Universidad de Buenos Aires

Maestría en Explotación de Datos
y Descubrimiento del Conocimiento

Data Mining en Economía y Finanzas

Edición 2021

Gustavo Denicolay
Alejandro Bolaños

última actualización 2021-08-18 19:43

Tabla de Contenidos

1 La Asignatura

Este documento se irá actualizando semanalmente, se agregaran los resultados de los principales experimentos. De igual forma sucederá con el repositorio GitHub. Es parte de la filosofía de la asignatura permitir la sorpresa que genera el descubrimiento del conocimiento.

1.1 Links fundamentales

Zulip https://dmeyf2021.zulip.rebelare.com/

repositorio Github https://github.com/dmecoyfin/dmeyf

Diccionario de Datos https://storage.googleapis.com/dmeyf/datasetsOri/DiccionarioDatos.ods

Invitación a la Primer Competencia Kaggle
https://www.kaggle.com/t/2410cfcc7d804b71a5b176231a442d39

Primer Competencia Kaggle https://www.kaggle.com/c/uba-dmeyf2021-primera/

Link Invitación a la Segunda Competencia Kaggle (se disponibilizará el lunes 18-oct)

1.2 Fechas Importantes

Comisión		Hora	Actividad
Lunes	**Jueves**		
18-ago		04:30	Envio por email de este documento a alumnos
18-ago	19-ago	19:00 a 22:00	Primera Clase
13-sep	09-sep	21:00	Lanzamiento Actividad sobre Feature Engineering "Dos Universidades"
20-sep	23-sep	19:00	Habilitación Scripts instalación Google Cloud
mie 29-sep		19:00	Lanzamiento "Desafío Cazatalentos"
dom 17-oct		23:59	Cierre automático de la Primer Competencia Kaggle
lun 18-oct		18:00	Lanzamiento Segunda Comptencia Kaggle
18-oct	21-oct	19:00	Análisis de resultados de la Primer Competencia, ¿Cómo evitar overfittear el Public Leaderboard? Lanzamiento lenguaje Julia
05-nov		19:00 a 22:00	Clase Conjunta Repensando el overfitting en 2021 con nuevo código para GBDT
mie 01-dic		23:58	Fecha límite entrega de los dos Videos y el brevísimo informe
		23:59	Cierre automático de la Segunda Competencia Kaggle
jue 02-dic		19:00 a 22:30	Presentación de equipos ganadores Segunda Competencia
jue 02-dic		20:30 a 22:00	Evaluación Individual Escrita con constancia para la CONEAU
vie 03-dic		23:00	Lanzamiento actividad Torneo de Videos
mie 08-dic		23:59	Fin actividad Torneo de Videos
06-dic	09-dic	19:00 a 22:00	Ultima Clase, Recapitulación, Lecciones Aprendidas, Mejores Prácticas, ¿Bala de Plata?
dom 12-dic		23:00	Entrega de notas por parte de los profesores. Se determina quienes pasan a recuperatorio.
mie 15-dic		23:00	Entrega de instrucciones, datasets y scripts para recuperatorios individuales
30-abr-2022		23:59	Fecha límite de entrega de recuperatorio

1.3 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 numero entero, la máxima nota final posible es 10.

Nota Ordinaria	
Contribución Nota	Actividades Obligatorias
60%	Segunda Competencia Kaggle más brevísimo reporte. Se utilizará el lenguaje de programación **Julia** Tarea grupal de una o dos personas.
15%	Primera Competencia Kaggle (sin reporte) Se utilizará el lenguaje de programación R . Tarea grupal de una o dos personas.
15%	Participación significativa en conversaciones el foro de la materia con herramienta *Zulip* (tipo Slack) en donde los alumnos deban reflexionar profundamente sobre el camino y obstaculos encontrados en la resolución del problema, intercambiando ideas y brindándose soluciones; con la participación de los docentes. Participación significativa en tareas de anotación colaborativa con herramienta *hypothes.is* anotando e intercambiando comentarios en artículos propuestos, papers, scripts y Jupyter notebooks
5%	Video Presentación de 5 minutos sobre lo más relevante de la confección del modelo de la competencia Kaggle, con storytelling dirigido al Gerente de Business Intelligence- Tarea Individual
5%	Video Presentación de 5 minutos con storytelling a la Directora Comercial explicando que clientes se dan de baja y proponiendo alguna acción Tarea Individual
100%	Total

Todas las actividades anteriores son obligatorias. La no participación/entrega en cualquiera de ellas implica el desaprobar la materia y pasar directamente a recuperatorio.

