



Instituto Tecnológico de Buenos Aires  
Especialización en Ciencia de Datos

Minería de Datos

**Libro de la Asignatura**

Edición 2024 A

Autor: Gustavo Denicolay

última actualización: 2024-05-31 23:05

# Tabla de Contenidos

1La Asignatura.....	6
1.1Sobre este documento.....	6
1.2Links fundamentales.....	7
1.3ITBA Cronograma de la Asignatura año 2024A.....	8
1.4Dedicación Estimada.....	9
1.5Modalidad y Criterios de Aprobación y Evaluación.....	10
1.5.1Modalidad y Criterios de Aprobación y Evaluación Competencia Kaggle.....	11
1.5.2Experimentos Colaborativos.....	12
1.5.3Evaluación participación en foro/chat Zulip.....	13
1.5.4Evaluación Google Sheet Colaborativa DURANTE las clases.....	13
1.5.5Evaluación Google Sheet Colaborativa en Tarea para el Hogar.....	14
1.6Misión de la Asignatura.....	15
1.6.1Objetivos de la Asignatura.....	15
1.7 Unidades Temáticas.....	16
1.7.1Tecnologías a enseñar y reforzar.....	17
1.8Bibliografía General de la Asignatura.....	18
2Metodología de Enseñanza.....	20
2.1Generalidades.....	20
2.2Manifiesto Pedagógico.....	20
2.3Pensamiento Creativo - Tinkering.....	21
2.4Dinámica Constructivista en el Aula.....	22
2.5Authentic Learning.....	22
2.6Flipped Classroom Forte.....	22
2.7Conectivismo.....	23
2.8Active Learning.....	24
3Arranque en Frío.....	25
3.1Alta en Plataformas.....	25
3.1.1Zulip herramienta de chat y foros discusión.....	25
3.1.2Google Sheet Colaborativa de la Asignatura.....	26
3.1.3Plataforma Hypothes.is.....	26
3.1.4Plataforma ChatGPT.....	26
3.1.5Plataforma Kaggle.....	27
3.1.6Plataforma GitHub.....	28
3.1.7Check In a la Asignatura en Zulip.....	30
3.2 Instalación de Herramientas.....	32
3.2.1Lenguaje de programación R.....	32
3.2.2Librerías lenguaje R.....	32
3.2.3Aplicación RStudio Desktop.....	32
3.2.4Estructura de Carpetas en su PC local.....	33
3.2.5Aplicación Git.....	33
3.2.6Clonado repositorio.....	34
3.2.7Prueba de RStudio y Kaggle.....	36
3.2.8Alta en GitHub Copilot.....	37
4Herramientas, Conceptos, Operación y Buenas Prácticas.....	38
4.1Zulip.....	38
4.1.1Buenas prácticas para el uso de Zulip.....	39
4.1.2Participaciones extraordinariamente significativas en Zulip.....	40
4.1.3Participaciones significativas en Zulip.....	40
4.1.4Mensajes Privados a los profesores Permitidos.....	41

4.2Git y GitHub.....	42
4.2.1conceptos y operación.....	42
4.3Librería data.table del lenguaje R.....	43
5Plataforma Kaggle.....	44
5.1La partición { Public, Private }.....	44
5.2 Cálculo de ganancias.....	45
5.3Elección manual de un submit.....	46
5.4Public Leaderboard.....	48
5.5Ganancia en el Public Leaderboard.....	49
5.6Ganancia en el Private Leaderboard.....	50
5.6.1Ningún submit elegido manualmente.....	50
5.6.2Submit elegido manualmente.....	51
5.7Operación de Kaggle.....	52
5.8Finalización de la competencia.....	52
6Experimentos Colaborativos.....	53
6.1Documento Colaborativo Google Slides.....	54
6.2Problemas disponibles y alumnos avanzados.....	55
6.3Información Preexistente.....	55
6.4Configuración de los grupos.....	55
6.5Forma de pre registración.....	56
6.6Negociación de Problemas : vacancia.....	56
6.7Negociación de Problemas : sobreoferta.....	56
6.8Asignación Grupo A vs Grupo B.....	56
6.9Antagonismo de los grupos.....	57
6.10Períodos de análisis de los grupos.....	57
6.10.1Período análisis Grupo B.....	59
6.11Aprobación de Diseños Experimentales.....	60
6.12Scripts.....	60
6.13Semillas generales en scripts.....	60
6.14Ceteris Paribus.....	60
6.15Capítulo Conclusiones.....	61
6.16Video Presentaciones Ejecutivas.....	61
6.17Luego de las Video Presentaciones.....	61
7¿Cómo seguir?.....	62
7.1Mantenerse actualizado.....	62
7.2Datasets Públicos.....	62
8Empiricismo en la Ciencia de Datos.....	63
8.1.1Bibliografía Introdutoria.....	63
9Las dos culturas en el modelado predictivo.....	64
10Estimando la Ganancia de un modelo predictivo.....	65
10.1Bibliografía Introdutoria.....	65
10.2Sesgo Varianza en la estadística clásica.....	65
10.2.1Bibliografía Inicial.....	65
10.3Sesgo Varianza en el moderno Machine Learning.....	66
10.3.1Bibliografía Inicial.....	66
11Análisis exploratorio de datos.....	67
11.1Consistencia longitudinal de los atributos.....	67
11.2Drifting de atributos.....	67
11.3Tratamiento de variables "rotas".....	67
11.4Tratamiento de nulos.....	67
11.5Tratamiento de outliers.....	67
12Algoritmo: Árbol de Decisión.....	68

12.1Bibliografía Inicial.....	68
12.2El hiperparámetro cp en la librería rpart.....	68
13Comparación de Modelos Predictivos.....	69
13.1Bibliografía Inicial.....	69
14Optimización de Hiperparámetros.....	70
14.1Bibliografía Introductoria.....	70
14.2Grid Search.....	70
14.3Bayesian Optimization.....	70
14.3.1Artículos de fácil Lectura.....	70
14.3.2Videos sencillos.....	70
14.3.3Video Lectures.....	70
14.3.4Libros de Texto.....	70
14.3.5Papers Técnicos.....	71
14.3.6Librería mlrMBO.....	71
15Curva ROC.....	72
15.1Bibliografía.....	72
15.2Curva ROC de un split en un árbol de decisión.....	72
15.3Curva ROC de un modelo predictivo.....	72
15.4Relación de la Curva ROC con la función de ganancia.....	72
16Controlando el False Discovery Rate.....	73
16.1Bibliografía Introductoria.....	73
17Feature Engineering.....	74
17.1Bibliografía Introductoria.....	74
17.2Cambiando la clase.....	74
17.3Variables Manuales.....	74
17.4Transformaciones tradicionales.....	74
17.5Variables históricas.....	75
17.6Lags y delta lags.....	75
17.7Tendencias.....	75
17.8Medias móviles.....	75
17.9Ajuste por Inflación.....	75
17.10Variables hojas de un Random Forest.....	75
17.11Reducción de la dimensionalidad para acelerar.....	75
18Ensembles de Modelos.....	76
18.1Bibliografía Introductoria.....	76
19Algoritmo: Random Forest.....	77
19.1Bibliografía Introductoria.....	77
20Algoritmo: Gradient Boosting of Decision Trees.....	78
20.1Bibliografía Introductoria.....	78
20.2XGBoost.....	78
20.2.1Bibliografía XGBoost.....	78
20.3Optimización hiperparámetros XGBoost.....	78
20.4Bibliografía LightGBM.....	78
20.5Optimización hiperparámetros LightGBM.....	78
21Estrategia de Entrenamiento.....	79
21.1Determinación de los mejores períodos para entrenar.....	79
21.2Acelerando el entrenamiento : evitando el cross validation.....	79
21.3Acelerando el entrenamiento : Undersampling de la clase mayoritaria.....	79
22Importancia de Variables.....	80
22.1Bibliografía Introductoria.....	80
22.2Valores Shapley.....	80
23Ensembles de Modelos Segunda Parte.....	81

23.1Semilleríos.....	81
23.2Hibridación de Semilleríos.....	81
24Acelerando el procesamiento.....	82
25Aprovechando el Public Leaderboard.....	83
25.1Bibliografía Inicial.....	83
25.2El problema de la calibración en Gradient Boosting.....	83
25.3Otra forma de cortar.....	83
26Clustering jerárquico basado en la distancia de Random Forest.....	84
27Storytelling.....	85
27.1Bibliografía Introductoria Storytelling.....	85
27.2Herramientas para grabar videos.....	85
27.3Grabación sencilla de videos con Microsoft Teams.....	85
27.4Torneo de Videos.....	86
28El Recuperatorio.....	86

# 1 La Asignatura

Esta asignatura en su versión 2024 será una **intensa** experiencia emocional, social y cognitiva, con una dinámica totalmente distinta a las que está acostumbrado; le demandará un gran involucramiento. Los instructores intentarán sorprenderlo en cada clase.

Aprenderá a resolver un problema real, con datos reales (millones de registros y cientos de columnas), procesará en la nube Google Cloud y utilizará los algoritmos estado del arte como ser la Bayesian Optimization y LightGBM. Será motivado a pensar *out of the box* en Feature Engineering.

La materia posee una competencia de ciencias de datos con el objetivo de construir el mejor modelo predictivo que resuelva el problema de retención de clientes financieros, utilizando la plataforma de competencias Kaggle.

La metodología es la de flipped classroom, deberá venir con los experimentos preparados a las clases y mantendrá contacto asincrónico con los profesores y compañeros con la herramienta de foro/chat Zulip. En las clases presenciales se desarrollarán dinámicas de *active learning* en contraposición a la tradicional dinámica transmisionista de clases magistrales.

El centro de esta asignatura es usted formando con sus pares una comunidad de aprendizaje colaborativo.

## 1.1 Sobre este documento

No se sienta abrumado por la extensión y detalle de este documento, en la primera clase se le explicará coloquialmente que se espera de usted.

**No imprima en papel este documento;** está solo completo tan solo hasta el capítulo 3, qué es lo que necesita para iniciar la materia. Se le irá agregando contenido los próximos días a los primeros capítulos, en particular la dinámica de Experimentos Colaborativos.

## 1.2 Links fundamentales

ITBA 2024A Data Mining Links Fundamentales	
Zulip	<a href="https://itba2024.zulip.rebelare.com/">https://itba2024.zulip.rebelare.com/</a>
repositorio GitHub oficial	<a href="https://github.com/itba-ecd/dm2024a">https://github.com/itba-ecd/dm2024a</a>
Invitación a la Competencia Kaggle	<a href="https://www.kaggle.com/t/ed204a0388755441e5908b5c121da3ec">https://www.kaggle.com/t/ed204a0388755441e5908b5c121da3ec</a>
Competencia Kaggle	<a href="https://www.kaggle.com/competitions/itba-data-mining-2024-a/">https://www.kaggle.com/competitions/itba-data-mining-2024-a/</a>
Diccionario de Datos	<a href="https://storage.googleapis.com/open-courses/itba2024-06b9/DiccionarioDatos_2024.ods">https://storage.googleapis.com/open-courses/itba2024-06b9/DiccionarioDatos_2024.ods</a>
Dataset Inicial Competencia	<a href="https://storage.googleapis.com/open-courses/itba2024-06b9/dataset_pequeno.csv">https://storage.googleapis.com/open-courses/itba2024-06b9/dataset_pequeno.csv</a>
Google Sheet Colaborativa	<a href="https://docs.google.com/spreadsheets/d/1rPIXzCbQNOufKkQub2zWtY18DPRdmXpT7naXeQ1Enig/edit?usp=sharing">https://docs.google.com/spreadsheets/d/1rPIXzCbQNOufKkQub2zWtY18DPRdmXpT7naXeQ1Enig/edit?usp=sharing</a>
Google Slides Experimentos Colaborativos	<a href="https://docs.google.com/presentation/d/1x_5H8AkNfT8qJ2coQxoYkNuHbJrIKLYH3ELE1xdIH8Y/edit?usp=sharing">https://docs.google.com/presentation/d/1x_5H8AkNfT8qJ2coQxoYkNuHbJrIKLYH3ELE1xdIH8Y/edit?usp=sharing</a>

### 1.3 ITBA Cronograma de la Asignatura año 2024A

02-may jue	10:01	email	Envío por email de este documento a alumnos, disponibilización en Blackboard.
06-may lun	17:30 a 22:00	Virtual	Clase 1 Presentación e Introducción, Metodología Pedagógica, Plataformas y Herramientas Básicas Planteo del Problema
09-may jue	17:30 a 22:00	Virtual	Clase 2 Arboles de Decisión, hiperparámetros, overfitting, optimización de hiperparámetros Grid Search
09-may jue	23:59	Zulip	Disponibilización Tarea para el Hogar UNO Instructivo para instalación entorno Google Cloud
14-may mar	17:30 a 22:00	Presencial	Clase 3 Optimización de Hiperparámetros : Bayesian Optimization
16-may jue	17:30 a 22:00	Presencial	Clase 4 Métodos de Ensemble : Arboles Azarosos, Random Forest Repensando el Overfitting "árboles que se defienden del overfitting" Lanzamiento <b>Experimentos Colaborativos</b>
16-may jue	23:59	Zulip	Disponibilización Tarea para el Hogar DOS
21-may mar	17:30 a 22:00	Presencial	Clase 5 Gradient Boosting of Decision Trees Lanzamiento de <i>Experimentos Colaborativos</i>
23-may jue	17:30 a 22:00	Presencial	Clase 6 Feature Engineering trabajando con datos históricos
24-may vie	08:00	Zulip	Disponibilización Tarea para el Hogar TRES
28-may mar	17:30 a 22:00	Presencial	Clase 7 Training Strategy
30-may jue	17:30 a 22:00	Presencial	Clase 8 Ensembles, Stacking, "Semilleros e Hibridación de Semilleros © "
04-jun mar	23:59	Zulip	Fecha límite entrega Video Experimentos Colaborativos
05-jun	09:00	Zulip	Inicio Torneo de Videos
09-jun	23:59	Zulip	Fin Torneo de Videos
10-jun	23:58	Zulip	Fecha límite disponibilización GitHub con solución reproducible.
	23:59	Zulip	Cierre automático de Competencia Kaggle
15-jun	23:59	Zulip	Entrega de Notas
17-jun	23:30	Zulip	Entrega de instrucciones, datasets y scripts para recuperatorios individuales
31-jul	23:59	Zulip	Fecha límite de entrega de recuperatorio



## **1.4 Dedicación Estimada**

La materia consta de 36 horas de clases oficiales sincrónicas presenciales más una media de 24 horas de trabajo asincrónico con un desvío estándar de 8 horas.

Usted deberá dedicar en promedio  $36 + 24 = 60$  horas a esta materia

Sin embargo, esas 60 horas son un promedio. Históricamente hay alumnos que llegan con un ímpetu extraordinario a la maestría, la asignatura les resulta de interés para su futuro laboral inmediato, disponen del tiempo, y dedican más de 80 horas a la asignatura.

Las 24 horas asincrónicas se dedican a :

- instalar, aprender a manejar herramientas como Git, GitHub, Google Cloud, Kaggle
- ver videos/lecturas previas para estar preparado para la siguiente clase
- diseñar, ejecutar, interpretar los resultados de experimentos que intenten mejorar los modelos predictivos. En función de esa interpretación, idear nuevos experimentos superadores
- compartir los resultados de los experimentos en documentos colaborativos
- leer los mensajes de sus compañeros en Zulip
- responder a sus compañeros en Zulip
- programar pequeños nuevos scripts, realizar modificaciones a los scripts oficiales de la materia
- buscar en Stack Overflow o su gran amigo Chat GPT alguna pista que lo ayude a solucionar un bug de su script (aquí va a pasar una interesante cantidad de horas ... sea bienvenido a la realidad del codeo )
- operar el entorno Google Cloud, administrando efectivamente las restricciones
- participar de "Experimentos Colaborativos" a los problemas que le sean asignados.
- armar la video presentación de Experimentos Colaborativos
- participar en el torneo de videos siendo jurado

Esperar a que una virtual machine de Google Cloud termine de procesar no cuenta para el cómputo de horas humanas trabajadas.

## 1.5 Modalidad y Criterios de Aprobación y Evaluación

La nota mínima de aprobación de la materia es 5.50 , la nota final oficial se redondea a un número entero, la máxima nota final posible es 10.

Cada alumno puede parametrizar como desea ser evaluado en la materia.

Deberá cargar los porcentajes de contribución de la nota, en la Planilla Colaborativa de la Asignatura, antes del lunes 20-mayo a las 23:59:59 . En caso de no cargar nada, se le asignarán los porcentajes default.

Nota Ordinaria			
Contribución Nota			Actividades Obligatorias
Min	Default	Max	
0%	5%	10%	Completado de la <i>Google Sheet Colaborativa</i> DURANTE las clases sincrónicas, tarea individual
0%	5%	10%	Completado de la <i>Google Sheet Colaborativa</i> para actividades de las Tarea para el Hogar, tarea individual
20%	25%	40%	Experimentos Colaborativos, presentado en el Google Slides Colaborativo, tarea grupal
0%	10%	15%	Video Presentación con resumen ejecutivo de Experimentos Colaborativos, tarea grupal
30%	49%	60%	Competencia en la plataforma Kaggle más scripts y experimentos en GitHub reproducibles. Tarea Grupal
0%	5%	15%	Participación <i>significativa</i> en conversaciones el foro de la materia con herramienta <i>Zulip</i> (tipo Slack) en donde los alumnos deban reflexionar profundamente sobre el camino y obstáculos encontrados en la resolución del problema, intercambiando ideas y brindándole soluciones; con la participación de los docentes.Tarea Individual
1%	1%	1%	Juez en Torneo de Videos tarea individual

Todas las actividades anteriores con una contribución mínima mayor a 0% son obligatorias, la no participación/entrega en cualquiera de ellas implica el desaprobación de la materia y pasar directamente a recuperatorio.

El total de los porcentajes elegidos debe sumar 100.

### 1.5.1 Modalidad y Criterios de Aprobación y Evaluación Competencia Kaggle

La asignatura está basada en la metodología de *Authentic Learning* donde se presenta a los alumnos un problema significativo del mundo real.

La competencia Kaggle de la asignatura se califica en función de la ganancia obtenida en el Private Leaderboard.

