Taller 1

Juan Sebastián Cañas

22 de Febrero de 2021

1 Presentación

1.1 Presentación

Soy estudiante de la Universidad Nacional de Colombia. Actualmente estoy terminando el pregrado en matemáticas y empezando a trabajar en una empresa que se dedica a resolver problemas de salud que van desde la salud poblacional hasta el ambiente clínico. Me interesa la matemática aplicada y la computación. Lo anterior, junto con la teoría y el conocimiento de dominio, son las herramientas que utilizo para resolver problemas. Los problemas que me gusta pensar y tratar de resolver están relacionados con aplicaciones en salud, sistemas biológicos y sistemas sociales. Los últimos años he estado involucrado en proyectos relacionados con políticas públicas y seguridad predictiva, epidemiología e informática biomédica. Creo que la computación tiene un rol social importante [Abe+20] y me gustaría dedicarme a ello.

1.2 Por qué me interesa la economía

La economía puede ser un lugar desde donde se puede pensar y repensar las discusiones sociales y utilizar herramientas teóricas que pueden modificar la compresión y el estado de las mismas.

1.3 Interés en el curso y objetivo de aprendizaje

En este curso tengo la expectativa de seguir aprendiendo del proceso investigativo desde la experiencia, ya que es una de las cosas que no he podido ejercer durante el pregrado. También me interesa conocer las herramientas teóricas y técnicas que se han desarrollado en economía del comportamiento pues estas tienen relaciones importantes con inteligencia artificial y ciencias de datos [AGG19]. Por último, un tema que me interesa mucho es el relacionado con el proceso de toma de decisiones. Es aquí donde los algoritmos más sofisticados suelen quedar cortos y donde más influencia pueden tener. Me gustaría entender el problema de las decisiones desde el punto de vista de la economía del comportamiento ya que mi trabajo de grado se está acercando al proceso de toma de decisiones. En específico, estoy pensando el trabajo dentro del problema de

las decisiones óptimas en el proceso de tratamientos clínicos utilizando datos observacionales [Sha20] [AW21] [Got+19] [Kom+18] [OS19].

2 Gentzkow and Shapiro

2.1 Resumen

El texto [GS14] reúne un conjunto de recomendaciones y buenas prácticas para llevar a cabo un proyecto en el que se usan herramientas de programación y datos. El texto está escrito pensando en los retos frecuentes que los investigadores sociales experimentales tienen al desarrollar proyectos que involucran programación sin tener una formación en computación. Entre los retos frecuentes están:

- Reproducibilidad: reutilización del código y los datos en el futuro, tanto por la persona que programó el código como por otros usuarios. La reproducción futura del proyecto debe obtener resultados similares.
- Transparencia: mostrar detalladamente las tareas que se ejecutaron en el código y las modificaciones de los datos.
- Ambigüedad: claridad con la que se organizan los diferentes archivos de un proyecto para evitar inconsistencias.
- Eficiencia: al utilizar nuevamente el código se puede perder tiempo corriendo todo por cualquier variación por ligera que sea. Como consecuencia, se debe invertir un esfuerzo grande en la modificación de los archivos para volver a obtener el resultado.

Las propuestas que se hacen para evitar los problemas anteriores son:

- Automatización: una vez instalados los requerimientos de software y paquetes, el código se debe ejecutar completo y fácilmente. Esto es importante porque hay un orden preciso de la ejecución del código de principio a fin para obtener los resultados.
- Versión de control: Permite hacer una trazabilidad de quién hizo cambios, cuándo los hizo y qué cambió. Esto evita la confusión al manejar un proyecto entre diferentes autores y no entrar en conflictos de software si se hacen modificaciones. Adicionalmente, permite unificar el proyecto en archivos estables que varían de acuerdo a los objetivos y avances del proyecto.
- Directorios: la organización de los archivos en carpetas hace más entendible la estructura del proyecto de acuerdo a las funciones y características de los archivos.

- Preprocesamiento: Normalizar los datos, identificar claramente los valores, entender la estructura de la base datos, el significado de los vacíos, la razón de los duplicados y la forma en que se recolectaron permite transformarlos para que puedan tener sentido dentro de un modelo.