Nota Extraordinaria se suma a la nota de la materia	
Nota	**Condición**
+1	ganador del *Desafío Cazatalentos 15k*
+2	ganador del *Desafío Cazatalentos 14k* (no es acumulable con el desafío 15k)

1.3.1 Evaluación Competencias Kaggle

Las competencias Kaggle de la materia se califican en función de la ganancia obtenida en el Private Leaderboard.

Debe quedar claro que 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 formando grupo de una sola persona, o grupos de dos personas; no están permitidos los grupos de 3 o más personas. Está permitido cambiar de grupo de la primer a la segunda competencia.

Las ganancias se normalizarán y se mapearán al intervalo [0,10] de notas.

1.3.2 Evaluación participación en foro Zulip

La idea de la herramienta Zulip es junto con las clases sincrónicas lograr una experiencia cognitiva y emocional de resonancia que amplifique el aprendizaje.

Rúbrica de participación en foro Zulip	
Porcentaje	**Concepto a Evaluar**
30%	Contenido. Propone ideas profundas, signficativas, claras, fáciles de implementar y con potencial para el proyecto de la asignatura. Compartiendo resultados logra atraer la atención y colaboración 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 algun 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. Todas las semanas posee relevantes participaciones.

Metafóricamente, la tarea de los profesores en Zulip, más allá de soporte y burócratas de la cartelera, será la de sacudir aún más el puente peatonal Millennium Bridge de Londres https://www.youtube.com/watch?v=t-_VPRCtiUg&t=915s

Cada alumno participa con su usuario en forma individual en Zulip

1.3.3 Evaluación participación en anotación colaborativa Hypothes.is

Todas las semanas se propondrán documentos para la anotación colaborativa que versarán sobre:

- Artículos de interés general que aunque no siempre están directamente relacionados con el problema a resolver en la asignatura, son relevantes al ejercicio profesional de la ciencia de datos. Algunos serán intencionalmente controversiales para provocar discusiones y ejercitar el pensamiento crítico. Estos artículos serán los más comunes las primeras semanas.
- Artículos técnicos relacionados con temas específicos requeridos para resolver el problema.
- Papers relacionados con temas necesarios para resolver el problema de la asignatura
- Código fuente, scripts y jupyter notebooks oficiales de la asignatura.

Rúbrica de participación en Anotacion Colaborativa	
Porcentaje	**Concepto a Evaluar**
30%	Profundidad y originalidad de la Interpretación, la mayoría de los comentarios del alumno revelan que ha evaluado el documento en detalle y reflexionado profundamente sobre el significado, aportando ideas originales
30%	Participación en conversaciones (responder anotaciones de otros alumnos), el alumno agrega sustancia a los comentarios de compañeros, más alla de simplemente indicar aprobación o desaprobación
20%	Cantidad de comentarios, el alumno agrega varios comentarios relevantes en cada uno de los documentos.
20%	Clarificación de conceptos, el alumno colabora activamente en clarificar conceptos relevantes, propone links interesantes, relaciona conceptos complejos.

Cada alumno participa con su usuario en forma individual en Hypothes.is , por supuesto, anotando colaborativamente con todos sus compañeros del curso.

1.3.4 Evaluación videos presentación

En el verano posterior a finalizar su primer año de la maestría usted ingresa como analista de ciencia de datos en la compañia. Pasaron ocho meses, le encargaron el proyecto del modelo predictivo de retención proactiva de clientes, lo ha terminado y ahora debe comunicarlo efectivamente a dos personajes muy distintos.

Se deberán confeccionar dos videos, uno dirigido al gerente de ciencia de datos y otro a la directora comercial. La duración de los videos debe estar en el entorno de los 5 minutos.

Juan Grande, 38 años, es el gerente de ciencia de datos de la compañía, es su jefe directo y fue quien lo contrató hace 8 meses.

