UNIVERSIDAD DE GRANADA

GRADO EN ESTADÍSTICA
PROPUESTA DE TEMA PARA TRABAJOS FIN DE GRADO
CURSO ACADÉMICO 2019/2020

Título del trabajo: Análisis de las herramientas de Machine Learning en el entorno estadístico de R: Comparativa experimental y casos de uso.

Tutor/a: Alberto Fernández Hilario

Cotutor/a:

Departamento responsable: Departamento de Ciencias de la Computación e Inteligencia Artificial

Perfil y número de estudiantes al que va dirigido (máximo 2):

(Rellenar sólo en caso de que la propuesta esté realizada a través de estudiante)

Estudiante que propone el trabajo (Nombre, Apellidos, DNI): José Manuel, Guzmán García,

Tipo de trabajo *8, 9, 1*

Competencias

Competencias generales: G02, G03, G04, G05, G06, G07, G08, G09.

Competencias específicas: E01, E02, E03, E04, E05, E08, E09, E10.

Resultados de aprendizaje

- Adquirir competencias globales ligadas al desarrollo y aplicación de los conocimientos del Grado.
- Adquirir competencias ligadas a la búsqueda y organización de información y documentación relevante sobre el tema objeto de estudio.
- Aplicar el "pensamiento estadístico" y tener capacidad para enfrentarse a las distintas etapas de un estudio estadístico (desde el planteamiento del problema hasta la exposición de resultados).
- Saber presentar, de forma escrita y oral, la memoria, los resultados y las conclusiones del trabajo realizado.

Antecedentes y resumen del tema propuesto:

En la actualidad, se maneja cada vez más información en formatos no estructurados o semiestructurados. Ejemplos claros son los mensajes de correo electrónico, notas de los centros de servicio al cliente, respuestas de encuestas con final abierto, fuentes de noticias, o formularios web, entre otros. Esta abundancia de información se presenta como un problema para muchas empresas a la hora de preguntarse cómo recopilar, explorar y aprovechar el conocimiento oculto detrás de esa cantidad ingente de datos.

Este es uno de los motivos por el que las técnicas basadas en Machine Learning se han vuelto tan populares tanto en academia como en industria. Consiste en utilizar algoritmos basados en inteligencia artificial para poder extraer conocimiento de manera automática a partir de grandes colecciones de datos. En concreto, el proceso completo de extracción de conocimiento en Ciencia de Datos incluye, entre otros, identificar la información sobre la cual se desea realizar la minería, preprocesar los datos para adaptarlos al aprendizaje, aplicar los algoritmos de modelado, y contrastar los resultados con herramientas de visualización.

Para agilizar el procedimiento anteriormente mencionado, se han desarrollado numerosos paquetes o herramientas en el software R, tales como "caret", "h20", "DMwR", "SuperLearner", o "mlr" entre otros. Pese a que esto se podría como una clara ventaja, son tantos los paquetes disponibles, y muchos de ellos incluyendo únicamente un número limitado de técnicas, que es muy complejo saber por dónde comenzar para un investigador interesado en la temática.

UNIVERSIDAD DE GRANADA

GRADO EN ESTADÍSTICA
PROPUESTA DE TEMA PARA TRABAJOS FIN DE GRADO
CURSO ACADÉMICO 2019/2020

De acuerdo a lo anterior, este Trabajo Fin de Grado tiene como finalidad la de realizar un amplio estudio comparativo sobre los principales paquetes y herramientas disponibles en R. Para ello, se realizará una descripción completa de los mismos, se analizará los algoritos implementados a todos niveles (preprocesamiento, modelado y visualización), sus principales ventajas e inconvenientes, y se determinarán diversos casos de estudio para contrastar la bondad de cada una de las soluciones desarrolladas.

Breve descripción de las actividades presenciales y no presenciales a realizar:

Actividades presenciales (15-30%)	Planteamiento, orientación y supervisión	50 horas
	Exposición del trabajo	1 horas
	Otras:	25 horas
Actividades no presenciales (70-85%)	Preparación del trabajo	160 horas
	Elaboración de la memoria	14 horas
	Otras:	50 horas
Total (12 ECTS)		300 horas

Objetivos que se pretenden alcanzar:

- Estudio y descripción del procedimiento de Ciencia de Datos en general y el uso de algoritmos de Machine Learning en particular.
- Estudio y aplicación de diferentes paquetes de Machine Learning disponibles para el software estadístico R. Realizar una descripción profunda de las distintas alternativas, definiendo sus pros y sus contras.
- Estudio comparativo de los paquetes de Machine Learning de acuerdo a diferentes casos de estudio planteados. Presentación de los resultados y conclusiones extraídos a partir del análisis realizado.
- Redacción de una memoria completa destacando los principales hitos alcanzados.

