Introducción.

Los contenidos del presente curso son de particular relevancia para la vida profesional del actuario. El análisis multivariante se ha convertido en una de las herramientas más potentes de análisis en cualquier rama del conocimiento, tanto teórico como práctico. La presencia en el mercado de software especializado y técnicas de aprendizaje automático permite tener acceso inmediato a potentes herramientas analíticas a bajo costo. El enfoque que tendremos en este curso será completamente aplicado, si bien revisaremos la base teórica suficiente que sustenta cada técnica, el objetivo es que el estudiante desarrolle sus capacidades analíticas y de aplicación en cualquier rama del conocimiento donde pretenda desempeñarse profesionalmente. En la materia de Análisis Multivariado, revisaremos desde el punto de vista estadístico, los problemas generales de predicción, clasificación y agrupamiento. En este contexto, la materia prima con la que contaremos serán datos en cantidad suficiente para poder determinar las relaciones entre una variable dependiente con un conjunto de predictoras, o en su defecto, los patrones y relaciones intrínsecas entre un conjunto de variables.

Plan de estudios

El programa de la asignatura ha quedado obsoleto con respecto a las necesidades del mercado laboral, revisemos los temas obligatorios, su uso anterior y el uso actual:

Tema	Uso Anterior	Uso Actual
Análisis Discriminante	Predicción de variables categóricas	Técnica individual en modelación supervisada
Análisis Factorial	Explicación de causas de un fenómeno mediante variables latentes	Reducción de dimensiones
Análisis de Conglomerados	Agrupación jerárquica de observaciones	Agrupaciones inteligentes, segmentación, creación de nuevos predictores.
Escalamiento Multidimensional	Localización de patrones por vista	Auxiliar en modelación no supervisada

Como podemos ver, ni siquiera se tiene el orden correcto en los contenidos y se han dejado de lado importantes temas en el estado del arte, por ejemplo, el temario solo contempla clustering jerárquico dejando de lado las técnicas difusas y de optimización mucho más socorridas actualmente. Escalamiento multidimensional se ha convertido en un auxiliar y no en un tópico

principal como se le trata en el programa. En cuanto al análisis discriminante, la técnica es muy vieja y ya solo se utiliza como una más dentro del machine learning. Por último, el análisis factorial puede ser sustituido fácilmente por técnicas más modernas, simples y robustas por lo que su obsolencia se confirma.

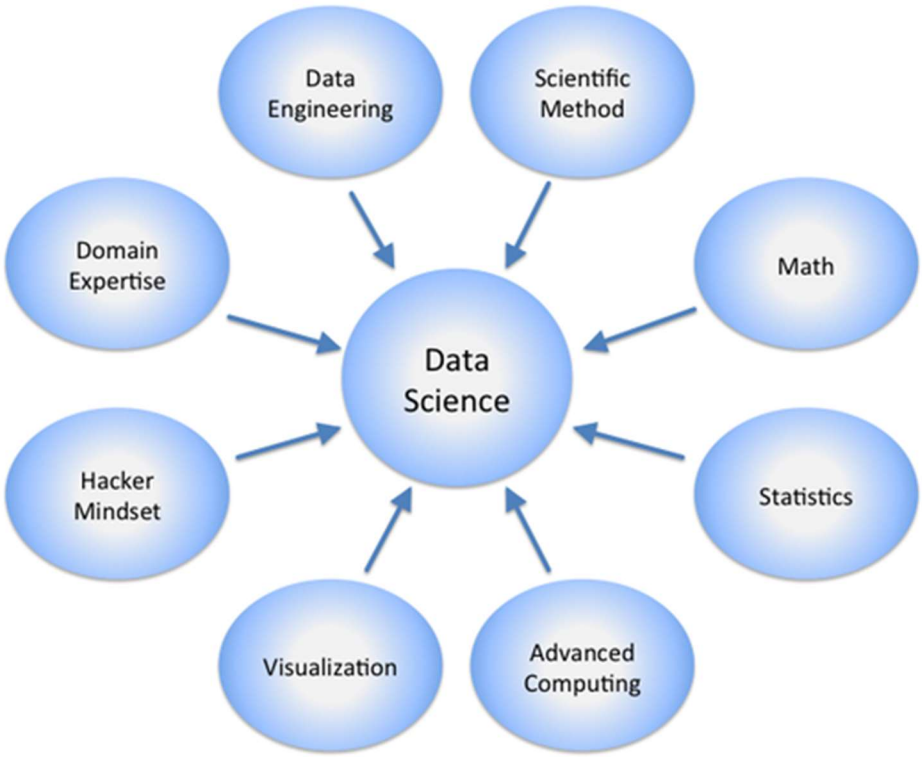
Se propone al grupo un temario mucho más útil y actualizado incorporando técnicas en el estado del arte y tecnología de punta sin dejar de revisar los temas que la cátedra exige.