Juan posee un título de grado de Actuario, desde su graduación de dedicó a análisis econométricos en riesgo crediticio, en el año 2015 cursó una maestría en ciencia de datos e ingresó a trabajar a la compañía en el año 2016, ya en 2019 lo ascendieron a gerente en medio de una reorganización general de toda el área comercial. Juan es una persona muy metódica y organizada, pausado en su hablar, elige sus palabras con gran precisión, reflexivo, considera muchas opciones antes de tomar una decisión, ante una situación difícil a resolver escribe un cuadro en su excel con las alternativas a las que les estima una probabilidad. Juan en las reuniones va escribiendo la minuta en tiempo real y la disponibiliza a todos ni bien termina. Juan es una persona muy focalizada y ninguna idea foránea lo distrae del problema que debe resolver. Prefiere aprender en forma estructurada y abordar los temas desde lo abstracto.

Juan posee un elevado sentido de la ética y la justicia.

Juan espera de usted un video con una breve presentación de alrededor de 5 minutos en donde con un storytelling le cuente la forma en que resolvió el problema, los hallazgos más importantes. Esta no es una tesis de maestría, usted no debe explicar el algoritmo árbol de decisión ni gradient boosting, usted debe ir al grano con Juan, pero sorprenderlo.

A Juan le reportan 20 personas; él le reporta a la directora comercial y es el "gerente de menos peso". Sus pares son las siguientes gerencias : sucursales (1000 comerciales), productos pasivos, productos activos, canales digitales, marketing, telemarketing y atención al cliente.

Miranda Wintour, argentina, 48 años, dos hijas gemelas pre adolescentes, es la directora comercial de la compañía desde hace dos años y medio, y en su meteórica carrera se pronostica que llegará a la gerencia general en dos años más.

Ambos padres de Miranda son nacidos fuera de Argentina dedicados en su momento a la actividad consular. Miranda de joven emigró, concluyó sus estudios secundarios en el UWC Atlantic College en Gales, se graduó con honores en Ciencias Políticas en la Sorbonne Université de París y cursó una maestría en economía en la London School of Economics and Political Science.

Miranda practica yachting desde su infancia, actividad fomentada por su padre quien le inculcó el trabajo en equipo y la competitividad. En su juventud participó de varias competencias internacionales, siendo un punto de inflexión en su vida la accidentada carrera de 1998 *54th Sydney to Hobart Yacht Race*. En su oficina posee un cuadro de muy importantes dimensiones con una fotografía de esa carrera en donde se aprecia a una joven Miranda formando parte de un numeroso equipo sobre una embarcación, al pie del cuadro reza una enorme leyenda “Las regatas se ganan en tierra”.

Miranda se unió desde muy joven a la compañía en Europa, estuvo a cargo del área Fintech europea y la

convencieron de hacerse cargo en argentina de una transformación radical. Su primera tarea fue desarticular el anquilosado club de amigos que había formado su antecesor.

Aunque va con una sonrisa y su tono de voz es muy bajo y sereno, todos tienen una especie de temor hacia ella. Se dice cuando luego de una exposición Miranda le dice al disertante "buen trabajo" sonriendo, antes de los tres meses esa persona ya no está más en la compañía.

Usted jamás ha participado en una reunión con ella y esta será su gran oportunidad de ser conocido y que lo empiece a tener en el radar, ya que rumores dicen que Miranda no está del todo convencida de la manifiesta aversión de Juan Grande a tomar riesgos.

Un total de 2500 personas dependen indirectamente de Miranda, estando el grueso en la red de sucursales. telemarketing y atención al cliente.

Usted debe realizar un video presentación a Miranda que no puede exceder 5 minutos en donde le explique los motivos por los que se dan de baja los clientes de paquete premium y proponga acciones para revertirlo. Miranda no sabe (ni le interesa saber) sobre Ciencia de Datos pero si está absolutamente convencida de las posibilidades que brinda la tecnología para tratar uno a uno a los clientes.

