Desarrollo de aplicación web con R y SHINY, para el tratamiento multivariado de números borrosos

César Matilde
Diplomatura en Métodos de Explotación Inteligente de Datos, Centro de Investigación CeReCoN, UTN - FRM
matilde.cesari@frm.utn.edu.ar

Resumen
En esta propuesta se busca comprobar que mediante una aplicación web, ágil, accesible y gráfica que permita transformar los datos cuantitativos y cualitativos, mediante la matemática borrosa, se facilita el uso y aprendizaje de un enfoque más apropiado para obtener conocimiento preciso a partir de valoraciones subjetivas; así como incorporar variables teóricas no directamente observables en el estudio. El enfoque difuso es una estrategia que permitirá modelar y manipular constructos, en estudios donde se abordan fenómenos que conducen al estudio de variables complejas; a diferencia de otras aproximaciones analíticas que solo manejan variables observadas [citas sobre celulares y vientos], mediante la definición de la variable lingüística a partir de referencias teóricas o de expertos y aplicando la misma a variables observables que permiten derivar la variable no observable directamente en los datos.

Palabras clave: aplicación web, Shiny R, matemática borrosa, análisis multivariado

Abstract
This doctoral thesis seeks to verify that through an agile, accessible and graphic web application that allows the transformation of quantitative and qualitative data, through fuzzy mathematics, the use and learning of a more appropriate approach is facilitated to obtain precise knowledge from subjective ratings; as well as incorporating theoretical variables not directly observable in the study. The fuzzy approach is a strategy that will allow modeling and manipulating constructs, in studies where phenomena that lead to the study of complex variables are addressed; Unlike other analytical approaches that only handle observed variables [citations about cell phones and winds], by defining the linguistic variable from theoretical or expert references and applying it to observable variables that allow deriving the non-observable variable directly in the data..

Keywords: web application, Shiny R, fuzzy math, multivariate analysis.

Propósito:
Proporcionar una solución computacional para el tratamiento de datos imprecisos mediante la lógica borrosa y técnicas de análisis multivariado de datos, a través de una aplicación web construida con el lenguaje R y el paquete Shiny

Descripción:
Se pretende aplicar una metodología de desarrollo de software basado en eventos, y demostrar la capacidad que tiene el lenguaje R y el paquete Shiny para proporcionar una solución completa y fácil de usar en el mundo de la ciencia de datos.
La denominada Ciencia de Datos (Data Science; también denominada Science of Learning) se ha vuelto muy popular hoy en día. Se trata de un campo multidisciplinar, con importantes aportaciones estadísticas e informáticas, dentro del que se incluirían disciplinas como Minería de Datos (Data Mining), Aprendizaje Automático (Machine Learning), Aprendizaje Profundo (Deep Learning), Modelado Predictivo (Predictive Modeling), Extracción de Conocimiento (Knowlegde Discovery) y, también, el Aprendizaje Estadístico (Statistical Learning).
Se podría definir la Ciencia de Datos como el conjunto de conocimientos y herramientas utilizado en las distintas etapas del análisis de datos. Esta ciencia incluiría también la gestión (sin olvidarnos del proceso de obtención), y la manipulación de los datos.
La elección del tema propuesto surge de la convergencia de las áreas de Ciencia de Datos y Lógica Difusa, ámbitos que son de especial interés en la actualidad debido a que, sea para descubrir tendencias, patrones en los datos o modelar una realidad, para resolver un problema de predicción o clasificación, se necesitan manipular datos de distinta naturaleza e incertidumbre.
Siguiendo con las propuestas en el área de análisis de datos sensorial, propuestas en la tesis para optar al título de Doctor en Ciencia de los Alimentos, (Césari, 2018), le brindará al tesista la posibilidad de indagar metodologías, para crear herramientas computacionales que permitan utilizar este enfoque en el análisis exploratorio de datos, no solo en estudios con datos sensoriales, sino también en ingeniería, para cualquier tipo de estudio, a partir de diversos tipos de datos.
En el marco del área de Ciencia de Datos cuando se habla tanto de la construcción de modelos o del estudio exploratorio de los datos imprecisos, se presentan restricciones, puestas en evidencia por Ávila-de Hernández *et al.* (2011).
La presencia de variables que presentan elevado número de valores perdidos, implica, por un lado, pérdida de eficiencia en el análisis y, por otro lado, frente a la posibilidad de que los valores perdidos sigan un patrón no aleatorio, tanto ignorarlos como estimarlos, mediante alguno de los sistemas de imputación, implica un sesgo, inconveniente en la obtención de patrones mediante el análisis factorial. Barda (2011) señala que los valores medios y la dispersión alertan de

