SYLLABUS ACADEMICO

POSTITULO INTRODUCCION AL DATA SCIENCE PARA EL SECTOR PUBLICO

2022

ASIGNATURA	**: Introducción al Lenguaje de Programación RStudio para la Manipulación de Data Estructurada.**
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 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, innovación y creación de valor público se encuentra, en primera instancia, determinado por estos yacimientos de datos, siendo estos poco a poco el activo estratégico principal de las instituciones para posteriormente eficientar, a través de procesos automatizados, el monitoreo, seguimiento y prospectiva adecuada de ciertos fenómenos para la prevención de diversos escenarios indeseables y por ende costes, en la provisión eficaz de servicios y bienes que dispone la administración del estado para la ciudadanía, usuarios y clientes, sean externos o internos.

En una segunda instancia; ¿Con sólo colectar y administran grandes volúmenes de datos es suficiente?

Para colocar en valor y dominio toda la data captada y almacenada, las instituciones requieren i) acelerar la incorporación de competencias en sus funcionarios, ii) como también establecer una nueva cultura basada en la gobernanza de datos. Respondiendo así a la creación o formación de este nuevo funcionario que requiere con apremio la crisis de legitimidad de las instituciones, la ciudadanía y las nuevas dimensiones posmodernas de la información y las comunicaciones, los sistemas tipo metaverso; un "Data tecnócrata y/o tecnoburócrata" ("Data Technocrat Scientist", "Data Bureaucrat Scientist" or "Bureaucrat Tecnoanalitic). En consecuencia, no bastando sólo con la adquisición de nuevas tecnologías y herramientas,

Entonces, como un primer paso (mas no perpetuo y/o estático), 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] (si se le pudiese identificar de alguna manera al pasado reciente, ligeramente posterior 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 commodit que todas las empresas y organizaciones pueden y deben tener. Por lo tanto, la existencia de una arquitectura de datos, tecnológica y hardwares (por ejemplo, computadores, servidores, data warehouse, etc), no es un componente exclusivo de una empresa u organización, ergo no generará por si sola una ventaja estratégica, comparativa o valor agregado en el negocio o en un equipo de trabajo x.

En un segundo paso (mas no perpetuo y/o estático), fue la recolección de datos ya que poseer datos en el mundo de la información es poseer "poder"[ii]. No obstante, nuevamente se 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 mismo valor en el mercado, sector, industria o en la empresa, entidad, institución u organización determinada.

En consecuencia, como tercer paso, camino o estadio, se requiere el aprendizaje e instalación de **nuevas técnicas** para analizar inteligentemente grandes volúmenes de datos para convertirlos así en **conocimiento valioso, que agrega valor al negocio y/o activo estratégico. En definitiva, en una inversión. De lo contrario será un pasivo o bien depreciable.**

Para esto, el módulo **"Introducción al Lenguaje de Programación en RStudio para la Manipulación de Data Estructurada"** entrega los conocimientos necesarios para que el alumno sea **introducido** en el uso de un lenguaje de programación ad hoc, orientado a objetos, para manipular grandes volúmenes de datos, con características auditables y trazables, es decir, será para el alumno; un punto de partida para la utilización de herramientas de referencia o de vanguardia en la programación para, automatizaciones, análisis exploratorios y predictivos.

Para lo anterior, este curso es inherentemente práctico y aplicado. Demanda la participación sistemática del alumno en clases; "aprender-haciendo". Incluye demostraciones metodológicas, aplicados y talleres prácticos durante todo el desarrollo del módulo, incorporando ejercicios y los elementos conceptuales inherentes para tales fines.

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 INAP y la Universidad de Chile, denominado u-cursos, en los plazos determinados e informados por el profesor para la resolución de dudas y consultas, que es desde que comienza el módulo hasta una hora antes de la evaluación.

Este módulo permitirá que el alumno amplíe su nivel de abstracción. Comenzará desde cero. Al terminar el curso, el alumno será capaz de poseer nociones introductorias sólidas de lenguaje de programación orientado a objetos, que en este caso será a través de Rstudio, para manipular a nivel básico grandes volúmenes de datos, objetos, y así comenzar a optimizar el tiempo y costos en sus procesos de colaboración en resolver diferentes problemas en su organización/institución.

2. OBJETIVO GENERAL DE LA ASIGNATURA

El objetivo general del módulo "Introducción al Lenguaje RStudio para la Manipulación de Data Estructurada" es que los alumnos adquieran un primer know how, background, framework, approach aplicado, en un primer acercamiento a un lenguaje de programación orientado a objetos, a través de RStudio. Diseñado para la Analítica Avanzada, que es la Naturaleza de este postítulo, para las funciones que desempeñará/a el alumno en su organización**. Implantar en ellos los conocimientos necesarios básicos para tareas de preprocesamiento,** tareas básicas pero que impliquen la manipulación de grandes volúmenes de datos, y con esto generar las transformaciones incipientes y adecuadas de un conjunto de datos durante todo un primer proceso de comprensión de la data que tengo enfrente.

3. OBJETIVOS ESPECÍFICOS DE LA ASIGNATURA

Al finalizar el modulo el alumno estará capacitado para:

- Aprender a programar un script básico en Rstudio con los tipos de datos, métodos y estructuras de control que permitan realizar la manipulación de un conjunto masivo de datos x.
- Conocer a través de una primera inmersión contextual, el lenguaje de Programación RStudio orientado a objetos.
- Conocer la interfaz, mecanismos y distintas funciones, para futuros métodos de trabajo con RStudio.
- Importación de datos.
- Cambio de idioma para los outputs.
- Expandir la memoria virtual (hardware) para el uso de R.
- Trabajo en Disco Externo (hardware).
- Eliminación y Adición de Variables.