Rúbrica Videos Presentación	
Porcentaje	**Concepto a Evaluar**
10%	Entretenimiento. El video presentación es apasionante y no hay parte que aburra, son 5 minutos de pura adrenalina.
10%	Audiencia. El video presentación tiene totalmente en cuenta la audiencia para la cual está dirigido y saca provecho de las características únicas de esa audiencia.
30%	Historia . La presentación narra una historia, hay una clara introducción con un "gancho" que invita a ver el video, un desarrollo adecuado con una continuidad argumental lógica y un desenlace concreto. La narrativa está organizada en torno a las etapas de la pirámide de Freytag o estructura de similar complejidad donde la emoción juega un papel fundamental.
25%	Consistencia del contenido. Lo presentado refleja fielmente el conocimiento descubierto en la Segunda Competencia Kaggle y las conclusiones están sustentadas en datos que aparecen presentados adecuadamente.
25%	Originalidad del contenido. Las ideas presentadas son originales, ingeniosas, basadas en una profunda comprensión del problema.

Cada alumno debe hacer dos video presentaciones, es una tarea individual. Por más que se haya formado grupo de dos personas para la segunda competencia los videos deben ser distintos así como el material que los soporta.

La entrega de los videos se hace por Zulip, enviando un mensaje al Stream **z-entregafinal** que se disponibilizará a su debido tiempo. El mensaje deberá contenir un link de acceso público que no haga falta ni usuario ni password ni estar registrado en ninguna plataforma para quienes lo vean. Podrá estar en Twitch, YouTube, Vimeo, o directamente en Google Drive, Microsoft Dropbox, etc No se debe adjuntar ningun archivo.

Es parte fundamental de la tarea que los alumnos investiguen la forma de hacer una video presentación efectiva e intercambien ideas en Zulip.

1.3.5 Instancia Evaluación Individual escrita obligatoria

Esta instancia de evaluación individual escrita es obligatoria para todos los alumnos, y se presenta en formato de examen escrito de una hora duración, en donde los alumnos deberán responder preguntas personalizadas sobre cómo hicieron la predicción de la Segunda Competencia Kaggle.

Debe ser escrita de puño y letra en papel, en caso de presencialidad debe entregar las hojas; en caso de virtualidad debe escanear/fotografiar las hojas y subirlas al stream correspondiente de la plataforma Zulip.

El objetivo es garantizar que el alumno realmente sea autor del trabajo de la Competencia Kaggle.

En caso que la prueba escrita presente dudas a criterio del profesor titular, éste tomará una extensa evaluación oral al alumno.

Ejemplos de preguntas personalizadas del año 2019 efectuadas a distintos alumnos;

En el comienzo de su informe usted hace un muy interesante y original desarrollo

“El modelo entregado analiza las principales variables que retienen al cliente, por un lado la actividad bancaria y de visa; y por el otro la existencia de una mora, que implica una extensión de la permanencia en el banco. La conclusión general es que debe analizarse en mayor profundidad en cuáles de los casos de morosos podemos encontrar un verdadero usuario del crédito, que paga aunque con demoras, y con los correspondientes intereses. “

¿De qué forma concreta eso afectó a la creación de los modelos? ¿Creó una nueva clase? ¿Creó campos nuevos? ¿Dividió el dataset en distintas partes y construyó modelos distintos para cada uno de ellos?

Explique en gran detalle esto, ya que es muy novedoso y ningún otro grupo descubrió este camino.

La frase del informe “En cuanto a limpieza de datos si bien hicimos tratamientos de nulos con rpart, no lo convalidamos con la ganancia por lo que lo desestimamos. En una primera instancia nos dio resultados positivos pero repitiendo las pruebas esto no sucedió.“ no me queda clara ¿Podría explicar en gran detalle ?

En su informe menciona “ Posterior a esto, utilizamos un modelo de XGBoost con parámetros estándar de dicho algoritmo, considerando nuevamente que a partir de realizar la optimización bayesiana teníamos resultados menores de ganancia que utilizando los valores preestablecidos”

A este profesor le cuesta mucho creer que dicha afirmación sea verdadera. Explique los experimentos que hizo, de qué forma fueron hechos, y qué resultados obtenia.

Explique en gran detalle como hizo lo siguiente “El stacking consistió en utilizar los IDs resultado del modelo principal (Xgboost - ‘modelitos’ - undersampling - hiperoptimización con ventana móvil) por encima del punto de corte 0.025, a los cuáles se le adicionaron los IDs resultado de los modelos secundarios (‘línea de muerte’ original, y su variación utilizando exhist) que se encontraran en los primeros cuatro deciles de cada uno de dichos modelos.