posibles deficiencias de los datos observados, en relación con problemas de la realidad estudiada, e incluso la propia naturaleza subjetiva de las mediciones.
El conjunto de las técnicas de recolección y análisis de datos, generados en la evaluación sensorial, en los estudios de análisis sensorial de alimentos, constituye una rama de la estadística llamada sensometría. Si se tiene en cuenta la vaguedad e incertidumbre con la que se manifiestan las percepciones humanas, la estrategia más apropiada para obtener conocimiento preciso a partir de valoraciones subjetivas, consiste en transformar los datos sensoriales mediante la matemática borrosa [Cesari M *et al.,* 2018].
No obstante, y a pesar del potencial que muestra el razonamiento borroso, en la actualidad existen pocas referencias y herramientas computacionales sobre la aplicación de la lógica difusa en la evaluación sensorial, y de cualquier otro estudio.
Para la gestión y superación de estas restricciones y problemas, es indispensable el empleo de métodos y herramientas robustos, como los que proveen el análisis multivariado y la aritmética borrosa, combinado con el diseño de una estrategia, que permita minimizar el efecto de valoraciones atípicas (outlier) y faltantes.
En este contexto, y dentro de la visión y alcance que se observa en las nuevas tendencias de análisis de datos, y el poder y utilidad de las diversas herramientas tecnológicas que se usan en la actualidad, para conseguir el ordenamiento y procesamiento de grandes volúmenes de información, se ha encontrado que existe una vacancia de nuevas herramientas y plataformas de análisis de datos.
La disponibilidad de una nueva herramienta es sumamente útil para analistas e investigadores que no cuenten con conocimientos específicos para desarrollarlas por su propia cuenta, y para que puedan acceder a varias funcionalidades útiles integradas en la misma, que les permita obtener gráficas o tablas, y adaptarlas a sus necesidades y requerimientos.
Esto, también, resulta en un desafío sumamente interesante para el tesista, ya que le brindará herramientas en ambas áreas, de la Ciencia de Datos y de la Lógica Difusa, para su desarrollo académico y profesional.

Metodología:

- **PRIMERAS PRUEBAS CON R Y SHINY**

Mediante una Investigación exploratoria se buscará diferentes metodologías para desarrollo de aplicaciones web, mediante métodos ágil
Se busca identificar las herramientas más efectivas y eficientes para el procesamiento de datos borrosos y la exploración de relaciones complejas entre variables en tablas de contingencia.
Se utilizará el lenguaje R y los paquetes encontrados para implementar la estrategia de conversión de los datos a números borrosos, y su posterior tratamiento para obtener conocimiento.
Mediante una Investigación exploratoria se buscará los paquetes y funciones relacionadas a la aritmética borrosa, métodos multivariados para el tratamiento de tablas de contingencia e inferencia estadística para validación estadística.

El resultado será una guía práctica para investigadores y profesionales que deseen utilizar el lenguaje R para el análisis de datos borrosos y la exploración de relaciones complejas en tablas de contingencia.
Se busca identificar las mejores prácticas y recomendaciones para el diseño y desarrollo de aplicaciones web con Shiny, así como el desarrollo de una aplicación web con inferencia con lógica difusa que permita la valoración de rúbricas de evaluación.
Mediante una Investigación exploratoria se estudiará las distintas funcionalidades y elementos del paquete Shiny para la implementación de aplicaciones web.
Se desarrollará una aplicación simple utilizando los paquetes de lógica borrosa (sets), para valoración de rúbricas en un modelo de evaluación, en particular utilizaremos modelos jerárquicos con variables lingüísticas borrosas y reglas lógica e interfaz.