1. Introducción: Breve contextualización y uso de software de última generación.
2. Análisis Factorial y Componentes principales: Técnicas de reducción de dimensiones
3. Análisis Clúster: Modelación no supervisada por medio de técnicas jerárquicas, de optimización y difusas.
4. Modelación Supervisada: Técnicas de machine learning aplicadas desde el punto de vista estadístico: análisis discriminante, árboles de decisión, regresión logística, redes neuronales, máquinas vector soporte, k vecinos más cercanos, gradiente estocástico, Bayes Ingenuo, ensambles.

Contextualización.

En el mundo actual (2017), el análisis multivariante ha evolucionado al convertirse en el soporte matemático (estadístico) de las técnicas de aprendizaje máquina (machine learning) en contraste con su otrora función de componente aislado y extensión a varias dimensiones de la estadística descriptiva e inferencial. En este entendido, se ha conformado una nueva ola de disciplinas de la ciencia que ha pasado por una vertiginosa transformación. Lo anterior fue posible gracias a la amplia disponibilidad de datos digitales como consecuencia inmediata del proceso de digitalización de las compañías. Una vez que los datos estuvieron disponibles, surgió la disciplina denominada **inteligencia de negocios** (Business Intelligence: BI) cuyo propósito fundamental es convertir el dato en información relevante mediante herramientas de visualización, es análogo a la **estadística descriptiva** donde resumimos información y la presentamos por medio de gráficos o estadígrafos (media, moda, varianza, cuantiles, etc) para tener un mejor entendimiento de los datos que nos han sido proporcionados. En el siguiente nivel se encuentra la disciplina conocida como **minería de datos** (surgido a finales de los 90 y principios del milenio), cuyo propósito es la obtención de patrones e información no trivial dentro de grandes volúmenes de datos, aquí surge una duda común, ¿en qué se diferencia con respecto al BI?, de igual forma que la estadística descriptiva con la estadística inferencial, el propósito está enfocado en la **predicción** (Donde se necesitará matemática más sofisticada). Durante los últimos años se han acuñado nuevas definiciones y se han extendido las disciplinas. Una de estas es el término **Analytics** (analítica en español) que consiste en el descubrimiento, interpretación y comunicación de patrones dentro de los datos, como vemos, es parecida a minería de datos, con el añadido de la comunicación e interpretación y no solo la extracción. El término sigue siendo ampliamente utilizado, sin embargo una nueva confusión surgió con la popularización del término **Data Science**. Ciencia de datos se considera una evolución multidisciplinaria en los campos de análisis de negocio, ciencias de la

computación, modelación matemática, estadística, analítica y minería de datos. El siguiente diagrama presenta un buen resumen de los tópicos asociados a la ciencia de datos:



Es decir, un data scientist debe poseer además de las habilidades de analítica, un fuerte componente de habilidades computacionales, inventivas y científicas.

