APRENDIZAJE BASADO EN RETOS PARA LA BIOLOGÍA COMPUTACIONAL Y LA CIENCIA DE DATOS

Emilio Serrano[1], Juan Carlos García, Martin Molina, Daniel Manrique y Javier Bajo

Departamento de Inteligencia Artificial
Universidad Politécnica de Madrid
{emilioserra, mmolina, dmanrique, jbajo}@fi.upm.es,
juancarlos.garcia.torrecilla@alumnos.upm.es

Resumen. *El Aprendizaje Basado en Retos (ABR) es un método docente con sus orígenes en la compañía Apple, Inc. que está ganando creciente popularidad en universidades y empresas alrededor del mundo. ABR ofrece un marco de trabajo efectivo y eficiente para aprender cómo abordar retos y cómo crear soluciones sostenibles. Por otro lado, la Ciencia de datos (DS) es un campo interdisciplinar que se encarga de la extracción del conocimiento a partir de datos y que ha revolucionado áreas tan dispares como los negocios, la educación, la sanidad, o la biología. Este proyecto contribuye con: (1) una adaptación del método del Aprendizaje Basado en Retos para la docencia de la Ciencia de Datos; (2) el diseño de quince retos específicos en el dominio de la Biología Computacional, i.e. el modelado y análisis de datos biológicos mediante tecnologías computacionales; y, (3) la revisión de distintas alternativas software para la realización de retos.*

Palabras clave: Aprendizaje Activo, Aprendizaje Basado en Retos (ABR), Big Data, Lengua inglesa, Máster

1. Introducción

La *Ciencia de Datos* o Data Science (DS) es un campo interdisciplinar que se encarga de la extracción del conocimiento a partir de datos en bruto. Esta ciencia se ha convertido en una revolución que ha cambiado la forma de hacer negocios, la sanidad, la política, la educación y la biología. En este último campo, la DS tiene un rol fundamental en la *Biología Computacional*, i.e. el modelado y análisis de datos biológicos mediante tecnologías computacionales.

El *Aprendizaje Basado en Retos* (ABR) es un método docente con sus orígenes en la compañía Apple, Inc. [1] que está ganando creciente popularidad en universidades y empresas alrededor del mundo. ABR ofrece un marco de trabajo efectivo y eficiente para aprender cómo abordar retos y cómo crear soluciones sostenibles [9]. En este proyecto de innovación educativa se explora el Aprendizaje Basado en Retos para la docencia de la Ciencia de Datos y más concretamente en el dominio de la Biología Computacional.

La Ciencia de Datos y la Biología Computacional están en una posición privilegiada respecto a otras ramas del conocimiento para articular su aprendizaje mediante retos. De esta manera, la plataforma Kaggle [2] libera periódicamente una serie de competiciones sobre problemas reales como “Predicting a Biological Response” [4]; que ofreció 20.000 dólares al mejor modelo predictivo que enlazase una respuesta biológica de moléculas con sus propiedades químicas. Estas competiciones públicas tienen el potencial de involucrar activamente al estudiante en

[1] Coordinador del PIE.

una situación problemática real, significativa, y relacionada con su entorno; incluyendo un marco de trabajo para la implementación de una solución para el reto.

Este proyecto contribuye con: (1) una adaptación del método del Aprendizaje Basado en Retos para la docencia de la Ciencia de Datos; (2) el diseño de quince retos específicos en el dominio de la Biología Computacional, así como la preparación de conjuntos de datos asociados a estos retos para ser cargados directamente en una herramienta de DS; y, finalmente, (3) la revisión de distintas alternativas software para articular un curso de aprendizaje basado en retos, incluyendo plataformas de gestión de aprendizaje, herramientas de DS, y software para la generación de vídeos.

Los resultados del proyecto están destinado a estudiantes del máster en Biología Computacional impartido de la UPM [3]; y más concretamente en la asignatura de "*Knowledge representation and acquisition*". El perfil de estos estudiantes es multidisciplinar, y, en la resolución del reto, tendrán la oportunidad de aplicar a casos reales el conocimiento adquirido en otras asignaturas del máster como "*Statistical Analysis and Data Visualization*" y "*Machine Learning*". Además, los resultados de este proyecto son directamente aplicables a otras asignaturas del máster, de otras titulaciones de postgrado, como el "*Master in Data Science (EIT Digital Master School)*", y también en asignaturas de grado como "*Minería de Datos*" en el "Grado en Ingeniería Informática".

