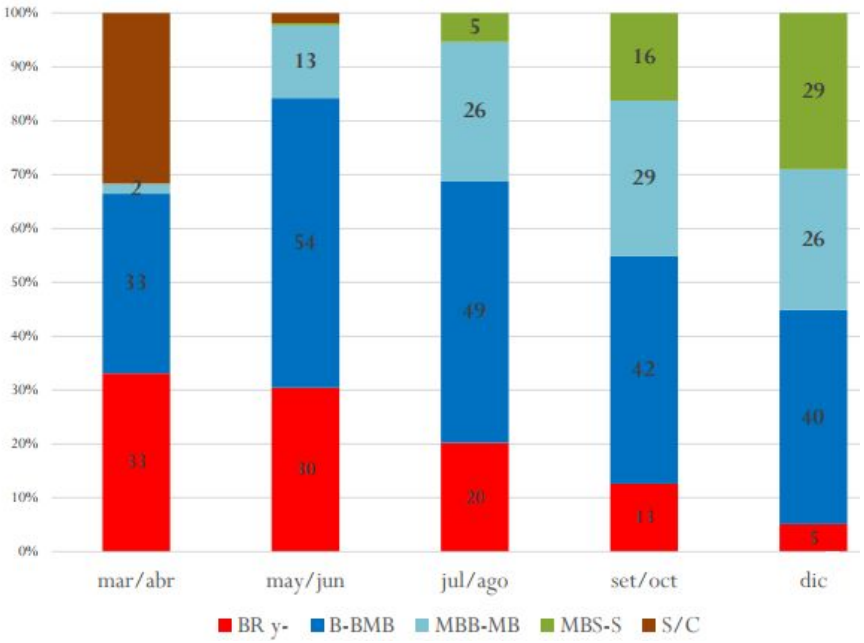
- Maduración del niño
- Rendimiento del niño en su primer carné de primer año
- Inasistencias en inicial
- Inasistencias en primer año



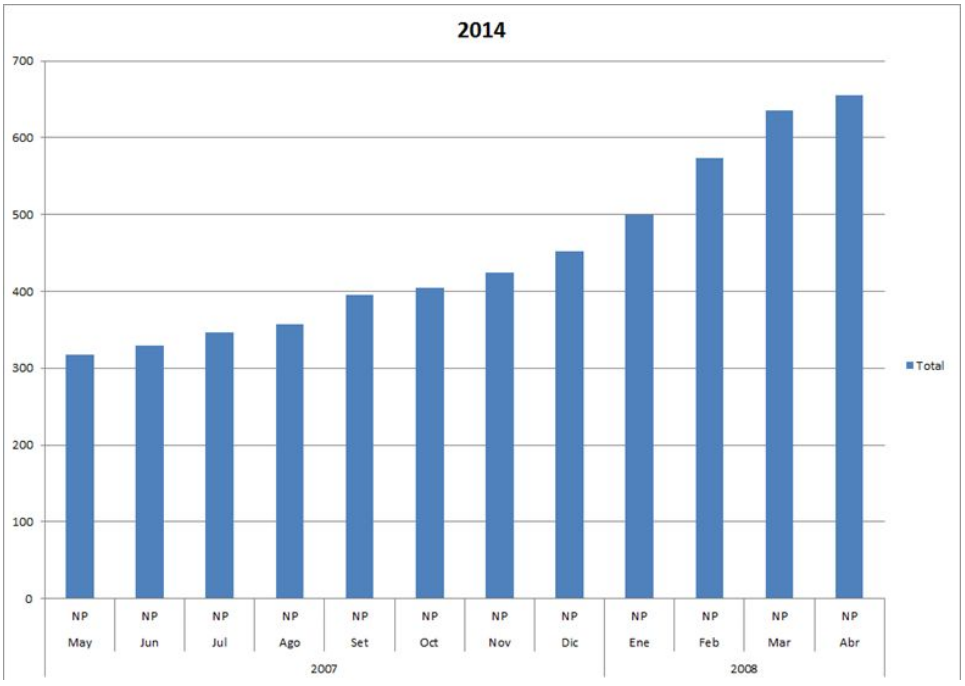
Porcentaje de repetición en 1er año según riesgos

Estos riesgos son acumulativos en el ciclo escolar y son independientes entre sí, sin embargo juntos pueden ser perjudiciales para el aprendizaje y en consecuencia afectar a la repetición hasta en casi un 50% como muestra la figura.

