Información General

Ciencia de Datos basada en programación en R para economistas

Justificación de la Propuesta:

Los seres humanos han adoptado masivamente el uso de dispositivos electrónicos en sus actividades diarias, teniendo de ejemplos los teléfonos móviles, los sensores para mediciones de salud, entre otros. En particular, resulta necesario resaltar que estos aparatos almacenan registros digitales de los valores que en ellos se levantan, quedando reflejadas las actividades que hace su portador. Igualmente, la sociedad ha llevado al plano digital una serie de procesos y actividades que antes se hacían en el mundo físico, quedando también los registros de los contenidos generados y de las distintas interacciones que ocurren. Como consecuencia de ambos factores, se encuentran disponibles grandes volúmenes de datos que pueden ser procesados parcial o totalmente por medio de computadores, posteriormente ser transformados en información y con esta se puede hacer el modelado que permite representar conductas, comportamientos sociales, así como la determinación de tendencias, patrones o análisis predictivos.

Dentro de este contexto se tiene que han surgido nuevas áreas de estudio, que desde distintos ángulos, fundamentan la generación de conocimiento en el uso de los datos. Algunas de estas son:

- **Ciencia de Datos:** es un campo interdisciplinario que utiliza métodos científicos, estadísticos y computacionales para extraer conocimientos y tomar decisiones informadas a partir de grandes conjuntos de datos, donde se recopilan, limpian, analizan y visualizan datos para identificar patrones, tendencias y relaciones que no serían evidentes a simple vista.
- **Ciencias Sociales Computacionales:** es un campo de estudio donde las computadoras se utilizan para modelar, simular y analizar fenómenos sociales, por ejemplo, estudiando los comportamientos de los individuos en las redes sociales, la influencia de estas redes en la toma de decisiones, la búsqueda de correlaciones o relaciones causales entre distintos fenómenos socioeconómicos.

Saber que existen estas nuevas áreas de conocimiento, algunas ya formalizadas como carreras de estudio en universidades del mundo, e incluso en latinoamérica, hacen ver claramente que la manipulación o el modelado de los datos no sólo debe quedar en manos de los computistas, sino se hace necesario que los científicos sociales conozcan y dominen distintos métodos para interactuar con tales cantidades de información, yendo más allá del uso de softwares tradicionales como lo son las hojas de cálculo tipo Excel.

Por tal motivo, en este documento se expone la propuesta de la materia de estudio “Ciencia de Datos basada en programación en R para economistas” dirigido a los estudiantes de pregrado de la Escuela de Economía.

El contenido de la materia, que será detallado más adelante, está diseñado para que el aprendiz pase por un proceso de aprendizaje y de práctica continua, que le permitan dominar diversos métodos de programación que sirvan como herramientas para sustentar y validar las hipótesis que tengan al realizar las investigaciones inherentes al eje central de la carrera Economía, así como también para coadyuvar a que puedan comunicar de manera efectiva los resultados de las investigaciones.

Igualmente, es necesario resaltar que a lo largo de todas las clases y actividades, se hace énfasis en el desarrollo de las capacidades analíticas del participante, mediante la interpretación de los resultados obtenidos al procesar los datos. De esta forma, se introducen conceptos y métodos de la ciencia de datos y de las ciencias sociales computacionales.

Justificación del Uso de R:

Para programar en informática se cuenta con distintos lenguajes, los cuales se clasifican principalmente según dos características. La primera viene dada en el propósito o el uso, que tendrá el programa informático que se esté codificando, ya que por ejemplo, son distintos los objetivos que se persiguen al programar las rutinas de un software que da soporte a un horno microondas a los que se tendrán, como es en este caso, al hacer la manipulación y el procesamiento de los datos. Es así, que dentro de la taxonomía de los lenguajes de programación, uno de los criterios de clasificación que se aplica, es determinar si son de "propósito general", equivalente a indicar que pueden ser usados para distintos fines y por otro lado están aquellos lenguajes que son de "propósito específico", en los que se delimita de forma taxativa, a qué están destinados los programas que se pueden codificar con tal lenguaje.

Adicionalmente, la otra característica que diferencia a los lenguajes, es lo que se denomina el "nivel" en el que se considera que están cuando se hace el enunciado y la ejecución de las instrucciones que el computador ejecutará. Para esto existe una escala entre los extremos "bajo nivel" y "alto nivel", catalogando como lenguajes de "bajo nivel", aquellos en los que se programa en instrucciones que son mucho más abstractas y cercanas al lenguaje intrínseco del computador, el cual tiende a ser bastante complejo. Por otra lado, están los lenguajes que se consideran de "alto nivel", en los que se programa con enunciados que son de más fácil comprensión para el ser humano, ya que las instrucciones que se codifican son más cercanas en sintaxis al lenguaje natural con que habitualmente nos expresamos.

