INSTITUTO DE

SYLLABUS ACADEMICO

POSTITULO INTRODUCCION AL DATA SCIENCE PARA EL SECTOR PUBLICO

2021

ASIGNATURA : **Clustering**
NOMBRE PROFESOR : **Adrian Armando Araneda Toro.**
EMAIL : **adrianaranedat@ug.uchile.cl**

NOMBRE PROFESOR : **Alex Meléndez Suazo.**
EMAIL : **alexmelend@gmail.com**

1. INTRODUCCIÓN

En diversas carteras y sectores de las instituciones y organizaciones que componen la administración del estado (y por cierto, de la sociedad civil), para la buena gobernanza y gobernabilidad se recogen día a día grandes volúmenes de datos, en diversas escalas y magnitudes. Hoy en día, la creación e innovación de valor público no solo se encuentra en estos yacimientos de datos. Estos se han convertido poco a poco en el activo estratégico principal de las instituciones para eficientar a través de, por ejemplos, procesos automatizados, el monitoreo, seguimiento, análisis exploratorio, y prospectiva adecuada de fenómenos para la prevención de diversos escenarios y costes en la provisión eficaz de servicios y bienes, tanto para la ciudadanía como para los usuarios o clientes internos.

¿Colectar y administrar grandes volúmenes de datos es suficiente?

Para colocar en valor y dominio toda la data recolectada, las instituciones requieren acelerar la incorporación de competencias en sus funcionarios, nuevas tecnologías y herramientas, así como establecer una nueva cultura basada en los datos, para así responder a la creación o formación de ese nuevo funcionario: "Data tecnócrata y/o burócrata" ("Data Technocrat Scientist" or "Data Bureaucrat Scientist").

Por lo general, poseer la mejor y más moderna infraestructura tecnológica se veía como una forma de lograr una ventaja estratégica y comparativa "en el mundo antiguo"[i], sin embargo, luego se demostró que no es efectivo, ya que la infraestructura tecnológica hoy en día es un commodit y que todas las empresas y organizaciones pueden y deben tenerlo, por lo tanto, la infraestructura tecnológica y hardwares (por ejemplo computadores, servidores, data warehouse, etc), no es algo exclusivo de una empresa u organización, ergo no generan una ventaja estratégica o comparativa o característica ventajosa de negocio. El segundo paso. Luego de superar el sesgo de poseer la "mejor infraestructura y harwares" fue la recolección de datos, ya que poseer datos es poseer información o "poder"[ii]. Pero nuevamente, se observa que todas las empresas y organizaciones, también colectan datos, y por lo tanto, tampoco es exclusivo y no confiere liderazgo en el mercado o sector a una empresa u organización de manera a priori.

Para abordar estas tareas, se requiere previamente de **nuevas técnicas** para analizar inteligentemente grandes volúmenes de datos, técnicas exploratorias, directamente tareas tales como el razonamiento bajo incertidumbre, identificación de patrones de comportamiento, representación no supervisada de un fenómeno conocido como desconocido. Optimización, predicción de fenómenos, identificación de factores determinantes en un negocio, área o sector. Detección de tendencias, segmentación y caracterización de grupos, etc. Convirtiéndolos así en **conocimiento valioso, activo, inversión y no en un pasivo o bien depreciable.**

Para todo lo anterior, el modulo de **"Clustering"** entrega los conocimientos necesarios para que el alumno adquiera técnicas introductorias de caracterización de datos, poder profundizar en aquellas.

Es imprescindible para cualquier organización o negocio el conocer a sus clientes o usuarios, sean estos internos o externos, para así responder oportunamente a las demandas que ellos requieren. **Para esto la caracterización** de estos clientes, usuarios, ciudadanos, agentes (o en otros contextos, objetos) **es fundamental para estudiar su comportamiento presente o a través del tiempo, sea para entender** su conducta o predecir su comportamiento. Las técnicas de Clustering responden a la forma en que se caracterizan a esos clientes o usuarios, y cuál sería la mejor técnica (óptimos) para representar aquellas conductas o comportamientos. Enfatizar, que no responde a un objetivo predictivo o supervisado.

