Universidad de Oviedo
Universidá d'Uviéu
University of Oviedo

Desarrollo de análisis de datos de los experimentos del *Large Hadron Collider* usando la herramienta REANA y la nube del CERN (PINN-19-A-107)

Convocatoria de los Proyectos de Innovación Docente 2019

José Enrique Palencia Cortezon – enrique.palencia@uniovi.es- Departamento de Física
Javier Cuevas Maestro – fjcuevas@uniovi.es- Departamento de Física
Víctor Rodríguez Bouza – vrbouza@uniovi.es- Departamento de Física

Palabras clave:

Tipo de proyecto

Tipo A (PINN-18-A)	**X**

Tipo B (PINN-18-B)	

En este apartado decir el tipo de proyecto (Tipo A o Tipo B) y únicamente en caso de ser de tipo B, describir las ampliaciones y novedades con respecto a los proyectos anteriores de los cuales es continuación y la referencia al proyecto previo.

Resumen / Abstract

En este proyecto se ha estudiado el uso de herramientas computacionales en la nube para la docencia de física experimental de altas energías. La computación es parte fundamental de este campo de la ciencia y permite enseñar competencias transversales informáticas al estudiantado. En el proyecto hemos empleado las herramientas REANA y SWAN, desarrolladas por el CERN, como soporte para que los estudiantes realicen un pequeño proyecto de iniciación al aprendizaje automático (*machine learning*) aplicado a la física. A pesar del reducido número de estudiantes y los condicionantes externos (pandemia de COVID-19), los resultados son en general positivos, e indican que este tipo de herramientas permiten mejorar el aprendizaje de los contenidos de física de la asignatura, así como conocer las herramientas en sí.

1 Contribución del proyecto a la consecución de los objetivos específicos y de los objetivos de la convocatoria

1.1 Objetivos específicos del proyecto conseguidos. Indicar y valorar el grado de consecución de cada uno.

Los objetivos planteados originalmente han sido conseguidos. Cabe resaltar especialmente el objetivo #3, ya que la mayor parte de la documentación utilizada sobre *machine learning* estaba en inglés, y el objetivo #4, porque, aunque se suministraron pautas generales, en todo momento se animó/sugirió a los estudiantes que miraran las cosas ellos y/o que las pensaran.

1.2 Objetivos de la convocatoria a los que se dirigía el proyecto conseguidos. Indicar valoración del grado de consecución.

De nuevo los cuatro objetivos planteados han sido conseguidos. En particular, nos gustaría destacar el objetivo #2, pues los alumnos trabajaron, en la mayor parte de los casos por primera vez, con aprendizaje automático/IA y en el uno que usaron herramientas destinadas a cursos en línea desarrollados por el CERN.

2 Contribución del proyecto al plan estratégico de la Universidad y repercusiones en la docencia. *Para la elaboración de este apartado describir el grado de cumplimiento de los compromisos adquiridos del punto 5 de la solicitud del proyecto.*

2.1 Alineamiento del Proyecto de Innovación Docente con el Plan Estratégico 2018-2022 de la Universidad de Oviedo en materia docente.

De acuerdo con el Plan Estratégico 2018-2022, este proyecto se alinea con los ítems FAE 5 y 6. En el caso del quinto, debido a la implementación de nuevas técnicas para la docencia innovadoras y en el caso del sexto porque las técnicas que se emplean se usan digitalmente y pueden usarse a distancia si es preciso. Esto quedó claramente de manifiesto este curso especialmente, debido a la pandemia por COVID-19: el estudiantado no tuvo problema en usar la herramienta finalmente escogida porque solo requería de un ordenador con navegador web y conexión a Internet.

2.2 Grado de consecución de las repercusiones esperadas del proyecto (en la docencia específica y en el entorno docente)

De nuevo, las repercusiones esperadas (“Prevemos que aumente el interés por los contenidos de la asignatura, al ser presentados a través de esta herramienta, ya que esperamos que facilite el desarrollo de análisis físicos (comparado a cómo se hacen en las prácticas de laboratorio) y que permita usar de forma más accesible y sencilla conceptos más complejos. De esta forma, esperamos que el aprendizaje de contenidos mejore (y, con él, la evaluación curricular del alumnado)”) se han visto satisfactoriamente cumplidas, los alumnos expresaron abiertamente (y también en la encuesta anónima proporcionada) su grado de satisfacción.