Debe quedar claro que para la calificación no se tiene en cuenta el Public Leaderboard en absoluto.

Kaggle por default elige la predicción que más ganancia obtiene en el Public Leaderboard, pero esto puede y con altísima probabilidad deberá ser modificado por el alumno, el que elegirá el modelo que a pesar de no ser el de más ganancia en el Public Leaderboard a su entender es el que más ganancia obtendrá en el privado.

Se puede participar en forma individual o en grupos de dos personas. No se permiten grupos de tres o más personas.

Para la nota final las ganancias del Private Leaderboard se normalizarán y se transformarán al intervalo  $[0,10]$  de nota.

Es parte de la filosofía de la materia la reproducibilidad de los experimentos. Su repositorio GitHub de la materia deberá permanecer completamente público y abierto, accesible para los profesores y sus compañeros, durante todo el transcurso de la asignatura hasta la entrega de las notas finales. Usted irá recibiendo feedback público en Zulip sobre sus scripts y parametrizaciones.

Como parte de la entrega usted deberá disponibilizar en su repositorio GitHub, en forma pública bajo una carpeta creará exclusivamente a ese propósito, los scripts y archivos de parámetros que permitan generar la solución definitiva que está entregando en Kaggle a partir del dataset original. Los profesores tienen que ser capaces de correr esos scripts y generar exactamente el archivo que usted ha subido a Kaggle como entrega final.

Su solución definitiva será analizada minuciosamente por los profesores para:

- garantizar que está haciendo modificaciones sustantivas a la solución oficial de la cátedra.
- garantizar que la solución de su equipo es original y sustancialmente distinta a la de los otros grupos.

En caso que los profesores posean inquietudes sobre su trabajo, usted pasará a una etapa de evaluación oral individual.

## 1.5.2 Experimentos Colaborativos

Para tomar buenas decisiones en la construcción de un excelente modelo predictivo se deben llevar a cabo una gran cantidad de experimentos para entender qué es lo que mejor funciona, en algunos casos solo para este dataset, en otros para la mayoría de los datasets.

Muchas veces esa necesaria cantidad de experimentos puede resultar abrumadora para quienes están enfrentando por primera vez un problema de dimensiones reales. Otras veces resolver una pregunta sobre la mejor decisión, debido a la naturaleza probabilística de los modelos predictivos requiere de repetir el experimento cambiando las semillas de los generadores de números pseudoaleatorios. En la asignatura vamos a repartir la carga de esta experimentación de forma colaborativa entre todos los alumnos, obteniendo conclusiones más robustas y reproducidas por pares.

Los instructores plantean problemas del modelado predictivo, de capital relevancia para el problema concreto propuesto en la asignatura, propondrán hipótesis experimentales y asignan a distintos grupos de alumnos a las tareas de diseñar y ejecutar, reproducir, presentar y revisar los experimentos, los que se presentarán en formato de video y se discutirán en clase y Zulip.

Los experimentos se vuelcan en un documento compartido accesible a todos los alumnos y constituyen la base para la toma de decisiones en la confección de un buen modelo predictivo final.

Es muy importante realizar una completa búsqueda bibliográfica para el experimento; lo que resulta un verdadero desafío para algunos experimentos.

Rúbrica de Experimentos Colaborativos	
Porcentaje	Concepto a Evaluar
10%	Hipótesis Experimental, será brindada por los profesores Los alumnos buscarán bibliografía que soporta, o no, la hipótesis experimental.
30%	Diseño experimental y materiales. El diseño del experimento muestra una acabada comprensión del problema a resolver, los experimentos están precisamente especificados y detallados, los scripts son adecuados.
20%	Procedimientos y recopilación de resultados La ejecución de los experimentos es correcta, los parámetros iniciales y resultados capturados son los adecuados.
20%	Análisis de los resultados. Se interpretan correctamente los resultados obtenidos
20%	Resultados, conclusiones y discusión con pares La conclusión determina claramente si la hipótesis experimental se cumple o no. Se presentan los resultados en forma clara y convincente, se debaten los mismos con el resto de los alumnos. Se relacionan los resultados con la bibliografía original. Se proponen nuevos experimentos en caso de cuestionamientos externos. Se responden adecuadamente las preguntas de los pares

El Grupo A debe hacer un video, el B debe hacer otro video distinto al del Grupo A; consiste en una presentación Resumen Ejecutivo del Experimento Colaborativo, que no supere los 5 minutos.

Los formatos preferidos para los videos son: YouTube, Prezi, Loom y Twitch. No se acepta la entrega de links a Google Drive o Dropbox. El video debe ser accesible para todos por lo menos hasta seis meses después de haber sido entregadas las notas, para que tengan acceso a todos ellos los alumnos que deben

realizar el recuperatorio de la materia. No es necesario que el video sea público, solamente podremos acceder los que tengamos el link y no podrá ser buscado en las redes.

Es parte fundamental de la tarea que los alumnos investiguen la forma de hacer una video presentación efectiva e intercambien ideas en Zulip. Ese descubrimiento será asistido por los profesores.

### 1.5.3 Evaluación participación en foro/chat Zulip

Zulip es el sistema nervioso de la asignatura ya que contribuye a generar una comunidad. Pasan más cosas en Zulip que en las clases sincrónicas.

Ver el capítulo Zulip dentro de Herramientas, Conceptos y Operación el detalle de lo que se espera de usted en Zulip

Rúbrica de participación en foro Zulip	
Porcentaje	Concepto a Evaluar
30%	Contenido. Genera preguntas o respuestas profundas, significativas, claras, fáciles de entender y con potencial para el proyecto de la asignatura. <u>Compartiendo</u> resultados logra atraer la atención y <u>colaboración</u> de sus compañeros a sus posts. Idealmente se transforma en un líder conceptual del grupo.
30%	Contribución al dinamismo de la comunidad. Plantea preguntas interesantes y relevantes, intenta motivar las discusiones grupales sobre tópicos relevantes a la asignatura, debate positivamente, participa de conversaciones iniciadas por otros.
20%	Colaboración en dudas operativas, colabora rápida y acertadamente con compañeros que requieren algún tipo de asistencia operativa en alguna de las herramientas del curso, lenguaje de programación, etc Idealmente se transforma en un referente en un tema específico del grupo.
20%	Frecuencia. <b>Al final del curso deberá tener como mínimo treinta participaciones.</b> Una consulta técnica cuenta como participación !

Cada alumno participa con su usuario en forma individual en Zulip

### 1.5.4 Evaluación *Google Sheet Colaborativa* DURANTE las clases

Completar durante la clase la Google Sheet colaborativa según la actividades solicitadas por el profesor con resultados razonables, no se acepta la carga posterior al fin de la clase.

### **1.5.5 Evaluación *Google Sheet Colaborativa en Tarea para el Hogar***

Completar la Google Sheet colaborativa según las actividades solicitadas por el profesor en la Tarea para el Hogar de la semana, con resultados razonables; el plazo máximo es el inicio de la siguiente clase.

## **1.6 Misión de la Asignatura**

Lograr que los alumnos sean capaces de resolver un problema de dimensiones reales del mercado argentino con una *excelencia* que sorprenda a sus pares.

Despertar la llama de la pasión sostenible por explorar con espíritu crítico nuevos saberes.

### **1.6.1 Objetivos de la Asignatura**

1. Resolver un problema de dimensiones reales del mercado local utilizando las herramientas tecnológicas para manejar grandes volúmenes de datos y ser capaz de generar una predicción competitiva.
2. Desarrollar una fuerte cultura de la experimentación siendo capaz de cuestionar todo lo que le han enseñado.
3. Desarrollar la creatividad para el Feature Engineering
4. Conocer y utilizar efectivamente las técnicas “estado del arte” en cuanto a
  1. Algoritmos y librerías de última generación de modelado predictivo sobre datos estructurados
  2. Optimización de hiperparámetros
  3. Interpretación de modelos predictivos
  4. Procesamiento de grandes volúmenes de datos en la nube

## **1.7 Unidades Temáticas**

1. Nociones elementales de la actividad bancaria, ciclo de vida de un cliente, valor vida de un cliente, retención de clientes, campañas de marketing directo.
2. Metodologías CRISP, Six sigma y SEMMA.
3. Comparación de modelos predictivos
  1. El problema de las múltiples comparaciones
  2. Estimación Montecarlo
  3. Validación Cruzada
4. Curva ROC, concepto de área bajo la curva AUC
5. Breve reseña de Árboles de Decisión y de Árboles de Estimación de Probabilidad.
6. Tratamiento de datasets desbalanceados
7. Optimización de hiperparámetros
  1. Red de Búsqueda (Grid Search)
  2. Optimización Bayesiana
8. Métodos de Ensamblado de Modelos Predictivos
  1. Voting
  2. Bagging
  3. Boosting
  4. Stacking
9. Algoritmos de Ensemble
  1. Random Forest
  2. Gradient Boosting of Decision Trees
    1. XGBoost (año 2016)
    2. LightGBM (año 2017)
10. Ingeniería de Atributos ( Feature Engineering )
  1. Tratamiento de nulos
  2. Selección de variables
  3. Variables derivadas, el problema de los cortes paralelos de los árboles de decisión.
  4. Incorporación de variables históricas
11. Metodología de Walk Forward Validation
12. Interpretación de predicciones por el método “Explicaciones Locales Aditivas de valores Shapley”
14. Nociones básicas de Storytelling
  1. armado de una narración, Pirámide de Freytag, Camino del Héroe de Joseph Campbell



2. buenas prácticas para el armado de gráficos
3. buenas prácticas para el armado de video presentaciones

### 1.7.1 Tecnologías a enseñar y reforzar

- Lenguaje Estadístico R
  - Librería [data.table](#) para el manejo de grandes volúmenes de datos
  - Librerías particulares de los distintos algoritmos ( rpart, ranger, xgboost, lightgbm, catboost)
- Plataformas RStudio y Jupyter Lab
- Control de versiones de código Git y GitHub
- Procesamiento en Google Cloud, creación de máquinas virtuales, imágenes del sistema operativo, instancias spot
- Plataforma de competencias Kaggle
- Rudimentos de la terminal de Linux

## 1.8 Bibliografía General de la Asignatura

La siguiente es la bibliografía general; en cada clase se brindará bibliografía específica sobre los temas vistos.

Libro de cabecera <https://hastie.su.domains/Papers/ESLII.pdf>

Hastie, T, Tibshirani, R. Friedman, J. *The Elements of Statistical Learning*, Second Edition, Springer Series in Statistics, Springer, 2017

Sitios con muy buenos artículos

- <https://machinelearningmastery.com/> muy precisos los artículos de Jason Brownlee
- <https://www.kdnuggets.com/>

Sitios populares fáciles de leer que hay que tomar con pinzas sus artículos ya que por lo general carecen de profundidad, y lo que es muy grave hay muchos con enormes errores conceptuales y prácticos:

- <https://towardsdatascience.com/>
- <https://medium.com/tag/data-science>

Papers y otros Libros

Chen, Tianqi, XGBoost: A Scalable Tree Boosting System, *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.

Duan, Tony, NGBoost: Natural Gradient Boosting for Probabilistic Prediction, *Proceedings of the 37th International Conference on Machine Learning*, 2020

Fawcett, Tom, ROC Graphs: Notes and Practical Considerations for Data Mining Researchers. *Technical Report HPL-2003-4*, HP Labs, 2003.

Flach, Peter Putting Things in Order, On the fundamental role of ranking in classification and probability estimation. , *18th European Conference on Machine Learning*, 2007

Gareth, James, An Introduction to Statistical Learning with Applications in R, *Springer Texts in Statistics*, Springer, 2017

Guo, P, Kim, Jm Rubin, R, How video production affects student engagement: An empirical study of MOOC videos, *Proceedings of the first ACM conference on Learning @ scale conference*, 2014

Hastie, T, Tibshirani, R. Friedman, J. *The Elements of Statistical Learning*, Second Edition, Springer Series in Statistics, Springer, 2017

Ke, Guolin, LightGBM: A Highly Efficient Gradient Boosting Decision Tree, *NIPS'17: Advances in Neural Information Processing Systems Conference*, 3148-3156 , 2017

Knaflitz, C. (2015). *Storytelling with data : a data visualization guide for business professionals* . Hoboken, New Jersey: Wiley.

Kuhn, Max, *Feature Engineering and Selection: A Practical Approach for Predictive Models*

Chapman&Hall/CRC Data Science Series, 2019

Lundberg, Scott, A Unified Approach to Interpreting Model Predictions, *31<sup>st</sup> Conference on Neural Information Processing Systems*, 2017

Pang-Ning Tan, Introduction to Data Mining second edition, *What's New in Data mining Series*, Pearson, 2018

Pearl, Judea, Causality, Models, Reasoning and Inference, *Cambridge University Press*, 2009

Provost, Foster, Robust Classification for Imprecise Environments. *Proceedings of the Fifteenth National Conference on Artificial Intelligence* (pp. 706-713). Menlo Park, CA: AAAI Press, 1998.

Prokhorenkova, Liudmila, CatBoost: unbiased boosting with categorical features, *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018

Raj, E, Engineering MLOps: Rapidly build, test, and manage production-ready machine learning life cycles at scale, Packt Publishing, 2021

Reshef, David, Detecting Novel Associations in Large Datasets, *Science* 2011, Vol. 334 no. 6062 pp. 1518-1524, 2011

Rokach, L, Ensemble Learning: Pattern Classification Using Ensemble Methods, Second Edition, Series in Machine Perception and Artificial Intelligence, WSPC, 2019

Salzberg, Steven, On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach. *Data Mining and Knowledge Discovery Journal*, Kluwer Academic Publishers, 1, 317-327. 1997.

Seni, Giovanni, Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions *Synthesis Lectures on Data Mining and Knowledge Discovery*, Morgan and Claypool Publishers, 2010

Para la búsqueda de papers se recomienda primero buscarlos en Google Scholar <https://scholar.google.com/> y si no está disponible bajarlo del inframundo <https://sci-hub.se/>

Página para bajar libros <https://libgen.rs/>

Herramienta para visualizar conexiones entre papers <https://www.connectedpapers.com/>

## 2 Metodología de Enseñanza

### 2.1 Generalidades

Basados en el [authentic learning](#) se plantea un problema de negocios de la vida real tal cual un alumno podría encontrar en un trabajo en Argentina, con una cantidad de datos real, y una forma de evaluación igual a la que sería evaluado en una empresa : ganancia en pesos argentinos de la campaña de retención de clientes y exposición con storytelling de los resultados.

La asignatura se dicta con el espíritu de la [teoría del conectivismo](#), donde los alumnos deben crear colaborativamente contenido, reflexionar sobre contenido de sus pares, discutirlo y generar nuevo contenido.

La dinámica de las clases es la del [flipped classroom](#) (aula invertida) donde los alumnos ven los materiales *más simples* asincrónicamente **antes** de la clase sincrónica, y una vez en la clase sincrónica se utiliza una dinámica de [active learning](#) basado en pares con algunas exposiciones del profesor para los temas más complejos. Dado que los alumnos pueden consultar por Zulip en forma asincrónica antes de la clase, técnicamente el nombre correcto de la metodología es *active learning forte*.

### 2.2 Manifiesto Pedagógico

Si una sola frase debiera resumir la pedagogía de esta asignatura sería la expresión de Plutarco:

**Enseñar, más que llenar un recipiente es encender un fuego.**

Las filosofías pedagógicas se pueden dividir a grandes rasgos en estas teorías antagónicas:

- el conductismo- instructivismo de Edward Thorndike y B. F. Skinner que es la predominante en el mundo y Argentina
- el constructivismo social de John Dewey y Seymour Papert que será utilizada en esta asignatura.

El estilo en esta asignatura es el del *aprendizaje basado en pares*, en donde a pesar de existir un sendero diseñado por el profesor se alienta continuamente a los alumnos a recorrer su propio camino, observar y aprender de sus pares, diseñar creativamente sus propios experimentos reflexionando profundamente sobre los resultados que se van obteniendo, cuestionando al establishment. Como referencia se puede tener la filosofía y comunidad que existe alrededor del lenguaje de programación Scratch y los conceptos de experiencias educativas de Mitchel Resnick.

La asignatura es una experiencia de gran intensidad intelectual y emocional que busca constantemente alentar a que cada alumno supere lo que se espera de él; se intenta lograr un efecto de resonancia grupal que amplifique el aprendizaje.

El estilo de dictado está basado en la frase "la emoción es el timón de la razón" donde cada tema es presentado de forma que genere gran emocionalidad en los alumnos, por momentos usted sentirá profundo amor y fascinación por las ideas que se estarán viendo, en otros literalmente sentirá odio.

Finalmente, aunque no menos importante, la asignatura busca la reflexión profunda y el cuestionamiento crítico de lo enseñado en otras asignaturas e incluso en ella misma. Citando a Richard Feynman

"The problem is not people being uneducated. The problem is that people are educated just enough to believe what they have been taught, and not educated enough to question anything from what they have been taught".

Dos videos y un artículo que reflejan aspectos de la forma de enseñanza

<https://www.youtube.com/watch?v=uRxD-pe3PN0>

Mitchel Resnick, Lifelong Kindergarten Group,

Media Lab, MIT

<https://web.media.mit.edu/~mres/papers/constructionism-2014.pdf>

Mitchel Resnick, Constructionism

<https://www.youtube.com/watch?v=9vp824ZbksI>

George Siemens, Connectivism

## 2.3 Pensamiento Creativo - Tinkering

Para ilustrar lo que se espera del alumno en esta asignatura, van unos párrafos del libro Resnick, Mitchel , *Lifelong Kindergarten: Cultivating Creativity through Projects, Passion, Peers, and Play*, The MIT Press, 2018

One of the students, named Nicky, started by building a car out of LEGO bricks. After racing the car down a ramp several times, Nicky added a motor to the car and connected it to the computer. When he programmed the motor to turn on, the car moved forward a bit—but then the motor fell off the body of the car and began vibrating across the table on its own.

Rather than trying to repair the car, Nicky became intrigued with the vibration of the motor. He played and experimented with the vibrating motor, and began to wonder whether he might be able to use the vibrations to power a vehicle. Nicky mounted the motor on a platform atop four “legs” (LEGO axles).

