



SYLLABUS ACADEMICO
**POSTITULO INTRODUCCION AL
DATA SCIENCE PARA EL SECTOR
PUBLICO
2022**

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

1. INTRODUCCIÓN

En diversas carteras y sectores de las instituciones y organizaciones que componen la administración del estado y en general de la sociedad civil, para la buena gobernanza y gobernabilidad se recogen y administran día a día grandes volúmenes de datos en diversas escalas y magnitudes.

Hoy en día la innovación y creación de valor público se encuentra determinada en primera instancia por la adquisición de infraestructura crítica (hardware, data warehouse, servidores, sistemas, redes, etc.). Luego o en segunda instancia, se encuentra determinada por yacimientos de datos. Así, el proceso de obtención/ “excavación” de estos se ha constituido poco a poco en el activo estratégico y principal, de las instituciones para distintos fines.

Fines tales como eficientar a través de técnicas automatizadas el monitoreo y seguimiento de macros y micros procesos.

O también, para otros fines tales como la prevención, la estimación y prospectiva adecuada de fenómenos, estimación cuyo objetivo sea la no ocurrencia de estos, anteponiéndonos de esta manera a diversos escenarios indeseables favoreciendo en consecuencia la disminución de costes en la cadena de provisión eficaz de servicios y bienes que dispone la administración del estado para la ciudadanía, usuarios y clientes, etc., sean externos o internos.

Pero la excavación u obtención no lo es todo. Entonces se enarbola una tercera instancia, ante sala o requisito; ¿con sólo coleccionar y administrar grandes volúmenes de datos es suficiente?

Para colocar en valor y dominio toda la data captada y almacenada, y tal como los estudiantes de este programa ya lo han estudiado en módulos anteriores, las instituciones requieren i) acelerar la incorporación de competencias en sus funcionarios y trabajadores, y ii) establecer una nueva cultura basada en la gobernanza de datos.

Una vez instaladas estas dos condiciones se estará en frente o en una situación potencialmente favorable para la aparición, creación o formación de un nuevo funcionario que requieren las complejidades de la realidad social con sus respectivos sistemas. Que requiere con apremio la crisis de legitimidad de nuestras instituciones. Que requiere la ciudadanía y las nuevas dimensiones posmodernas del tratamiento y procesamiento de la información, las comunicaciones y los sistemas tipo metaverso; es decir, para **cualquier disciplina** que trabaje con grandes volúmenes de datos contar con trabajadores o funcionarios ad hoc, un "Data tecnócrata y/o tecnoburócrata". Data Technocrat Scientist", "Data Bureaucrat Scientist" or "Bureaucrat Tecnoanalytic.

Repasando entonces. Como un primer paso o instancia, poseer la mejor y más moderna infraestructura tecnológica se veía como una forma de lograr una ventaja estratégica y comparativa “en el mundo antiguo”, en el mundo de los primeros 60 años posteriores a la revolución industrial. Sin embargo, luego se demostró que tal premisa no es efectiva ya que la infraestructura tecnológica hoy en día es un commodity que todas las empresas y organizaciones pueden y deben tener.

Por lo tanto, la existencia de una arquitectura de datos, tecnologías y hardwares (por ejemplo, computadores, servidores, data warehouse, etc), no es un componente exclusivo de una empresa u organización. Ergo no generará por sí sola una ventaja estratégica comparativa o valor agregado en el negocio o en algún equipo de trabajo X.

Así, un segundo paso o instancia fue la recolección de datos ya que poseer datos en el mundo de la información y globalizado es poseer “poder”. No obstante, y como se dijo, nos encontramos que si esta condición se perpetúa en el tiempo deja de ser relevante y se transforma en insuficiente, ya que todas las empresas y organizaciones también colectan datos. Por lo tanto, tampoco es atributo de exclusividad, no confiere liderazgo ni crea por sí sólo valor en el mercado, sector, industria, empresa, entidad, institución u organización determinada.

Por ende, como tercer paso, camino o estadio, se ha requerido continuamente el aprendizaje e instalación de nuevas técnicas para **cualquier nicho o disciplina** que trabaje con grandes volúmenes de datos, para ser analizados inteligentemente, convirtiéndolos así en conocimiento valioso y/o activo estratégico. De esta manera, agregando valor al negocio. En definitiva, se constituye en una inversión. De lo contrario, lo inerte se convierte rápidamente en un pasivo o bien depreciable.