Visto que es una tarea comprometedora para el docente calificar al niño en su primer carné gran parte de los docentes no ingresan una calificación tabulada sino que ingresan un juicio en texto plano.



Evolución de las calificaciones en el correr del año lectivo



Notas según maduración del alumno

Es de interés de este proyecto verificar si existe una correlación entre los juicios de los maestros y las variables anteriormente dichas.

Herramientas

Weka (Waikato Environment for Knowledge Analysis)

Weka [1] fue desarrollado en la Universidad de Waikato. Es una herramienta, implementada en Java, que permite la experimentación de análisis de datos mediante la aplicación, análisis y evaluación de las técnicas más relevantes de análisis de datos, principalmente las provenientes del aprendizaje automático, sobre cualquier conjunto de datos del usuario.

Contiene una colección de herramientas de visualización y algoritmos para análisis de datos y modelado predictivo, unidos a una interfaz gráfica de usuario para acceder fácilmente a sus funcionalidades.

Utiliza los métodos de la inteligencia artificial, aprendizaje automático, estadística y sistemas de bases de datos.
El objetivo general consiste en extraer información de un conjunto de datos y hacer predicciones automáticas que ayudan en la toma de decisiones.

El panel Preprocess dispone de opciones para importar datos de una base de datos, de un fichero CSV, entre otros, y para pre-procesar estos datos utilizando los denominados algoritmos de filtrado.
Weka permite realizar manipulaciones sobre los datos aplicando filtros. Se pueden aplicar en dos niveles: atributos e instancias. Además las operaciones de filtrado pueden aplicarse en cascada, de forma que la entrada de cada filtro es la salida de haber aplicado el anterior filtro. El resultado de estos filtros nos va a servir de ayuda para el resto de aplicaciones de la herramienta. Estos filtros se pueden utilizar para transformar los datos (por ejemplo convirtiendo datos numéricos en valores discretos) y para eliminar registros o atributos según ciertos criterios previamente especificados.

El panel **Classify** permite al usuario aplicar algoritmos de clasificación estadística y análisis de regresión a los conjuntos de datos resultantes. También permite estimar la exactitud del modelo predictivo resultante. Tiene utilidades para visualizar el propio modelo, en aquellos casos en que esto sea posible, como por ejemplo un árbol de decisión.

El panel **Associate** proporciona acceso a las reglas de asociación aprendidas que intentan identificar todas las interrelaciones importantes entre los atributos de los datos.

El panel **Cluster** da acceso a los algoritmos de clustering o agrupamiento de Weka.

El panel Selected attributes proporciona algoritmos para identificar los atributos más predictivos en un conjunto de datos.

El panel Visualize muestra una matriz de puntos dispersos (scatterplot) donde cada punto individual puede seleccionarse y agrandarse para ser analizados en detalle usando varios operadores de selección.

Análisis de datos

Pre-procesamiento

A partir de los datos recabados, en una instancia previa a poder aplicar el procesamiento de los datos por el algoritmo de aprendizaje automático se debe hacer un pre-procesamiento de los datos por los siguientes motivos:

1. Se deben transformar los textos de los juicios de alguna manera que los haga procesables.
2. Se deben transformar los atributos numéricos que puedan tomar una gran cantidad de valores posibles.
3. Eliminar atributos que no aporten al problema de clasificación, por ejemplo el identificador del estudiante.

Pre-procesamiento de juicios

Los juicios van a ser representados mediante vectores de palabras (bolsa de palabras). El modelo bolsa de palabras (del inglés, Bag of Words) es un método que se utiliza en el procesado del lenguaje para representar documentos ignorando el orden de las palabras. Con este modelo podemos tener una representación de cada documento, en función de las palabras que este contiene.

Veamos un ejemplo simplificado del procedimiento [2].

Dados dos mensajes msg_a y msg_b:

msg_a: ***"Hola, cómo estás hoy? Quería saber si te interesa comprar mi producto. Mi número es 4241421"***

msg_b: ***"Hola, te queria avisar que no tengo noticias todavía. Te llamo, avisa"***

1. Se extraen todas las palabras presentes en ambos mensajes.
2. Se eliminan los símbolos "[?\\.!,]"
3. Se pasan todas las palabras a minúsculas.
4. Se genera el conjunto de palabras presentes en cada mensaje que consta de un total de 23 palabras.

```
##         bolsa_de_palabras
##  [1,] "hola"
##  [2,] "como"
##  [3,] "estas"
##  [4,] "hoy"
##  [5,] "queria"
```

```
##  [6,] "saber"
##  [7,] "si"
##  [8,] "te"
##  [9,] "interesa"
## [10,] "comprar"
## [11,] "mi"
## [12,] "producto"
## [13,] "numero"
## [14,] "es"
## [15,] "4241421"
## [16,] "avisar"
## [17,] "que"
## [18,] "no"
## [19,] "tengo"
## [20,] "noticias"
## [21,] "todavia"
## [22,] "llamo"
## [23,] "avisa"
```

