

**Caso de estudio. Perspectivas del uso de herramientas de aprendizaje automático y cómputo de alto rendimiento en investigación científica por parte de estudiantes de pregrado en una universidad del Ecuador**

**Autores:**

Miguel Méndez ([mmendez@usfq.edu.ec](mailto:mmendez@usfq.edu.ec))

Samara Oña ([smona@estud.usfq.edu.ec](mailto:smona@estud.usfq.edu.ec))

Sebastián Ayala ([asayala@estud.usfq.edu.ec](mailto:asayala@estud.usfq.edu.ec))

**Institución:**

Universidad San Francisco de Quito

COCIBA

**Área del conocimiento:** Retos y Responsabilidad Social de la Investigación en la Educación

## Resumen

Hay un gran vacío en investigación, innovación y desarrollo en Ecuador. Una de las estrategias a seguir es incrementar la exposición a investigación de los estudiantes de pregrado del sistema nacional de universidades. El uso de aplicaciones digitales en educación está consolidándose en Ecuador. En el país por ejemplo ya es común el uso de ambientes de aprendizaje virtual. Cada vez más, la colaboración interdisciplinaria prueba que es el mejor enfoque para la resolución de problemas de investigación.

El presente estudio pretende evidenciar la oportunidad de utilizar a nivel educativo herramientas del aprendizaje automático y el cómputo del alto rendimiento en diversas áreas del conocimiento, aplicado en la investigación y ejecución de proyectos por parte de estudiantes de pregrado. Para esto se tomó como caso de estudio a estudiantes de pregrado que han desarrollado proyectos en el Grupo de Química Computacional y Teórica (QCT) y el Instituto de Simulación Computacional (ISC) de la Universidad San Francisco de Quito (USFQ).

Se evaluó una muestra de quince estudiantes que durante sus estudios de tercer nivel realizaron actividades de investigación en el grupo QCT-ISC USFQ. Los resultados obtenidos demuestran que los estudiantes percibieron claros beneficios en el uso de herramientas de aprendizaje automático y cómputo de alto rendimiento. Además, consideraron la experiencia como una parte importante en su desarrollo profesional y académico.

**Palabras clave:** aprendizaje automático, cómputo de alto rendimiento, supercomputación, biología computacional, bioinformática, investigación en la educación de tercer nivel

## Abstract

There is a great gap in research, innovation and development in Ecuador. One of the strategies to follow is to increase the exposure to research of undergraduate students of the national university system. The use of digital applications in education is consolidating in Ecuador. In the country, for example, the use of virtual learning environments is common. Increasingly, interdisciplinary collaboration proves that it is the best approach to solving research problems.

The present study aims to demonstrate the opportunity to use at the educational level tools of automatic learning and computation of high performance in various areas of knowledge, applied in the research and execution of projects by undergraduate students. For this, undergraduate students who have developed projects in the Computational and Theoretical Chemistry Group (QCT) and the Computational Simulation Institute (ISC) of the University of San Francisco de Quito (USFQ) were taken as case studies.

We evaluated a sample of fifteen students who during their third level studies carried out research activities in the QCT-ISC USFQ group. The results show that students perceived clear benefits in the use of high-performance computer and automated learning tools. In addition, they considered the experience as an important part in their professional and academic development.

**Keywords:** automatic learning, high performance computation, supercomputing, computational biology, bioinformatics, research in third level education

## **Introducción**

De acuerdo a la *World Intellectual Property Organization*, en el Ecuador durante el período 2001-2010 se concedieron solo 76 patentes. En comparación en Perú, un país que en cuanto a población y ubicación geográfica es aproximadamente comparable al Ecuador, en el mismo período se concedieron 437 patentes.

Lo cual evidencia, que hay un importante vacío en la producción en investigación, innovación y desarrollo. Se ha observado que una de las estrategias para contribuir al desarrollo de la ciencia e innovación es incrementar la investigación a nivel de las universidades (Masfield, 1991; Acs et al, 1992; Etzkowitz, 2014).

Por lo tanto, estrategias que ayuden a que se generen nuevos grupos de investigación, o que más estudiantes se puedan involucrar en grupos ya existentes, contribuyéndose de manera importante a mejorar la situación de la investigación e innovación en Ecuador.

Dado el contexto ecuatoriano donde existe un número limitado de programas de postgrado en investigación. Por ejemplo, una búsqueda en el Sistema Nacional de Información de Educación Superior del Ecuador, arroja que la oferta académica para el régimen 2013 en Quito, modalidad presencial fue de 172 programas registrados de cuarto nivel y 202 programas de tercer nivel (búsqueda en septiembre 2017). De los programas de cuarto nivel solo el 24 % están registrados como Maestrías específicamente de investigación, y tan solo 5 (2.9% de la oferta de cuarto nivel)

doctorados (de 20 doctorados en total) son en las áreas de ciencias e ingenierías. Esto sugiere que la gran mayoría de profesionales formados en Ecuador no reciben formación específicamente enfocada en investigación. Con lo cual, se sugiere, que una alternativa para empezar a fomentar la ciencia e innovación en la academia en Ecuador es una mucho mayor exposición a investigación de los estudiantes de pregrado del sistema nacional de universidades (Hajdarpasic, et al 2015).