After some experimentation. He realized that he needed some way to amplify the motor vibrations.

To do that, he drew upon some personal experiences. Nicky enjoyed riding a skateboard, and he remembered that swinging his arms gave him an extra push on the skateboard. He figured that a swinging arm might accentuate the vibrations of the motor as well, so he connected two LEGO axles with a hinged joint to create an arm and attached it to the motor. As the motor turned, the arm whipped around—and amplified the motor vibrations, just as Nicky had hoped. In fact, the system vibrated so strongly that it frequently tipped over.

A classmate suggested that Nicky create a more stable base by placing a LEGO tire horizontally at the bottom of each leg. Nicky made the revision, and his “vibrating walker” worked perfectly. Nicky was even able to steer the walker. When he programmed the motor to turn in one direction, the walker vibrated forward and to the right. When he programmed the motor to turn in the other direction, the walker vibrated forward and to the left.

I was impressed with Nicky’s vibrating walker—but even more impressed by the strategies he used in creating it. As Nicky worked on his project, he was constantly tinkering. Throughout the process, he was playfully experimenting, trying out new ideas, reassessing his goals, making refinements, and imagining new possibilities.

Like all good tinkerers, Nicky was:

- **Taking advantage of the unexpected.** When the motor fell off of his car, Nicky didn’t see it as a sign of failure; he saw it as an opportunity for new explorations.
- **Drawing on personal experience.** When Nicky needed to amplify the vibrations of the motor, he relied on his experiences as a skateboarder and knowledge of his own body.
- **Using familiar materials in unfamiliar ways.** Most people don’t imagine LEGO axles as arms or legs, nor do they imagine LEGO wheels as feet—but Nicky was able to look at objects in the world around him and see them in new ways.

Usted no jugará con bloques de un LEGO , sino que configurará y reconectará scripts y bloques de código brindados por la cátedra, y si tiene la fortuna de ser atrapado por el entusiasmo y motivación de Nicky, desarmará scripts y creará algo totalmente innovador.

## 2.4 *Dinámica Constructivista en el Aula*

En caso de preguntarse cómo concretamente difieren el conductismo del constructivismo, aquí va un ejemplo de dinámica sobre como medir para niños de 6 años

- [https://storage.googleapis.com/open-courses/videos-d0a6/whale\\_measuring.mp4](https://storage.googleapis.com/open-courses/videos-d0a6/whale_measuring.mp4) (4 minutos)
- paper <http://www.gphillymath.org/ConstructivistLearn/Constructivist.pdf>

Sin embargo esta dinámica en el aula puede ser muy difícil para algunos alumnos adulto, ya que suele estar acostumbrado a la "Educación Colonial" [https://storage.googleapis.com/open-courses/videos-d0a6/colonial\\_education.mp4](https://storage.googleapis.com/open-courses/videos-d0a6/colonial_education.mp4) (1 minuto) y esperar "recetas" del profesor.

## 2.5 *Authentic Learning*

El enfoque pedagógico de esta asignatura está fuertemente relacionado con el *Authentic Learning* e indirectamente al Constructivismo Social en cuanto a:

- El aprendizaje comienza con el planteo de un problema del mundo real, de dimensiones reales, que la mayoría de alumnos muy probablemente deban enfrentar su actividad profesional.
- Algunos aspectos del problema están intencionalmente definidos de forma difusa para que los alumnos reflexionen profundamente sobre las mejores soluciones posibles.
- La evaluación de la asignatura refleja la evaluación del mundo real profesional, basada en la rentabilidad, comunicación de resultados y colaboración con pares.
- La teoría está al servicio de la práctica, y aparece justo a tiempo; jamás a la inversa.
- El profesor colabora con *Instructional Scaffolding* ayudando a escalar los distintos niveles de complejidad conceptual.

## 2.6 *Flipped Classroom Forte*

La idea es que el alumno reciba la ayuda directa del profesor en los momentos de aprendizaje que más lo necesita que generalmente es cuando el alumno se enfrenta a resolver un problema en soledad (no en el aula); para generar esa disponibilidad de ese tiempo lo fácil, la transmisión de información, se pasa para antes de la clase en modalidad asincrónica, cuando los profesores no están disponibles en tiempo real.

Además se evalúa lo asincrónico con quizzes y con las subidas a Kaggle, permitiendo a los profesores responder por el foro/chat Zulip preguntas simples, y dejar lo que no ha sido entendido para reforzarlo en la siguiente clase presencial. Los quizzes muestran que se está entendiendo y que no, por más que el alumno crea que está entendiendo todo lo que lee.

Una presentación sobre Flipped Learning Forte clara, concisa y carente del usual laberinto literario de las ciencias sociales, es <https://www.slideshare.net/alfredo.prietomartin/intercambio-hispano-chileno-de-experiencias-de-flipped-learning-en-enseanza-blended-y-online> junto con este artículo

<https://revistaventanaabierta.es/flipped-classroom-sus-limitaciones-y-aparicion-de-una-nueva-variante-flipped-learning-forte/>

## 2.7 Conectivismo

Las ideas de teoría del aprendizaje del conectivismo tienen gran semejanza con el deep learning, <https://www.youtube.com/watch?v=yx5VHpaW8sQ> (3 min)

- El aprendizaje ocurre a través de la red neuronal del cerebro, redes de humanos y redes de entidades no humanas.
- La red de conocimiento se desarrolla a partir de las interacciones con otras entidades de la red y el mundo en general. El objetivo de la enseñanza es estimular dichas interacciones.
- La red crece:
  - nodos
    - incorporando nuevos nodos
    - eliminando nodos ya no útiles
  - conexiones entre nodos
    - creando conexiones entre nodos que antes no estaban conectados
    - reforzando la conexión entre dos nodos, al ser transitado frecuentemente
    - disminuyendo el peso de la conexión, al ser poco transitado ese link
- El conocimiento reside en los pesos de las conexiones
- Aprender es el proceso de conectar nodos específicos o fuentes de información
- **La capacidad de aprender cosas nuevas es más importante de lo que se ya se conoce.**
- Tomar decisiones es un proceso de aprendizaje, de transitar reiteradamente y en formas novedosas la red

## 2.8 Active Learning

Active Learning significa que se llevará al mínimo lo que en la jerga se denomina "clase magistral" que es básicamente un profesor hablando y los alumnos escuchando; los alumnos se deberán involucrar activamente durante la clase con actividades diseñadas y guiadas por el profesor, en donde deban reflexionar profundamente sobre el tema que se está tratando.

Active Learning es *la metodología* en la que toda la comunidad educativa está de acuerdo que brinda resultados superadores. Puede verse este artículo "Active learning increases student performance in science, engineering, and mathematics" <https://www.pnas.org/doi/full/10.1073/pnas.1319030111>

Es tan distinto a la mecánica tradicional que muchos alumnos se sentirán incómodos al comienzo porque jamás han sido expuestos a la misma. No se confunda, active learning no tiene que ver con ser extrovertido, sino con ser reflexivo !

Videos cortos que explican los fundamentos del active learning:

[https://www.youtube.com/watch?v=z0a2pKYp\\_fk](https://www.youtube.com/watch?v=z0a2pKYp_fk) (5 min) Janet Rankin, MIT

<https://www.youtube.com/watch?v=Z9orbxoRofI> (14 min) Eric Mazur, Harvard University

Un artículo sobre aulas de active learning <https://er.educause.edu/articles/2017/12/creating-active-learning-classrooms-is-not-enough-lessons-from-two-case-studies>

Esta es una lista de ejemplo *no vinculante* del tipo de actividades que se llevarán a cabo (ir a Large Group ) [https://www.queensu.ca/teachingandlearning/modules/active/12\\_exmples\\_of\\_active\\_learning\\_activities.html](https://www.queensu.ca/teachingandlearning/modules/active/12_exmples_of_active_learning_activities.html) siendo la actividad más tradicional el Think Pair Share que es la que habla Eric Mazur

Al siguiente extenso video es una clase en el MIT donde en un posgrado de educación la profesora demuestra a los alumnos diversas dinámicas de active learning <https://www.youtube.com/watch?v=hGBNi4P9OfA> , no hace falta que lo vea, ya lo experimentará.



### 3 Arranque en Frío

En la materia trabajaremos con una gran variedad de herramientas, plataformas y conjuntos de datos, las que deben ser instaladas, configuradas, interconectadas y finalmente usted deberá aprender a utilizarlas. No se sienta desbordado por el shock inicial, aprenderá a utilizarlas. En palabras de un ex alumno "todas las semanas tuve en simultáneo la frustración de no entender alguna implementación y la satisfacción de dominar lo que una semana atrás parecía imposible."

Intente avanzar por su cuenta con los siguientes pasos de instalación, aunque no entienda en detalle lo que está haciendo, ya se le explicará el uso de las herramientas en clase.

En caso de tener dificultades con algún punto, o consúltelo en Zulip o espere a la primera clase en donde se verán conceptualmente estos pasos.

Palabras de un ex alumno : "Unos pocos empezamos la maestría con PC de Empresa con restricciones, bloqueos y rigideces. La fuimos "piloteando" pero cada vez se fue haciendo más cuesta arriba... Creo que en esta materia finalmente todos terminamos con PC libre. En mi caso fue un salto cuántico de comodidad."

#### 3.1 Alta en Plataformas

##### 3.1.1 Zulip herramienta de chat y foros discusión

[Registrarse](#) En la materia no se utilizará el email, toda la comunicación y discusión se llevará a cabo en Zulip. Los profesores serán por lejos quienes más participen en las conversaciones de Zulip.

Para darse de alta en el Zulip de la materia usted debe ingresar al link de Zulip de la sección 1.1 Links Fundamentales y presionar el link de [Sign Up](#) o Regístrese (abajo a la derecha del recuadro)

Usted puede acceder a Zulip de tres formas:

- En su PC desde algún browser
- En su PC utilizando la app, instálela desde <https://zulip.com/apps/>
- Desde su smartphone, busque la app en su *App Store* ( alerta : no es buena la app)

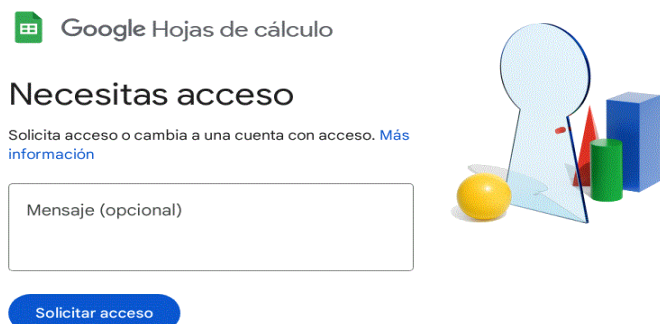
En el capítulo Herramientas, Conceptos y Operación se le enseñará cómo utilizar efectivamente Zulip a lo largo de la asignatura, ya que parte de la nota final depende de ello.

Una vez dentro de Zulip, vaya a la rueda dentada ubicada arriba a la derecha, Manage streams, elija el stream cartelera, y a la derecha marque Email notifications

### 3.1.2 Google Sheet Colaborativa de la Asignatura

Para proceder con este paso es necesario que usted deberá contar con una cuenta de gmail la que además le será indispensable para trabajar con la plataforma Google Cloud.

Siga el link de Google Sheet Colaborativa de la sección 1.2 Links Fundamentales de este documento. Le aparecerá una pantalla de este estilo



presione el botón de Solicitar acceso , a las pocas horas recibirá un email en su gmail informándole que ya tiene acceso a la planilla colaborativa, en donde deberá ir cargando resultados de las corridas que realice a lo largo de la asignatura y podrá comparar sus resultados con los de sus compañeros.

### 3.1.3 Plataforma Hypothes.is

Hypothes.is es una herramienta de anotación colaborativa de páginas web y documentos .pdf, que también utilizaremos para anotar scripts y Jupyter Notebooks . Le permitirá ver anotaciones que el profesor ha realizado en papers a utilizar en la asignatura, y usted mismo podrá agregar anotaciones.

La siguiente es una razonable introducción a Hypothes.is <https://www.youtube.com/watch?v=87h0nYi-i9o> presta atención a cómo se hacen las anotaciones y a cómo se participa en una discusión sobre una anotación.

Ir a la página <https://web.hypothes.is/start/> y seguir los pasos para crear un usuario, se recomienda enfáticamente crear un usuario con el formato que sea <su nombre> punto <su apellido> . Adicionalmente instale la extensión para el navegador Google Chrome, que le será indispensable para la anotación de documentos .pdf

No haremos públicas las anotaciones en la asignatura, sino que las acotaremos al grupo privado [itba-dm](#), para sumarse a estos grupos estando logueado a Hypothes . is en su browser y luego seguir este link <https://hypothes.is/groups/BRPv86ZM/itba-dm>

### 3.1.4 Plataforma ChatGPT

Utilizaremos el novedoso ChatGPT para que escriba scripts simples por nosotros.

Ingresa a <https://chat.openai.com/auth/login> y cree un usuario gratuito. Dada la alta demanda de ChatGPT es posible que deba probar en distintos horarios hasta lograrlo.

### 3.1.5 Plataforma Kaggle

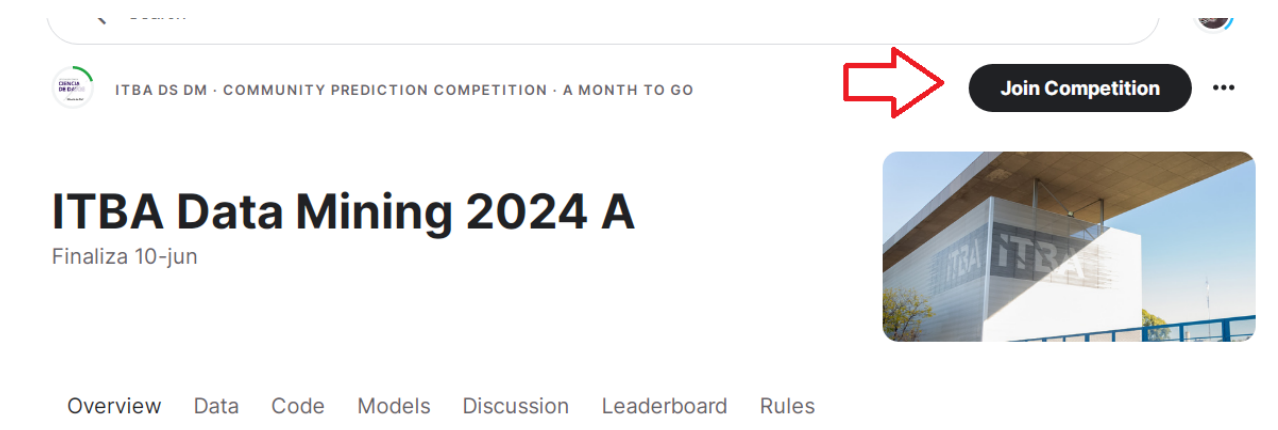
Utilizaremos la plataforma Kaggle para alojar la Competencia Kaggle de la Materia la que será cerrada, limitada a los alumnos de la materia.

Kaggle es, entre otras cosas, una plataforma que permite participar de competencias mundiales de ciencias de datos. Es muy útil para el dictado de cursos ya que genera un notable involucramiento de los alumnos, los que todo el tiempo van teniendo feedback de qué tan buenos son los modelos predictivos que van probando en comparación a los compañeros.

Si no posee un usuario de Kaggle deberá registrarse a la plataforma Kaggle utilizando este link <https://www.kaggle.com/account/login?phase=startRegisterTab> se sugiere que para ello utilice su verdadero nombre (no un pseudónimo).

Anote el nombre de usuario ya que lo requerimos más adelante.

Una vez que ya posea un usuario de Kaggle, deberá registrarse en la competencia, para ello debe seguir el link de invitación a la competencia del capítulo 1.2 *Links Fundamentales* y luego presionar el botón negro que dice Join Competition.



Preste atención a que hay dos links de Kaggle, uno es la invitación a la competencia, que será necesario solo la primera vez, y el otro link es con el que ingresará asiduamente.

Lea con atención la sección Overview del menú principal, que contiene las subsecciones Description, Evaluation, Evaluation Details.

Una vez registrado en Kaggle con su número telefónico verificado, vaya a <https://www.kaggle.com/settings/account> y presione

# Settings

Control over your Kaggle account and all communications

Account   Notifications

## Your email address

gustavo.denicolay@gmail.com

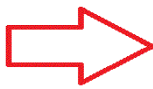
Change email

## Phone verification

Verified

## API

Using Kaggle's beta API, you can interact with Competitions and Datasets and more via the command line. [Read the docs](#)



Create New Token

Expire Token

## Quotas

Guarde en un lugar seguro el archivo que se bajará, contiene su usuario Kaggle y una key. Lo necesitará en la instalación de Google Cloud, para poder subir en forma automática submits.

### 3.1.6 Plataforma GitHub

Para poder compartir nuestros scripts y experimentos habrá un repositorio oficial de la asignatura y uno suyo, ambos residirán en la nube. Para ello usted dará de alta una cuenta en los servicios de la plataforma GitHub

Registrarse en [https://github.com/signup?ref\\_cta=Sign+up](https://github.com/signup?ref_cta=Sign+up)

Una ayuda simple es <https://www.classicpress.net/github-desktop-a-really-really-simple-tutorial/>

Una vez ya registrado y dentro de su sesión de GitHub, es momento de crear su Personal Access Token que será algo así como su usuario y clave para acceder desde su PC local y desde Google Cloud.

Dentro de su sesión de GitHub vaya a la página <https://github.com/settings/tokens/new> ,

- donde dice Note escriba "mi primer token",
- en Expiration marque Custom y elija como fecha el 31-dic-2024 (realmente es muy recomendable que ponga una fecha de expiración)
- luego en Select Scopes marque TODOS los recuadros que le aparecen

- al final de la página presione el botón **Generate Token**

Le aparecerá una pantalla como esta:


## Personal access tokens

Generate new token

Revoke all

Tokens you have generated that can be used to access the [GitHub API](#).

Make sure to copy your personal access token now. You won't be able to see it again!