Luego, cada mensaje puede ser representado utilizando esta bolsa de palabras. En donde por cada mensaje se construye un vector que contabiliza el número de veces que una palabra está presente en dicho mensaje. En este caso primero creamos una matriz de 2x23, donde cada columna está asociada a una palabra y cada fila está asociada a cada mensaje.
Luego en cada fila asignamos el vector que contiene la frecuencia absoluta de cada la palabra para obtener una matriz de palabras.

	hola	como	estas	hoy	queria	saber	si	te	interesa	comprar	mi
msg_1	1	1	1	1	1	1	1	1	1	1	2
msg_2	1	0	0	0	1	0	0	2	0	0	0

	producto	numero	es	4241421	avisar	que	no	tengo	noticias	todavia	llamo	avisa
msg_1	1	1	1	1	0	0	0	0	0	0	0	0
msg_2	0	0	0	0	1	1	1	1	1	1	1	1

De esta manera vemos que en este caso, cada mensaje se encuentra representado en la matriz de palabras a partir de la distribución de frecuencia de las palabras de la bolsa de palabras.

Pre-procesamiento de atributos numéricos

Algunos de los algoritmos que se utilizan para crear modelos de minería de datos requieren tipos de contenido específicos para poder funcionar correctamente. En estos casos, se puede discretizar los datos en las columnas de modo que pueda utilizar los algoritmos para producir un modelo de minería de datos.
La discretización es el proceso mediante el cual los valores se incluyen en depósitos para que haya un número limitado de estados posibles. Los depósitos se tratan como si fueran valores ordenados y discretos.

Se le aplicó discretización (para convertir de números a intervalos) a los siguientes datos:
La edad al 30 de abril en meses del alumno.
La cantidad de faltas del alumno en primer año.
La cantidad de faltas del alumno en nivel inicial 5.

Eliminación atributos irrelevantes

- Número de escuela
- Zona
- Área
- Nivel
- aluid
- perfchnac
- Nota Promoción
- Faltas 2013
- Faltas 2014

Algoritmos de aprendizaje automático

Los algoritmos de aprendizaje automático obtienen buenos resultados o no según el tipo de problema que se intente resolver y de los datos que se dispongan. Es por esto que de forma de poder obtener los mejores resultados posibles se realizarán varias ejecuciones con diferentes algoritmos de aprendizaje.

Para el problema en cuestión (la clasificación del riesgo de repetición de alumnos de primaria) se utilizarán los siguientes algoritmos:

- Naive Bayes
- K-NN
- Árbol de decisión

Naive Bayes

Este algoritmo es un clasificador probabilístico basado en el teorema de Bayes con algunas hipótesis simplificadoras (por eso “naive” que significa ingenuo) que ha demostrado muy buenos resultados. [3]

K-NN (K-Nearest Neighbour)

Como lo dice su nombre, este algoritmo se basa en clasificar una instancia según la clasificación de sus K vecinos más cercanos o similares. Por lo general se eligen valores impares y pequeños como 1, 3 o 5, de forma que no pueda haber empate. [4]

Árbol de decisión

Este algoritmo primero genera un modelo de clasificación en forma de árbol, donde los nodos son los atributos de las instancias, las ramas son los valores posibles de cada una, y las hojas se etiquetan con la clasificación más probable para los valores desde la raíz a la hoja. [5]

Resultados

A continuación se presentan los resultados obtenidos de la ejecución de los distintos algoritmos de aprendizaje sobre el conjunto de datos con las siguientes características:

- **Cantidad de instancias: 2370**
- **Cantidad de promovidos: 1382 (58%)**
- **Cantidad de repetidores: 988 (42%)**

Para la clasificación se realizó un particionamiento en 70% de las instancias para entrenar, y 30% de las instancias para validar, es decir:

- **Cantidad de instancias para entrenar: 1659**
- **Cantidad de instancias para validar: 711**

Naive Bayes

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Prom.	0.764	0.282	0.783	0.764	0.773	0.480	0.815	0.841
Rep.	0.718	0.236	0.695	0.718	0.706	0.480	0.815	0.767
Total	0.744	0.262	0.745	0.744	0.744	0.480	0.815	0.809