Dado que la infraestructura necesaria para realizar investigación de alto impacto puede ser muy costosa especialmente en ramas de ciencias e ingenierías, aquí se presenta un caso de estudio donde las principales herramientas son cómputo de alto rendimiento y métodos de aprendizaje automático cuya aplicación en Ecuador será más factible en considerando los costos de mantenimiento y/o creación de laboratorios e infraestructura de investigación más tradicionales. Adicionalmente, los mismos métodos pueden ser aplicados en áreas fuera de las ciencias e ingenierías, así tanto en áreas sociales, económicas y de la salud solo para mencionar unos pocos ejemplos.

## **Desarrollo**

### **Estado del Arte**

En los últimos años se es testigo de un crecimiento vertiginoso de la información digital, cuyo origen se remonta a los inicios del internet y la consecuente creación de páginas web, bases de datos, redes *sociales* y demás formas de almacenamiento digital, así como el desarrollo de innumerables dispositivos electrónicos, que permiten a una persona estar “conectada” durante cada segundo del día, generando más información durante el proceso, por lo que se piensa que estamos inmersos en la era de los “grandes datos” (*Big Data*) (Chen, Mao, & Liu, 2014).

Estos datos, de libre acceso permiten que se pueda llevar a cabo investigación de alta calidad por parte de estudiantes de pregrado si conocen el funcionamiento de herramientas de análisis de datos. Adicionalmente, la información puede ser procesada y convertida en conocimiento utilizando procesamiento computacional (*i.e.* modelación). El presente estudio, presenta un análisis de caso en que estas herramientas han ayudado a desarrollar varios proyectos de investigación por estudiantes de pregrado.

Dos de las herramientas más versátiles que se pueden hacer uso en investigación aprovechando bases de datos públicas y herramientas computacionales son el aprendizaje automático y modelación utilizando el cómputo de alto rendimiento.

Dado la gran cantidad de datos disponibles es imprescindible la búsqueda de herramientas que permitan manejar tantos datos, entre los que destacan los métodos de aprendizaje automático. Estos métodos pueden aplicarse a múltiples y variadas áreas del conocimiento como la gestión empresarial, el derecho, la psicología, el diseño y la pedagogía entre otras (Walker, 2014).

El aprendizaje automático es un campo de estudio multidisciplinario que emplea conceptos de estadística, ciencias de la computación y áreas afines. De forma general busca que un proceso mejore automáticamente a través de la experiencia en la ejecución de una tarea.

En la rama del aprendizaje automático supervisado, la experiencia implica usar resultados conocidos asociados a una gama de datos (propiedades) con la finalidad de predecir eventos de resultado desconocido en base a la información disponible. La forma en que se logra este tipo de “aprendizaje” es a través de un algoritmo, que consiste en un conjunto de instrucciones que permiten a la computadora resolver un problema determinado (Jordan & Mitchell, 2015). Sin embargo, no hay un solo problema por resolver, lo que hace posible la existencia de muchos algoritmos que cambian dependiendo de la tarea en cuestión, y brinda a este tipo de herramientas un amplio espectro de aplicaciones con un alto grado de eficiencia en su proceso (Domingos, 2012).

Es posible encontrar aplicaciones de aprendizaje automático en cualquier área del conocimiento. Por ejemplo, en gestión empresarial existen diversas aplicaciones adoptadas por las empresas para mejorar su rendimiento y ser competitivas en el mercado, entre estas se tiene a SAP-HANA, que sirve para gestionar bases de datos de la información colectada como transacciones, información de los clientes, trabajadores y otras. Por ejemplo, puede sugerir qué cupones de descuentos darle a un cliente en base a las compras que hace regularmente. Esta herramienta es empleada por empresas como Walmart.

En diversos escenarios, no solo necesitamos analizar datos para encontrar tendencias o aprender de los datos; en investigación hay escenarios nuevos o donde los experimentos podrían ser costosos, demorados o simplemente no viables. En

estos casos, la simulación computacional es una opción sin embargo los recursos computacionales necesarios son elevados. El cómputo de alto rendimiento o supercomputación es el uso de computadoras en la resolución de problemas que involucran grandes cantidades de datos y/o muchos cálculos. Estas máquinas de supercómputo se pueden ver como una colección de computadoras conectadas por una red (clústers) que ejecutan los programas mediante procesamiento paralelo, que se refiere a la división de las instrucciones del programa en distintos procesadores con el objetivo de aminorar el tiempo de la tarea (Verma, Ahuja, & Neogi, 2008).

La ciencia e ingeniería han sido los campos que más se han beneficiado del cómputo de alto rendimiento, pero esta herramienta también se puede implementar en otras áreas del conocimiento como por ejemplo en psicología para probar modelos de procesos cognitivos, entender lo que ocurre en la mente de un individuo y la explicación de su comportamiento (Ark, et. al., 2015). Además, en el derecho y áreas afines se brinda la posibilidad de trabajar con cantidades de datos inabarcables de otro modo, lo que permite obtener la información requerida para cualquier fin de un modo extremadamente rápido y eficiente (Howe, 2012).