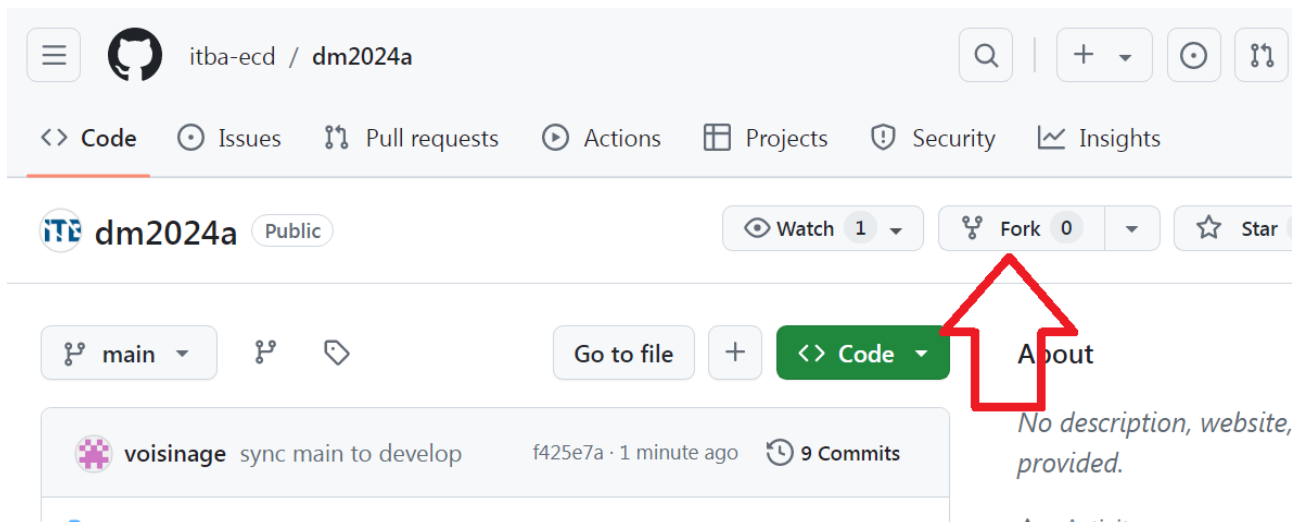
✓ ghp\_bG7T3t7vMRrUHlbrMYIVHzJnOJFDs11pCGxz 

Delete

allí copie y guarde en un lugar seguro el token, que en el ejemplo de la imagen es el string `ghp_bG7T3t7vMRrUHlbrMYIVHzJnOJFDs11pCGxz`

usted volverá a utilizar ese token cuando instale Google Cloud, guárdelo en un lugar seguro por favor

Una vez hecho el login a su GitHub usted no tendrá ningún repositorio creado inicialmente, ir al repositorio GitHub de la materia cuyo link está en el capítulo *Links Fundamentales* y hacer un fork del repositorio presionando el botón que está arriba a la derecha de la página de GitHub .



Le aparecerá un formulario para crear el fork

De esta forma usted pasará a tener en su cuenta de GitHub una copia del repositorio oficial de la asignatura. Ambos están en la nube, pero usted no puede trabajar directamente con esos archivos en la nube sino que debe generar una copia en su computadora de su repositorio GitHub local.

### 3.1.7 Check In a la Asignatura en Zulip

Para los profesores es muy importante conocer quién es usted de forma de personalizar los mensajes en Zulip, recomendar lecturas personalizadas acordes a su formación, experiencia e intereses.

Enviar un gran mensaje Zulip al stream `z-CheckIn` (recuerde que es público)

1. Nombre y Apellido
2. LinkedIn
3. Usuario Kaggle
4. Usuario GitHub.com
5. Edad
6. Carrera de grado, Universidad, año de graduación
7. Posgrados realizados o en curso (no incluya a este posgrado...)
8. ¿Posee experiencia docente? explíquela
9. Educación previa en ciencia de datos ( Coursera, edX, Udemy, Digital House, etc ) detalle
10. ¿Qué lo motivó a realizar este posgrado en Ciencia de Datos?
11. ¿A que se dedica en su trabajo actual? ¿Trabaja en ciencia de datos o en algo relacionado? En caso afirmativo detalle.
12. ¿En su trabajo actual tiene profesionales ( o equivalente) que le reportan directamente a usted?
13. ¿Cómo prefiere aprender usted, como el hijo o como la hija de Feynman? Vea el siguiente video de Richard Feynman, ganador del Premio Nobel por sus aportes en Física Cuántica, considerado el mejor profesor del siglo XX <https://www.youtube.com/watch?v=BY6VntTmtIo>
14. ¿Cuántas horas ha dedicado a programar en algún lenguaje en lo que va del año 2024, sin tener en cuenta este posgrado? Relate lenguajes y tareas.
15. ¿Tiene experiencia en la metodología y realización de experimentos, ya sea de ciencias naturales, físicas, o simulaciones informáticas?
16. Primero lea en detalle <https://subscription.packtpub.com/book/business-and-other/9781787287037/1/ch01lvl1sec13/12-developer-learning-curve-why-learning-how-to-code-takes-so-long>

y luego relate su experiencia con lenguajes/entornos de programación/algoritmos

1. lenguaje R
2. Tidyverse y dplyr en R
3. data.table en R
4. Git / GitHub
5. Cloud Computing (Google Cloud, Azure, AWS, ...)
6. lenguaje Python
7. lenguaje Julia
8. SQL

9. Sistema operativo Linux
10. Árboles de Decisión
11. Gradient Boosting (XGBoost/LightGBM)
17. Utiliza algún entorno en particular de R ( Rstudio, R Base, etc )
18. Utiliza algún entorno en particular de Python ( VS Code, PyCharm )
19. ¿Ha participado en alguna competencia Kaggle *fuera* de este posgrado ? relate
20. Elija CINCO números primos,  $100003 < p < 999983$  , intentando que sus números no se solapen con los que elijan sus compañeros de curso. Serán utilizados como *sus* semillas de generadores números pseudoaleatorios a lo largo de la asignatura.
21. Suba una imagen que represente como ha sido su vivencia de la maestría, si lo desea puede generarla con inteligencia artificial.

## 3.2 Instalación de Herramientas

Para proceder con los siguientes pasos usted deberá ya haber dado en alta en todas las plataformas

### 3.2.1 Lenguaje de programación R

En la materia se trabajará con el lenguaje estadístico R, el cual en las primeras clases lo utilizará en su propia computadora y luego en máquinas virtuales de gran tamaño corriendo en la nube Google Cloud.

En caso de no tenerlo instalado en su computadora proceder a la página <https://cran.r-project.org/> , y según su sistema operativo ( Windows, MacOS, Linux ) seguir las instrucciones. Elija SIEMPRE la versión de 64 bits.

Trabajaremos con la versión 4.4.0 liberada el 24-abril-2024 , **no utilice versiones anteriores, evítese problemas de compatibilidad de librerías.**

### 3.2.2 Librerías lenguaje R

Finalmente, desde la consola de R correr lo siguiente ( no ingrese a RStudio )

```
install.packages(c('repr', 'IRdisplay', 'evaluate', 'crayon') )
install.packages(c('pbdZMQ', 'devtools', 'uuid', 'digest'), dependencies=TRUE )

install.packages('languageserver')
library('devtools')
devtools::install_github("ManuelHentschel/vscDebugger")
install.packages('IRkernel')
```

finalmente, salir de la consola de R con el comando `quit()`

### 3.2.3 Aplicación RStudio Desktop

Es muy recomendable instalar RStudio Desktop, trabajaremos con la última versión, que al momento de creación de este documento es la 2024.04.0+735 liberada el 29-abril-2024

En caso de no tenerlo instalado proceder a instalarlo de <https://www.rstudio.com/products/rstudio/download/#download>

Una vez instalado el entorno de desarrollo RStudio Desktop verifique que puede ingresar al mismo.

En RStudio Desktop trabajaremos con proyectos en lugar de scripts sueltos, de forma de poder tener versionado de código fuente.



### 3.2.4 Estructura de Carpetas en su PC local

Cree en su computadora una carpeta exclusiva para la asignatura, elija un nombre corto y en lo posible que en toda la ruta no tenga espacios.

Dentro de esa carpeta crear la siguiente estructura de carpetas

- datasets
- exp
- TareasHogar

Bajar a la carpeta `datasets` el archivo señalado en Dataset Inicial del capítulo *1.2 Links Fundamentales*. Este archivo llamado `dataset_pequeno.csv` es el que utilizaremos en las primeras clases.

### 3.2.5 Aplicación Git

En ciencia de datos, dado un script inicial es una práctica usual generar múltiples versiones con pequeños cambios ya sea de parámetros o de funcionalidad en el código para experimentar alternativas que mejoren la métrica que se intenta optimizar. Para hacerlo en forma ordenada trabajaremos con control de versiones, la aplicación Git y usaremos el repositorio en la nube Github

Instalar Git en su computadora local siguiendo estas instrucciones <http://git-scm.com/downloads>

Videos introductorios a Git

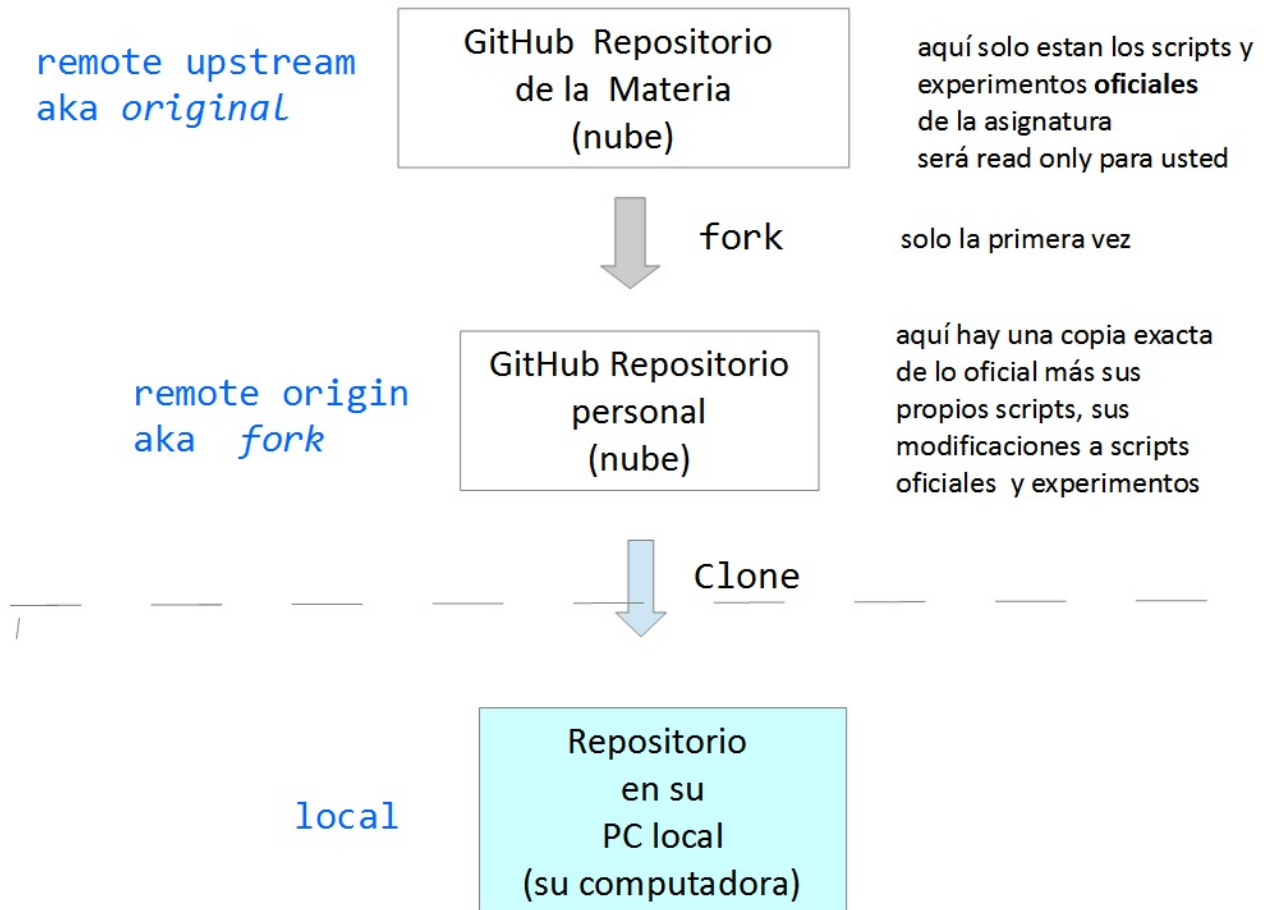
- <https://www.youtube.com/watch?v=2mxh3tgx71c>
- <https://www.youtube.com/watch?v=kEPF-MWGq1w>

Tutorial en slides

- <http://rogerdudler.github.io/git-guide/index.es.html>

### 3.2.6 Clonado repositorio

Conceptualmente esto es lo que estamos haciendo esta primera vez



ahora, usted debe crear una copia en su computadora de su Repositorio GitHub personal.

Para ello vaya por la línea de comando **cmd** hasta la carpeta de la materia que creó en el punto 3.2.6 y ejecute el siguiente comando

```
git clone https://MariaLopez:token@github.com/MariaLopez/dm2024a.git
```

debe reemplazar

- **MariaLopez** por su nombre de usuario en GitHub
- **token** por el token que acaba de generar en el página anterior

verá que se crea en su computadora local una carpeta llamada dm2024a, que contendrá por ahora muy pocos archivos.

Luego, desde la línea de comando ingrese a la carpeta dm2024a con el comando `cd dm2024a` y una vez allí dentro ejecute lo siguiente

```
git remote add upstream https://github.com//itba-ecd/dm2024a
git config user.email "su-email"
git config user.name "MariaLopez"
```

debe reemplazar

- `MariaLopez` por su nombre de usuario en GitHub ( va con las comillas )
- `su-email` por su dirección de email ( va con las comillas )

Algo importante a tener en cuenta es que en este momento usted posee una copia del repositorio oficial de la materia tal cual está en este instante. Dicho repositorio será actualizado decenas de veces por los profesores a lo largo de la materia, con lo cual usted deberá cada vez que se lo notifique, "volver a traer" esos cambios, y no será ni con un fork ni con un clone.

En el capítulo Herramientas, Conceptos y Operación se le enseñará como utilizar Git & GitHub a un nivel básico.

### 3.2.7 Prueba de RStudio y Kaggle

El objetivo de este paso es probar parte del entorno que acaba de instalar. Correrá un script que genera un árbol de decisión en los datos de entrenamiento, y luego lo aplica a los datos del futuro, que no tiene clase, y finalmente a esa predicción la deberá subir a Kaggle. El problema se explicará en la primera clase.

Ingresa al RStudio e instala las tres librerías: `data.table`, `rpart`, `rpart.plot`

Desde RStudio abre el archivo `./dm2024a/src/rpart/z101_PrimerModelo.R`

En la línea 10 del script cambia la ruta a la ruta de su carpeta de la materia, esta tarea suele resultar desafiante a gran cantidad de alumnos, NO se desanime si no lo logra, en última instancia recibirá ayuda personalizada en la primera clase.

Ejecutar el script línea a línea.

Finalmente deberá generar en la carpeta kaggle el archivo `./exp/KA2001/K101_001.csv`

Básicamente ha corrido un experimento de tipo Kaggle por eso las letras iniciales son KA

Sus números de experimento comenzarán a partir del 2001 en adelante

Ingresa a Kaggle a la página de la competencia, y presiona el botón negro con letras blancas en el menú superior izquierdo que dice "Submit Predictions", una vez allí:

1. ir al Step 1 que consiste en hacer el upload del archivo `K101_001.csv`
2. luego el Step 2 póngale un nombre significativo al submit
3. Finalmente presione el botón que está en el centro inferior de la pantalla y dice "Make Submission".

Verificar que ha sido exitosa.

Felicitaciones, usted ha corrido RStudio y subido su primera predicción a Kaggle; si visita el Leaderboard verá que está obteniendo una ganancia de más de 30 millones

Si se le ha presentado algún error que no pudo solucionar, consúltelo en Zulip, que un profesor o compañero se lo contestará.

### 3.2.8 Alta en GitHub Copilot

GitHub Copilot tiene una versión 100% gratis para estudiantes mientras sean alumnos regulares ( no confundir con la versión que es gratis el primer mes y luego USD 10 mensuales)

Estado conectado desde el browser a su cuenta de GitHub, vaya a [https://education.github.com/discount\\_requests/application](https://education.github.com/discount_requests/application) y siga cuidadosamente todos los pasos que le piden, donde deberá proveer una imagen que compruebe que es alumno regular, generalmente una constancia de alumno regular de ITBA es suficiente.

Recién a los cuatro días tendrá acceso a GitHub Copilot, lo podrá ver en <https://github.com/settings/copilot> Luego, en clase aprenderá a conectar su VSCode con GitHub Copilot

Finalmente, lea <https://docs.github.com/en/copilot/overview-of-github-copilot/about-github-copilot-for-individuals>

## 4 Herramientas, Conceptos, Operación y Buenas Prácticas

### 4.1 Zulip

Zulip es el sistema nervioso de la asignatura. Pasan más cosas en Zulip que en las clases presenciales.

Utilizaremos Zulip para

- Solicitar y brindar soporte técnico.
- Mantener discusiones en foros, asincrónicos fuera de las clases.
- Intercambiar mensajes, links, artículos, archivos entre alumnos entre sí, alumnos y profesores
- quizzes en la clase sincrónica y también en la tarea para el hogar asincrónica
- Cartelera de información de la asignatura
- resultados de Experimentos Colaborativos
- Entregas de la materia

Zulip es una aplicación, open source, de chat y foros de discusión con videoconferencia integrada, que permite organizar las conversaciones en *streams* y *topics* posibilitando mantener intercambio de ideas en paralelo, lo que es una gran ventaja si no se está online todo el tiempo. La existencia de hilos de conversación (*streams*) y tópicos permite una eficiente conversación no lineal y asincrónica.

Zulip es open source y está corriendo en un servidor propio, Slack no es open source. Zulip no está tan pulido y es menos intuitivo que Slack. Sin embargo Slack tiene la contra que revisar canales con mucha actividad se vuelve muy tedioso ya que no se dividen en tópicos.