Matriz de confusión:

PRO	REP	← clasificados como
310	96	**PRO**
86	219	**REP**

1-NN

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Prom.	0.700	0.407	0.696	0.700	0.698	0.293	0.663	0.685
Rep.	0.593	0.300	0.597	0.593	0.595	0.293	0.663	0.548
Total	0.654	0.361	0.654	0.654	0.654	0.293	0.663	0.626

Matriz de confusión:

PRO	REP	← clasificados como
284	122	**PRO**
124	181	**REP**

Árbol de decisión

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Prom.	0.677	0.407	0.689	0.677	0.683	0.270	0.608	0.629
Rep.	0.593	0.323	0.580	0.593	0.587	0.270	0.608	0.520
Total	0.641	0.371	0.642	0.641	0.642	0.270	0.608	0.852

Matriz de confusión:

PRO	REP	← clasificados como
275	131	**PRO**
124	181	**REP**

Mejor resultado

A partir de los resultados obtenidos se determinó que el mejor algoritmo de aprendizaje fue la de Naive Bayes que obtuvo:

- **75% de precisión**
- **74% de recall**
- **76% de verdaderos promovidos**
- **72% de verdaderos repetidores**
- **24% de repetidores clasificados como promovidos**
- **28% de promovidos clasificados como repetidores**

Conclusiones

Como bien se planteó en el comienzo de este estudio, el objetivo consistía en poder predecir el riesgo de repetición de alumnos de primer año, a partir de un conjunto de datos entre los cuales se encontraban los juicios de sus calificaciones, y además poder determinar si existen correlaciones entre los atributos de los datos y el resultado de promoción.

Si bien no se han encontrado correlaciones de forma directa entre los datos y el resultado de promoción, se logró implementar un sistema de aprendizaje automático que mediante un conjunto relativamente pequeño de alumnos logra un acierto del resultado en 3 de cada 4 de ellos.

Las causas a las que atribuímos no haber logrado un mayor acierto (cercano al 90% quizás) son las siguientes:

- El conjunto de datos era pequeño respecto a las necesidades de algunos algoritmos de aprendizaje. Esto en parte se debe a que el porcentaje real de repetición en primer año es de aproximadamente 12%, y de forma de tener un conjunto de entrenamiento con cantidad equilibrada de clasificaciones, entonces se tuvo que reducir el conjunto original.
- La incapacidad de ejecución y desarrollo del sistema en un hardware potente también restringía la cantidad de datos con las que se podía entrenar.
- La herramienta utilizada (Weka) no contaba con variantes para procesar los juicios de los alumnos. Alguna posibilidad hubiera sido realizar un procesamiento del lenguaje natural de los mismos, de forma de poder determinar palabras clave o analizar su “sentimiento”, pero esto no era compatible con los juicios en castellano.

Un sistema de estas características tiene el potencial de detectar de forma temprana a aquellos alumnos con problemas en los aprendizajes, y de esta forma poder brindarles apoyo mediante los recursos disponibles. De esta forma se lograría el objetivo final de realizar un aporte a la educación del país.

Referencias

[1] Weka https://www.cs.waikato.ac.nz/ml/weka/ (accedido en noviembre de 2017)

[2] Modelo Bolsa de Palabras. http://rstudio-pubs-static.s3.amazonaws.com/268824_161580c9cae441cf85adb95122ee659e.html (accedido en noviembre de 2017)

[3] “Clasificadores Bayesianos”. http://www.sc.ehu.es/ccwbayes/docencia/mmcc/docs/t6bayesianos.pdf . Departamento de Ciencias de la Computación e Inteligencia Artificial Universidad del País Vasco–Euskal Herriko Unibertsitatea (accedido en noviembre de 2017)

[4] “Clasificadores K-NN”. http://www.sc.ehu.es/ccwbayes/docencia/mmcc/docs/t6bayesianos.pdf . Departamento de Ciencias de la Computación e Inteligencia Artificial Universidad del País Vasco–Euskal Herriko Unibertsitatea (accedido en noviembre de 2017)

[5] “Modelos clasificadores”. http://ocw.uv.es/ingenieria-y-arquitectura/2/classificacio.pdf . Curso Procesado y Análisis de Datos Ambientales de la Universidad de Valencia (accedido en noviembre de 2017)

[6] Monitor Educativo de Primaria http://www.anep.edu.uy/portalmonitor/servlet/acercade . Portal ANEP (accedido en noviembre de 2017)