El uso de aplicaciones digitales en educación está consolidándose en Ecuador (Ramirez 2006). En Ecuador por ejemplo, ya es común el uso de ambientes de aprendizaje virtual (D2L, Moodle, blackboard) en muchas instituciones de educación superior. También hay un incremento en el uso de ambientes de educación personalizada (Ramos et al, 2014). Sin embargo, aún hay oportunidad de implementar nuevas herramientas en el desarrollo de investigación en la educación de tercer nivel en Ecuador. Por ejemplo, actualmente ya varias universidades ecuatorianas cuentan con sistemas de cómputo de alto rendimiento para investigación (por ejemplo, ESPE, USFQ, UTPL, entre otras) (USFQ, 2017; RED CEDIA, 2017; ESPE, 2017).

El presente estudio evidencia la oportunidad de utilizar a nivel educativo herramientas del aprendizaje automático y el cómputo del alto rendimiento en diversas áreas del conocimiento, aplicado en la investigación y ejecución de proyectos u otras tareas de estudiantes de pregrado. Con la finalidad de utilizar a nivel educativo herramientas del aprendizaje automático. Asimismo, el cómputo del alto rendimiento se analizó como caso de estudio la experiencia en el área de un grupo de investigación en una universidad ecuatoriana, donde la actividad se centró

principalmente en ejecución de proyectos con la participación activa de estudiantes de pregrado.

El grupo de Química Computacional y Teórica (QCT) y el Instituto de simulación computacional (ISC) de la Universidad San Francisco de Quito forma profesionales a nivel de pregrado y posgrado para el estudio computacional y teórico de las propiedades fisicoquímicas de sistemas a nanoescala entre otros.

Uno de los principales pilares de investigación de este grupo es el modelado molecular de sistemas biológicos utilizando herramientas de bioinformática y biología computacional, dentro de las cuales, se destacan aprendizaje automatizado y cómputo de alto rendimiento. Muchos de los proyectos realizados son de carácter interdisciplinario y, también, involucran a otras instituciones universitarias a nivel nacional como Escuela Politécnica Nacional, Escuela Politécnica del Litoral, Escuela Politécnica del Chimborazo, Universidad Central del Ecuador, Universidad Técnica Particular de Loja e instituciones a nivel internacional como Université de Bordeaux y Università degli Studi Torino.

Los proyectos de investigación, son ejemplos que se analizan en el presente estudio. Están principalmente centrados, en el entendimiento o la resolución de problemas en la interface entre las ciencias biológicas y químicas. El uso de este tipo de herramientas está bastante consolidado en estas áreas.

Por ejemplo, el aprendizaje de máquina aplicado a la resolución de problemas biológicos comenzó en los años 60 con el trabajo de Frank Rosenblatt como una propuesta para modelar redes neurales mediante un conocido algoritmo llamado perceptrón, lo cual dio paso, a la rama de la ciencia conocida como inteligencia artificial (Rosenblatt, 1958). En los años posteriores, debido al desarrollo de modelos matemáticos enfocados en los procesos biológicos y el rápido crecimiento del poder de procesamiento computacional el aprendizaje automático se ha vuelto una de las mejores alternativas para analizar grandes y complejas cantidades de información biológica (Clote & Backofen, 2000).

Los sujetos del presente estudio han participado en el desarrollo de investigación dentro del QCT-USFQ en distintas modalidades donde la principal categoría es de “aprendiz de investigación”. Un *Aprendiz de investigación* puede ser definido como un estudiante cuya dedicación a un proyecto incluye tanto su participación durante vacaciones (receso académico de fin de semestre/año académico), así como

participación regular durante todo el año. Esta modalidad está ampliamente establecida en Universidades de artes liberales (cita).

No existió limitación en admisión a un proyecto en cuanto al año de estudio que se encontraban cursando los estudiantes. El profesor sirvió de mentor a un grupo pequeño de estudiantes (2-8) en cualquier momento dado. Los tópicos en los que realizaron sus proyectos de investigación, aunque se promovió la apropiación de los proyectos, en su gran mayoría son proyectos ramificados de proyectos del profesor/mentor. Aunque han existido varios casos en que los mismos estudiantes han propuesto proyectos propios. De varios de los trabajos resultaron publicaciones en revistas internacionales revisadas por pares.

### **Metodología**

Para este estudio se utilizó tanto encuestas como entrevistas personalizadas a los sujetos, con la finalidad de identificar las razones/motivaciones de los estudiantes para participar en los proyectos de investigación, así como los beneficios percibidos por estos estudiantes. Aunque muchas de las preguntas son genéricas enmarcadas en “experiencia de investigación a nivel de pregrado”, se incluyeron preguntas específicas concernientes al uso de las dos herramientas: aprendizaje automático, y uso de cómputo de alto rendimiento. Finalmente, se colectó información sobre ejemplos específicos de proyectos con la finalidad de dar una visión general de la aplicación específica de las herramientas en este grupo de investigación. También se tomaron unos pocos ejemplos de los productos tangibles, producidos por los estudiantes en forma de presentaciones, participación en congresos y publicaciones. En la primera parte del estudio, se realizó una encuesta a una muestra de 15 personas que pertenecen o pertenecieron al QCT o ISC para conocer su experiencia y opinión sobre el uso de herramientas computacionales en el desarrollo de sus proyectos. La muestra contiene tanto jóvenes que aún son estudiantes universitarios, así como jóvenes profesionales graduados en los últimos tres años (al momento del estudio se encontraban trabajando o cursando estudios de postgrado).