De esta manera, dentro de los lenguajes de programación se encuentra R, el cual es un lenguaje de propósito específico, orientado a la estadística que ha sido adoptado ampliamente para realizar la manipulación y al procesamiento de datos, que inicialmente a principios del año 2000, estuvo destinado a correr modelos estadísticos como una opción de código abierto alternativa a programas como Stata o EViews.

Adicionalmente, el lenguaje R también se asume que es de "alto nivel", facilitando al usuario la comprensión del lenguaje y de cómo hacer uso de las instrucciones que en él se pueden ejecutar, teniendo una curva de aprendizaje con una pendiente más suave a la de otros lenguajes de programación, lo que lo hace ideal para ser aprendido por aquellos que no tienen formación en

ciencias de la computación sino que provienen de otras áreas del conocimiento como lo son las ciencias sociales.

No obstante, como en todo proceso en que se hace la selección de un recurso sobre otro, existen beneficios y sacrificios, no siendo la selección de R ajena a estos contrapesos. Las opciones de lenguajes disponibles para los investigadores cercanos a las ciencias que basan sus estudios en el uso de los datos, tienen como otras opciones los lenguajes Python y Julia.

Python es un lenguaje de programación de propósito general, que ha sido adoptado por la comunidad científica por su facilidad de uso y por la cantidad de librerías que se han desarrollado para el procesamiento de datos, mientras que Julia es un lenguaje de programación de alto rendimiento, que ha sido adoptado por la comunidad científica por su velocidad de ejecución y por la facilidad de escribir código que se asemeja a las matemáticas.

Sin embargo, dentro de estas opciones el lenguaje que se ha seleccionado para la propuesta de esta materia es ***R***, ya que es un lenguaje que ha sido adoptado ampliamente por la comunidad de científicos sociales por su facilidad de uso, la integración con el entorno de desarrollo RStudio que facilita la configuración inicial del ambiente de trabajo dentro del computador, así como también por la gran cantidad de librerías que se han desarrollado para el procesamiento de datos asociados a dominios de las ciencias sociales.

En líneas generales, a continuación se mencionan algunos de los beneficios que se obtienen al programar en R:

1. **Análisis de datos:** R es una herramienta poderosa para el análisis de datos, lo que permite a los economistas manejar grandes conjuntos de datos de manera eficiente. Esto es crucial en economía, donde los datos pueden ser complejos y abundantes.
2. **Visualización de datos:** R ofrece una amplia gama de paquetes para visualización de datos, lo que permite a los economistas crear gráficos y visualizaciones claras, interactivas y efectivas, lo cual es útil para comunicar los resultados de las investigaciones de manera más impactante y comprensible.
3. **Modelado estadístico:** R cuenta con una gran variedad de librerías para realizar análisis estadísticos avanzados y modelado, lo que permite a los economistas desarrollar modelos complejos para comprender mejor los fenómenos sociales y económicos.
4. **Reproducibilidad, comunicar resultados y transparencia:** Programar en R permite a los economistas documentar y reproducir fácilmente sus análisis, lo que aumenta la transparencia y la credibilidad de la investigación. Igualmente los investigadores puedan comunicar de manera efectiva los resultados de sus investigaciones, generando reportes, gráficos, tablas, mapas, aplicaciones web interactivas, entre otros, que pueden ser publicadas con cadenas de trabajo automatizadas que minimizan los esfuerzos y el tiempo que se requiere para la generación y actualización de estos productos.

Sin embargo, la propuesta acá detallada, no aborda en profundidad los aspectos mencionados en el punto 3 sobre el modelado estadístico, ya que al existir materias especializadas dentro de la

carrera de Economía para abordar estos tópicos, se considera que la materia propuesta se enfoca en la enseñanza del resto de los beneficios enumerados.