Para todo lo anterior, el módulo de **“Clustering”** entrega los conocimientos necesarios para que el alumno adquiera competencias y **técnicas introductorias** para la caracterización desagregada de personas, objetos y/o fenómenos para posteriormente poder profundizar en la constitución de estos **y su focalización**.

Es imprescindible para cualquier organización o negocio **el conocer a sus clientes, ciudadanos y/o usuarios**, sean estos internos o externos, para así responder oportunamente a las demandas que ellos requieren y hacer buen uso (inteligente) de los recursos. **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 sea para luego predecir su comportamiento.

Las técnicas de Clustering responden I. a la forma en que se caracterizan a clientes, usuarios u objetos, **y II., al estudio de cuál sería la mejor técnica** (óptimos y distancias aritméticas) para representar aquellas conductas o comportamientos.

Dado lo indicado, cabe enfatizar que estas técnicas que se enseñarán y practicarán no responden a un objetivo predictivo o supervisado. **Para lo anterior, este curso es inherentemente introductorio, práctico y aplicado. Demanda la participación sistemática y activa del alumno en clases; “aprender-haciendo”**. Incluye demostraciones metodológicas, aplicados y talleres prácticos durante todo el desarrollo del módulo si el docente así lo requiriese, incorporando ejercicios, interrogatorios y los elementos conceptuales inherentes para tales fines. Pero más importante aún, requiere que el alumno haga uso de sus 7 hrs. no sincrónicas de estudio individual.

También requiere de la participación del alumno fuera de clases o fuera de las horas sincrónicas a partir de las plataformas formales e institucionales que dispone la Facultad de Gobierno (ex INAP) y la Universidad de Chile, denominada U-cursos, en los plazos determinados e informados por el profesor para la resolución de dudas y consultas; se abre el plazo de dudas desde que comienza el módulo hasta 18 horas antes de la evaluación.

Este módulo permitirá que el alumno amplíe su nivel de abstracción. Y requiere imprescindiblemente que el alumnos haya aprobado los módulos anteriores para cumplir los objetivos de aprendizaje y otros, tanto actitudinales como aptitudinales.

2. OBJETIVO GENERAL DE LA ASIGNATURA

Adquirir y desarrollar las habilidades metodológicas y teóricas introductorias para la comprensión y entendimiento de un problema o fenómeno del Mundo o Universo "No Supervisado", y por ende, la adquisición de técnicas analíticas introductorias de caracterización (clusterización) para grandes volúmenes de datos que permitan a los estudiantes resolver problemas donde es necesario **entender de manera diferenciada y especificada** el comportamiento de un universo o muestra de sujetos, agentes entidades u objetos para focalizar bienes, servicios y/o productos. Con el fin de observar el buen uso de los recursos y no perjudicar así la asignación de beneficios a usuarios o ciudadanos que realmente lo necesiten.

Para lo anterior es apremiante que los alumnos puedan entender y diferenciar cognitivamente los conceptos de aprendizaje supervisado contra el de aprendizaje no supervisado entre otros conocimientos y técnicas, es decir, que hayan aprobado y cumplido con los requerimientos de los módulos anteriores para así distinguir y discernir qué características debe tener un fenómeno a estudiar **para abordarlo desde un enfoque predictivo versus un enfoque exploratorio**, aplicando dicho enfoque y posteriores análisis a través de la utilización de las herramientas ad hoc (motor de datos) y técnicas adecuadas para aquello que el profesor enseñará.

3. OBJETIVOS ESPECÍFICOS DE LA ASIGNATURA

Al finalizar el módulo el alumno estará capacitado para abordar a través de un script básico en RStudio desarrollado por él, el estudio introductorio de clusters y proponer criterios técnicos de intervención o focalización. En donde al alumno adquirirá:

- 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 de clustering (los más frecuentes en su uso): K-Means, Pam y Herarquichal.
- Aplicar técnicas de selección del “óptimo” número de particiones o clusters para un fenómeno determinado en estudio.
- Evaluar el mejor algoritmo o la técnica de clustering más apropiada para un fenómeno determinado a estudiar.
- Visualización o ploteo de las Técnicas de Clusterización.
- Interpretación y limitaciones de los resultados de las técnicas de clusterización para grandes volúmenes de datos y sus óptimos.
- Estrategias de selección y rendimiento de algoritmos de clustering: Ventajas y Desventajas de los algoritmos que se enseñaran.