3 Memoria del Proyecto

3.1 Marco Teórico del Proyecto

Hoy en día muchas herramientas informáticas empleadas tanto en ciencia y tecnología, como en casi cualquier ámbito profesional, están «deslocalizándose» de nuestros ordenadores personales para ubicarse en la nube. Este movimiento tiene sus ventajas (e inconvenientes), entre las que podemos contar la eliminación de la necesidad de poseer la herramienta individualmente, a cambio de tener una conexión a Internet. Esto presenta oportunidades no solo para el día a día, en el uso productivo de estos instrumentos, sino también para, por ejemplo, la docencia y la divulgación científica.

El CERN (Organización Europea para la Investigación Nuclear) posee una estrategia de ciencia abierta (*open science*), como muestra la publicación de los datos recogidos en sus experimentos (CERN Open Data Portal) [1], el hecho de que todo usuario del CERN se espera que publique los resultados de sus investigaciones en acceso abierto (*CERN Open Access Policy*), [2] o la apuesta por el software abierto (*open source*) que se puede ver ejemplificada en el proyecto MAlt (*Microsoft Alternatives*) [3].

Entre estas iniciativas, el CERN ha desarrollado REANA ([4]), una herramienta FOSS (*Free and Open Source Software*) que permite desarrollar análisis físicos empleando una infraestructura de contenedores de Docker organizados con Kubernetes basada en la nube. La finalidad principal de REANA es la de servir de soporte no solo para realizar y ejecutar análisis de datos en sí, sino para poner estructurarlos y que sean rehechos, lo que se consigue a través de contenedores: estos permiten fácilmente ofrecer el mismo entorno en el que se realizó y programó originalmente el análisis.

A la hora de realizar los análisis, todos deben estructurarse de forma similar, describiendo su flujo de trabajo usando algún lenguaje para tal fin, como por ejemplo *Common Workflow Language* (multidisciplinar, usado también para finalidades de bioinformática o astronomía, [5]) o *Yadage* [6]. Después, se debe realizar el análisis en sí empleando lenguajes comunes (Python, C++). REANA permite también ejecutar análisis con los datos de acceso abierto del CERN: hay varios ejemplos en su página web (uno de ellos: [7]).

La herramienta ofrece la posibilidad de ejecutar sesiones interactivas basadas en cuadernos Jupyter [8], que, debido a su facilidad para combinar código con ejecución inmediata, así como texto explicativo, resultan efectivos para labores pedagógicas y docentes. También son eficientes a la hora de experimentar con nuevos conceptos o ideas de programación (e.g. probar distintas funciones de ajuste para un conjunto de datos).

En este sentido, otra herramienta que también desarrolló el CERN es SWAN (*Service for Web based ANalysis*) [9], que vía cuadernos de Jupyter permite realizar análisis de física a través de

un navegador, y el acceso en el caso de la instalación del CERN a los *clusters* computacionales del mismo.

El acceso a los datos en abierto del CERN permite realizar análisis sobre datos experimentales reales, comprendiendo en el proceso aspectos fundamentales de la física experimental de partículas (etiquetado de jets provenientes de quarks b, estrategias de análisis, cálculo de secciones eficaces, medición de otros observables como el espín, etc.), que casan con el contenido de la asignatura para el que se propone este proyecto. Para ser precisos, contenidos de todos temas según la Guía Docente (del curso 2019-2020), en especial de los temas siete (técnicas de análisis de datos y tratamiento estadístico avanzado de datos) y ocho (herramientas de cálculo distribuido y herramientas GRID).

Además, en la realización de estos análisis pueden emplearse técnicas usadas no solo ya en el ámbito científico o académico, sino en el de la industria (empresa privada), que se han popularizado recientemente y son de actualidad. Por ejemplo, el aprendizaje automático (*machine learning*) o técnicas de manejo de grandes datos *(big data*): directamente relacionadas con varias salidas laborales de los grados (en Física y en Matemáticas, puesto que los bigraduados obtienen ambos).

Nuestra propuesta de proyecto consiste en emplear SWAN y REANA para trabajar con un análisis de física experimental de altas energías a lo largo de la asignatura. Para ello, el estudiantado que voluntariamente quiera participar en el proyecto será guiado por el profesorado de la asignatura en el uso de esta y en el entendimiento del análisis y del trabajo realizado, evaluando a la par la adecuación de la herramienta a la labor docente. Más allá de la novedad de la herramienta, el profesorado se esforzará en presentar atractiva e interesante la iniciativa para el estudiantado.