Bibliografía básica para la puesta en marcha del trabajo:

- Cathy O'Neil and Rachel Schutt. Doing Data Science. O'Reilly Media, 2013
- Scott V. Burger.Introduction to Machine Learning with R. Rigorous Mathematical Modeling. O'Reilly Media 2018 (1st Edition)
- Michael R. Berthold, Christian Borgelt, Frank Hppner, and Frank Klawonn. 2010. Guide
- T. Hastie, R. Tibshirani, J. Friedman. The elements of statistical Learning. Data Mining inference and prediction. 2001 Springer (2nd Edition)
- V. Cherkassky, F.M. Mulier Learning from Data: Concepts, Theory, and Methods, 2nd Edition Wiley-IEEE Press, 2007
- G. Grolemund, H. Wickham. R for Data Science: Import, tidy, transform, visualize and model data. 1st Edition. O'Reilly (2016)
- B. Lantz. Machine Learning with R: Expert techniques for predictive modeling to solve all your data analysis problems, 2nd Edition. PACKT (2015)
- Scott V. Burger. Introduction to Machine Learning with R. O'Reilly (2018)

UNIVERSIDAD DE GRANADA

GRADO EN ESTADÍSTICA
PROPUESTA DE TEMA PARA TRABAJOS FIN DE GRADO
CURSO ACADÉMICO 2019/2020

Tipo de trabajo (*):

1. Estudio de profundización en algún tema concreto de Estadística, o como proyecto de aplicación de la misma a estudios o problemas de otros ámbitos científicos o sociales.
2. Realización completa de todas las fases de un proyecto estadístico, bien con auxilio de prácticas en empresas o con prácticas propuestas y dirigidas por el tutor.
3. Estudio de casos, teóricos o prácticos, relacionados con la Estadística.
4. Elaboración de un informe o un proyecto de naturaleza profesional.
5. Elaboración de un plan de empresa.
6. Simulación de encargos profesionales.
7. Trabajos bibliográficos sobre el estado actual de una temática relacionada con la Estadística.
8. Creación y/o empleo de herramientas informáticas para su uso en Estadística.
9. Trabajos de inicio a la investigación.
10. Trabajos cuya finalidad sea la divulgación de la Estadística en diversos contextos.
11. Trabajos sobre Historia de la Estadística.
12. Trabajos relacionados con la docencia de la Estadística.

Competencias (**)

Competencias generales:

G01. Poseer los conocimientos básicos de los distintos módulos que, partiendo de la base de la educación secundaria general, y apoyándose en libros de texto avanzados, se desarrollan en la propuesta de título de Grado en Estadística que se presenta.

G02. Saber aplicar los conocimientos básicos de cada módulo a su trabajo o vocación de una forma profesional y poseer las competencias que suelen demostrarse por medio de la elaboración y defensa de argumentos y la resolución de problemas dentro de la Estadística y ámbitos en que esta se aplica directamente.

G03. Saber reunir e interpretar datos relevantes para emitir juicios que incluyan una reflexión sobre temas relevantes de índole social, científica o ética.

G04. Poder transmitir información, ideas, problemas y sus soluciones, de forma escrita u oral, a un público tanto especializado como no especializado.

G05. Haber desarrollado aquellas habilidades de aprendizaje necesarias para emprender estudios posteriores con un alto grado de autonomía.

G06. Saber utilizar herramientas de búsqueda de recursos bibliográficos.

G07. Poder comunicarse en otra lengua de relevancia en el ámbito científico.

G08. Poseer habilidades y aptitudes que favorezcan el espíritu emprendedor en el ámbito de aplicación y desarrollo de su formación académica.

G09. Fomentar y garantizar el respeto a los Derechos Humanos, a los principios de accesibilidad universal, igualdad, y no discriminación; y los valores democráticos, de la cultura de la paz y de igualdad de género.

Competencias específicas:

E01. Conocer los fundamentos básicos del razonamiento estadístico, en el diseño de estudios, en la recogida de información, en el análisis de datos y en la extracción de conclusiones.

E02. Conocer, saber seleccionar y saber aplicar, técnicas de adquisición de datos para su tratamiento estadístico.

E03. Conocer los fundamentos teóricos y saber aplicar modelos y técnicas estadísticas en estudios y problemas reales en diversos ámbitos científicos y sociales.

E04. Saber seleccionar los modelos o técnicas estadísticas para su aplicación en estudios y problemas reales en diversos ámbitos científicos y sociales, así como conocer herramientas de validación de los mismos.

E05. Comprender la importancia de la Investigación Operativa como metodología de optimización, toma de decisiones y diseño de modelos particulares para la resolución de problemas en situaciones específicas.

E06. Comprender y utilizar básicamente el lenguaje matemático.

E07. Conocer los conceptos y herramientas matemáticas necesarias para el estudio de los aspectos teóricos y prácticos de la Probabilidad, la Estadística y la Investigación Operativa.

E08. Conocer y saber utilizar aplicaciones informáticas de análisis estadístico, cálculo numérico y simbólico, bases de datos, visualización gráfica y optimización, que sean útiles para la aplicación y desarrollo de las técnicas estadísticas.

E09. Conocer los conceptos básicos y habilidades propias de un ámbito científico o social en el que la Estadística o la Investigación operativa sean una herramienta fundamental.

E10. Tomar conciencia de la necesidad de asumir las normas de ética profesional y las relativas a la protección de datos y del secreto estadístico, como premisas que deben guiar la actividad profesional como profesionales de la Estadística.