2. OBJETIVO GENERAL DE LA ASIGNATURA

Adquirir y desarrollar las habilidades, metodológicas y teóricas introductorias, para la comprensión o el entendimiento de un problema o fenómeno del Mundo, o Universo No Supervisado, y por ende utilizar técnicas de caracterización (clusterización) de grandes volúmenes de datos, que permitan a los estudiantes resolver problemas donde se es necesario **entender de manera diferenciada** y especificada el comportamiento de un universo o muestra de agentes u objetos.

Para lo anterior es apremiante que los alumnos puedan entender y diferenciar cognitivamente los conceptos de aprendizaje supervisado contra el aprendizaje no supervisado, para así distinguir que características debe tener un fenómeno a estudiar para abordarlo desde un enfoque predictivo versus un enfoque exploratorio, aplicando dicho enfoque y posterior análisis a través de la utilización de una herramienta (motor de datos) adecuada para aquello.

3. OBJETIVOS ESPECÍFICOS DE LA ASIGNATURA

Al finalizar del modulo el alumno estará capacitado para:

- Aprender a utilizar un script básico en RStudio para el estudio introductoria de clustering y optimización previa, de los mismo (y posterior también) En donde se aprenderá:

- El dominio de técnicas de clusterización de datos, que en este caso será a través de RStudio con sus respectivas librerías.

- Aplicar y resolver problemas con diferentes algoritmos, para la Clusterización, esencialmente los algoritmos más frecuentemente utilizados, K-Means, Pam y Herarquichal.

- Aplicar técnicas de selección del "óptimo" número de particiones o lusters.

- Evaluar el algoritmo o técnica de clustering apropiada para un fenómeno determinado a estudiar.

- Visualización o ploteo de las Técnicas de Clusterización.

- Interpretar los resultados de las técnicas de clusterización para grandes volúmenes de datos y sus óptimos.

- Estrategias de selección y rendimiento: Ventajas y Desventajas de los algoritmos.

4. METODOLOGIA

- Este módulo está diseñado para aprender, implementar y utilizar, diversas herramientas de segmentación de datos, haciendo uso de métodos pedagógicos, que incluyen: clases exponenciales, discusiones interactivas, elementos audiovisuales y ejercicios prácticos en clases. El profesor desarrollará clases expositivas donde mostrará, entregará y explicará aspectos teóricos y prácticos de la materia en estudio. Los contenidos serán entregados de una forma sistemática. Además, mediante el desarrollo de ejercicios prácticos en conjuntos (Script) y individuales y grupales, se motivará el autoaprendizaje y el estudio continuo. Las clases expositivas estarán apoyadas con lecturas de artículos de revistas y trabajos de investigación si el profesor lo requiriese. Además, se destinará tiempo para discutir y guiar las actividades desarrolladas.

- Para sacar el mayor provecho al curso, el alumno debe invertir horas de autoestudio y/o trabajo en equipo/grupo.

- El profesor desarrollará clases expositivas donde mostrará, entregará y explicará aspectos teóricos y prácticos de la materia en estudio, ejercitando las sintaxis y códigos, con ellos.

- Los contenidos serán entregados de una forma sistemática.

- Además mediante el desarrollo de ejercicios prácticos en conjuntos, individuales y grupales se motivara el autoaprendizaje y el estudio continuo.

- Se destinará tiempo para discutir y guiar los ejemplos desarrollados.

- **Se utilizará una metodología de taller, activo-participativa que articula clases expositivas con realización de actividades en ordenadores.**

Para la correcta implementación de la metodología, el alumno deberá asegurarse de poseer:

- El alumno necesitará una conexión estable a Internet sea por wifi o punto de red.

- El hardware con el que deberá contar el estudiante es un Notebook o pc ordenador.

- La sesión o perfil que utilice el alumno en dicho notebook u ordenador debe estar liberada, desbloqueada, o contar con privilegios de "administrador de equipo" o contar con los permisos correspondientes para no impedir la instalación de softwares nuevos y su libre utilización.