Aparte de mejorar la adquisición de conocimientos del temario de la asignatura, entendemos que este proyecto mejora la adquisición de sus competencias. Especialmente, las competencias CE8 (capacidad para utilizar herramientas informáticas para resolver y modelar problemas y para presentar sus resultados), CE6 (capacidad para elaborar proyectos de desarrollo tecnológico y de iniciación a la investigación científica), CE4 (capacidad de medida, interpretación y diseño de experiencias en el laboratorio), CT8 (razonamiento crítico), CT9 (aprendizaje autónomo) y CT10 (creatividad). Al estar ambas herramientas desarrollada en inglés, y con ella toda su documentación, se incentiva al alumnado a familiarizarse con la lengua (franca, en ciencia e informática) y con sus tecnicismos.

3.2 Metodología utilizada

3.2.1 Plan de Trabajo desarrollado

Cronológicamente y en el contexto del alumnado, el PINN se desarrolló de la siguiente forma.

- Se presentó a comienzo de curso el PINN, ofreciendo la oportunidad de apuntarse o no al mismo.

- Una vez los detalles técnicos estuvieron listos, el profesorado dio un contexto de las herramientas a usar y de los ejercicios a hacer. Todo se compartió vía campus virtual y también correo electrónico.
- Se iteró con los estudiantes para resolver las problemáticas técnicas que hubo y también las dudas de conceptos / uso (e.g. dudas de la física de los ejercicios, de las herramientas de inteligencia artificial, etc.).
- Finalmente, se hizo una sesión de exposiciones con las charlas de los estudiantes. Se recogieron las presentaciones en el campus virtual en formato PDF.
- Después, se habilitó una encuesta breve sobre el proyecto en el campus virtual.

3.2.2 Descripción de la Metodología

El proyecto se presentó al estudiantado como una alternativa voluntaria en sustitución de parte de los trabajos (presentaciones orales y prácticas de laboratorio) que de normal se realizan durante el curso. Su valoración es la misma que la parte sustituida (un 30% de la nota final, parte de las prácticas y parte de los trabajos) y consistió en una única presentación oral de máximo 10-15 minutos de duración y las preguntas del profesorado a esta (el 80% de la nota) y el seguimiento del trabajo realizado durante el curso (el 20%), evaluado a partir de la consecución de los objetivos marcados por este para cada estudiante y su dificultad. No se entregó ningún documento escrito, aunque sí la presentación (las diapositivas).

El trabajo a realizar por el alumnado consistió en tres ejercicios cortos, bastante abiertos en los que se trabajaba con un mismo análisis de física que empleaba datos abiertos del CERN y en el que se intentaba, empleando SWAN y REANA, profundizar en el mismo. En los ejercicios se agregó una introducción a herramientas de aprendizaje automático o inteligencia artificial para hacer más atractivos los contenidos del PINN. También, teniendo en cuenta el contexto global por la pandemia, se dieron facilidades y flexibilidades a la hora de entregar los ejercicios, así como unos cuadernos de Jupyter base que contenían buena parte de lo necesario para hacer los propios ejercicios.

En cuanto a las herramientas, se decidió priorizar el uso de SWAN al entender que sus ventajas, tras analizar en profundidad SWAN y REANA, eran mucho más interesantes que las características de REANA para el PINN. Ambas dos se usaron de todas formas.

Técnicamente optamos definitivamente por la plataforma openUp2U [10], codesarrollada por el CERN, que nos permitía emplear SWAN en remoto con poco esfuerzo técnico por nuestra parte. A esta conclusión llegamos tras bregar largamente por la posibilidad de soportar esto en primera instancia en una máquina virtual en el propio CERN, y después tratar de hacerlo funcionar en nuestros (de nuestro grupo de investigación) propios servidores de computación. Esto, mezclado con las vicisitudes derivadas de la pandemia, nos hizo finalmente replegarnos a la opción de openUp2U que resultó funcionar de forma sencilla y óptima para nuestros intereses.