Como grupo de control, se realizó una encuesta a una muestra de 15 estudiantes de pregrado que no han estado involucrados con las actividades del grupo del caso en estudio. La muestra control, incluye estudiantes en el mismo grupo de edad (en el rango entre 18-23 años), y de programas académicos a los que usualmente pertenecen los estudiantes que entran a realizar investigación en el grupo de QCT-USFQ/ISC-USFQ.



Para la selección del grupo control se realizó una invitación a participar en el estudio a los estudiantes de dos paralelos/cursos. El primero con estudiantes de diversas carreras en ciencias de la salud y ciencias biológicas, y el segundo un curso ofertado exclusivamente a estudiantes de Ingeniería en procesos biotecnológicos de la USFQ. De 53 estudiantes, 15 respondieron en el tiempo proporcionado para responder a la encuesta (72h). Para la encuesta tenían un link a un servidor de internet donde podían llenar su encuesta de manera anónima.

La encuesta consistió de 10 preguntas de opción múltiple, en las cuales se preguntó al estudiante datos generales, como: carrera, año en el que comenzó a utilizar herramientas computacionales, tipo de proyecto en el que las aplicó, nivel de satisfacción con las mismas y, finalmente, perspectivas futuras sobre la aplicación de este tipo de técnicas en su vida profesional.

La encuesta en el grupo del caso de estudio (estudiantes que participan o han participado en proyectos de investigación en el grupo QCT-USFQ) reveló que el 80% de los estudiantes estudian o estudiaron Ingeniería en Procesos Biotecnológicos. El otro 20% representa a estudiantes de Química, Física y Ecología Marina. A pesar del marcado interés de los estudiantes de biotecnología en el modelado computacional y uso de herramientas de bioinformática, la enseñanza en biotecnología en el país rara vez incluye elementos académicos relacionados a este tema.

La educación en ciencias biológicas tiene en el país un enfoque mayoritariamente experimental. Las clases de modelado matemático, programación y análisis de datos que son propias de otras ingenierías como sistemas o química no forman parte de la malla académica del ingeniero en biotecnología lo cual es una desventaja para estos estudiantes quienes tienen que aprender por su propia cuenta. Por otra parte, esto también puede darse en proyectos en otras áreas de conocimiento especialmente fuera de las ramas de ciencias e ingenierías y sugiere que a pesar de que un estudiante en particular no tenga en su pensum materias específicas para este tipo de investigación (cómputo de alto rendimiento y/o aprendizaje automático), los estudiantes de pregrado con la motivación adecuada están dispuestos a aprender “desde cero”.

En cuanto al año de la carrera en el cual el estudiante comenzó a usar cómputo de alto rendimiento y aprendizaje de máquina, se observa que en el primer caso: los alumnos tuvieron sus aproximaciones iniciales al tema principalmente en los primeros años de estudio. En cambio, se observa que las primeras aproximaciones

del estudiante al uso de técnicas de aprendizaje de máquina se dan a lo largo de todos los años de la carrera y no en un año determinado. Adicionalmente, se les preguntó para qué utilizan estas herramientas a lo cual respondieron en primer lugar para la realización de proyectos de investigación, en segundo lugar los proyectos de titulación y, en tercer lugar, curiosidad personal. Ninguno de los estudiantes escogió la respuesta “Por obligación”.

Los resultados de las preguntas anteriores indican que los encuestados tienen niveles de conocimiento variado sobre las herramientas de cómputo descritas. Además, es importante resaltar que la mayor parte de estudiantes están involucrados en proyectos de investigación que no son requisitos para graduarse. Esto destaca, la curiosidad científica de los estudiantes y sus deseos de involucrarse en el avance de la biología computacional. Otro aspecto interesante, es que algunos de los miembros del grupo de investigación optan por realizar su trabajo de titulación utilizando cómputo de alto rendimiento y/o aprendizaje de máquina; esto implica que los alumnos conocen el valor de estas herramientas y se sienten confiados en utilizarlas para uno de los trabajos más importantes de su carrera.

En un tercer tipo de preguntas, se pidió a los estudiantes que expresen su satisfacción con el uso de cómputo de alto rendimiento y aprendizaje automatizado. Ante lo cual, más de dos tercios de la muestra respondió que estaba altamente satisfecha con el uso de estas herramientas y el resto afirmó estar medianamente satisfecho. Finalmente, se preguntó a los estudiantes sobre la posibilidad de seguir una carrera profesional que involucre proyectos relacionados con estas herramientas computacionales. De forma sorprendente, 14 de los 15 encuestados respondió que estaría dispuesto a hacerlo.

En la muestra control, se obtuvo una respuesta de 8 estudiantes de la carrera de Ing. en procesos biotecnológicos, 3 de odontología, 2 de biología, 1 de nutrición humana y 1 de medicina veterinaria. Un 29% de los estudiantes afirmó que empezó a utilizar herramientas de cómputo de alto rendimiento durante su segundo año de estudios, todos los demás encuestados respondieron no haber utilizado nunca cómputo de alto rendimiento. De los que respondieron, haber utilizado cómputo de alto rendimiento, 1 afirma que lo utilizó por curiosidad personal y, los otros 2 por obligación. En cuanto a la percepción de la utilidad de cómputo de alto rendimiento en ciencias exactas, un 43% afirmó que si es importante y el 57% considera que no lo es. Recuérdese que tenían la opción de no responder (solo 1 estudiante tomo esta

opción). Algo que llama la atención es que un 92 % respondió que le gustaría trabajar en proyectos que involucren uso de cómputo de alto rendimiento, con solo 1 respuesta negativa.