- Concatenar, tratamiento de minúsculas a Mayúsculas, y Separador de caracteres.
- Clases de variables y conversión de ellas.
- Indexación de Row y creación de id correlativo.
- Tratamiento de N/A y reemplazo de observaciones.
- Selección con algunas Condicionantes.
- Tratamiento de Duplicidades.
- Renombrar Variables.
- Frecuencia de observaciones.
- Agrupación de Información.
- Unión de Dataframes.
- Creación de Columnas.
- Conexión Remota ODBC.
- Consulta SQL desde R.
- Transponer una matriz.
- Exportación de datos.
- Seteo de directorios.
- Estadística Descriptiva (Medidas de Tendencia Central).
- Escalar Variables.

Bonus Track si el tiempo lo amerita: Creación de una librería en disco local, para sus propias funciones que su imaginación le permita.

El alumno después de este módulo no estará habilitado para realizar minería de texto, de sentimientos ni procesamiento complejo de caracteres. Tampoco para el análisis supervisado (predictivo) y automatizaciones. Estas técnicas se verán en los módulos posteriores.

4. METODOLOGIA

- El módulo es principalmente práctico y consiste de 6 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 horas de autoestudio – hacer uso de sus horas asincrónicas- y/o trabajo en equipo/grupo, fuera del horario de clases (sincrónicas).
- Los contenidos serán entregados bajo un esquema sistemático.
- 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. Estas preguntas podrán ser consideradas para sumar décimas para la evaluación correspondiente.

Para la correcta implementación de la metodología, o de todo lo dicho anteriormente, el alumno deberá asegurarse de poseer:

- 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, contar con privilegios de "administrador de equipo", o en su defecto 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 corta fuegos o bloqueos que interrumpan la normalidad de la instalación de un nuevo software con sus respectivos componentes (paquetes y librerías).

- 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, problemas en la tarjeta de sonido, etc.

- Se facilitará a los alumnos con anticipación al comienzo de este curso, los manuales y/o tutoriales correspondientes para la instalación liberada del softwar que ocuparemos y que será imprescindible para este diplomado; R y RStudio. Tanto para dos sistemas operativos distintos, Windows y MacIOS.

- Los alumnos deberán llegar con el Software citado ya instalado a la primera o segunda clase de este curso, es decir, a la fecha que lo requiera el profesor. A partir de este Módulo los alumnos aprenderán a hablar, comenzarán a hacer uso de un nuevo lenguaje (orientado a objetos, de programación; R), el cuál será el que se utilizará para los siguientes módulos. Dicho software es Open Source o Código Abierto, es decir, gratuito.

5. EVALUACION

- La asistencia tiene una obligatoriedad del 75%. Ante dudas consultar con el coordinador de este postítulo ya que esta métrica es fijada por INAP y computada automáticamente por el sistema o plataforma U-cursos.

- 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 que el profesor realice.

- 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 o más evaluaciones.

- 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: 4,5 (cuatro y cinco décimas).

- La fecha de la evaluación será el sábado 03 de septiembre. La evaluación se abrirá a las 16 hrs. y se cerrará a las 20 hrs. No recibiendo evaluación desde las 20.01 hrs.

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 subirán al sistema las rubricas correspondientes al momento de cargar la(s) evaluación(es).**

- **En caso de reprobar (Nota Final < 4.5), el alumno tendrá derecho a una instancia de examen recuperativo. La nota del examen recuperativo se promediará con la nota de la primera evaluación reprobada, optando así a la calificación mínima de aprobación del curso.**

  La fecha de dicho examen recuperativo será en la semana de exámenes recuperativos, la que comenzará una semana después del último módulo del postítulo, previa coordinación del profesor con el coordinador. El profesor comunicará la fecha y hora oportunamente.

- Los alumnos no pueden repetir los módulos reprobados en una siguiente versión.

  Los alumnos que desean retomar el programa en una versión posterior, pueden hacerlo **sólo en aquellos casos que han presentando de manera anticipada antecedentes que permitan justificar "congelar" su participación en la versión actual que necesitan congelar**.

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

INASISTENCIA A EVALUACIONES

Los alumnos deben conectarse a las clases desde su cuenta institucional entregada por INAP.

Los alumnos que se conectaron a las clases desde otra cuenta, no siendo desde la institucional entregada por INAP, se registrarán como ausentes. El único mecanismo para justificar una inasistencia es licencia o certificado médico y/o laboral. Este punto no es apelable.

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 por la obtenida en el examen de repetición o recuperación. Este punto no es apelable.

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

6. BIBLIOGRAFÍA

Contenidos y Apuntes de las Presentaciones Vistas en Clases.

- "Introducción a R" – R Development Core Team. Disponible en: https://cran.r-project.org/doc/contrib/R-intro-1.1.0-espanol.1.pdf

- Material Docente.

- Papers y experiencia comparada subidas por el profesor si fuera el caso.

7. 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 Púbico. 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.

- Lugar Actual de Trabajo: 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.

- E-mail: **adrianaranedat@ug.uchile.cl**

- Áreas de Investigación: Data Science. Modelo Micro Econométrico Clustering. Modelos Predictivos (Machinme 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.

8. SESIONES Y FECHAS

SESION 1:	**JUEVES**	**25-08-22**	**19.30 – 21.45 HORAS**
SESION 2:	**SABADO**	**27-08-22**	**09.00 – 13.30 HORAS**
SESION 3:	**MARTES**	**30-08-22**	**18.30 – 21.45 HORAS**
SESION 4:	**JUEVES**	**01-09-22**	**18.30 – 21.45 HORAS**

[i] Adrian Armando Araneda Toro utiliza esta frase parea referirse a condiciones de "subdesarrollo" en el estadio en el uso de los datos.

[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 y categorías de Émile Durkheim y Pierre Bourdieu).