4. METODOLOGIA

- El módulo es principalmente práctico y consiste en 5 sesiones en que se exponen los contenidos y al mismo tiempo se llevan a la práctica con ejercicios en clases (y tareas si fuese el caso).
- El alumno “aprende haciendo”, practicando y ejercitando.
- El profesor también entregará elementos teóricos que serán consultados en la respectiva evaluación.
- El profesor entregará elementos prácticos que serán consultados en la respectiva evaluación.
- El profesor entregará Material Docente contextual que podrá ser consultada en la evaluación correspondiente.
- Para sacar el mayor provecho al curso el alumno debe invertir 7 horas de autoestudio – hacer uso de sus horas asincrónicas- del horario de clases (sincrónicas) y realizar consultas al profesor fuera de las horas sincrónicas como en ellas hasta 18 hrs. antes de la evaluación Cualquiera sea esta.
- Los contenidos serán entregados bajo un esquema sistemático a través del desarrollo de Scripts o líneas de códigos con los alumnos.
- Se destinará tiempo para discutir y guiar los ejemplos desarrollados en clases como fuera de clases, por los canales que el profesor en la primera clase indique; el cuál será preferentemente a través del correo electrónico u-cursos.

- El profesor hará preguntas en clases a los alumnos mediante un código de aleatoriedad programado por él, para obtener retroalimentación del aprendizaje y estudio continuo que las horas no sincrónicas requieren. Estas preguntas podrán ser consideradas para sumar décimas para la evaluación correspondiente.
- **Se utilizará una metodología de taller, activo-participativa que articula clases expositivas con realización de actividades en ordenadores. Ante esto es vital que los alumnos se encuentren “conectados de manera activa” y con “cámara encendida” en las clases a realizar. Esto es esencial para la participación y retroalimentación.**
- Los conocimientos requeridos para la comprensión y avance del módulo de Clustering con los que deberán contar los alumnos son: La Gestión Pública y Ética en los Datos. Neuroderechos para el uso Ético de los Datos. Introducción a la Manipulación y Procesamiento de Base de Datos con RStudio. Condicionantes y Automatización (Ciclos). Estadística Descriptiva e Inferencial. Visualización. Análisis Exploratorio de Datos. Introducción a la Econometría y Regresiones. E Introducción al Machine Learning. Es decir, los módulos que precedieron.

5. CONTENIDOS DEL MODULO

- Presentación del Syllabus.
- 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 Clusters.
- 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 de un mundo no supervisado a un mundo supervisado a través de las etiquetas de salida o de control.

6. EVALUACION

- La asistencia tiene una obligatoriedad del 75%. Esta métrica es fijada por la Facultad De Gobierno y computada automáticamente por el sistema o plataforma U-cursos. Deben conectarse a clases a través de la plataforma U-cursos. Se computará como ausente si no lo hacen desde esta plataforma.
- Se solicita a los alumnos que participen de manera activa si se encuentran conectados a las clases con el fin de sumar décimas ante una eventual pregunta sorpresa que el profesor realice aleatoriamente.
- La evaluación del módulo corresponde a una prueba individual, práctica sobre cada una de las técnicas y contenidos teóricos-contextuales vistos en clases. 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).
- La(s) Fecha(s) de la Tarea(s) o evaluación a entregar, así como el formato de entrega se anunciarán y especificará(n) oportunamente por el profesor. Y las fechas serán definitivas, así como los horarios de entrega. Ante el atraso de un minuto, se considera no recibida la evaluación.

Escenario 1:

NOTA FINAL = Evaluación Individual 100%

Escenario 2:

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

- **Tanto para el escenario 1 como para el escenario 2 se cargarán al sistema las rubricas correspondientes al momento de cargar la(s) evaluación(es).**
- En caso de reprobación (Nota Final < 4.0), el alumno tendrá derecho a una instancia de examen recuperativo a fecha a estipular por el profesor, optando a la calificación mínima de aprobación del curso. La nota del examen recuperativo se promediará con la nota de la primera evaluación reprobada.

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 a reemplazar la nota 1.0 por la obtenida en el examen de repetición o recuperación.

Si el alumno se ausenta al examen de recuperación, se entenderá como reprobado con nota 1.0.

Si el alumno reprueba este Módulo no podrá optar a la aprobación de este diplomado.

La evaluación 1 o única será cargada el miércoles 28 de diciembre a las 18.30 hrs. y se cerrará su entrega a las 23.59 hrs. No recibiendo evaluaciones a partir de las 00.00 hrs. o medianoche. Dicha evaluación deberá ser entregada respetando estrictamente el horario señalado, bajo las condiciones que el profesor indicará y tomando las providencias del caso desde el día de hoy, 09 de diciembre de 2022, con casi 20 días de anticipación, en que se le es dada esta información a través de este Syllabus.