3.3 Resultados alcanzados

3.3.1 ***Valoración de indicadores*** *detallando los instrumentos utilizados para recoger la información, se valora la inclusión de tablas o figuras que faciliten la comprensión de lo expuesto. Al menos un indicador se vinculará con el grado de satisfacción del alumnado que participe en el proyecto.*

En la tabla resumen se incluyen los indicadores fijados en la propuesta aprobada del proyecto, aunque aquí comentamos algunos más, fruto de la encuesta anónima que se ofreció a los estudiantes. Todos la contestaron, pero vaya por delante para cualquier conclusión que se extraiga de estos números que la cantidad de alumnos (cinco) no permite aseverar con garantías ninguna conclusión de los observables estadísticos aquí mostrados.

Tabla resumen (a incluir obligatoriamente)

Nº	Indicador	Modo de evaluación	Rangos fijados y obtenidos
1	Porcentaje de aprobados en convocatoria ordinaria	Notas finales de la convocatoria ordinaria (mayo).	<60%: bajo, 60-75%: aceptable, >75%: bueno. **Obtenido: 100%**
2	Media de notas del proyecto (con fines curriculares)	Notas finales de la convocatoria ordinaria (mayo).	< 4: malo, 4-7: aceptable, >7 bueno. **Obtenido: 8.5**
3	Media de notas del proyecto (de la encuesta)	Notas de las encuestas realizadas por los estudiantes involucrados tras la evaluación.	<4: malo, 4-7: aceptable, >7 bueno. **Obtenido: 6.3**

En el caso de los tres observables incluidos en la tabla, vemos que obtenemos tanto para el porcentaje de aprobados como el de las notas asignadas en el proyecto un resultado bueno, y un resultado aceptable para el proyecto según los intervalos marcados previamente.

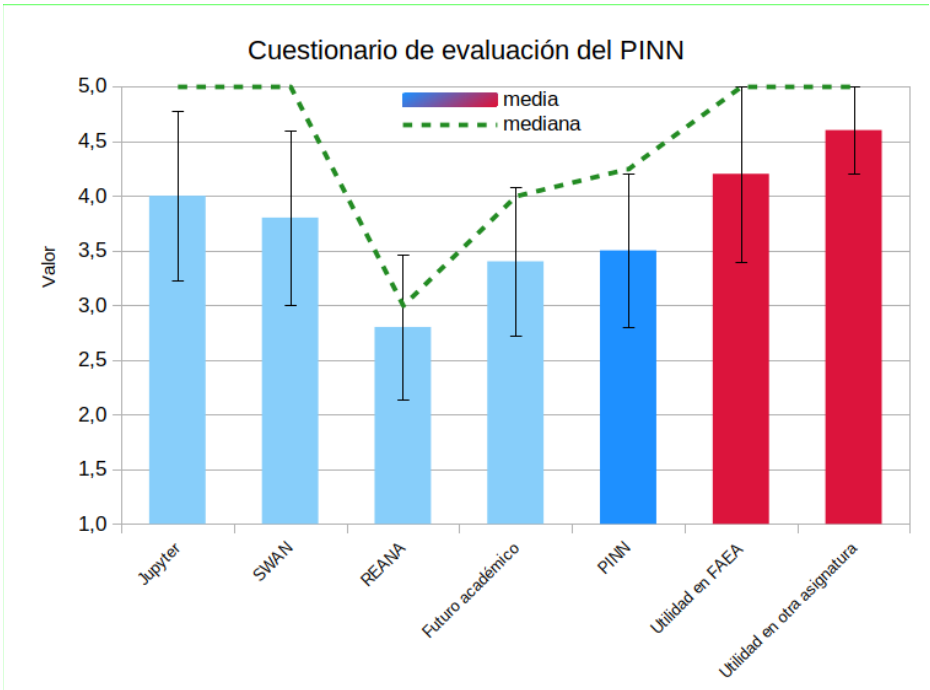
La encuesta nos permite extraer más resultados. En la siguiente tabla se pueden consultar las medias y medianas muestrales y la desviación típica del estimador de las primeras de las respuestas a las distintas preguntas. La valoración del proyecto al completo en la encuesta se extrae del promedio de las medias de las primeras cuatro respuestas.

