# MIND: Minería de Datos
# Proyecto Final

Octubre 15 de 2020

## Objetivos:

1. Aplicar conceptos y técnicas relacionadas con la minería de datos.
2. Aplicar las fases de la metodología CRISP-DM a un caso real de estudio.
3. Hacer uso de la herramienta `Jupyter Notebook` para dar solución al problema.
4. Elaborar un trabajo escrito con altos estándares de calidad, reflejando todo lo realizado en su proyecto.

## Fechas Importantes:

- Noviembre 22, fecha máxima para cargar el .zip del proyecto hasta 11:55pm en Moodle. Debe cargar un .zip que incluya: el trabajo escrito, el cuadernillo en Jupyter y el dataset relacionado con su proyecto.
- La nota del proyecto se compone de: Trabajo Escrito: 45%, Trabajo Computacional: 35%, y Presentación del Trabajo: 20%.
- Noviembre 25, durante la sesión de clase se desarrollará el segundo parcial teórico.
- Noviembre 25, se entrega el segundo parcial práctico para ser desarrollado de forma individual bajo la modalidad TakeHome.
- La presentación del proyecto se desarrollará en la semana de parciales, en el espacio programado por la oficina de registro para el parcial final de MIND.

## Reglas de juego:

- Al tratarse de un proyecto final y en especial de un curso de Maestría, lo más probable es que usted tenga que hacer **búsqueda y lectura de artículos científicos**, y material relevante que le permita profundizar y complementar temas relacionados con el proyecto.
- El trabajo escrito debe tener los **más altos estándares de calidad**: ortografía, puntuación, redacción, coherencia, justificación, argumentación, profundización, entre otras. El reporte debe contener los 6 capítulos presentados en el siguiente link: Project Writeup
- Se sugiere adiciona en el documento el workflow utilizado durante el análisis, con la intención de brindar una ilustración y dar claridad al proceso realizado.
- Adicional al trabajo debe entregar el cuadernillo en Jupyter relacionado con su proyecto. En caso de utilizar o adaptar algún fragmento de código basado en Internet, no olvide dar crédito al autor o sitio web, en caso contrario, **se considerará plagio**, con las implicaciones estipuladas en el reglamento académico. No olvide al comienzo del cuadernillo indicar las librerías requeridas.
- El proyecto se puede desarrollar de a dos personas, pero la elección del compañero está sujeta a la siguiente condición: su pareja de proyecto debe tener una nota de MIND **más o menos de** 0.2 **a la suya.** Por ejemplo: Juanito con nota de 1.9, solamente podrá hacer grupo con una pareja cuya nota oscile entre 1.7 y 2.1. Hago referencia a las notas temporales, al 45% enviado días atrás.
- Si usted va a realizar el proyecto en grupo, entre los dos deben elegir el dataset a utilizar en el proyecto, la elección es a partir de lo indicado semanas atrás por cada uno, es decir, deben elegir una de las dos opciones indicadas semanas atrás.

- Para todos: en el siguiente archivo debe indicar el estudiante o los estudiantes que conforman el grupo de proyecto. Por favor, tenga **presente la condición indicada** anteriormente Project Teams.
- El proyecto también se puede desarrollar de **forma individual** y **tendrá el mismo alcance**, es decir, debe hacer lo mismo si está en grupo o individual.
- Durante el desarrollo del proyecto se resolverán dudas por medio del **foro** y también el **monitor** será de gran ayuda. El monitor **NO** tiene la función de desarrollar proyectos, ni tampoco hacer código.
- Todos deben realizar **presentación gerencial del proyecto**, la duración será indicada más adelante.

## Enunciado:

El desarollo del proyecto implica tareas relacionadas con la minería de datos enmarcadas en la metodología **CRISP-DM**: comprensión del negocio, estudio y entendimiento de los datos, preparación de los datos, modelado, evaluación de resultados y despliegue. En el proceso de modelado-evaluación se realizarán tareas de predicción y descripción a partir de los datos elegidos. A continuación el detalle de la funcionalidad:

1. **Datos Faltantes**. Si su dataset NO tiene valores faltantes debe generar un nuevo dataset con el 10% de valores faltantes. Si tiene valores faltantes, debe hacer imputación por la técnica que le resulte más eficiente, y posteriormente incorporar el 10% de NA. Tenga presente que la clase no debe contener valores faltantes. A partir de ese dataset con valores faltantes debe desarrollar su proyecto.
2. **Exploratory Data Analysis**. EDA sobre las variables relevantes de su dataset, lo más importante es tener su explicación e interpretación de los gráficos, las solas imágenes no sirven de nada.
3. **Preprocesamiento**. A partir del dataset con valores faltantes debe aplicar los métodos que considere necesarios previo al análisis de datos, recuerde que hay diversas técnicas para: imputar, normalizar, discretizar, reducir, entre otras tareas relacionadas con esta fase relevante para el proceso de análisis. Las técnicas elegidas y aplicadas deben estar justificadas y soportadas de forma coherente.
4. **Clasificación Supervisada**. Durante el análisis se aplicarán tres métodos diferentes de clasificación supervisada, por lo tanto, el concepto de cada uno, la forma de aplicación y los resultados obtenidos en cada uno pueden ser diversos, al final, ustedes verán las bondades y dificultades sobre cada uno de los métodos aplicados.

   Los métodos son: Logistic regression, Naive Bayes, Fisher's linear, Support vector machines, k-nearest neighbor, Decision trees, Boosted Trees, Random forests y Neural Networks. Algunos de estos se explican en el curso, y otros **deben ser profundizados e investigados por ustedes**.

   **Importante**: la asignación de los tres métodos la realizará el Profesor luego de la consolidación de los grupos del proyecto, es decir, de los nueve métodos se le asignará los tres a aplicar en su proyecto.
5. **Evaluación**. Para los tres métodos aplicados debe mostrar la respectiva evaluación y explicar el rendimiento de cada modelo, de igual manera, debe indicar y justificar el modelo elegido, la argumentación debe ser clara y coherente. Adicionalmente, debe presentar de manera clara la comparación entre las tres técnicas de clasificación aplicadas y sus resultados.
6. **Clasificación No Supervisada**. Aplicar al menos una técnica de clasificación no supervisada sobre el dataset inicial/imputado. Como resultado, debe explicar lo encontrado en el conjunto de datos y la interpretació a los hallazgos. En caso de tener valores faltantes en su dataset original recuerde que primero debe hacer imputación.
7. **Gráficos Interactivos**. Su cuadernillo debe tener al menos un gráfico interactivo (widgets) para presentar visualmente, de forma gráfica, y en tiempo real algún efecto luego de ajustar en la interfaz gráfica parámetro(s) para algún método de clasificación asignado. A continuación, algunos links sobre gráficos interactivos: Jupyter Widgets, Notebook Widgets, además del extenso material en Internet.