Zulip permite enviar estos tipos de mensajes :

- privados a otro usuario o grupo de usuarios o a un stream privado
- públicos, a streams públicos, que actúan como salas de chat/foros de discusión

**Todas las preguntas a profesores, sobre cualquier tema de la asignatura, deben ser siempre públicas.** Es mandatoria esta transparencia. Escribir mensajes privados es caer en el oscurantismo de los emails, es dejar de compartir con el resto y evita que se pueda leer lo que otros están pensando.

Lo más probable es que usted acceda a Zulip desde su computadora y además tenga instalada la app en su smartphone (Android y iOS están soportados)

Zulip como aplicación permite esto <https://zulip.com/features/> , una relativa "falta" de Zulip es que no permite enviar mensajes de voz en forma nativa.

Un video que muestra como utilizar Zulip es <https://www.youtube.com/watch?v=xWa56KdgYZM>

Stream	Utilidad
z-CheckIn	CheckIn , será utilizado solamente para enviar el check in a la materia, la información es pública, y servirá de consulta constante a los profesores para personalizar la comunicación con cada alumno.
cartelera	alumnos no pueden enviar mensajes, profesores anuncian <ul style="list-style-type: none"> <li>• recordatorios fechas importantes</li> <li>• commits del repositorio oficial</li> <li>• actualizaciones Libro de la Asignatura</li> <li>• bibliografía nueva</li> <li>• comunicaciones generales</li> </ul>
Clase $n$	Cada clase tendrá su correspondiente stream, y estos dos tópicos estarán siempre presentes: <ul style="list-style-type: none"> <li>• <b>antes</b> : de que tratará la clase y que debo traer hecho/leído</li> <li>• <b>despues</b>: resumen de lo que fue visto en clase, de las conclusiones</li> </ul>
arranque en frio	todos los problemas que surgen en las instalaciones de Arranque en Frío
SoporteTecnico	Soporte técnico sobre R, Rstudio, scripts oficiales, librerías de R generales, no incluidos en streams específicos
data.table	Temas específicos y soporte técnico sobre data.table
Google Cloud	Temas específicos y soporte técnico de Google Cloud
Git & GitHub	Temas específicos y soporte técnico de Git y GitHub
Feature Engineering	Intercambio de ideas sobre la creación de atributos nuevos
Estrategias	Discusión de estrategias sobre cómo mejorar los modelos predictivos. Temas principalmente conceptuales, pero también aplicados.
Exp Colaborativos	Intercambio Experimentos Colaborativos
general	Diálogos generales.
off-topíc	estados de ánimo generales, memes generales, todo lo que no valdría la pena ir a buscar a los threads oficiales
zero2hero	temas sobre From Zero to Hero
z-entregafinal	será utilizado para entregar la video presentación al Gerente de Ciencia de Datos
mejoras2025	discusión y propuestas de mejoras para la edición 2024 de la materia, aparecerá justo antes de finalizar la materia.

Los streams en fondo amarillo serán habilitados a su debido momento.

Si usted desea que un profesor preste atención a su mensaje, simplemente incluya @profesores en el texto de su mensaje.

Alumnos de promociones anteriores han manifestado la importancia que tiene elegir correctamente el thread y tópico de cada mensaje : poder encontrarlos en el futuro, tener lo de un mismo tema todo en un solo lugar.

Según la dinámica del curso es posible que se agreguen threads o que jamás se activen algunos de la tabla.

#### 4.1.1 Buenas prácticas para el uso de Zulip

Zulip es una herramienta asincrónica, sus compañeros y profesores pueden escribir en cualquier momento del día o la noche, incluso durante el fin de semana.

En caso que usted no pueda acceder en forma diaria a Zulip, al ingresar se sentirá abrumado por la enorme cantidad de mensajes que han enviado sus compañeros y se planteará : “¿Qué es lo importante y que es lo que puedo obviar?

El stream cartelera es fundamental, ya que es ahí donde los profesores pasan unos pocos anuncios vitales.

Todo mensaje privado que reciba de compañeros, también será vital que los responda cuanto antes. Toda respuesta a un previo mensaje suyo también será importante que la vea cuanto antes. Si configuró bien Zulip, en estos dos casos usted recibirá un email para avisarle que tiene algo importante.

Los streams clase *n* y sus tópicos antes y después también son vitales y debe leer lo antes posible todo mensaje que aparezca, ya que lo orientarán.

Es una maravillosa práctica que usted cree un stream personal en Zulip con su nombre completo, que solamente usted podrá ver y escribir en él, totalmente invisible a todos, incluso invisible a los profesores. En ese stream personal usted podrá escribir sus ideas, y también guardarse links (punteros) a otros mensajes que encontró interesantes.

Para todo el resto, le recomendamos esta estrategia que le ha resultado muy útil a alumnos de años previos:

1. Empezar a leer superficialmente todos los mensajes no leídos
2. Si el mensaje vale la pena, marcarlo con la estrella “Star this message Ctrl + s”
3. Luego, cuando disponga un momento de tranquilidad, ir a la carpeta de ★ Starred Messages y leer cada mensaje en detalle, quizás hasta preguntando algo a su autor. Una vez leído a conciencia el mensaje, es muy importante guardar un link del mismo en su stream privado (el que tiene su nombre) y lo desmarca.
4. Con este método, ★ Starred Messages contiene los mensajes que temporalmente están pendientes de revision, y su stream privado contiene punteros a los mensajes que le interesaron.

#### **4.1.2 Participaciones extraordinariamente significativas en Zulip**

Son participaciones extraordinariamente significativas :

- Responder una pregunta técnica a un compañero.
- Realizar una pregunta conceptual (no de soporte técnico).
- Recomendar bibliografía junto con un comentario por haber sido leída previamente.
- Compartir resultados de análisis exploratorio de datos.
- Compartir un problema que se tuvo y la forma en que se soluciona.
- Compartir tips sobre como operar eficientemente alguna aplicación.

#### **4.1.3 Participaciones significativas en Zulip**

Son participaciones significativas en Zulip:

- Una pregunta técnica que aún no fue realizada por un compañero

La cátedra prefiere que usted realice preguntas técnicas por Zulip en lugar del grupo privado de whatsapp



para así poder entender los temas que no están quedando claros y reforzarlos ya sea por Zulip mismo o en la clase.

Es preferible curar a ese paciente aún con vida, antes que una vez finalizada la materia realizar una autopsia para entender qué salió mal y salvar vidas futuras.

#### **4.1.4 Mensajes Privados a los profesores Permitidos**

**Todas las preguntas a profesores, sobre cualquier tema de la asignatura, deben ser siempre públicas.**

La heterogeneidad de alumnos en las maestrías de ciencias de datos genera que por un lado nos encontramos con alumnos que ya saben programar, se encuentran trabajando en ciencia de datos y realizan preguntas avanzadas en clase, mientras que otros recién están aprendiendo lo básico.

Por lo general, varios alumnos se sienten tentados a preguntar en forma privada al profesor, manifiestan tener miedo que su pregunta sea elemental, o incluso mal formulada, que pueda ser mal vista por el resto. Se sienten incómodos con dicha exposición. Le doy una maravillosa noticia, el 65% de los alumnos está en su misma situación, con lo que al hacer en forma pública la pregunta que considera elemental, dará una alentadora señal al resto y se producirá la propiedad emergente que muchos más se animen a preguntar, y tal cual lo expresa Eric Mazur será muy posible que obtenga una clara respuesta de un compañero que acaba de aprender el tema hace muy poco.

Estos son ejemplos de mensajes privados dirigidos a los profesores, que por no hablar de la asignatura, y ser sobre temas privados, si están permitidos.

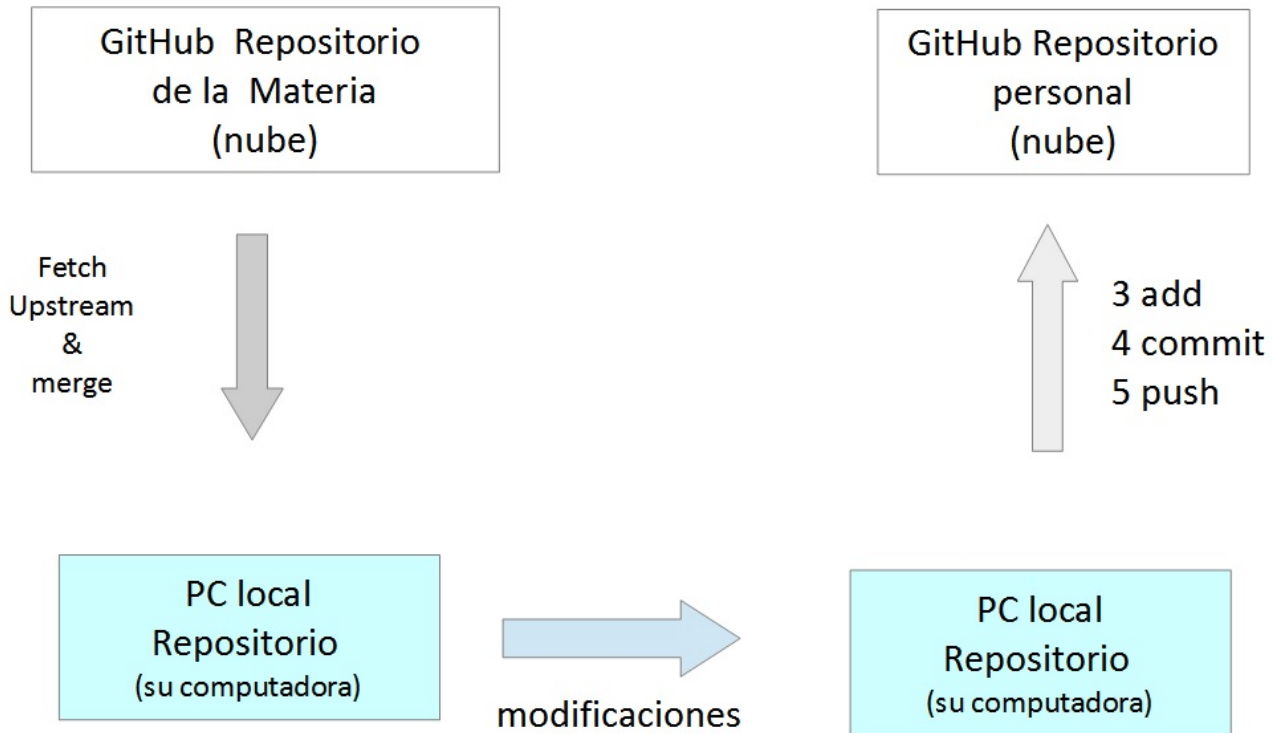
"Me surgió una posibilidad laboral en Madrid y la acepté, por lo que me voy a estar yendo en las próximas semanas y estoy haciendo gestiones a destajo. Como te podrás imaginar voy a tener que dejar el posgrado, aunque pienso empezar otro allá. Y en este punto, si sabés y tenés ganas de pasarme info, me vendría muy bien recomendaciones de buenos programas de maestría en ciencia de datos."

"Más adelante... si es posible y no es molestia, quisiera consultarte si podrías orientarme en encontrar una aplicación (PROYECTO DE TESIS) en la Industria Petroquímica / Generación de Energía....que es donde me desempeño actualmente.

Por supuesto que encontrar la aplicación es mi tarea, pero quizás algún exalumno hizo o está haciendo algo dentro del campo...o hay algún paper que se me esté escapando... . Cualquier dato es bienvenido. "

## 4.2 Git y GitHub

### 4.2.1 conceptos y operación



o desde la línea de comando de la PC local

```
git checkout main
git pull

git fetch upstream
git checkout main
git merge -Xtheirs upstream/main -m "Z manda"

git checkout main
git add .
git commit -m "otro commit"
git push
```

### 4.3 Librería *data.table* del lenguaje R

En el lenguaje R tradicional se utiliza el objeto `dataframe` que posee severas deficiencias en cuanto a performance y rebuscada complicada. En esta asignatura utilizamos la librería `data.table` que es ampliamente superadora y que permite manejar grandes volúmenes de datos, tiene una sintaxis más simple PERO muy distinta a la de `dataframes`, por lo que es necesario aprenderla.

Leer los siguientes artículos

- <https://towardsdatascience.com/data-table-vs-best-data-object-c95b7d5f0104>
- <https://towardsdatascience.com/blazing-fast-data-wrangling-with-r-data-table-de5045cc4b4d>
- introduction to data table <https://cran.r-project.org/web/packages/data.table/vignettes/datatable-intro.html>
- data.table cheat sheet [https://www.beoptimized.be/pdf/R\\_Data\\_Transformation.pdf](https://www.beoptimized.be/pdf/R_Data_Transformation.pdf)

## 5 Plataforma Kaggle

### 5.1 La partición { Public, Private }

Los datos del mes a predecir 202109 han sido particionados por Kaggle en forma aleatoria, estratificada en el campo `clase_ternaria`, en dos partes, un 30% que es llamado Public y el restante 70% es llamado Private

Particion	registros	%
Public	49530	30.00%
Private	115563	70.00%
<b>Total</b>	<b>165093</b>	<b>100.00%</b>

Que la partición fue realizada en forma estratificada en la clase ternaria significa que en cada partición se mantienen las proporciones de cada clase

Particion	registros	clase_ternaria		
		BAJA+1	BAJA+2	CONTINUA
Public	49530	311	345	48874
Private	115563	722	804	114037
<b>Total</b>	<b>165093</b>	<b>1033</b>	<b>1149</b>	<b>162911</b>

Notar como para la proporción de BAJA+2 se mantiene lo más fiel posible

Public  $345/1149 = 30.02611\%$

Private  $804/1149 = 69.97389\%$

Esta división se ha hecho con total honestidad y sin ningún tipo de trampa, téngalo bien presente ya que se sorprenderá de lo distinto que dan las ganancias de un mismo submit en el Public y Private cuando empiece a subir varios modelos y los profesores en algunas ocasiones, en las primeras semanas de la asignatura, le muestren por breves momentos el Private Leaderboard.

La razón de esto es simplemente la varianza de la distribución binomial por más que la partición es estratificada en la `clase_ternaria`.

## 5.2 Cálculo de ganancias

El archivo .csv que se sube a Kaggle tiene la siguiente forma, donde Predicted=1 significa “se ha tomado la decisión de enviar estímulo a ese cliente” tomando valores { 0, 1 }

numero_de_cliente	Predicted
31116053	0
31116803	0
31117730	1
...	...

Al hacer un submit, Kaggle calcula en primera instancia la “ganancia cruda” para la partición Public y también para la partición Private.

La ganancia cruda se calcula sumando para cada registro

Predicted	clase_ternaria	ganancia
1	BAJA+1	-3000
1	BAJA+2	117000
1	CONTINUA	-3000
0	BAJA+1	0
0	BAJA+2	0
0	CONTINUA	0

La ganancia normalizada en el Public es  $\text{ganancia\_cruda} / 0.3$

La ganancia normalizada en el Private es  $\text{ganancia\_cruda} / 0.7$

### 5.3 Elección manual de un submit

Es posible, y deseable, que usted elija manualmente un submit particular de todos los que ha subido a Kaggle.

Esto se realiza de la siguiente marcando el uno de los recuadros que aparecen en la columna Select (si uno ya estuviera marcado, se debe desmarcarlo primero antes de marcar el nuevo)

## Submissions

1/1

Select up to 1 submissions that will count towards your final leaderboard score. If less than 1 are selected, Kaggle will automatically select from your best scoring submissions. [Learn More](#)

☒ Auto-selection candidates 








All

Successful

Selected

Errors

Recent ▼

Submission and Description		Public Score 	Select
	<b>ZZ7958_02_019_11000pr_003.csv</b> Complete · 4m ago	<b>58.22894</b>	<input type="checkbox"/>
	<b>ZZ7958_02_019_10500pr_014.csv</b> Complete · 5m ago	<b>56.39898</b>	<input checked="" type="checkbox"/>
	<b>ZZ7990_01_013_10000p.csv</b> Complete · 7m ago	<b>53.78902</b>	<input type="checkbox"/>
	<b>KA3210_100.csv</b> Complete · 15m ago	<b>47.70913</b>	<input type="checkbox"/>
	<b>KA3210_005.csv</b> Complete · 16m ago	<b>44.1192</b>	<input type="checkbox"/>
	<b>K101_001.csv</b> Complete · 17m ago	<b>34.00938</b>	<input type="checkbox"/>

Durante la competencia usted solo podrá ver el Public Score, los profesores podrán ver los Public y Private Scores.

Para el ejemplo anterior los valores reales son:

Submission	Public	Private
ZZ7958_02_019_11000pr_003.csv	58.22894	56.30187
ZZ7958_02_019_10500pr_014.csv	57.68895	56.87616
ZZ7990_01_013_10000p.csv	53.78902	54.43328
KA3210_100.csv	47.70913	48.22323
KA3210_005.csv	44.11920	41.91032
K101_001.csv	34.00938	35.05313

Aquí debe ser notado algo que será uno de los puntos principales de la asignatura : elegir el submit que obtiene mayor ganancia en el Public casi nunca (menos del 2% de las veces) es el que obtiene mayor ganancia en el Private.

¿Es posible mirando las ganancias del Public determinar cual va a ser el mejor submit en el Private?

En la literatura se ha alertado sobre overfitear el Public Leaderboard:

Overfitting the Leaderboard in Ernst & Young Data Science Competition 2019

And subsequently losing 8000 USD + a ticket to New York.

<https://medium.com/hmif-itb/overfitting-the-leaderboard-da25172ac62e>

The dangers of overfitting: a Kaggle postmortem

<https://gregpark.io/blog/Kaggle-Psychopathy-Postmortem/>

### 5.4 Public Leaderboard

Este es un ejemplo de las primeras posiciones de un Public Leaderboard de una competencia de años anteriores, con otro dataset, en otra universidad (note que las ganancias son muy distintas a las que se obtendrán en esta competencia)

Leaderboard

Raw Data





Refresh

Search leaderboard

Public

Private

This leaderboard is calculated with approximately 30% of the test data. The final results will be based on the other 70%, so the final standings may be different.