Pregunta	$\bar{x}$	$\bar{s}/\sqrt{n}$	$M_e(x)$
En una escala del uno al cinco, (siendo uno lo peor y cinco lo mejor), ¿cuán útiles te parecen, globalmente, las siguientes herramientas para hacer prácticas de clase? -> Cuadernos de Jupyter	**4.0**	**0.8**	**5**
-> SWAN (en conjunto)	**3.8**	**0.8**	**5**
-> REANA	**2.8**	**0.7**	**3**
Del uno al cinco, ¿cuán útil piensas que es lo que aprendiste en este proyecto para tu futuro académico o profesional?	**3.4**	**0.7**	**4**
Proyecto (promedio de medias de las anteriores respuestas)	**3.5**	**0.7**	**4**

Pregunta	$\bar{x}$	$\bar{s}/\sqrt{n}$	$M_e(x)$
¿Cuán positivo o útil crees que enseñemos herramientas de aprendizaje automático o inteligencia artificial… -> ... en esta asignatura?	**4.2**	**0.8**	**5**
-> ... en alguna asignatura optativa / de último curso del Grado en Física?	**4.6**	**0.4**	**5**

De estos resultados podemos intentar ver alguna tendencia, aunque como ya se indicó, con tan poca muestra de estudiantes es complicado extraer conclusiones sólidas.

En primer lugar, vemos que en lo que concierne a la valoración del proyecto (las cuatro primeras preguntas) hay valoraciones moderadamente positivas, a excepción del apartado de REANA. Esto parece ir en la línea ya comentada previamente de que el estudiantado también opina que resulta más útil SWAN que REANA como herramienta. También podemos apreciar que sistemáticamente existe una desviación de la media respecto a la mediana, quedando esta última siempre por encima. Esto se puede apreciar gráficamente en la siguiente figura, que representa los mismos números (la desv. típica del estimador de la media se representa a través de las barras de error).



Como se observa al comparar media y mediana, esta desviación proviene de una sola respuesta que es globalmente muy negativa, lo que desbalancea las opiniones. Eliminando este valor extremo de la muestra se puede observar que la valoración del proyecto para la mayoría de los estudiantes es bastante mejor en algunos aspectos.

Finalmente, la inclusión de herramientas de aprendizaje automático o inteligencia artificial en los ejercicios surtieron efecto en los estudiantes, pues valoran positivamente su utilidad bien sea en esta asignatura, y aún más en otra hipotética asignatura de cuarto año de carrera.

Además de las preguntas de valoración (entre uno y cinco), se dejó un espacio para insertar comentarios o sugerencias. De las cinco respuestas, se recibieron cuatro comentarios que aquí se replican.

> **i)** *Me hubiera gustado que la práctica tuviera un apartado más donde después de entrenar el algoritmo de inteligencia artificial, lo usáramos para aplicarlo sobre los datos del análisis. Y con ello, ver como quedaba finalmente el discriminante y el acuerdo datos/MC etc..*
>
> **ii)** *En general el PINN resulta muy útil para pensar un poco menos en la programación y mucho más en la física detrás de los procesos con los que se trabaja. Además los cuadernos no son muy complicados de entender si lees la documentación tranquilamente y son además sencillos de utilizar.*
>
> **iii)** *En mi opinión, no me ha sido útil lo que he aprendido en esta parte de la asignatura (PINN). El problema creo que es que le dedicamos una parte muy pequeña del semestre cuando realmente no estamos familiarizados con ello y necesitamos (al menos yo) más tiempo para asimilar los conceptos. Creo que si se le dedicara más tiempo o incluso como pone en la pregunta 3 se diera en una asignatura optativa, sería muy útil de cara al futuro porque tendríamos más control sobre este tipo de programación que a día de hoy, es muy importante.*
>
> **iv)** *Me parece muy ''refrescante'' que al final de esta carrera, donde a veces casi tienes la noción de que eres un oficinista, redactando informes en todos los cursos y para gran parte de las asignaturas, se haya propuesto sustituir el enésimo informe que haríamos por algo distinto y más liviano. A pesar de que se exigía poco en este PINN, me ha sorprendido que he acabado dedicando bastante tiempo a leer sobre ello solo movido por la propia curiosidad, cosa que llevaba mucho tiempo sin ocurrirme, y he conseguido comprender conceptos fundamentales a la par que importantes simplemente por el afán de saber. Además, al tratarse de una asignatura optativa, que uno la selecciona más bien por el interés que le pueda suscitar, pero también teniendo en cuenta que no quiere complicarse la vida demasiado*

De los comentarios podemos extraer algunas reflexiones. La primera, que varias personas inciden en la duración del proyecto. Esto probablemente se pueda explicar con que, debido a las dificultades técnicas para tener la infraestructura computacional necesaria y a la propia pandemia, el inicio del PINN propiamente dicho se retrasó bastante en el curso, y eso afectó notoriamente al tiempo que se le pudo dedicar (se redujo la cantidad de ejercicios, la complejidad, los requisitos, de forma acorde a esto) por parte del alumnado.