- En el mismo sentido que lo anterior, los respectivos antivirus que se encuentren instalados en los ordenadores también deben contar con la configuración correspondiente para no presentar un corta fuego o bloqueo que interrumpa la instalación de un nuevo software y su uso con normalidad.

- Así también, los equipos de los alumnos deberán encontrarse sin problemas de rendimiento, velocidad, procesamiento, problemas en la tarjeta de video o gráfica, discos duros copados o llenos, problemas de memoria, etc.

- Se facilitará a los alumnos con anticipación al módulo 4, los manuales y/o tutoriales correspondientes para dos sistemas operativos (Windows y Mac), para la instalación liberada del softwar que ocuparemos y que será imprescindible para este diplomado, RStudio. Software que se utilizará para cada alumno, especialmente para este módulo 10.

- Previamente al comienzo de este módulo el profesor enviará a los alumnos a través de correo electrónico, las instrucciones y la versión del software de programación R, denominada, versión "R-4.0.4-win " para que sea instalada por el alumnado. Sin importar la versión de la consola o interfaz "RStudio" que posea.

- Los Conocimientos requeridos para la comprensión y avance del Módulo 10 con los que deberán contar los alumnos, son: La gestión Pública y Ética en los Datos, Base de datos e Introducción al Machine Learning, Estadística Descriptiva e Inferencial, Manipulación de Base de Datos e Introducción a Rstudio, Fundamentos de Programación y Automatización, Análisis Exploratorio de Datos, Visualización, Introducción a la econometría y regresiones y Modelamiento Predictivo. Es decir, los módulos que precedieron.

5. CONTENIDOS DEL MODULO

- Presentación del Syllabus.
- La revolución de los datos, Big Data.
- Diferencia entre aprendizaje supervisado y no supervisado (Análisis Exploratorio).
- Caracterización versus Clasificación.
- Tipologías de Clustering: Partitioning Clustering, Hierarchical Clustering y Combinados.
- Medidas de Distancia (Euclidea y Manhattan).
- Evaluación del Algoritmo de clustering apropiado para la Master Data en estudio: Óptimo de K.
- Introducción a la evaluación de Cclusters.
- Determinación del Número de Clusters.
- Algoritmo K-Medias
- Algoritmo PAM

- Algoritmo Hierarchical
- Algoritmo DBSCAN.
- Visualización de Clusters.
- Visualización animada de Clusters.
- Enfoque basado en clustering para el tránsito del mundo no supervisado al supervisado a través de las etiquetas de salida o de control.

6. EVALUACION

- La asistencia tiene una obligatoriedad del 80%.

- La evaluación del módulo corresponde a un control teórico y práctico individual.

- No obstante, el profesor podrá considerar la adición de una segunda evaluación

- La evaluación(es) se expresará(n) en una escala de 1,0 (uno y cero décimas) como mínimo a 7,0 (siete y cero décimas) como máximo.

- Nota final mínima para aprobar el curso : 4,0 (cuatro y cero décimas).

- La(s) Fecha(s) de la Evaluación(s) a entregar así como su formato de entrega se anunciarán y especificará(n) oportunamente por el profesor.

- Las fechas serán definitivas.

Escenario 1:

NOTA FINAL = Evaluación Individual 100%

Escenario 2:

NOTA FINAL = Evaluación Individual 1 x 50% + Evaluación Individual 1 x 50%

Tanto para el escenario 1 como para el escenario 2 se subirán al sistema las rubricas correspondientes.

En caso de reprobar (Nota Final < 3.95), el alumno tendrá derecho a una instancia de examen recuperativo, a fecha a estipular por la Coordinación Académica o profesor, optando a la calificación mínima de aprobación del curso.

Si el alumno, después de realizar el examen recuperativo, no cumple con alguno de estos dos requisitos, tendrá que cursar la actividad curricular en una segunda oportunidad cuando se dicte nuevamente el curso.