2. El aprendizaje basado en retos para la ciencia de datos

El primer objetivo del proyecto ha sido el desarrollo de métodos para el ABR de Ciencia de Datos. Este objetivo comprende la instanciación de metodologías generales de ABR al campo específico a tratar. Destaca el marco de trabajo para el ABR propuesto por Apple Inc [1] en 2011 y la actualización propuesta por Nichols et al. [9] en 2016. A continuación, se listan los entregables planteados en el ABR y su adaptación en el contexto de un proyecto típico de DS; ya sea usando la estructura clásica de un informe de investigación [8], o metodologías de DS como *Cross-industry standard process for data mining* (CRISP-DM) [6].

Reporte de Big Ideas: Si se usa el repositorio de retos proporcionado en este proyecto, ver sección 3, este apartado no se entregará.

Propuesta de reto: Video, documento o presentación en la que se plantee el reto del proyecto. Aunque se seleccione un reto de repositorio, la importancia e interés de este deberá ser presentada por los estudiantes. Existe cierta correspondencia con la fase "Business understanding" de CRISP-DM donde se describe lo qué se quiere cumplir en el proyecto de DS.

Conjunto de preguntas guía: Reporte con las preguntas guía identificadas en la fase *Guiding Questions*. Categorizadas y ordenadas por prioridad. En DS, se puede usar esta sección para describir el conjunto de datos e incluso su análisis exploratorio que llevará a preguntas a plantear para el reto. De esta manera, se incluiría la fase de "Data understanding" de la metodología CRISP-DM.

Plan de aprendizaje/investigación y timeline: Plan de trabajo dividido en fases para determinar cómo se va a realizar la investigación que dé respuesta a las preguntas guía planteadas. Para DS, se puede dar una planificación para todas las fases de CRISP-DM. De esta manera, se incluiría la metodología de la investigación, es decir, el procedimiento que los experimentos siguen [8].

Reporte de investigación: Documento que recoja el conocimiento adquirido durante la investigación. En DS, se incluirán los resultados y discusión de las fases de "Data preparation" y "Modeling" de CRISP-DM. Note que la fase de "Modeling" termina con una evaluación del modelo aprendido en términos de precisión, pero quedando una fase completa de "Evaluation".

Propuesta de solución/Diseño: Presentación en la que se propone una solución a partir del reporte de investigación. Pude incluir prototipos. Existe un salto entre la obtención de un modelo de DS y el uso de este modelo que se cubre en esta propuesta.

Plan de implementación y evaluación: Plan para implementar la solución planteada, con fechas, costes, responsabilidades de cada miembro, etc. Y plan de evaluación para medir el impacto de la solución, con planes de beta-testing y parámetros a medir. En DS, se incluirían la fase de "Evaluation" de CRISP-DM, que incluye una evaluación más profunda de los modelos que la cubierta en el reporte de investigación. También se incluirán puntos de la fase de "Deployment" de CRISP-DM como el plan de desarrollo, de monitorización, y de mantenimiento.

Resultados de la evaluación: Reporte de los datos recogidos en la evaluación, desarrollando un plan de mejora para la solución si el tiempo lo permite.

Presentaciones finales: Documento, presentación o video describiendo el proceso completo, indicando información del grupo, planteamiento del reto, importancia del mismo, y describiendo la solución, implementación y evaluación. En DS, CRISP-DM describe apartados específicos en la fase final de "Deployment" con una clara correspondencia con este entregable del ABR: "produce final report" y "review project".

Diarios, videos de reflexiones finales, y portfolios: Durante la experiencia los estudiantes pueden mantener diarios sobre su experiencia personal o en grupo. También se pueden dar vídeos de reflexiones finales tras la conclusión del reto. Finalmente, se pueden generar portafolios con los entregables o productos del proyecto. La sección 4 describe las recomendaciones de entregables y herramientas en un curso de DS.

3. Retos para la ciencia de datos en biología computacional

En el caso de la DS, si bien se puede dar libertad al estudiante para que elija su propio reto, encontrar datos asociados a este podría llevar años. Incluso si se encuentra un reto de interés en Kaggle [2] con datos asociados, el procesamiento de estos hasta cargarlos a una herramienta concreta de DS para realizar tareas más avanzadas puede ser muy tedioso para los estudiantes. Por ello, dejando abiertas estas posibilidades, se ha elaborado un repositorio de retos para la DS en el campo de la biología computacional.

El repositorio contiene 15 restos distintos y para cada uno se incluye: un id, título, temática, número de variables, número de casos o instancias, si contiene valores desconocidos, el año de generación, una descripción del reto base, otros retos a plantear sobre esos datos, la presencia del conjunto de datos en la literatura con referencias, y una pequeña descripción de los datos.