En cuanto al uso de aprendizaje automático, se obtuvieron similares resultados que para cómputo de alto rendimiento, con un 77 % en el grupo control que afirma no haber utilizado estas herramientas. En cuanto a la percepción de si se tratan de herramientas útiles para resolver problemas en ciencias exactas, un 43 % respondió afirmativamente y el restante consideró que no era útil.

Con la encuesta, se pudo tener una mejor valoración sobre la enseñanza de técnicas computacionales (que incluye tanto el uso de cómputo de alto rendimiento como de herramientas de aprendizaje automático) a estudiantes de pregrado. Las respuestas demuestran que existe interés y compromiso por parte de los estudiantes una vez que ya se han involucrado en proyectos en estas áreas.

La mayoría de encuestados del grupo del caso de estudio sienten que el aprendizaje de estas técnicas es importante para su desarrollo profesional y piensan seguir con el estudio de las mismas en el futuro. En el grupo control, se encontró que la mayoría indicó tener un interés por utilizar estas técnicas, aunque no las hayan utilizado nunca.

Con la finalidad de tener una idea más clara de los beneficios específicos para los estudiantes de su experiencia en investigación, se realizaron entrevistas individualizadas utilizando como preguntas un set de posibles beneficios ante los que los estudiantes podían identificarse o no como receptores del beneficio. Adicionalmente, en el caso afirmativo, podían calificar el grado entre “muy satisfactorio”, “satisfactorio”, y “poco satisfactorio”.

La lista de beneficios se encuentra en la tabla 2 y está basado en un estudio publicado en 2003 que tenía como fin específico describir los beneficios identificados por estudiantes de pregrado cuando realizan actividades de investigación en cuatro Universidades de Artes Liberales (Seymour et al, 2002). Se realizaron 5 entrevistas donde dos estudiantes llevaban un periodo corto de participación 4 semanas – 10 semanas, dos estudiantes que realizaban investigación en el grupo por lo menos dos años, y un estudiante ya graduado (en 2015).

<b>¿Cuál(es) de los siguientes beneficios tu percibes que los obtuviste al realizar investigación utilizando modelación (HPC) y/o aprendizaje</b>
---

**automático? (si/no)**

**¿En qué grado? Muy satisfactorio/satisfactorio/poco satisfactorio**

**Beneficios Personales/ profesionales**

Aumento en la confianza en tu habilidad de:

Realizar investigación

De contribuir a la ciencia

De presentar ¡/defender investigación

“Sentirse como un científico”

Establecer relaciones de trabajo a nivel de colega con tu mentor y compañeros

**Trabajando y pensando como un científico**

Incremento en la aplicación del conocimiento y habilidades de:

- pensamiento crítico
- resolución-de-problemas
- entendimiento del diseño de investigación
- la naturaleza del conocimiento científico

Incremento en el conocimiento y entendimiento de la ciencia y el trabajo de investigación

Teoría

Conceptos

Conexiones entre y dentro de las ciencias

Profundización del conocimiento vía presentaciones

Profundización del conocimiento vía enseñar a pares

Desarrollar el temperamento necesario para realizar investigación

**Habilidades**

Mejorar la comunicación (principalmente en presentación/argumentación oral)

Mejorar la comunicación (en edición de texto y redacción)

Técnicas de laboratorio (simulación, análisis de datos)

Organización de trabajo

Uso de computadoras

Lectura comprensiva

Trabajo colaborativo

Búsqueda de información

**Clarificación, confirmación y refinamiento de caminos de carrera y**

**educación**

Validación de tu disciplina principal de interés

Intención de realizar un postgrado

Incremento en el interés/entusiasmo por el campo de aprendizaje automático

Incremento en el interés/entusiasmo por el campo de simulación con el uso de HPC

Mayor conocimiento de las opciones de educación y carreras

**Mejora en la preparación para tu carrera y/o estudios de postgrados**

Experiencia de investigación auténtica

Oportunidades de colaboración y networking con profesores, compañeros y otros científicos

Mejora en tu resume

**Nuevas experiencias profesionales**

Cambio de actitudes hacia el aprendizaje y el trabajo como investigador

Tomar mayor responsabilidad en un proyecto

Incremento en la independencia en el proceso de toma de decisiones

Control del trabajo e intereses intrínsecos en el aprendizaje

**Otros beneficios**

Un buen trabajo de verano

Acceso a buenos equipos, software, técnicas de análisis de datos

En general los estudiantes entrevistados se identificaron con la mayoría de los beneficios con un nivel “muy satisfactorio” o “satisfactorio”. Entre los otros beneficios que precibieron se mencionaron: “Aprender sobre la forma de realizar una investigación para su posible publicación”, “Impulsa la creatividad en innovaciones”, “Aprendizaje del manejo de problemas que involucran gran cantidad de datos”, “se pueden hacer cosas que nunca te hubieras imaginado. No haces solo cosas de tu carrera, sino otras cosas que van más allá”

Uno de los entrevistados, manifestó un nivel “Poco satisfactorio” en cuanto a clarificación, confirmación y refinamiento de caminos de carrera y educación; el “sentirse como un científico”; así como también, expuso que no tiene interés en realizar un postgrado.