#	Team	Members	Score	Entries	Last Solution
1	Pablo Sack		26.37193	65	9mo
2	Noelia Garro		25.75197	52	9mo
3	Rafael Fran Sanchez		25.69864	96	9mo
4	Torres Diego Juan		25.59197	132	9mo



## 5.5 Ganancia en el Public Leaderboard

Para un usuario, Kaggle muestra en el Public Leaderboard de la competencia la mayor ganancia Public de todos los submits que ha subido el usuario.

Esto es independiente de si el usuario ha marcado manualmente algún submit o no. No importa lo que el usuario marque en forma manual, siempre mostrará para ese usuario la mayor ganancia del Public de ese usuario.

En este ejemplo, SIN IMPORTAR lo que se elija, o incluso si no se elige nada, SIEMPRE la ganancia que se mostrará en el Public Leaderboard será de 58.22894 ya que es para el ejemplo, la mayor ganancia. Por favor, note que NO se muestra la ganancia Public de el submit elegido !!

### Submissions







Select up to 1 submissions that will count towards your final leaderboard score. If less than 1 are selected, Kaggle will automatically select from your best scoring submissions. [Learn More](#)

1/1

☒ Auto-selection candidates [?](#)

All Successful Selected Errors


Recent ▾

Submission and Description	Public Score <a href="#">?</a>	Select
 ZZ7958_02_019_11000pr_003.csv Complete · 4m ago	58.22894	<input type="checkbox"/>
 ZZ7958_02_019_10500pr_014.csv Complete · 5m ago	56.39898	<input checked="" type="checkbox"/>
 ZZ7990_01_013_10000p.csv Complete · 7m ago	53.78902	<input type="checkbox"/>
 KA3210_100.csv Complete · 15m ago	47.70913	<input type="checkbox"/>
 KA3210_005.csv Complete · 16m ago	44.1192	<input type="checkbox"/>
 K101_001.csv Complete · 17m ago	34.00938	<input type="checkbox"/>

es decir

Public Private

This leaderboard is calculated with approximately 30% of the test data. The final results will be based on the other 70%, so the final standings may be different.

#	Team	Members	Score	Entries	Last
1	voisinage		58.22894	6	1h
<div><input type="checkbox"/> Your Best Entry! Your most recent submission scored 58.22894, which is an improvement of your previous score of 56.39898. Great job!</div> <div>Tweet this</div>					







## 5.6 Ganancia en el Private Leaderboard

Hay dos casos posibles, que se ven a continuación

### 5.6.1 Ningún submit elegido manualmente

Si usted **no** ha elegido manualmente ningún submit, es decir están todos sin marcar, entonces Kaggle elegirá por usted como su submit final el submit que en el Public Leaderboard obtiene la mayor ganancia, y la ganancia del Private será la de ese submit.

Por ejemplo, en este caso su ganancia del Private será 56.30187







Submission and Description	Public Score ⓘ	Select
 <b>ZZ7958_02_019_11000pr_003.csv</b> Complete · 35m ago	<b>58.22894</b>	<input type="checkbox"/>
 <b>ZZ7958_02_019_10500pr_014.csv</b> Complete · 37m ago	<b>56.39898</b>	<input type="checkbox"/>
 <b>ZZ7990_01_013_10000p.csv</b> Complete · 39m ago	<b>53.78902</b>	<input type="checkbox"/>
 <b>KA3210_100.csv</b> Complete · 1h ago	<b>47.70913</b>	<input type="checkbox"/>
 <b>KA3210_005.csv</b> Complete · 1h ago	<b>44.1192</b>	<input type="checkbox"/>
 <b>K101_001.csv</b> Complete · 1h ago	<b>34.00938</b>	<input type="checkbox"/>

Submission	Public	Private
<b>ZZ7958_02_019_11000pr_003.csv</b>	58.22894	56.30187
ZZ7958_02_019_10500pr_014.csv	57.68895	56.87616
ZZ7990_01_013_10000p.csv	53.78902	54.43328
KA3210_100.csv	47.70913	48.22323
KA3210_005.csv	44.11920	41.91032
K101_001.csv	34.00938	35.05313

## 5.6.2 Submit elegido manualmente

Si usted ha elegido manualmente un submit y la ganancia del Private será la de ese submit.

Por ejemplo, en este caso su ganancia del Private será 56.87616

Submission and Description	Public Score ⓘ	Select
 <b>ZZ7958_02_019_11000pr_003.csv</b> Complete · 40m ago	<b>58.22894</b>	<input type="checkbox"/>
 <b>ZZ7958_02_019_10500pr_014.csv</b> Complete · 42m ago	<b>56.39898</b>	<input checked="" type="checkbox"/>
 <b>ZZ7990_01_013_10000p.csv</b> Complete · 44m ago	<b>53.78902</b>	<input type="checkbox"/>
 <b>KA3210_100.csv</b> Complete · 1h ago	<b>47.70913</b>	<input type="checkbox"/>
 <b>KA3210_005.csv</b> Complete · 1h ago	<b>44.1192</b>	<input type="checkbox"/>
 <b>K101_001.csv</b> Complete · 1h ago	<b>34.00938</b>	<input type="checkbox"/>

Submission	Public	Private
ZZ7958_02_019_11000pr_003.csv	58.22894	56.30187
ZZ7958_02_019_10500pr_014.csv	57.68895	56.87616
ZZ7990_01_013_10000p.csv	53.78902	54.43328
KA3210_100.csv	47.70913	48.22323
KA3210_005.csv	44.11920	41.91032
K101_001.csv	34.00938	35.05313

## **5.7 Operación de Kaggle**

Usted dispone de 20 submits diarios en Kaggle. El contador de submits se resetea cada día a las 21:00 , hora Buenos Aires.

## **5.8 Finalización de la competencia**

La competencia finaliza en forma automática en la fecha indicada en el cronograma.

Un segundo después de finalizada, si refresca el browser, podrá ver el Private Leaderboard y observar en qué posición quedó del ranking.

Para el ranking solamente se considera la ganancia del Private Leaderboard.

## 6 Experimentos Colaborativos

El gran objetivo es encontrar cuál es la mejor configuración de las etapas del workflow que genere la mayor ganancia en el Private Leaderboard de Kaggle.

Para ello se parte de una buena solución, la configuración workflow-inicial y basado en ella mediante la experimentación se prueban alternativas ya disponibles en los scripts provistos por la cátedra, o incluso, para el caso de alumnos avanzados, se agrega nueva funcionalidad a los scripts del workflow.

Aclaración importante : el objetivo de estos experimentos es el capitalismo salvaje, lograr ganar más dinero que el workflow-inicial. Decididamente NO nos son experimentos académicos por el amor al conocimiento abstracto.

Cada problema es abordado por dos grupos experimentadores **antagónicos**, el Grupo A y el Grupo B, que utilizando diseños experimentales distintos intentan responder la Hipótesis Experimental. Se puede leer lo siguiente para entender esta usual práctica de la reproducción de un experimento en el mundo científico <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5778115/>

Se ha observado que la etapa inicial de elección de un problema despierta en ocasiones ciertas batallas y malos entendidos entre los alumnos, para lo cual se han diseñado un conjunto de reglas simples orientadas a suavizar las fricciones.

Es el objetivo de la cátedra que los experimentos realizados sean auténticamente útiles para encontrar el mejor modelo predictivo. Lograr lo anterior demanda experimentar en el entorno de una buena configuración general de parámetros la que es “computacionalmente pesada” ya que está entrenando en 18 meses de historia.

Por la naturaleza de la ciencia de datos se cumple generalmente que utilizando algoritmos de Gradient Boosting of Decision Trees, entrenar en mayor cantidad de datos (más historia y más columnas) produce mejores modelos predictivos.

Si se entrena en datasets pequeños se llega a parámetros óptimos de las etapas del workflow que luego no se corresponden con los que se encuentran si el entrenamiento se hace en un dataset grande de varios meses.

Las corridas con grandes volúmenes de datos son naturalmente más lentas, más propensas a fallas de una máquina virtual, requieren de más control, y por lo general intentan ser evitadas por los alumnos.

Es tarea de la cátedra enseñarle a buscar y encontrar el procedimiento que produzca los mejores resultados y que usted sea capaz de ponderar el costo beneficio de la experimentación.

No intente buscar atajos subóptimos en la experimentación !

## **6.1 Documento Colaborativo Google Slides**

En el documento colaborativo de Google Slides se encuentran las siguientes secciones:

- Índice
- Plantilla de Problema, en donde hay un slide para cada uno de los capítulos
  - Portada
  - Hipótesis Experimental
  - Bibliografía
  - Sesgos Cognitivos
  - Diseño Experimental
  - Limitaciones
  - Resultados
  - Discusión de los resultados
  - Conclusiones
  - Futuros problemas y experimentos
  - Anexo
- Ejemplo completo de un Problema resuelto por un grupo A y la contraparte grupo B
- Potenciales Problemas a resolver, con su portada, hipótesis experimental y en algunos casos algo de bibliografía.

Los capítulos de la Plantilla del Problema y su estricto orden deben respetarse siempre; por supuesto algunos de los capítulos podrán demandar más de un slide.

Será en este mismo documento colaborativo donde los alumnos escriban sus presentaciones.

El Grupo A tendrá su presentación completa con todos los capítulos y a continuación vendrá la presentación del Grupo B la que también incluirá todos los capítulos, incluirá su propia Portada e Hipótesis Experimental que deberán ser idénticas a las del Grupo A

Lo más razonable es que usted elija uno de los Potenciales Problemas que ya aparecen en el documento, los que superan la cantidad de grupos posibles de esta cohorte. Si usted quisiera resolver un problema que no está planteado en el documento, deberá proponérselo a los profesores y éstos darle su autorización en función de los temas que fueron elegidos por el resto de los alumnos, que los problemas más interesantes tengan grupos asignados que los resuelvan, que usted tenga el know how suficiente en este momento para enfrentar el problema, etc

## 6.2 Problemas disponibles y alumnos avanzados

En el documento colaborativo de Google Slides la cátedra ha propuesto una serie de problemas cuyos experimentos son muy relevantes para la generación de un modelo que aumente la ganancia y factibles de realizar en el tiempo disponible.

Si usted es un/a alumna/o avanzado con conocimientos previos de Ciencia de Datos o de Estadística Clásica, se lo autoriza a que :

- Proponga a los profesores un nuevo Problema, junto con su diseño experimental
- Convenza a uno o dos alumnos más, no necesariamente estos avanzados, para poder conformar los grupos antagónicos A y B

## 6.3 Información Preexistente

En el documento compartido de Google Slides se proveen experimentos para algunos de los cuales hay información previa de alumnos de otras cohortes. Dicha información está simplemente como una ayuda, usted puede obviarla completamente, armar su propia bibliografía, etc. De ninguna forma se sienta limitado por dicha información, sea completamente libre.

## 6.4 Configuración de los grupos

A continuación las únicas dos configuraciones posibles en cuanto a la cantidad de integrantes de los grupos para un problema:

Configuraciones Válidas		
# Integrantes por equipos		
Configuración	Grupo A	Grupo B
Solitaria	1	1
Asimétrica	2	1

Por favor notar que siempre deben existir los dos grupos antagónicos para un Problema, los que realizan diseños experimentales distintos.

No pueden trabajar cuatro personas en un mismo problema, es un desperdicio de materia gris.

Si su computer literacy es bajo esta cátedra le implora que no forme un Grupo B, por favor estratégicamente forme un Grupo A con alguien que sepa más que usted.

## **6.5 Forma de pre registraci3n**

Simplemente anote su nombre en la portada del problema que es de su inter3s, directamente en el documento de Google Slides.

Podr3 anotar su nombre en un 3nico problema. Siempre podr3 cambiarlo por otro problema en la medida que se cumpla el proceso que sigue a continuaci3n.

## **6.6 Negociaci3n de Problemas : vacancia**

Se cumple lo siguiente:

- Cada problema requiere como m3nimo dos personas, una que integre el Equipo A y otra que integre el Equipo B.
- Hay una oferta de Potenciales Problemas mayor a la cantidad de equipos posibles

Las condiciones anteriores producen por lo general la situaci3n de personas que anotan su nombre en uno de los Potenciales Problemas y ese problema no le interesa a nadie m3s, pasan los d3as y nadie m3s anota su nombre en esa portada.

Esta persona deber3 :

- convencer a otras que se pasen a su problema, al menos una persona m3s
- o abandonar ese problema y unirse a alg3n otro problema/grupo que est3 en etapa de formaci3n.

## **6.7 Negociaci3n de Problemas : sobreoferta**

Esta es la situaci3n que produce m3s fricciones y sensaci3n de injusticia entre los alumnos.

Se cumple lo siguiente:

- A pesar de haber sobreabundancia de Potenciales Problemas, hay algunos muy populares en el imaginario de los alumnos que son pretendidos por m3s de tres alumnos.
- Hay algunos alumnos que ingresan al Google Slides varios d3as despu3s del lanzamiento de la registraci3n y desean unirse a un problema con sobreoferta de alumnos.

Negocie sin intentar sacar ventajas solo por haber llegado antes.

## **6.8 Asignaci3n Grupo A vs Grupo B**

El grupo A posee cierta ventaja ya que es quien primero establece el dise1o experimental, y una vez que este fue aprobado por los profesores via Zulip deber3 el Grupo B proponer su dise1o experimental, mandatoriamente distinto, el que tambi3n deber3 ser aprobado por los profesores, deber3n garantizar que los dise1os experimentales sean lo suficientemente disjuntos.

Dicha m3nima ventaja a favor del Grupo A tiene el potencial de generar alg3n tipo de conflicto. Se sugiere en este caso tirar una moneda al aire.



## 6.9 Antagonismo de los grupos

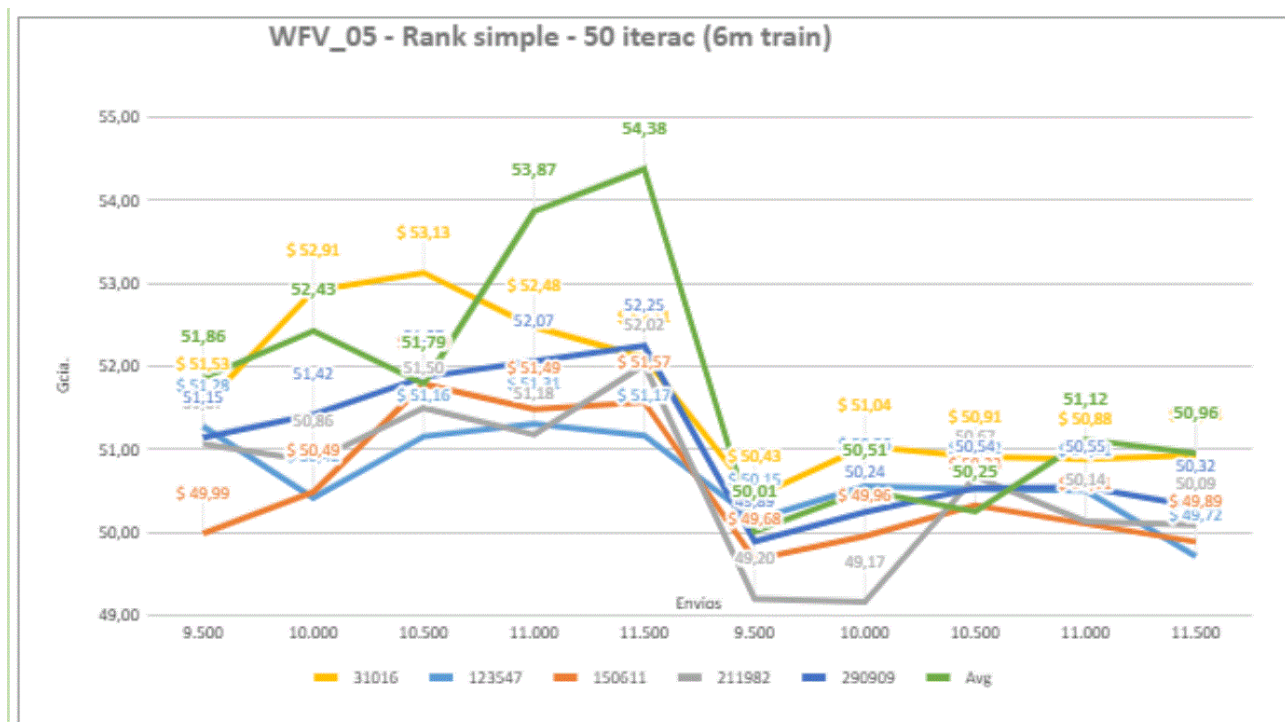
Para un problema dado, el Grupo A y el B son antagonísticos. Ellos intentan llegar a resultados distintos. Solo se comunican al principio, cuando el Grupo A establece su diseño experimental y el B está obligado a hacer uno completamente distinto. A partir de allí, nunca más interactúan, evitan por todos los medios de contaminar sus mentes con potenciales ideas equivocadas del otro grupo.

## 6.10 Períodos de análisis de los grupos

Los grupos A y B analizarán distintos períodos con el objetivo de determinar si lo que mejor funciona lo hace lamentablemente para un solo período o si por el contrario es una propiedad que generaliza.