- Abstracción: Una forma de evitar la redundancia y disminuir la probabilidad de errores es crear funciones propias que puedan hacer tareas recurrentes en diferentes archivos e incluso en diferentes proyectos.

- Documentación: Es importante que el código pueda ser entendido por cualquiera que lo lea, los comentarios pueden aclarar y comunicar el objetivo de hacer cierta tarea en el código. Una recomendación es que la documentación debe actualizarse tanto como el código. Por otro lado, una forma de documentación es el estilo del código, lo que Donald Knuth llama *programación literaria*, esto es, la forma como se escribe el código debe comunicar los razonamientos e ideas detrás de su ejecución y el estilo debe ser similar al lenguaje natural humano [Knu84].

- Manejo del proyecto: esto permite fijar objetivos y responsabilidades, monitorear tareas, discutir dificultades, mejorar la comunicación entre el equipo, focalizar esfuerzos y ver el estado de avance del proyecto.

2.2 Aplicación de estos principios en sus trabajos futuros y que acciones quiere adoptar

La gran mayoría de propuestas para llevar los proyectos ya las he usado en proyectos de software y académicos, sin embargo sigo siendo negligente en el uso de versiones de control. Desde el semestre pasado he estado usando en algunos cursos el versionador de control y en el trabajo se ha vuelto una necesidad. Mejor dicho, la primera semana antes de hacer cualquier cosa tuve que aprender bien a utilizar git. Ahora que soy un usuario diario de git puedo decir que es una herramienta que facilita y aclara mucho el trabajo tanto aplicado como investigativo. También deseo aprender a usar OSF para el trabajo de grado e incluirlo en los proyectos futuros de investigación.

Algo que me gustaría aprender y no está contemplado en el artículo es el uso de herramientas como Docker. En ocasiones el código abierto, organizado y con versiones de control no es suficiente para solucionar el problema de la reproducibilidad. Las configuraciones propias de la máquina que aloja el proyecto terminan afectando los resultados y cuando se intentan instalar las versiones de los paquetes necesarios no es suficiente para reproducir los resultados.

Finalmente, una práctica que actualmente uso en mis proyectos es la de no usar software privado o pago. Por mucho tiempo usé Matlab y Wolfram Mathematica porque era el software que usualmente se usaba en las clases y teníamos licencias. En algún momento esas licencias terminaron y ya no podía correr el código salvo con software de dudosa procedencia. De alguna forma esas licencias son una barrera para la reproducibilidad porque no todas las personas tienen acceso a estas. En este momento todo mi entorno de trabajo es software

libre y la filosofía colaborativa detrás es interesante. El cobro por el uso de un lenguaje de programación me hace sentir como si me cobraran por hablar inglés o cualquier otro idioma.

References

[Knu84] Donald Ervin Knuth. "Literate programming". In: *The Computer Journal* 27.2 (1984), pp. 97–111.

[GS14] Matthew Gentzkow and Jesse M Shapiro. "Code and data for the social sciences: A practitioner's guide". In: *Chicago, IL: University of Chicago* (2014).

[Kom+18] Matthieu Komorowski et al. "The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care". In: *Nature medicine* 24.11 (2018), pp. 1716–1720.

[AGG19] Ajay Agrawal, Joshua Gans, and Avi Goldfarb. *The economics of artificial intelligence: an agenda*. University of Chicago Press, 2019.

[Got+19] Omer Gottesman et al. "Guidelines for reinforcement learning in healthcare". In: *Nature medicine* 25.1 (2019), pp. 16–18.

[OS19] Michael Oberst and David Sontag. "Counterfactual off-policy evaluation with gumbel-max structural causal models". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 4881–4890.

[Abe+20] Rediet Abebe et al. "Roles for computing in social change". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 252–260.

[Sha20] Uri Shalit. "Can we learn individual-level treatment policies from clinical data?" In: *Biostatistics* 21.2 (2020), pp. 359–362.

[AW21] Susan Athey and Stefan Wager. "Policy learning with observational data". In: *Econometrica* 89.1 (2021), pp. 133–161.