Por último, el término **Big Data** se añade a la discusión, BigData se refiere a toda información (estructurada, no estructurada, semiestructurada) que no puede ser procesada o analizada usando procedimientos y herramientas tradicionales. La situación actual está enfocada en la recolección y procesamiento de esas enormes cantidades de información (terabytes en adelante) mediante la infraestructura adecuada de hardware y software, por tanto, es independiente de los términos revisados con anterioridad ya que se añaden únicamente los componentes de volumen, velocidad y variedad de datos que fungen como materia prima del científico de datos, el analista de información, el diseñador de tableros de BI, etc.

Herramientas de software

Anteriormente, el software estadístico estaba encasillado dentro de un nicho muy específico cuyo propósito principal era la automatización de los cálculos propios de la disciplina. Actualmente, los simples cálculos se obvian y se opta por empoderar las herramientas de software expandiendo su funcionalidad. En el mercado existen tres jugadores principales, sin embargo, existe una gran cantidad de alternativas (libres y comerciales) la mejor herramienta no existe, cada quien resolverá los problemas que se le planteen de acuerdo a su necesidad, comodidad y recursos disponibles.

En primera instancia, las herramientas deberán ser separadas por funcionalidad:

- Herramienta de manipulación de datos: Necesaria para construir tablas analítica de datos o alguna estructura de datos (resumen, consulta) que fungirá como entrada.
- Herramienta de visualización de datos: Aunque no indispensable, nos permite tener "primeras impresiones" de lo que muestran los datos.
- Software estadístico: Necesario para realizar análisis estadístico básico, descriptivo e inferencial.
- Software para machine learning: Necesario para generar modelos de soporte no supervisados y supervisados.
- Software para presentaciones: Necesario para comunicar resultados, proyectos, hallazgos, estrategias, etc.
- Hoja de cálculo: Herramienta hiperbásica de análisis, permite crear reportes, tablas pivote y gráficas de forma muy simple.
- Lenguaje de programación: No indispensable, pero si muy recomendable. Brinda la capacidad de automatizar tareas y expandir nuestras posibilidades mediante la creación de herramientas propias y enlace a API's

En el mercado hay tres suites que cubren las primeras 4 necesidades a cabalidad; dos son alternativas de software libre y una de ellas de software propietario: Python, R-Project y SAS respectivamente. En cuanto al lenguaje de programación, las alternativas libres lo poseen, mientras que SAS corresponde más a un lenguaje de manipulación de datos que a uno de programación. Se obviará la parte de ofimática.

No existe una "mejor" herramienta, todas poseen ventajas y desventajas, la elección dependerá de las circunstancias de cada empresa/persona.

En mi muy particular experiencia profesional, SAS ha sido el software que más presente ha estado. Esto es debido a la sencillez en su uso y poderosas capacidades (además de ser el indiscutible líder del mercado en analytics), en detrimento del mismo: su precio. SAS en un software que estará disponible solamente en grandes compañías debido al precio de las licencias y requerimientos de hardware, sin embargo, recientemente el SAS Institute proporciona una versión gratuita del

software (prácticamente idéntica aunque con capacidades limitadas de datos) llamada SAS University Edition con la que libre de precio se puede practicar y aprender esta tecnología proporciona una enorme ventaja competitiva.

R-Project es una opción muy popular (y muchas veces usado como alternativa para quien no puede pagar SAS) entre empresas medianas y startups. Posee una gran comunidad que le da soporte al software así como variedad de paquetes que incrementan su funcionalidad, sin embargo, le hacen falta capacidades de manipulación de grandes volúmenes de datos, además de que por ser de nicho en concepción(estadística) la integración con otras plataformas no es directa.

En cuanto al lenguaje de programación Pyhton, es un poderoso y muy simple lenguaje de programación que nos brinda posibilidades ilimitadas. Como inicio, la distribución Anaconda que proporciona todos los paquetes precargados para ciencia de datos. La desventaja es que requiere conocimiento más avanzado en temas de cómputo.

En este curso los ejemplos se realizarán en SAS y Phyton simultáneamente en aras de contar con una alternativa de cada bando (libre/propietario). Recuerde que el valor de un profesional no se relaciona con el software que utiliza, sino con su capacidad para GENERAR VALOR.