Algunos de los retos seleccionados incluyen: determinar qué proteínas están relacionadas con el síndrome de Down en ratones, diagnosticar enfermedades de la soja, estudiar los antecedentes genéticos que pueden causar diabetes, o identificar los micro ARNs (ácido ribonucleico) relacionados con el cáncer de cuello uterino.

Además, los datos de cada reto han sido convertidos al formato de una herramienta para DS, Weka [5]. Esta herramienta ha sido seleccionada para la asignatura por ser muy intuitiva para personas sin conocimientos de programación.

4. Revisión de software para el aprendizaje basado en retos

Este objetivo del proyecto comprende el análisis y documentación de herramientas disponibles para dar soporte al ABR en la asignatura concreta. Respecto a herramientas de DS, como se explicó se ha seleccionado Weka [5] por su usabilidad

si bien hay otras alternativas con una GUI intuitiva como RapidMiner. Kaggle [2] también dispone de un entorno de trabajo, *Kaggle Kernels*, pero requiere conocimientos de programación considerables.

Para plataformas de enseñanza o gestores de contenido de aprendizaje se han revisado 5 alternativas, sin encontrar motivos de peso para no usar *Moodle*; que es proporcionado y gestionado por el *Gabinete de Tele-Educación* (GATE) – UPM. También se han revisado 6 alternativas para el desarrollo de e-portafolios, habiéndose elegido *Google Sites* por no requerir servidor ni conocimientos de programación. Finalmente, se han considerado 15 herramientas para la generación de vídeos para la presentación final, habiéndonos decantado por *Power Point* para la realización de una presentación narrada donde se pueden usar herramientas como el puntero laser.

Finalmente, propondremos a los estudiantes: el análisis de datos con Weka, la generación de un e-portfolio con Google Sites accesible online y que contenga o enlace a todos los productos descritos en la sección 2, un vídeo a modo de presentación final con Power Point, y un vídeo grabado con smartphone para cada miembro del grupo con las experiencias personales a modo de diario.

5. Conclusiones

La investigación realizada en este proyecto ha contribuido en la metodología del Aprendizaje Basado en Retos (ABR) para el campo de la Ciencia de Datos. También se han diseñado y obtenido material para 15 retos en Biología Computacional. Adicionalmente, se han revisado numerosas herramientas software para articular un curso en el Máster Universitario en Biología Computacional con ABR. Si bien los retos todavía no se han presentado a los estudiantes, se han publicado resultados parciales en una conferencia internacional [10] y se realizará una ponencia invitada en el "XXVI Congreso Internacional sobre Aprendizaje". Además, esperamos una alta satisfacción entre los estudiantes como se consiguió en proyectos anteriores [11]. Consideramos que el ABR es uno de las mejores formas para quitar énfasis de la palabra "datos" de "Ciencia de Datos", de manera que los estudiantes puedan centrarse en la palabra "Ciencia" y en las preguntas científicas que la Ciencia de Datos puede resolver.

REFERENCIAS

[1] Challenge Based Learning. A Classroom Guide. https://goo.gl/vAwsg8, 2011[2].
[2] Kaggle: Academic Machine Learning Competitions. https://inclass.kaggle.com/.
[3] Máster Universitario en Biología Computacional. https://goo.gl/mHvQPF.
[4] Predicting a Biological Response. https://www.kaggle.com/c/bioresponse.
[5] Weka 3: Data Mining Software in Java. https://www.cs.waikato.ac.nz/ml/weka/.
[6] What is the CRISP-DM methodology?. https://goo.gl/pzuVr9.
[7] Edu Trends, Aprendizaje Basado en Retos. Editorial Instituto Tecnológico y de Estudios Superiores de Monterrey. https://goo.gl/dA3ux8, 2016.
[8] M. Carter et al. The Parts of a Laboratory Report. https://goo.gl/ct2Xzm.
[9] M. Nichols, K. Cator, and M. Torres. Challenge Based Learner User Guide. Redwood City, CA: Digital Promise, 2016.
[10] E. Serrano et al. Challenge-based learning in Computational Biology and Data Science. In ICTERI 2018: 14th International Conference on ICT in Education, Research, and Industrial Applications.
[11] E. Serrano et al. Métodos, experiencias, y herramientas para el aprendizaje experiencial de la Ciencia de Datos. https://goo.gl/Yy7XeT.

[2] Todos los sitios webs de las referencias han sido accedidos en octubre de 2018.