Igualmente resulta de interés hacer las siguientes consideraciones al usar un lenguaje de programación como R, también extrapolable a python:

1. **Lenguaje de código abierto:** R es un lenguaje *open source* (código abierto) y su uso no implica ningún pago de licencia y los recursos computacionales que se necesitan para ejecutarlo están optimizados, siendo un elemento muy beneficioso, por ejemplo para estudiantes e instituciones que pueden contar con equipos informáticos que no son de última generación, haciendo que los procesos que en él se corran puedan hacerse de una manera más eficiente en comparación a si se usarán softwares privativos.
2. **Comunidad de Usuarios:** la existencia de una comunidad de usuarios muy activa, donde en múltiples investigaciones el componente de programación se hace en este lenguaje, llegando algunas de estas investigaciones a formar parte del estado del arte en su dominio y en muchos casos los modelos implementados o aplicados, son publicados mediantes "librerías" o "paquetes", haciendo que los usuarios de R puedan acceder de forma libre a los más recientes y novedosos modelos, siendo nuevamente importante resaltar que esto lo pueden hacer de forma legal, sin realizar el desembolso de alguna suma de dinero.
3. **Acceso a Cómputos en la Nube:** los programadores de R pueden acceder a recursos de cómputo que se encuentran disponibles en la nube, donde mediante un navegador web se pueden ingresa a sitios de internet que permiten programar y ejecutar los códigos, disipando barreras que generalmente vienen dadas en que el investigador cuenta con computadoras obsoletas o de muy limitados recursos de procesamiento informático. Al usar la nube, el computador del programador pasa a ser un terminal y la contraparte de los cálculos ocurre en un servidor remoto, con robustos recursos que empresas tecnológicas ponen a disposicición de la comunidad científica de forma gratuita.

Método de Enseñanza Propuesto:

Trabajar por proyectos grupales donde tengan que resolver un un problema real y aprendan haciendo las codificaciones, asociado a algún tema de interés de la realidad económica venezolana y abarcando las distintas fases de un proyecto de investigación, desde la obtención de los datos, su manipulación y procesamiento, hasta la presentación de los resultados, garantizando que las investigaciones pueda ser reproducibles por la comunidad científica.

Ejemplos de esto puede ser la obtención de datos de forma periódica y automatizada para la generación de una canasta de precios de la cual se puedan hacer las mediciones de las variaciones de precios detectadas y así construir un indicador de inflación.

Posibles Fuentes de Datos:

A los fines de indicar algunas de las fuentes de datos a las cuales se puede acceder para realizar las investigaciones a desarrollar, se mencionan las siguientes:

- Páginas web o bases de datos de instituciones públicas y privadas.

- Páginas web de comercio electrónico como supermercados donde se publican precios.
- Información georreferenciada disponible en las API´s de Google Maps y de Open Street Maps.
- Conexiones a API´s de distintos sistemas de información.
- Datos de redes sociales.

Modalidad de las Clases:

Sesiones de clases y prácticas presenciales, ya que se considera que el aprendizaje de la programación se facilita cuando se cuenta con la presencia del instructor y se puede interactuar con los compañeros de clase y más aún, teniendo presente los problemas de conectividad a internet.

Requerimientos Técnicos del Laboratorio:

Las clases deben ser impartidas en un laboratorio que por participante debe contar con un equipo (computador) con al menos 4 GB de RAM; navegador web actualizado (Firefox o Google Chrome); 10 Gb de disco duro disponible; sistema operativo: windows 8 (o superior) , MacOS o Ubuntu 18 o superior (en caso de usar Ubuntu la memoria RAM puede ser mínima de 2 Gb); arquitectura 64 bits.

Consideraciones Adicionales:

El curso contará con un repositorio de acceso público mediante la web, con los códigos, datos, guías y demás materiales que sean usados y generados a lo largo de las sesiones de trabajo. Igualmente se adoptará el uso un grupo de mensajería como Discord para almacenar todas las preguntas y respuestas que vayan surgiendo a lo largo del curso, las cuales podrán servir como material de consulta para futuros asistentes al curso.

Los proyectos de programación que se generen a lo largo del curso por parte de los alumnos, deberán ser compartidos mediante el repositorio, garantizando el acceso por parte de todos los interesados. Igualmente se fomentará la creación de conjuntos de datos de libre acceso que sirvan para toda la comunidad de investigadores.

Consideraciones Finales:

En resumen, aprender a programar en R y familiarizarse con la ciencia de datos y las ciencias sociales computacionales, puede permitir a los economistas realizar análisis más sofisticados y rigurosos, así como abordar preguntas de investigación más complejas en el ámbito de las ciencias sociales, abriendo la posibilidad de que se obtengan y generen datos estructurados, en medio de una situación país donde no existe la cultura gubernamental ni privada de compartir y/o publicar conjuntos de datos.

Los participantes del curso podrán hacer investigaciones que resulten reproducibles y mediante procesamientos encadenados y se les facilitará comunicar los resultados mediante blogs, página web o reportes técnicos de fácil acceso mediante la web.

Igualmente, podrán hacer prototipos de aplicaciones web interactivas para el análisis exploratorio de los datos, su modelado mediante modelos econométricos o estadísticos y dispondrán de nociones de elementos introductorios a la ciencia de datos y sus métodos de trabajo.