El entrevistado que ya culminó sus estudios de postgrado, mencionó adicionalmente algunos otros beneficios, que con estas herramientas no se necesita centrarse en

algo muy específico si no que te brinda un amplio campo de aplicación. Adicionalmente mencionó, que considera implementar el mismo modelo pedagógico, así como de organización/manejo del grupo de investigación para manejar su grupo de investigación en un futuro. De la misma manera, al compararse con otros pares cuando estudiaba su maestría se dio cuenta que ahorro tiempo.

En comparación sus compañeros en la maestría, por primera vez, estaban involucrados en un proyecto de investigación. Al tener ya experiencia en investigación a nivel de pregrado le permitió comenzar su investigación más rápidamente y organizarse mejor. También mencionó que la experiencia de investigación con uso de cómputo de alto rendimiento influenció en su decisión de seguir estudios de postgrado, aun cuando la temática de postgrado fue netamente experimental.

### **Productos de los estudiantes utilizando herramientas de computacionales con sistemas de cómputo de alto rendimiento y/o herramientas de aprendizaje automático.**

En la Tabla 1, se encuentran ejemplos de trabajos publicados por estudiantes de pregrado empleando herramientas de cómputo de alto rendimiento y aprendizaje automático. Para ilustrar con más detalle el trabajo que los estudiantes realizaron se seleccionó dos proyectos de investigación en los cuales los alumnos utilizaron las herramientas ya mencionadas para dar solución a problemas relevantes y crear información valiosa para la continuación del estudio en estas temáticas. Se debe resaltar, que estos proyectos han sido publicados en revistas académicas pertinentes al área y han sido expuestos en varios congresos y charlas magistrales. Además, a partir de ellos se han derivado nuevos proyectos e ideas que se siguen desarrollándose al momento en el grupo QCT-USFQ.

**Tabla 1.** Ejemplos de trabajos publicados por estudiantes de pregrado empleando herramientas de cómputo de alto rendimiento y aprendizaje de máquina.

<b>Carrera</b>	<b>Trabajo publicado</b>	<b>Tipo herramienta</b>
Biología Marina	Zipper Like Structures Possible Intermediates to Assemble Duplex Mediated G-Quadruplex DNA	Cómputo de alto rendimiento
Biotecnología	Molecular Dynamics Studio of Poly (VINYL Alcohol) Mechanical Properties for its Incorporation in Bones Structures as a PVA-PLA Substrate for Tissue	Cómputo de alto rendimiento

	Regeneration	
Biotecnología	Molecular Dynamics Study of the Biodegradation Process of the Biopolymer Poly (LACTIC)/Poly (VINYL) Scaffold for Bone Tissue Engineering	Cómputo de alto rendimiento
Biotecnología	DNA Assembly of Nanofibers from Duplex-Quadruplex DNA	Cómputo de alto rendimiento
Biotecnología	Structure and sequence based functional annotation of Zika virus NS2b protein: Computational insights.	Aprendizaje de máquina
Biotecnología	Discovering key residues of dengue virus NS2b-NS3-protease: New binding sites for antiviral inhibitors design.	Aprendizaje de máquina

Fuente: elaboración propia

### **Proyecto ejemplo 1. Uso de cómputo de alto rendimiento en la clasificación de estructuras conformacionales estables del tipo G-cuádruple.**

Un ejemplo del trabajo interdisciplinario e interuniversitario en cómputo de alto rendimiento fue la investigación realizada por la biofísica Daniela Barragán quien fue estudiante de biofísica de la Escuela Superior Politécnica de Chimborazo y que llevó a cabo este proyecto bajo la supervisión de Miguel Ángel Méndez y Robert Cázar profesores del Colegio de Ciencias e Ingenierías, Politécnico de la Universidad San Francisco de Quito y de la Facultad de Ciencias, Escuela Superior Politécnica de Chimborazo, respectivamente. El estudio se enfocó en la evaluación de la dinámica conformacional de estructuras de conformación de ADN G-Cuádruple tetramoleculares paralelas, formadas a partir de una secuencia repetitiva telomérica humana (Barragán et al., 2016).

Las conformaciones de la molécula fueron evaluadas según la cantidad de flexibilidad y movilidad a lo largo de la simulación, obteniéndose como resultados conformaciones que confieren estabilidad a la estructura global debido la presencia de los residuos de Guanina en las estructuras G-Cuádruples (Barragán et al., 2016). Para poder realizar esta investigación se utilizó un clúster de computadoras (HPC-USFQ) dedicado para investigación en cómputo de alto rendimiento.

El equipo permitió el desarrollo de esta simulación computacional basado en el empleo de modelos matemáticos durante un tiempo de simulación de 200 ns (esto no corresponde al tiempo de cálculo de la computadora, pero a unidades usadas en

esta área de la simulación en específico) donde se muestrearon un total de 400 000 conformaciones (Barragán et al., 2016).