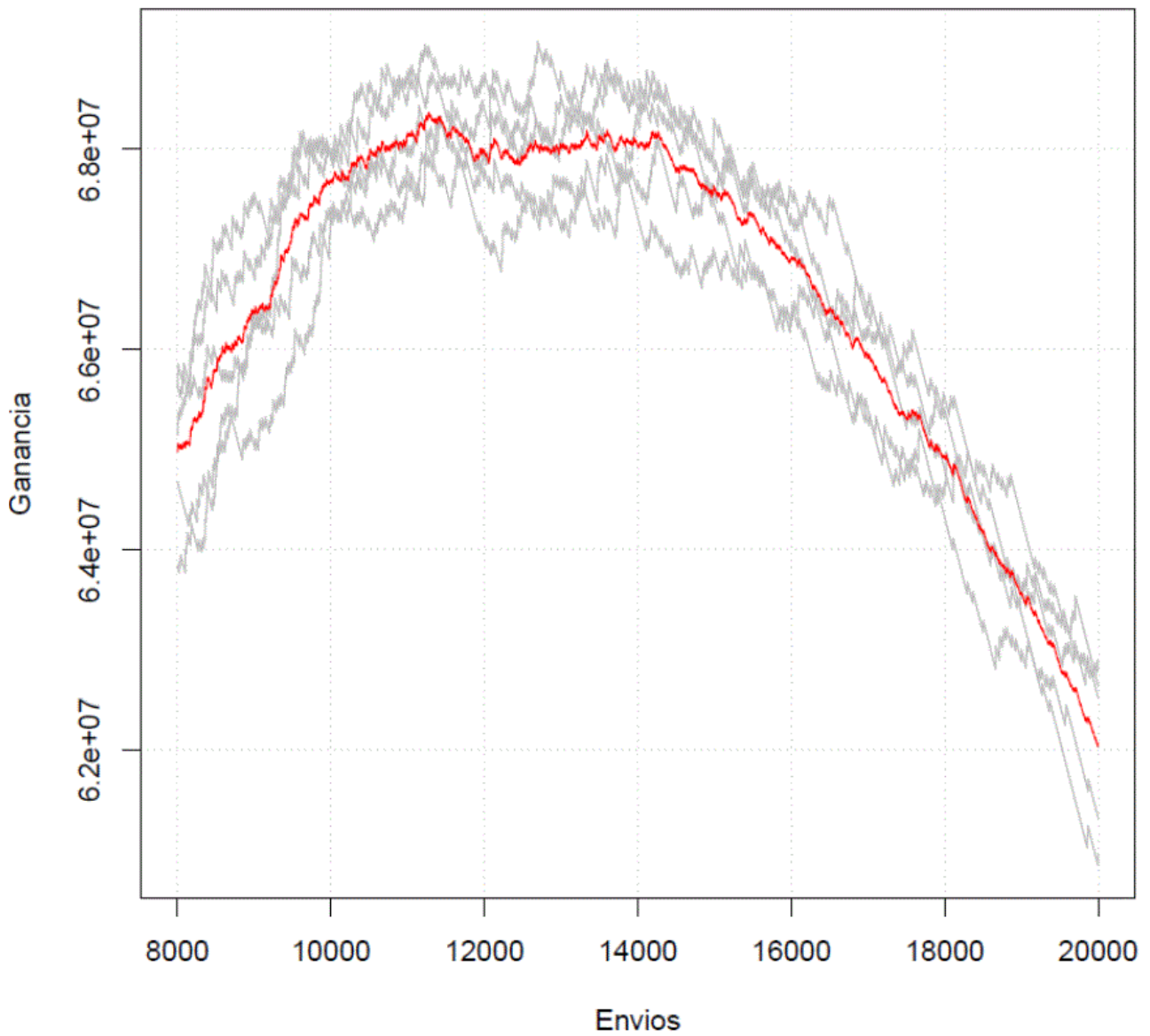
El grupo A tendrá como future el mes 202109 que son los datos sin `clase_ternaria`, con lo que las mediciones de la ganancia de los experimentos las deberán hacer con subidas a Kaggle. Dado que el usuario inicial de Kaggle no les será suficiente por su limitación a 20 submits diarios (él máximo que permite Kaggle) se autoriza a estas personas a crear usuarios fake en Kaggle.

En este caso usted deberá dibujar a mano las curvas de ganancia, como la que se muestra a continuación, note que las ganancias solamente han sido calculadas para la cantidad de envíos { 9500, 10000, 10500, 11000, 11500 }



El grupo B tendrá como future el mes 202107 que son datos para los que se conoce la `clase_ternaria`, en este caso los scripts ya están preparados para graficar las curvas de ganancia vs cantidad de estímulos como la que se muestra a continuación, note que conocer la `clase_ternaria` y no tener que subir a Kaggle permite calcular la ganancia para todos los números enteros en el rango [ 8000, 20000 ]

**Mejor gan prom = 68069781**



Período análisis Grupo A

```
PARAM$future <- c(202109)
PARAM$final_train <- c(
  202107, 202106, 202105, 202104, 202103, 202102,
  202101, 202012, 202011, 202010, 202009, 202008, 202002, 202001, 201912,
  201911, 201910, 201909
)

PARAM$train$training <- c(
```

```

202105, 202104, 202103, 202102, 202101,
202012, 202011, 202010, 202009, 202008, 202002, 202001, 201912, 201911,
201910, 201909, 201908, 201907
)

```

### 6.10.1 Período análisis Grupo B

```

PARAM$future <- c(202107)
PARAM$final_train <- c(
  202105, 202104, 202103, 202102,
  202101, 202012, 202011, 202010, 202009, 202008, 202002, 202001, 201912,
  201911, 201910, 201909, 201908, 201907
)

```

```

PARAM$train$training <- c(
  202103, 202102, 202101,
  202012, 202011, 202010, 202009, 202008, 202002, 202001, 201912, 201911,
  201910, 201909, 201908, 201907, 201906, 201905
)

```

```

PARAM$train$validation <- c(202104)
PARAM$train$testing <- c(202105)

```

## 6.11 Aprobación de Diseños Experimentales

El procedimiento a seguir es el siguiente

1. Al Grupo A deberá cuanto antes escribir su diseño experimental en el Google Slides colaborativo, y **solicitar rápidamente su revisión a los profesores, por medio de Zulip**. En esta etapa podrá haber una brevísima conference con el profesor.
2. Luego de un ida y vuelta, a la mayor brevedad el profesor, por Zulip, aprueba el diseño experimental del Grupo A. En ese momento el Grupo A ya puede comenzar la experimentación.
3. El Grupo B, viendo ese diseño experimental aprobado, construye un diseño experimental alternativo, y solicita la aprobación del profesor.
4. Luego de un ida y vuelta, el profesor, por Zulip, aprueba el diseño experimental del Grupo B, y éste puede comenzar la experimentación.

## 6.12 Scripts

Los scripts a utilizar para los Experimentos Colaborativos son los que están en la carpeta del repositorio oficial de la asignatura [src/workflow-inicial](#)

Recuerde que la forma de trabajo consiste en hacer siempre una copia del script oficial que comienza con la letra “z” a un nuevo nombre sin esa “z”.

## 6.13 Semillas generales en scripts

Cada alumno deberá utilizar sus propias semillas, jamás podrá utilizar ninguna semilla del profesor para sus experimentos.

En la Etapa Final ( el script `_ZZ_` ) cada persona deberá utilizar sus propias cinco semillas, incluso en caso pertenezca a un grupo que posee dos integrantes, en ese caso el experimento saldrá con 10 semillas para la etapa final.

## 6.14 Ceteris Paribus

Cada Potencial Problema hace referencia únicamente a una parte de un script.

Los grupos que intenten resolver una Hipótesis Experimental deberán hacer modificaciones a los parámetros del correspondiente script, o en caso de tener el conocimiento suficiente al código del script, pero deberán mantener sin modificación el resto de los scripts que no son objeto de su experimentación, excepto :

- las semillas generales de todos los scripts que siempre deben ser cambiadas a las propias.
- Las pruebas que se realicen en otros períodos “future”, donde se elige un mes para el que se conoce la clase y se pueden dibujar las curvas

Concretamente, en los scripts de workflow-inicial se entrena en 18 meses de historia ya que esto produce buenas ganancias. Excepto que su Problema sea el de probar la cantidad óptima de meses donde entrenar, usted deberá correr el workflow respetando esos 18 meses y no podrá tomar ningún atajo de utilizar menos meses para que sus procesos corran más rápido.

Una vez que haga las corridas esperadas con esos 18 meses, podrá agregar a los experimentos, como algo adicional, alguna corrida en menos meses.

### **6.15 Capítulo Conclusiones**

El capítulo de las conclusiones, en alguna parte se debe dar una receta bien concreta : ¿Cómo se deben setear los parámetros de esa etapa del Workflow? Incluso para el caso que los experimentos no sean concluyentes se debe dar una clara receta.

Por supuesto, quienes lean el experimento simplemente lo tomarán como un punto de partida a sus propios “últimos experimentos”.

### **6.16 Video Presentaciones Ejecutivas**

Los links a las video presentaciones de los distintos problemas deberán disponibilizar antes de la fecha indicada en el Cronograma de la Asignatura, y se harán directamente a partir del documento colaborativo de Google Slides.

Dado un problema, el Grupo A debe hacer un video y el Grupo B debe hacer otro video.

Todos los participantes de un grupo deben participar significativamente en el video de ese grupo.

Características que deberá tener el video:

- La duración esperada es de alrededor de 5:00 minutos; jamás podrá exceder los 5:30 minutos.
- La cara de la persona que habla, mostrando claramente sus gestos, debe aparecer durante al menos el 95% del tiempo, y debe ocupar al menos el 10% de la superficie del cuadro.
- Usted debe hablar mirando a la cámara.
- Por favor, evite leer frente a cámara.
- El video debe estar a velocidad normal, por favor no acelerarlo en ningún tramo.

### **6.17 Luego de las Video Presentaciones**

Luego de finalizados los Experimentos Colaborativos, aún queda la competencia Kaggle final.

Los resultados de los experimentos son “hombros de gigantes” donde deben pararse todos los alumnos para hacer unos últimos experimentos intentando superar la ganancia.

No está permitido que un alumno simplemente recoja las conclusiones de las todas las exposiciones, corra esa configuración, y esa sea su entrega final en Kaggle. Recuerde que usted debe proveer un link a una carpeta en GitHub con lo que corrió para la entrega final y eso debe ser replicable por los profesores, éstos deben ser capaces de generar su exacto submit a Kaggle.

## 7 ¿Cómo seguir?

### 7.1 *Mantenerse actualizado*

¿En que sitios es posible mantenerse actualizado luego de finalizada la asignatura ?

- <https://machinelearningmastery.com/>
- <https://www.kdnuggets.com/>
- <https://www.analyticsvidhya.com/>
- <https://neptune.ai/blog>
- <https://www.fast.ai/#category=technical>
- <https://openai.com/blog>
- <https://www.deepmind.com/blog>
- <https://blog.ml.cmu.edu/>
- <https://news.mit.edu/topic/machine-learning>

### 7.2 *Datasets Públicos*

¿Dónde encontrar datasets públicas para una Tesis de Maestría o practicar ?

- <https://cloud.google.com/datasets>
- <https://github.com/awesomedata/awesome-public-datasets>
- <https://www.kaggle.com/datasets>
- <https://www.datos.gob.ar/dataset>
- <https://data.buenosaires.gob.ar/dataset/>
- <https://data.gov/>
- <https://careerfoundry.com/en/blog/data-analytics/where-to-find-free-datasets/>
- <https://medium.com/analytics-vidhya/top-100-open-source-datasets-for-data-science-cd5a8d67cc3d>
- <https://www.v7labs.com/open-datasets>
- <https://azure.microsoft.com/en-us/products/open-datasets/>
- <https://registry.opendata.aws/>

## 8 Empiricismo en la Ciencia de Datos

En el Machine Learning muchos algoritmos son encontrados por prueba y error, al inicio se los prueba en decenas de datasets, se divulgan, los usa la comunidad, ganan competencias Kaggle, pasan a ser usados por practicantes en innumerables situaciones, pero carecen de demostraciones matemáticas sobre su robustez- Recién luego de varios años, o más de una década se demuestra la razón por la que funcionan, los efectos que se observan y las razones de su poder (por ejemplo que converjan o que el overfitting esté controlado).

Esta situación colisiona violentamente con quienes provienen de la estadística clásica.

### 8.1.1 Bibliografía Introductoria

- <https://www.kdnuggets.com/2015/07/deep-learning-triumph-empiricism-over-theoretical-mathematical-guarantees.html>
- <https://royalsocietypublishing.org/doi/pdf/10.1098/rsta.2016.0153>
- <https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124>
- [https://ris.utwente.nl/ws/files/177935074/Boon\\_2020\\_BioTechnoPractic\\_How\\_scientists\\_are\\_brought\\_back\\_into\\_science\\_preprint\\_May\\_9\\_2019.pdf](https://ris.utwente.nl/ws/files/177935074/Boon_2020_BioTechnoPractic_How_scientists_are_brought_back_into_science_preprint_May_9_2019.pdf)
- <https://hdsr.mitpress.mit.edu/pub/l39rpgyc/release/1>
- <https://www.youtube.com/watch?v=oONHlua2gBY>
- <https://royalsocietypublishing.org/doi/10.1098/rsta.2018.0145>

Arquetípico ejemplo de empiricismo, acontecido en la medicina

- <https://www.redalyc.org/journal/920/92044744013/html/> leer la sección **Anexo Semmelweis y la fiebre puerperal** (está después de la bibliografía)

## 9 Las dos culturas en el modelado predictivo

- El paper original de Leo Breiman Statistical Modeling, The Two Cultures  
<http://www2.math.uu.se/~thulin/mm/breiman.pdf>
- <https://towardsdatascience.com/thoughts-on-the-two-cultures-of-statistical-modeling-72d75a9e06c2>



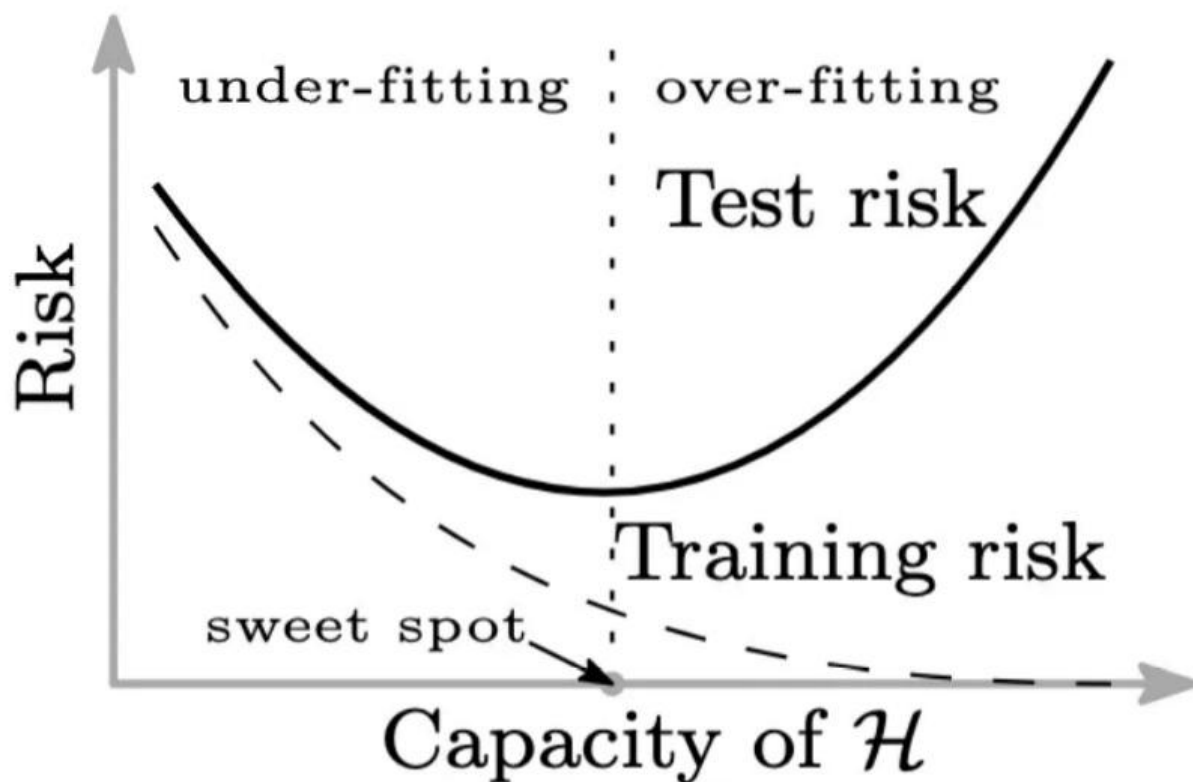
## 10 Estimando la Ganancia de un modelo predictivo

### 10.1 Bibliografía Introductoria

Bibliografía

- <https://machinelearningmastery.com/statistical-significance-tests-for-comparing-machine-learning-algorithms/>
- <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.183.534&rep=rep1&type=pdf>

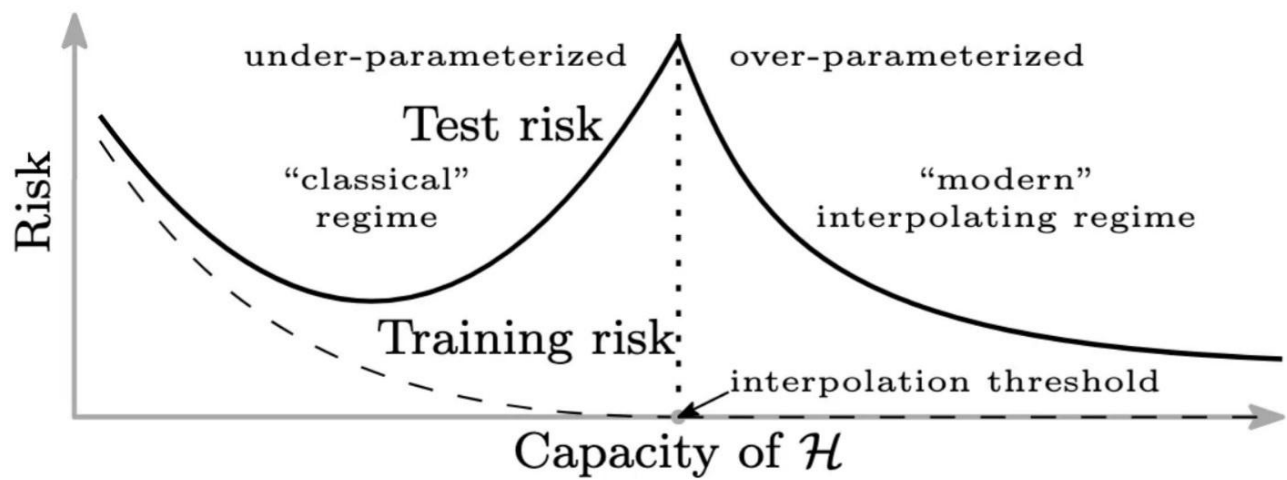
### 10.2 Sesgo Varianza en la estadística clásica



#### 10.2.1 Bibliografía Inicial

- <https://machinelearningmastery.com/gentle-introduction-to-the-bias-variance-trade-off-in-machine-learning/>
- <http://scott.fortmann-roe.com/docs/BiasVariance.html> (15 minutos)

### 10.3 Sesgo Varianza en el moderno Machine Learning



#### 10.3.1 Bibliografía Inicial

- <https://medium.com/mlearning-ai/double-descent-8f92dfdc442f>
- <https://arxiv.org/pdf/1812.11118.pdf>

## 11 Análisis exploratorio de datos

### 11.1 Consistencia longitudinal de los atributos

### 11.2 Drifting de atributos

En el período (mes) donde debo aplicar el modelo para hacer la predicción final, ¿existen variables con una distribución distinta a los períodos anteriores?