1.4 Dedicación Estimada

Horas Semanales Alumno *estándar*			
Tarea	**Semanas**		
	Iniciales	**Intermedias**	**Finales**
Clase Sincrónica (presencial/semi presencial/virtual) diálogo oral con profesor participación sincrónica en Zulip durante la clase	3	3	3
Participación en foros de Zulip sobre lo visto en clase y discusiones sobre ideas derivadas	0.2	0.2	0.4
Tarea para el Hogar Instalación, configuración, aprender a utilizar herramientas de la materia, ver tutoriales en video, interactuar en foros Zulip para resolver problemas	1	0.5	0
Tarea para el Hogar Lectura y anotación colaborativa con Hypothes.is de artículos de interés general	0.5	0.5	0
Tarea para el Hogar Lectura y anotación colaborativa con Hypothes.is de artículos técnicos	0	0.5	0.5
Tarea para el Hogar ejecución paso a paso de scripts, experimentación con cambios, anotación colaborativa de scripts oficiales con Hypothes.is, participación en chats en Zulip modificaciones a scripts oficiales tomar prestadas ideas de un compañero Feature Engineering	1.5	1	2
Operación de experimentos	0	0.25	0.6
Registracíón e interpretación de experimentos	0.2	0.25	1
Diseño conceptual de experimentos nuevos para mejorar el modelo predictivo, codificación a partir de scripts oficiales, debug de errores	0	1	2
TOTAL	6.4	7.2	9.5

Estos tiempos corresponden a la mediana de los alumnos, los dos deciles superiores llegan facilmente a un 50% mas de tiempo a las semanas finales de la asignatura. Lo mismo sucede con el primer decil de los alumnos con más dificultades.

Luego de cada clase sincrónica se disponibilizará una Tarea para el Hogar, que será una guía de lecturas de artículos de interés general, artículos técnicos, papers más repaso de lo visto en clase, propuestas de pequeñas modificaciones a scripts vistos en clase e interpretación de resultados.

En general no es la idea de la cátedra que los alumnos asistan a clase con lectura previa de los temas teóricos que serán vistos en la misma, sino que se espera sorprenderlos en clase.

Si es la idea que los alumnos asistan a la clase habiendo hecho los experimentos sugeridos y probado alternativas originales para mejorar el modelo predictivo, de forma de compartir los hallazgos y dificultades en la clase sincrónica y generar una tormenta de ideas grupal.

1.5 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 Objetivos de la Asignatura

1. Resolver un problema de dimensiones reales del mercado local utilizando las herramientas tecnológicas para manejar grandes volumenes de datos y ser capaz de generar una predicción competitiva en el mercado profesional laboral argentino.
2. Aplicar en forma práctica a un problema concreto con datos reales los conocimientos vistos en las materias teóricas de los cuatrimestres anteriores, enfrentando eficazmente todas las etapas del data mining.
3. 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. Metodologia Machine Learning Operations
    5. Procesamiento de grandes volúnes en 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. Metodología CRISP-DM
3. Comparación de Modelos Predictivos
    1. Training Testing
    2. Estimación Montecarlo
    3. Validación Cruzada
    4. Metodología Walk Forward Validation
    5. Curva ROC, concepto de área bajo la curva AUC
    6. El problema de las múltiples comparaciones y su relación fundamental con el sobreajuste *overfitting*
4. Algoritmos
    1. Breve reseña sobre Arboles de Decisión y Árboles de Estimacion de Probabilidad
    2. Algoritmos de Ensemble
        1. Random Forest
        2. Gradient Boosting of Decision Trees (XGBoost, LightGBM, CatBoost)
5. Métodos de Ensemble
    1. Voting
    2. Bagging
    3. Boosting
    4. Stacking
6. Optimización de Hiperparámetros
    1. Optimización Manual
    2. Red de Búsqueda (Grid Search)
    3. Optimización Bayesiana, teoría y práctica
7. Ingeniería de Atributos
    1. ¿Tratamiento de Nulos?
    2. ¿Selección de Variables?
    3. Variables derivadas, el problema de los cortes paralelos en los árboles de decisión.
    4. Incorporación de variables históricas
8. Interpretación de predicciones por el método “Explicaciones Locales Aditivas de valores Shapley”
9. Metodologías de MLOps, Machine Learning Operations, registración de experimentos
10. Narración de resultados, storytelling
11. Elementos del lenguaje estadístico R
12. Procesamiento en la nube, creación de máquinas virtuales, instancias “spot” o “preemptive”
13. Entornos y Herramientas
    1. RStudio, Jupyter Notebooks
    2. Kaggle