Por ejemplo, una computadora regular, o una laptop aunque podría usarse para tales cálculos se demoraría demasiado (semanas a meses de cálculo ininterrumpido), en su lugar con la utilización de un equipo dedicado a cómputo de alto rendimiento las mismas simulaciones pueden realizarse en cuestión de horas. El *software* utilizado es libre para uso académico y cumple con altos estándares pues fue desarrollado por un reconocido grupo de investigación en la Universidad de Illinois, Urbana Champagne donde el artículo que describe el programa ha recibido más de 6000 citas (Philips et al., 2005)

Adicionalmente, los datos generados en la simulación generaron 400 000 “casos” o conformaciones espaciales de las moléculas. Uno de las tareas que se necesitaba realizar fue encontrar grupos de conformaciones que se parezcan.

Para realizar esta tarea, se utilizó un método de aprendizaje automático no supervisado, denominado del “vecino cercano”. Para poder usar el método se requirió de dos etapas previas usando los datos crudos. En la primera, se generaron parámetros numéricos fáciles que resumiera alguna propiedad del sistema (Ej. el largo de la molécula), con estos parámetros se identificaron “ejemplos” de grupos (configuraciones) que serían utilizados por el algoritmo para clasificar automáticamente todo el resto de información obtenida de la simulación (las 400 000 conformaciones) (Barragán et al., 2016).

En resumen, en este proyecto en particular se hizo uso de los dos tipos de herramientas discutidas en este artículo. Ambas herramientas, facilitaron la ejecución del proyecto en un estudio que habría sido muy difícil de llevar a cabo sin el uso de estas herramientas. Específicamente, las simulaciones en una laptop personal habrían tomado varios meses de cálculo ininterrumpido. Por otra parte, la clasificación de tantos casos de manera automatizada habría sido difícil de conseguir sin hacer uso del aprendizaje automático.

### **Proyecto ejemplo 2: uso de aprendizaje de máquina para el descubrimiento de residuos clave de la proteasa NS2b-NS3 del virus del dengue.**

Diferentes metodologías basadas en la utilización de Aprendizaje automático pueden ser ventajosas para realizar predicciones relevantes acerca de la funcionalidad de moléculas biológicas. Un ejemplo de esto es el trabajo realizado por estudiantes de



biotecnología y química de la USFQ con la colaboración de profesores de la facultad de medicina y el Colegio de Ciencias e Ingenierías, Politécnico.

En este trabajo se construyó paso a paso un modelo basado en redes neuronales para predecir la importancia funcional de todos los aminoácidos de la proteasa NS2b-NS3 del Virus del Dengue (DENV) (Aguilera-Pesantes et al., 2017). Esta proteína es de vital importancia para el ciclo de replicación y por ende infectivo del virus. Es por esto, que conocer las secciones funcionalmente importantes de la proteína es útil para diseñar componentes activos que tengan actividad inhibitoria frente a la proteasa y consecuentemente, contra la replicación de DENV.

Para lograrlo, se alimentó el algoritmo con datos experimentales disponibles y diferentes tipos de descriptores moleculares que fueron obtenidos computacionalmente (Aguilera-Pesantes et al., 2017). Cabe recalcar, que varios de estos descriptores fueron obtenidos de igual manera por servidores de Aprendizaje Automático abiertos al público disponibles en internet. Muchas veces, los métodos de aprendizaje automático son preferidos ya que son mucho más rápidos que otras metodologías donde se emplean cálculos extensos.

Mediante la aplicación del aprendizaje automático se logró reconocer de una manera bastante rápida y precisa la característica de interés para el caso de estudio. Se identificaron cuatro nuevos sitios (grupos de aminoácidos) en la proteína con importancia funcional, que pueden constituir sitios alostéricos (sitios en la proteína donde no se une directamente el substrato de la enzima) (Aguilera-Pesantes et al., 2017).

De manera directa, la información generada en este estudio es de gran valor para investigadores que deseen utilizar estrategias de diseño de fármacos basados en estructura, en donde es necesario primero definir el receptor (sitios alostéricos) sobre el cual se desea trabajar y sus características. De manera indirecta, se describió paso a paso una metodología para predecir importancia funcional de los residuos de casi cualquier proteína, y que puede ser utilizada por cualquier usuario utilizando recursos de software libres.

## **Conclusiones**

En este estudio de caso, se evaluó una muestra de estudiantes de pregrado miembros del grupo de química computacional y teoría USFQ/Instituto de Simulación Computacional-USFQ. Los resultados obtenidos demuestran que el uso de herramientas de aprendizaje automático y cómputo de alto rendimiento es una parte

importante en su desarrollo profesional y académico. Para poder evidenciar esto se utilizó un modelo de entrevistas validado anteriormente para evaluar el impacto de las experiencias de investigación a nivel de pregrado (Seymour, 2017). Adicionalmente, la producción científica en formatos de artículos nacionales, internacionales, presentación en formato poster, presentaciones orales obtenidas directamente de estudiantes de pregrado demuestra que esta metodología es una alternativa viable para realizar investigación en el país.

La mayoría de estudiantes que trabajaron en estos proyectos se encuentran satisfechos con el desempeño de este enfoque de investigación y desean continuar con el uso de estas herramientas en el futuro. Sugerimos que similares experiencias se podrían implementar en varias áreas de conocimiento por grupos de investigación en instituciones de educación superior en Ecuador.