INASISTENCIA A EVALUACIONES

En caso de que el alumno no pueda asistir a una evaluación, deberá justificar con certificado médico o laboral y tendrá derecho reemplazar la nota por la obtenida en el examen recuperativo.

Si el alumno se ausenta al examen de repetición, éste quedará pendiente hasta la próxima ocasión en que la asignatura o módulo se dicte.

7. BIBLIOGRAFÍA

- Contenidos y Apuntes de las Presentaciones Vistas en Clases.

- clValid , an R package for cluster validation. Department of Bioinformatics and Biostatistics, University of Louisville. Guy Brock, Vasyl Pihur, Susmita Datta, and Somnath Datta.

- Sesgos Cognitivos. Recomendaciones de Política para guiar reformas Económicas en América Latina: Una Mirada desde la Moderna Teoría del Comportamiento Económico. Juan Carlos Lerda. Doctor en Economía, Universidad de Harvard, Estados Unidos.

7. CURRICULUM RESUMIDO DEL PROFESOR

- Título profesional: Administrador Público. Universidad de Chile (2011)

- Grado Académico: Licenciado en Ciencias Políticas y Gubernamentales. Mención Gestión Pública. Universidad de Chile (2009)

- Postítulos:

  - Diplomado en Inteligencia de Negocios y Minería de Datos. Facultad de Ciencias Físicas y Matemáticas. Universidad de Chile (2017).

  - Diplomado en Análisis Masivo de Datos. Escuela de Gobierno. Universidad Adolfo Ibáñez (2018-2019).

  - Diplomado en Data Sciencie. Facultad de Ingeniería. Universidad Adolfo Ibáñez (2019).

- Master en Data Science. Facultad de Ingeniería. Universidad Adolfo Ibáñez (2019-2021 en curso).

- Selección Paper "Modelo Microeconométrico" para BAFI 2020 para Investigadores y Desarrolladores de la ciencia de los datos. Facultad de Ciencias Físicas y Matemáticas. Universidad de Chile.

  - https://drive.google.com/file/d/1ORIVFc9Vwosnyd2V0DpziOaRGIT-jMZD/view?usp=sharing

  - https://baficonference.cl/20/default/inicio

- Lugar de Trabajo Actual: Equipo de Análisis de Datos. Unidad de Análisis de Declaraciones de Intereses y Patrimonio. División de Auditoría. Contraloría General de la República.

- Horario de Atención: Previa coordinación por correo electrónico.

- Lugar de Atención: Previa coordinación por correo electrónico.

- E-mail: adrianaranedat@gmail.com

- Áreas de Investigación: Data Science. Modelo Micro Econométrico. Modelos Predictivos para la detección de Anomalías Patrimoniales. Paper BAFI 2020-

Profesor Alex Meléndez Suazo. Ingeniero Civil Industrial, Licenciado en Ingeniería Aplicadas, Universidad de Santiago de Chile. Magíster Economía Financiera, Universidad de Santiago de Chile.

- Profesor Ayudante de Econometría 1 y 2 para el Magister en Economía Financiera de la Universidad de Santiago de Chile.

- Experiencia profesional en Telecomunicaciones "Claro Chile, Mercado Empresarial", y en consultorías técnicas "Contac Ingenieros" e "Invensys". Actualmente miembro del equipo de Análisis de Datos de la Unidad de Análisis de Declaraciones de Intereses y Patrimonio de Contraloría General de la República.

8. SESIONES Y FECHAS

SESION 1:	**JUEVES 05-08-2021**	**18.30 – 21.45 HORAS**

SESION 2:	**SABADO 07-08-2021**	**09.00 – .14.15 HORAS**

[i] Se utiliza esta frase para referirse a condiciones de "subdesarrollo" que influyen en la cultura en la valoración del uso de los datos (mal uso por ejemplo).

[ii] El poder de la información en el contexto de los Tipos de poderes y su relación con los universos simbólicos; poder coercitivo, económico y simbólico (según las referencias a Émile Durkheim y Pierre Bourdieu.