1.8 Desarrollo de una semana

Desarrollo de una semana	
Momento	**Actividad**
entre semana	Alumnos leen y comentan colaborativamente artículos de interés general , técnicos, papers, y código oficial provisto en clase.
entre semana	Alumnos realizan experimentos propuestos en la Tarea para el Hogar, hacen modificaciones al código para probar sus propios experimentos, intentan mejorar el modelo predictivo. **Esta es la parte más importante de la materia, construir creativamente con sus propias manos, tener los momentos "eureka".**
entre semana	Profesores participan activamente del chat Zulip y orientan la anotación colaborativa en Hypothes.is
clase	Profesor realiza brevísima síntesis de las discusiones que se presentaron en Zulip e Hypothes.is
clase	Alumnos y profesores analizan e interpretan los resultados de los experimentos corridos en la semana. Análisis de nuevas ideas traidas por los alumnos; grupalmente se analizan caminos alternativos. Apuntamos a que esta parte de la clase sincrónica ocupe al menos un tercio del tiempo. **Esta es la parte más importante de la clase sincrónica.**
clase	El profesor hace un repaso de temas anteriores, establece en el lugar del sendero se encuentra el proyecto., los problemas ya resueltos, los problemos resueltos a medias, y lo que sigue en el corto plazo.
clase	El profesor muestra slide con los temas del programa que seran tratados en clase en las próximas horas.
clase	El profesor comenta slide con la bibliografía específica de esa clase .
clase	El profesor plantea con gran emocionalidad el problema que se intenta resolver, la motivación y el objetivo pedagógico
clase	El profesor desarrolla algunas soluciónes al problema planteado
clase	Se establece una dinámica en clase, los alumnos corren algun pequeño script en clase y ven los resultados ahí mismo. Alumnos modifican creativamente el script. Se interpretan colectivamente los resultados del experimento. **Esta es la segunda parte más importante de la clase sincrónica.**
clase	Se alienta todo el tiempo a los alumnos a participar oralmente y en tiempo real por Zulip
clase	El profesor, cerrando la clase, resume lo que se ha podido tratar en la misma.
en la semana	Los profesores envian la correspondiente Tarea para el Hogar, donde incluyen a la plantilla cuestiones puntuales que surgieron durante la clase.

1.9 Bibliografía General

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

Libro de cabecera

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

Papers y otros Libros

Breiman, L. *Random Forests*, Journal of Machine Learning, Volume 45, Issue 1, 2001
Chen, T. and Guestrin, C., XGBoost: A Scalable Tree Boosting System, *Proceedings of the 22Nd ACM SIGKDD InternationalConference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.
Fawcett, Tom, ROC Graphs: Notes and Practical Considerations for Data Mining Researchers. *Technical Report HPL-2003-4*, HP Labs, 2003.
Flach, Peter The Geometry of ROC Space: Understanding Machine Learning Metrics though ROC Isometrics, *Proceedings of the Twentieth International Conference on Machine Learning (ICML 2003)*, 2003
Ke, Guolin, LightGBM: A Highly Efficient Gradient Boosting Decision Tree, *NIPS'17: Advances in Neural Information Processing Systems Conference*, 3148-3156 , 2017
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, *31st 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, Second Edition, *Cambridge University Press*, 2013
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
Reshef, David, Detecting Novel Associations in Large Datasets, *Science* 2011, Vol. 334 no. 6062 pp. 1518-1524, 2011
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
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

1.10 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.

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.

El enfoque pedagógico de esta asignatura está fuertemente relacionado al *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 la 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.

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.

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". "*Be undisciplined. Be irreverent. Be original*. Work hard. And focus on what you love. It doesn't get any simpler than that."

2 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 se debe aprender a utilizarlas.