### **Referencias Bibliográficas**

- Acs, Z. J., Audretsch, D. B., & Feldman, M. P. (1992). Real effects of academic research: comment. *The American Economic Review*, 82(1), 363-367
- Aguilera-Pesantes, D., & Méndez, M. A. (2017). Structure and sequence based functional annotation of Zika virus NS2b protein: Computational insights. *Biochemical and Biophysical Research Communications*.
- Aguilera-Pesantes, D., Robayo, L., Méndez, P., Mollocana, D., Marrero-Ponce, Y., Torres, F., & Méndez, M. (2017). Discovering key residues of dengue virus NS2b-NS3-protease: New binding sites for antiviral inhibitors design. *Biochemical and Biophysical Research Communications*. doi:10.1016/j.bbrc.2017.03.107
- Ark, L. A. van der, Bolt, D. M., Wang, W.-C., Douglas, J. A., & Chow, S.-M. (2015). Quantitative Psychology Research: The 79th Annual Meeting of the Psychometric Society, Madison, Wisconsin, 2014. Springer.
- Barragán D., Cazar R., Méndez M. (2016). Clasificación vía aprendizaje automático de conformaciones moleculares en estructuras teloméricas. Chimborazo-Ecuador: *Repositorio ESPOCH*.
- Big Data: A Revolution That Will Transform How We Live, Work, and Think. (2014). *International Journal of Advertising*, 33(1), 181–183. Recuperado el 21 de Agosto del 2016 de <https://doi.org/10.2501/IJA-33-1-181-183>

- Chen, M., Mao, S., & Liu, Y. (2014). Big Data: A Survey. *Mobile Networks and Applications*, 19(2), 171–209. Recuperado el 21 de Agosto del 2016 de <https://doi.org/10.1007/s11036-013-0489-0>
- Clote, P., & Backofen, R. (2000). *Computational molecular biology: an introduction* (Vol. 1). Wiley.
- Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78–87.
- Etzkowitz, H., & Leydesdorff, L. (2014). The endless transition: a'Triple Helix'of university industry government relations.
- ESPE (2017). Clúster de supercomputación ESPE. Recuperado el 22 de Septiembre de 2017 de <http://ugi.espe.edu.ec/ugi/cluster-de-supercomputacion-espe/>
- Hajdarpasic, A., Brew, A., & Popenici, S. (2015). The contribution of academics' engagement in research to undergraduate education. *Studies in Higher Education*, 40(4), 644-657.
- Howe, B. (2012, January 11). High-Performance Computing Meets Databases | Intel Science & Technology Center for Big Data. Recuperado el 21 de Agosto del 2016 de <http://istc-bigdata.org/index.php/high-performance-computing-meets-databases/>
- Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., ... & Schulten, K. (2005). Scalable molecular dynamics with NAMD. *Journal of computational chemistry*, 26(16), 1781-1802.
- Mansfield, E. (1991). Academic research and industrial innovation. *Research policy*, 20(1), 1-12.
- Ramírez Romero, J. L. (2006). Las tecnologías de la información y de la comunicación en la educación en cuatro países latinoamericanos. *Revista Mexicana de investigación educativa*, 11(28).
- Ramos, P. R. H., Conde, M. Á., & Peñalvo, F. J. G. (2014, June). Differences and similarities in use and acceptance of PLEs between universities in Ecuador and Spain. In *Proceedings of the 2014 Workshop on Interaction Design in Educational Environments* (p. 70). ACM.
- RED CEDIA (2017). Ampliación cluster UTPL. Recuperado el 22 de Septiembre de 2017 de

<https://cedia.edu.ec/es/proyectos-cluster/proyectos-antiores-cluster/ampliacion-cluster-utpl>

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6), 386.

Ruth, J.-P. (2017, May 8). 6 Examples of AI in Business Intelligence Applications -. Recuperado el 18 de Agosto del 2016 de <https://www.techemergence.com/ai-in-business-intelligence-applications/>

Seymour, E., Hunter, A. B., Laursen, S. L., & DeAntoni, T. (2004). Establishing the benefits of research experiences for undergraduates in the sciences: First findings from a three-year study. *Science education*, 88(4), 493-534.

Smith, G. (2017, July 24). Inside the Artist's Headset – Sterling Crispin's Cyber Paint. Recuperado el 18 de Agosto del 2016 de <http://www.creativeapplications.net/unity-3d/inside-the-artists-headset-sterling-crispins-cyber-paint/>

USFQ (2017). USFQ's High Performance Computing System. Recuperado el 22 de Septiembre de 2017 de [http://www.usfq.edu.ec/investigacion\\_y\\_creatividad/hpc-usfq/Paginas/default.aspx](http://www.usfq.edu.ec/investigacion_y_creatividad/hpc-usfq/Paginas/default.aspx)

Verma, A., Ahuja, P., & Neogi, A. (2008). Power-aware dynamic placement of hpc applications. In *Proceedings of the 22nd annual international conference on Supercomputing* (pp. 175–184). ACM. Recuperado el 20 de Agosto del 2016 de <http://dl.acm.org/citation.cfm?id=1375555>