- <https://towardsdatascience.com/data-drift-part-1-types-of-data-drift-16b3eb175006>
- <https://towardsdatascience.com/data-drift-part-2-how-to-detect-data-drift-1f8bfa1a893e>
- <https://towardsdatascience.com/machine-learning-in-production-why-you-should-care-about-data-and-concept-drift-d96d0bc907fb>
- <https://medium.com/data-from-the-trenches/a-primer-on-data-drift-18789ef252a6>
- 
- <https://arxiv.org/pdf/2004.05785.pdf>
- <https://wires.onlinelibrary.wiley.com/doi/epdf/10.1002/widm.1381>

### 11.3 Tratamiento de variables "rotas"

¿Qué funciona mejor con una variable "rota" para un mes? ¿Imputar su valor interpolando, o asignarle nulo? sí, ha leído bien, estamos proponiendo asignar NA a una variable !

### 11.4 Tratamiento de nulos

¿Debo tratar a los nulos de alguna forma en especial?

¿Es razonable eliminar un registro donde algún campo importante sea nulo?

¿Qué funciona mejor, imputar nulos con alguna metodología o permitir que XGBoost/LightGBM trabajen con ellos?

### 11.5 Tratamiento de outliers

¿Debo tratar a los outliers de alguna forma en especial?

¿Qué funciona mejor, tratar los outliers con alguna metodología o permitir que XGBoost/LightGBM trabajen con ellos?

¿Qué pasa con los outliers en el binning que realiza LightGBM?

## 12 Algoritmo: Árbol de Decisión

### 12.1 Bibliografía Inicial

Capítulo 9.2 de <https://hastie.su.domains/Papers/ESLII.pdf>

Optimización de hiperparámetros <https://bradleyboehmke.github.io/HOML/DT.html>

Explicación sobre rpart <https://www.learnbymarketing.com/tutorials/rpart-decision-trees-in-r/>

### 12.2 El hiperparámetro *cp* en la librería *rpart*

En esta materia "un experimento no se le niega a nadie", con lo cual independientemente de lo que digan papers, libros, otras materias, profesores con gran cantidad de tiras en su uniforme, towardsdatascience e incluso el infame Medium, haremos el experimento, y que una optimización de hiperparámetros diga si el hiperparámetro *cp* óptimo es positivo o negativo.

Luego, ante la evidencia experimental, si un **cp** negativo es el óptimo, buscaremos una explicación, aquellos que tengan ganas y tiempo harán nuevos experimentos.

Seamos fieles a "The problem is that people are educated just enough to believe what they have been taught, and not educated enough to **question anything from what they have been taught**" y maravillarnos ante nuestra propia ignorancia y tendencia a ponernos cómodos techos.

"I am **tormented** with an everlasting itch for things remote. I love to sail **forbidden** seas. " Moby Dick

## 13 Comparación de Modelos Predictivos

### 13.1 Bibliografía Inicial

Cross Validation <https://machinelearningmastery.com/k-fold-cross-validation/>

Receta <https://machinelearningmastery.com/statistical-significance-tests-for-comparing-machine-learning-algorithms/>

Paper <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.183.534&rep=rep1&type=pdf>

## 14 Optimización de Hiperparámetros

### 14.1 Bibliografía Introductoria

### 14.2 Grid Search

### 14.3 Bayesian Optimization

El algoritmo de Optimización Bayesiana posee una compleja base matemática. La librería `mlrMBO` que utilizamos en R posee extensas opciones de parametrización, las que producen mejoras muy marginales.

#### 14.3.1 Artículos de fácil Lectura

- <https://towardsdatascience.com/bayesian-optimization-concept-explained-in-layman-terms-1d2bcdeaf12f>
- <https://towardsdatascience.com/the-beauty-of-bayesian-optimization-explained-in-simple-terms-81f3ee13b10f>
- <https://machinelearningmastery.com/what-is-bayesian-optimization/> (los ejemplos son en Python)

#### 14.3.2 Videos sencillos

- <https://www.youtube.com/watch?v=41gKFmKQDIg>

#### 14.3.3 Video Lectures

- <https://www.youtube.com/watch?v=C5nqEHpdyoE> Optimización Bayesiana
- <https://www.youtube.com/watch?v=EnXxO3BAgYk>
- <https://www.youtube.com/watch?v=MfHKW5z-OOA> Métodos Gaussianos

#### 14.3.4 Libros de Texto

Este es un interesante capítulo de un libro de texto [https://www.automl.org/wp-content/uploads/2019/05/AutoML\\_Book\\_Chapter1.pdf](https://www.automl.org/wp-content/uploads/2019/05/AutoML_Book_Chapter1.pdf)

Si usted desea entender en total profundidad la técnica matemática de Kriging, creada para la geología, debe leer este libro:

Interpolation of Spatial Data, Some Theory for Kriging , Michael L. Stein, 1999

[https://storage.googleapis.com/lab02023v/\(Springer%20Series%20in%20Statistics\)%20Michael%20L.%20Stein%20\(auth.\)%20-%20Interpolation%20of%20Spatial%20Data\\_%20Some%20Theory%20for%20Kriging-Springer-Verlag%20New%20York%20\(1999\).pdf](https://storage.googleapis.com/lab02023v/(Springer%20Series%20in%20Statistics)%20Michael%20L.%20Stein%20(auth.)%20-%20Interpolation%20of%20Spatial%20Data_%20Some%20Theory%20for%20Kriging-Springer-Verlag%20New%20York%20(1999).pdf)

### 14.3.5 Papers Técnicos

- [Taking the Human out of the Loop](#) (para quienes quieran leer en profundidad en algún momento tranquilo de la vida)
- [A tutorial on Bayesian Optimization](#)
- [mlrMBO: A modular framework for model-based optimization of expensive black-box functions](#)

### 14.3.6 Librería mlrMBO

- Repositorio GitHub (con ejemplos) <https://github.com/mlr-org/mlrMBO>
- Documentación Oficial de la Librería <https://cran.r-project.org/web/packages/mlrMBO/mlrMBO.pdf>
- Ejemplos de uso de la librería
  - <https://cran.r-project.org/web/packages/mlrMBO/vignettes/mlrMBO.html>
  - [https://rdrr.io/cran/mlrMBO/man/mlrMBO\\_examples.html](https://rdrr.io/cran/mlrMBO/man/mlrMBO_examples.html)
  - <https://www.kaggle.com/code/xanderhorn/train-r-ml-models-efficiently-with-mlr/notebook>
  - <https://r-craft.org/r-news/stepwise-bayesian-optimization-with-mlrmbo/>
  - <https://bikeactuary.com/datasci/bayesian-mbo>
  - [https://mlrmbo.mlr-org.com/articles/supplementary/machine\\_learning\\_with\\_mlrmbo.html](https://mlrmbo.mlr-org.com/articles/supplementary/machine_learning_with_mlrmbo.html)

## **15 Curva ROC**

### ***15.1 Bibliografía***

- Ling, C. X., Huang, J., & Zhang, H. (2003, August). AUC: a statistically consistent and more discriminating measure than accuracy. In Ijcai (Vol. 3, pp. 519-524). <https://www.ijcai.org/Proceedings/03/Papers/077.pdf>

### ***15.2 Curva ROC de un split en un árbol de decisión***

### ***15.3 Curva ROC de un modelo predictivo***

### ***15.4 Relación de la Curva ROC con la función de ganancia***



## 16 Controlando el False Discovery Rate

### 16.1 Bibliografía Introductoria

- <https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124>
- <https://www.youtube.com/watch?v=6ZxlzVjV1DE> (subtítulos en inglés activables)

## 17 Feature Engineering

### 17.1 Bibliografía Introductoria

- <https://www.youtube.com/watch?v=X4pWmkxEikM>
- <https://www.youtube.com/watch?v=06-AZXmwHjo> video Andrew Ng

### 17.2 Cambiando la clase

La clase a predecir es  $POS = \{ BAJA+2 \}$  ,  $NEG = \{ BAJA+1, CONTINUA \}$

La forma natural de predecir es con  $POS = \{ BAJA+2 \}$  , sin embargo ¿Es esa la mejor forma, o me convendría entrenar con otra clase?

### 17.3 Variables Manuales

Artículos donde resolvieron un problema similar

- [https://www.researchgate.net/profile/Xingsen-Li/publication/224759968\\_The\\_Analysis\\_on\\_the\\_Customers\\_Churn\\_of\\_Charge\\_Email\\_Based\\_on\\_Data\\_Mining\\_Take\\_One\\_Internet\\_Company\\_for\\_Example/links/5891e87a92851cda256a0358/The-Analysis-on-the-Customers-Churn-of-Charge-Email-Based-on-Data-Mining-Take-One-Internet-Company-for-Example.pdf](https://www.researchgate.net/profile/Xingsen-Li/publication/224759968_The_Analysis_on_the_Customers_Churn_of_Charge_Email_Based_on_Data_Mining_Take_One_Internet_Company_for_Example/links/5891e87a92851cda256a0358/The-Analysis-on-the-Customers-Churn-of-Charge-Email-Based-on-Data-Mining-Take-One-Internet-Company-for-Example.pdf)

Tesis de maestría donde se solucionó un problema parecido.

- <https://core.ac.uk/download/pdf/83461632.pdf>
- [http://research.sabanciuniv.edu/39116/1/AneelaTanveer\\_10236886.pdf](http://research.sabanciuniv.edu/39116/1/AneelaTanveer_10236886.pdf)
- <https://arrow.tudublin.ie/cgi/viewcontent.cgi?article=1218&context=scschcomdis>

### 17.4 Transformaciones tradicionales

¿Qué efectos tienen en un árbol de decisión y algoritmos derivados, transformaciones del tipo: ?

- transformación lineal  $x' = ax + b$
- $x' = \log(x)$
- normalización

***17.5 Variables históricas***

***17.6 Lags y delta lags***

***17.7 Tendencias***

***17.8 Medias móviles***

***17.9 Ajuste por Inflación***

***17.10 Variables hojas de un Random Forest***

***17.11 Reducción de la dimensionalidad para acelerar***

## 18 Ensembles de Modelos

### 18.1 Bibliografía Introductoria

- BBC - The Code - The Wisdom of the Crowd <https://www.youtube.com/watch?v=iOucwX7Z1HU> (5 minutos)
- <https://www.all-about-psychology.com/the-wisdom-of-crowds.html> (10 minutos)
- <https://machinelearningmastery.com/what-is-ensemble-learning/> (10 minutos)

## 19 Algoritmo: Random Forest

### 19.1 Bibliografía Introductoria

- [https://www.youtube.com/watch?v=J4Wdy0Wc\\_xQ](https://www.youtube.com/watch?v=J4Wdy0Wc_xQ)

## 20 Algoritmo: Gradient Boosting of Decision Trees

### 20.1 Bibliografía Introductoria

### 20.2 XGBoost

#### 20.2.1 Bibliografía XGBoost

Inicial

- <https://towardsdatascience.com/xgboost-regression-explain-it-to-me-like-im-10-2cf324b0bbdb>
- <https://towardsdatascience.com/https-medium-com-vishalorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d>
- <https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/>

Avanzada

- paper original de XGBoost <https://dl.acm.org/doi/pdf/10.1145/2939672.2939785>
- 

### 20.3 Optimización hiperparámetros XGBoost

- <https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/beginners-tutorial-on-xgboost-parameter-tuning-r/tutorial/>
- [Narrowing the Search: Which Hyperparameters Really Matter?](#)

### 20.4 Bibliografía LightGBM

Avanzada

- Paper original de LightGBM <https://papers.nips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf>
- 

### 20.5 Optimización hiperparámetros LightGBM

- <https://lightgbm.readthedocs.io/en/latest/Parameters-Tuning.html>
- <http://devdoc.net/bigdata/LightGBM-doc-2.2.2/Parameters-Tuning.html>
- <https://neptune.ai/blog/lightgbm-parameters-guide>

## **21 Estrategia de Entrenamiento**

***21.1 Determinación de los mejores períodos para entrenar***

***21.2 Acelerando el entrenamiento : evitando el cross validation***

***21.3 Acelerando el entrenamiento : Undersampling de la clase mayoritaria***

## **22 Importancia de Variables**

### **22.1 Bibliografía Introductoria**

### **22.2 Valores Shapley**

- <https://www.youtube.com/watch?v=ngOBhhINWb8>
- <https://www.youtube.com/watch?v=B-c8tlgchu0>



## 23 Ensembles de Modelos Segunda Parte

### Bibliografía de Model Stacking

- <https://www.youtube.com/watch?v=DCrcoh7cMHU>
- <https://www.youtube.com/watch?v=8T2emza6g80>
- <https://www.kdnuggets.com/2017/02/stacking-models-improved-predictions.html>
- <https://medium.com/ml-research-lab/stacking-ensemble-meta-algorithms-for-improve-predictions-f4b4cf3b9237>
- <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.56.1533&rep=rep1&type=pdf> el paper original del año 1992

### 23.1 Semilleríos

### 23.2 Hibridación de Semilleríos

## **24 Acelerando el procesamiento**

## 25 Aprovechando el Public Leaderboard

### 25.1 *Bibliografía Inicial*

- <https://medium.com/hmif-itb/overfitting-the-leaderboard-da25172ac62e>
- <https://www.kaggle.com/general/54610>
- <http://gregpark.io/blog/Kaggle-Psychopathy-Postmortem/>

### 25.2 *El problema de la calibración en Gradient Boosting*

### 25.3 *Otra forma de cortar*

## 26 Clustering jerárquico basado en la distancia de Random Forest

Un modelo de Random Forest es un ensamble de árboles de decisión. Ese ensamble puede servirnos para definir una distancia entre registros del dataset. Esta distancia que decididamente no es la euclídea, es insensible a los outliers, a los nulos y a las diferencias de escala de las variables; no hace falta normalizar los atributos del dataset.

Utilizando esa distancia podemos realizar una partición por el método de clusters jerárquicos.

## 27 Storytelling

### 27.1 Bibliografía Introductoria Storytelling

- <https://www.youtube.com/watch?v=3KQaWHesWwc>
- <https://itappler.medium.com/effective-data-storytelling-and-visualization-81ad24229745> (5 minutos)
- <https://www.youtube.com/watch?v=9UNIHfy4hs> (4 minutos)
- <https://www.themuse.com/advice/public-speaking-tips> (10 minutos)
- <https://www.researchgate.net/publication/262393281> How video production affects student engagement An empirical study of MOOC videos (15 minutos)
- <https://www.youtube.com/watch?v=ba-CB6wVuvQ> (15 minutos)

### 27.2 Herramientas para grabar videos

- Herramientas online, all included
  - <https://prezi.com/>
  - <https://www.loom.com/>
- Herramientas locales
  - grabación sencilla ( ver siguiente capítulo )
  - <https://obsproject.com/> (es la que utiliza el profesor)

### 27.3 Grabación sencilla de videos con Microsoft Teams

Esta forma de grabar videos en Windows fue propuesta por un alumno y se hizo popular entre sus compañeros

1. descargar microsoft teams [1. https://www.microsoft.com/es-ww/microsoft-teams/download-app](https://www.microsoft.com/es-ww/microsoft-teams/download-app)
2. abrir teams, e ir a “Meet Now” para crear una sala vacía con tu cámara.
3. abrir la meet estará usted solo y nadie más
4. dar click a “Share” y elegir tu tipo de pantalla: Standout, side-by-side o Reporter, y elegir Window para que muestre una pantalla y al lado su cámara
5. elegir tamaño de cámara (su imagen) y posición left-right de la cámara, recordar que debe ser al menos el 10% de la superficie ( se puede elegir otro tipo de composición, que no recorte el fondo, dando un resultado más formal)
6. Presionar las teclas Windows + G, para abrir consola de grabación de Windows
7. Elegir bien la fuente (su pantalla compartida del Microsoft Teams) y darle a REC (el botón redondito que debería ser rojo pero ahí es blanco)
8. No olvidarse de activar el audio para que grabe con audio
9. también se puede grabar directamente usando el shortcut ALT+WIN+R y arranca a grabar directamente
10. Arrancar el speech pitch del video con la ppt

11. A finalizar, detener el video
12. Volver a abrir la consola de grabación con windows + G para encontrar el video guardado: poner open file location:
13. ingresar a youtube y subir el video
14. detalle: revisar que cuando se inicia la grabación con Windows G, el recording sea SOBRE el meeting, así graba el video que está sucediendo dentro de la ventana del meeting.

## 27.4 Torneo de Videos

Los videos se calificarán mediante la metodología *Swiss-system tournament* con el método de apareo *Dutch system* ( ver [https://en.wikipedia.org/wiki/Swiss-system\\_tournament](https://en.wikipedia.org/wiki/Swiss-system_tournament) ) este método implica la existencia de rondas y que en cada ronda hay duelos de videos donde se intenta que compitan videos de similar calidad. Solo se pasa a la ronda siguiente una vez que todos los duelos hayan sido calificados.

Los jueces de cada duelo serán los mismos alumnos y también los profesores. Se espera que cada alumno sea juez en al menos 6 duelos de cada torneo. Esto significa que cada alumno deberá ver  $2 * \log_2(\text{cantidad\_alumnos})$  videos dirigidos a Juan Grande. Nadie será juez de sus propios videos ni de compañeros. La votación en los duelos de videos es secreta.

La nota que se asigna por ser juez en el torneo de videos es simbólica ya que apenas representa el 1% de la nota final de la asignatura y está regida por el cumplimiento en su rol como jurado. Se pretende que TODOS los alumnos participen en todas las rondas como jueces.

## 28 El Recuperatorio

El recuperatorio está destinado a personas que habiendo asistido al 75% de las clases, o no cumplieron con la entrega de todas las actividades obligatorias o la suma de las notas es insuficiente.