7. BIBLIOGRAFÍA

- Contenidos y Apuntes de las Presentaciones Vistas en Clases.
- cIValid , 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.

8. CURRICULUM RESUMIDO DEL PROFESOR

- Magíster en “Data Science”. Universidad Adolfo Ibáñez. 2019-2022.
- Postítulo Diplomado en “Data Science”. Universidad Adolfo Ibáñez. 2019.
- Postítulo Diplomado en “Estadística y Análisis Masivo de Datos”. Universidad Adolfo Ibáñez. 2018-2019.
- Postítulo Diplomado en “Inteligencia de Negocios”. Facultad de Ciencias Físicas y Matemáticas. Universidad de Chile. 2017.
- Administrador Público. Escuela de Gobierno y Gestión Pública. Universidad de Chile. 2011.
- Licenciado en Cs Políticas y Gubernamentales con mención en Gestión Pública. Escuela de Gobierno y Gestión Pública. Universidad de Chile. 2009.
- Autor del Modelo “Microeconométrico”, seleccionado para su exposición en BAFI 2020 para Investigadores y Desarrolladores de la ciencia de los datos. Facultad de Ciencias Físicas y Matemáticas. Universidad de Chile. 2020.
 - <https://drive.google.com/file/d/1ORIVFc9Vwosnyd2V0DpziOaRGIT-jMZD/view?usp=sharing>
 - <https://baficonference.cl/20/default/inicio>
- Autor del Modelo de Deep Learning y Machine Learning “ALQUIMIA”: Modelo predictivo para la detección de anomalías patrimoniales y lavado de activos, actualmente en uso por el equipo de “Análisis de Datos” de la “Unidad de Análisis de Declaraciones de Intereses y Patrimonio” de Contraloría General de la República. Operativo desde diciembre de 2021 hasta la actualidad. Los primeros testeos comenzaron el año 2018-2019.

- 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. 2018-2022.
- Profesor de "Introducción a la Programación en RStudio" y "Clustering" en el postítulo "Introducción al Data Science para el Sector Público", versión 1, 2 y 3, en el Instituto de Asuntos Públicos (INAP). Universidad de Chile. 2021-2022.
- Profesor de "Introducción a la Programación en RStudio" para el Módulo de "Inteligencia Artificial" en el postítulo "Gestión de Procesos, Innovación, Excelencia Operacional e Inteligencia Artificial", en la Facultad de Ciencias Químicas y Farmacéuticas. Universidad de Chile. 2021-2022.
- Experiencia profesional en el Sector Privado como encargado del Sistema de Gestión de Calidad ISO 9001:2008, en la Red de Empresas Fidegroup S.A. 2013.
- Experiencia profesional en materia de Levantamiento de Procesos, Tratamiento de Acciones Correctivas, No Conformidades y Sociometría, para el Departamento de Estadísticas e Información de Salud (DEIS), de Ministerio de Salud. 2011-2013.
- Horario de Atención: Previa coordinación por correo electrónico.
- E-mail: adrianaranedat@ug.uchile.cl
- Áreas de Investigación Profesional: Data Science. Modelo Micro Económico Clustering. Modelos Predictivos de Machine Learning y Deep Learning para la Detección de Anomalías Patrimoniales.
- Otras Herramientas de Dominio:

Curso "WEB SCRAPING, EXTRACCIÓN DE DATOS EN LA WEB", Ciencia de Datos. Facultad de Ingeniería. Universidad de Santiago de Chile. 2021.

Curso "APLICACIÓN DE SQL", Ciencia de Datos. Facultad de Ingeniería. Universidad de Santiago de Chile. 2021.

Curso "HERRAMIENTAS DE PROGRAMACIÓN EN PYTHON". Universidad de Santiago de Chile. 2020.

Curso de Especialización: "GOBIERNO CORPORATIVO, RIESGO, PLANIFICACIÓN, SANCIONES Y FRAUDE". Facultad de Economía y Negocios, Universidad de Chile. 2019.

Curso "SOFTWARE I2" para el levantamiento de Redes Familiares y Mallas Societarias. IBM. 2018.

9. SESIONES Y FECHAS

SESION 1:	Martes 13-12-2022	20.15 – 21.15 HORAS
SESION 2:	Jueves 15-12-2022	18.30 – 21.15 HORAS
SESION 3:	Sábado 17-12-2022	09.00 – 13.30 HORAS
SESION 4:	Martes 20-12-2022	18.30 – 21.15 HORAS
SESION 5:	Martes 27-12-2022	18.30 – 19.00 HORAS