- **ANÁLISIS DISEÑO E IMPLEMENTACIÓN DE LA APLICACIÓN WEB**

Asimismo, se implementará la aplicación a través del lenguaje R y el paquete Shiny, con el fin de facilitar el acceso, la visualización y el procesamiento de los datos.
Dentro de los análisis y diseño de la aplicación, está la decisión de qué metodologías ágiles de desarrollo de software convienen más para el desarrollo; por ende, se realizará una investigación por medio de Google académico, usando las palabras metodología ágil de desarrollo de software y programación dirigida por eventos.
En el diseño de sistemas implementados con programación dirigida por eventos se utilizan metodologías estructuradas. Para el desarrollo de la aplicación se seguirá una metodología que combina distintos marcos muy habituales en el desarrollo de software, cada uno afectando a una parte diferente del proceso (Torres 2018).
En primer lugar, se seguirá los principios de la metodología Agile para la gestión del proyecto y la trazabilidad del trabajo. Se trata de una metodología muy extensa y que no aplica en su totalidad a un proyecto de este calado, pero se tomarán algunos de sus principios muy en cuenta. Trabajar así supone seguir un proceso iterativo, y con la intención de emitir progresos y desarrollos en incrementos pequeños, pero asumibles. Así, se establecerán sprints de dos semanas en los que en cada uno se desplegará al menos una funcionalidad o parte de la aplicación. Los resultados se evaluarán con un alto estándar de calidad técnica y diseño.
De entre las diferentes metodologías o lenguajes de diseño que existen en la actualidad se analizará el UML (lenguaje de modelado unificado) UML (Unified Modeling Language). Es un lenguaje que sirve para especificar,visualizar, construir y documentar sistemas de software (Larman 2002).

El UML es un lenguaje de modelado gráfico que usa una variedad de elementos visuales para mostrar los elementos semánticos de manera que sean fácilmente aprovechables y manipulables por un modelador para especificar o describir métodos o procesos.

Se utilizará Git y Github muy extensamente para la trazabilidad de los cambios producidos durante el desarrollo. Estas herramientas permiten tener visibilidad de todos los cambios, mantener la gobernanza sobre el código y la posibilidad de publicarlo en un repositorio remoto

Se busca crear una imagen en Docker sobre Ubuntu, de tal forma que acciones como la instalación de paquetes, la configuración de aspectos del sistema operativo, o la instalación del mismo servicio que permite publicar una aplicación Shiny, se encuentren todos almacenados bajo una imagen en Docker, y asegurarse que el entorno es replicable de tal forma que la aplicación se encuentre disponible en cualquier momento que la imagen de Docker se ejecuta sobre cualquier otra máquina con Ubuntu

Luego del desarrollo de la parte visual y lógica de la aplicación web, así como de preparar el conjunto de datos que se carga por defecto para que la información esté limpia, se procederá a la publicación de la aplicación en un servidor web. Como la mejor forma para llevar a cabo las pruebas con los investigadores es a través de la publicación de la aplicación en algún servidor web que permita el acceso fácil a los investigadores desde cualquier parte; con el conocimiento de que la el laboratorio ReAVi posee un servidor web que se puede utilizar para publicar la aplicación y que ésta sea accesible desde internet, se preparará la aplicación para que pueda ser publicada en el entorno del servidor

El servidor, además, contará con el servidor de Shiny instalado como servicio, de forma que la publicación de la aplicación consistirá en compartir el repositorio público de GitHub con el administrador del servidor para que descargue el código de la aplicación al servidor y aloje el código en el directorio que el servidor de Shiny utiliza para la publicación de aplicaciones.

Considerando que puede haber personas que deseen probar el código de la aplicación pero que no tengan ya instalado un servidor Shiny, por lo que para probar la aplicación fácilmente necesitan antes haber realizado la instalación y configuración del servicio de Shiny. Como aporte adicional a este proyecto de tesis, se creará una imagen con Docker para el sistema operativo Debian, con el propósito de dar la posibilidad de que una persona pueda descargar dicha imagen y, con la ejecución de un par de comandos en la línea de comandos de Linux, tenga la aplicación funcionando.

Conclusiones

Las aplicaciones Shiny/R, proporciona un modelo de distribución sin carga y de código abierto que representan un aporte significativo para los desarrolladores de software en cualquier ámbito

Citas

- Ávila de Hernández, R. M.; González-Torrivilla, C. C. 2011. La evaluación sensorial de bebidas a base de fruta: una aproximación difusa. Universidad, Ciencia y Tecnología. 15 (60): 171-182
- Barda, N. 2011. Análisis sensorial de los alimentos. Fruticultura & Diversificación. Disponible en: http://www.biblioteca.org.ar/libros/210470.pdf
- César, M. I., Ventrera, N. B., & Gámbaro, A. (2018). Análisis de datos sensoriales de tomate triturado con lógica difusa y técnicas multivariadas. Revista de la Facultad de Ciencias Agrarias. Universidad Nacional de Cuyo, 50(1), 233-248.
- Docker. (s/f). Dockerfile reference. Docker docs. Recuperado el 2 de abril de 2021, de https://docs.docker.com/engine/reference/builder/ Docker Inc. (s/f).
- Larman, C. (2002) Applying UML and Patterns: An Introduction to Object-Oriented Analysis and Design and the Unified Process, Second Edition. Prentice Hall, Englewood Cliffs, NJ
- Torres, M. 2018, Pasos para diseñar una aplicación web exitosa