Intente avanzar por su cuenta con los siguientes pasos, en caso de tener dificultades espere a la primera clase en donde se verán estos pasos y luego podrá consultar temas puntuales en Zulip.

2.1 Zulip herramienta de chat y foros discusión

Utilizaremos Zulip para

- Mantener discusiones en foros, sincrónicos durante la clase y asincrónicos luego.
- Intercambiar mensajes y archivos entre alumnos entre sí, alumnos y profesores
- Cartelera de información

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

En la materia no se utilizará el email, toda la comunicación y discusión se llevará a cabo en Zulip. Los profesores participarán activamente en las conversaciones de Zulip.

Adicionalmente las máquinas virtuales en la nube que se utilizarán para procesar los modelos avanzados le enviarán notificaciones a Zulip.

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

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

Zulip como aplicación permite esto https://zulip.com/features/ , una "falencia" de Zulip es que no permite enviar mensajes de voz.

Para darse de alta en el Zulip de la materia usted debe ingresar al link de Zulip de la sección 1.1 Links Fundamentales https://dmeyf2021.zulip.rebelare.com/ y presionar el botón de register.

Un video que muestra como utilizarlo 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.
general	Dialogos generales. Será el stream más utilizado.
cartelera	Profesores anuncian, alumnos no pueden enviar mensajes
clasesAlejandro	temas puntuales sobre las clases de Alejandro
clasesGustavo	temas puntuales sobre las clases de Gustavo
z-dosUniv	Tarea DosUniversidades
z-cazatalentos	Desafío Cazatalentos
z-entregafinal	Para ser usado al final de la materia para entregar: • Breve Reporte de lo hecho en la Competencia • Video presentacion al Gerente de Ciencia de Datos • Video presentacion al Director Comercial
z-torneovideos	Para el funcionamiento del Bot que permitirá alos alumnos rankear los videos

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

Hay un grupo creado llamado `profesores` , si usted desea que un profesor preste atención a su mensaje, simplemente incluya @profesores en el texto de su mensaje.

Toda comunicación con los profesores deberá ser pública, de forma que todos los alumnos puedan nutrirse de dichas interacciones.

2.2 Plataforma Kaggle

Kaggle es, entre otras cosas, una plataforma que permite participar de competencias de ciencias de datos.

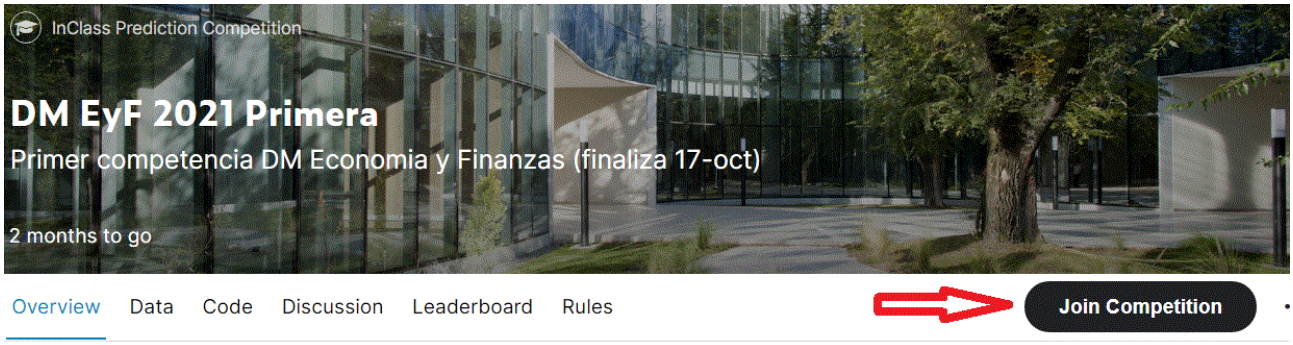
Utilizaremos la plataforma Kaggle para hostear las Comptencias Kaggle de la Materia.

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 email personal, no el de que le provee la universidad y utilice su verdadero nombre (no un pseudónimo).

Anote el nombre de usuario ya que lo requeriremos 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 https://www.kaggle.com/t/2410cfcc7d804b71a5b176231a442d39

y unirse a la competencia



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 luego ingresará asiduamente.