También se puede ver cómo se incide en lo interesante de las herramientas (SWAN, REANA) al ser diferentes de lo que usualmente se emplea en la carrera, y que las herramientas de aprendizaje automático podrían ser ellas mismas constitutivas de una asignatura optativa entera. Además, se valoró positivamente cómo ayudan a explicar los conceptos de la asignatura («resulta muy útil para pensar un poco menos en la programación y mucho más en la física detrás de los procesos»).

3.3.2 ***Observaciones más importantes sobre la experiencia*** *relacionando los resultados con los objetivos del proyecto evitando afirmaciones que no estén fundamentadas en lo realizado, redundancias o reiteraciones.*

- La inclusión de estas herramientas como añadido al aprendizaje de habilidades de programación tradicional incide positivamente en el interés del estudiantado.
- También es una nueva forma de aprehender los conocimientos de la asignatura (la física). Para algunas cuestiones, como entender cómo se traduce la participación de distintos procesos físicos en los resultados experimentales que se observan en los detectores de partículas, muy especialmente.
- Entre las dos herramientas empleadas, no solo pedagógicamente identificamos más positivamente a SWAN frente a las potencialidades de REANA, sino que el alumnado también lo vio así.
- Existe aparente interés por técnicas de aprendizaje automático o inteligencia artificial (herramientas de ciencia o ingeniería de datos, en definitiva; también usadas en física) entre el estudiantado.

3.3.3 ***Información online, publicaciones o materiales en abierto derivados de los resultados del proyecto (se valorará especialmente que se proporcionen los enlaces a los mismos)***

3.4 Conclusiones, discusión y valoración global del proyecto. Se destacarán los puntos fuertes y débiles del proyecto contrastándolas con los resultados de otros estudios referenciados en el apartado 3.1 sin reiterar los datos ya comentados en otros apartados.

Hemos conseguido llevar a cabo con éxito el plan de trabajo establecido, bajo unas condiciones especiales debido a la pandemia por la COVID-19, en el que presentamos a los estudiantes de FAEA nuevas herramientas para aprender los conocimientos de la asignatura, y permitirles ahondar un poco más en los conocimientos transversales de estas. Como se detalló en los puntos 1 y 2 de esta memoria, el proyecto llevó a cabo los objetivos planteados para el mismo, alineados con los del Plan Estratégico de la Universidad de Oviedo 2018-2022 y los de la convocatoria asignada a este plan.

Como punto menos positivo, en el que coincidimos el profesorado y los alumnos, es el poco uso de la herramienta REANA frente a SWAN, esto nos indica un camino para el futuro donde tendremos que reconsiderar el uso de esta herramienta. Por contra, las herramientas usadas

(SWAN esencialmente) para entender el contenido de los ejercicios y también de los conceptos de *machine learning* que les enseñamos han tenido una acogida excelente lo que, dadas sus características no solo tecnológicas sino también conceptuales, son potencialmente de gran utilidad para el futuro. Finalmente, nos gustaría incidir en las buenas valoraciones de las puntuaciones en la encuesta anónima, lo que nos anima a seguir en esta dirección con las modificaciones/correcciones ya mencionadas.

4 Bibliografía

La inclusión de la bibliografía de referencia utilizada para la elaboración del proyecto es obligada. Las citas bibliográficas deberán extraerse de los documentos originales indicando siempre la página inicial y final del trabajo del cual proceden, a excepción de obras completas. No debe incluirse bibliografía no citada en el texto. Su número ha de ser ajustado, y se presentarán alfabéticamente por el apellido primero del autor (agregando el segundo sólo en caso de que el primero sea de uso muy común). Se valorará la correcta citación conforme a normativas estandarizadas tipo APA o similares, también se valorará positivamente que haya referencias no sólo a trabajos nacionales, sino también internacionales.

[1] http://opendata.cern.ch/

[2] https://cds.cern.ch/record/1955574

[3] https://malt.web.cern.ch/malt/

[4] http://www.reanahub.io/

[5] https://www.commonwl.org/

[6] https://github.com/yadage/yadage

[7] https://github.com/reanahub/reana-demo-cms-h4l

[8] https://jupyter.org/

[9] https://swan.web.cern.ch/

[10] https://up2university.eu/open/