Lea con atención la sección Overview del menu principal, que contiene las subsecciones Description, Evaluation, Evaluation Details, Q&A y Past Opinions.

Lea con atención la sección Data del menu principal.

2.3 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 4.1.1 lanzada el 10-ago-2021

Una vez instalado el lenguaje R verifique que puede ingresar a la consola de R.

2.4 Librerías de R

Para la etapa inicial harán falta al menos las siguientes librerías, proceda a su instalación.

Utilidad	**Librerias**
misceláneas	Matrix, Hmisc , rlist , yaml , parallel , primes , bit64 , IRdisplay , repr , vioplot , DT , ROCR , R.utils
bajo nivel	Rcpp, devtools
gráficos	ggplot2, gganimate , transformr , DiagrammeR
Manejo datasets	data.table
Árboles de decisión	rpart , rpart.plot , treeClust
Bagging	ranger , randomForest
Boosting	xgboost , lightgbm
Optimización Bayesiana	DiceKriging , mlrMBO

Se deberán instalar TODAS las librerías anteriores

Usted notará que NO estaremos utilizando la librería *dplyr* pero que si usamos *data.table*

La instalación de paquetes en R presenta algunas sutiles diferencias entre los distintos sistemas operativos requiriendo algunas veces tener que instalar librerías adicionales.

La forma más recomendable de instalar es la siguiente (ejemplo con “data.table”)

```
install.packages(  "data.table", dependencies=TRUE )
```

Finalmente, desde la consola de R correr lo siguiente

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

2.5 Entorno de desarrollo RStudio Desktop

Es muy recomendable instalar RStudio Desktop, trabajaremos con la versión 1.4.1717 lanzada el 01-jun-2021

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.

2.6 Lenguaje de programación Python

Instalar la última versión del lenguaje Python de la página https://www.python.org/downloads/

Si usted posee Windows 10 podrá instalar la nueva versión 3.9.6 lanzada el 28-jun-2021

Si posee el antiguo Windows 7, deberá instalar la versión 3.8.10

Seguir los pasos y aceptar todo lo default, MARCANDO al inicio la opción "Add Python 3.9 to PATH" que se encuentra en la base de la primera ventana.

2.7 Entorno Jupyter Lab

En Microsoft Windows, desde la línea de comando ejecutar

```
pip install  jupyterlab   jupyterlab-git
```

Si se tiene otro sistema operativo, seguir las instrucciones de

https://jupyterlab.readthedocs.io/en/stable/getting_started/installation.html

2.8 Acceso de Jupyter Lab a R

Ingresar a la consola de R y correr lo siguiente

```
library( "IRkernel" )
IRkernel::installspec()
```

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

2.9 Estructura de Carpetas en su PC local

Cree en su computadora una carpeta exclusiva para la materia Data Mining en Economía y Finanzas (elijan un nombre corto y en lo posible sin espacios)

Dentro de esa carpeta crear la siguiente estructura de carpetas

- `datasets`
- `datasetsOri`
- `kaggle`
- `work`
- `TareasHogar`

Bajar estos tres archivos a la carpeta `datasetsOri`

https://storage.googleapis.com/dmeyf/datasetsOri/paquete_premium_202009.csv

https://storage.googleapis.com/dmeyf/datasetsOri/paquete_premium_202011.csv

https://storage.googleapis.com/dmeyf/datasetsOri/DiccionarioDatos.ods

Estos archivos son los primeros que utilizaremos para la Primer Competencia Kaggle

El archivo `paquete_premium_202009.csv` tiene el campo `clase_ternaria` que es nuestro objetivo de predicción, y es con el cual entrenaremos los primeros modelos.

El archivo `paquete_premium_202011.csv` es donde aplicaremos los modelos predictivos y generaremos la salida para Kaggle.

En la carpeta `kaggle` los scripts dejarán los archivos ya preparados para subir a Kaggle.

En la carpeta `work` los scripts dejarán resultados intermedios del procesamiento, algunos tan importantes como los resultados de la búsqueda de los hiperparámetros óptimos de los algoritmos.

2.10 Git para control de versiones

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 estamos optimizando. Para hacerlo en forma ordenada trabajaremos con control de versiones, la aplicación Git y usaremos el repositorio Github

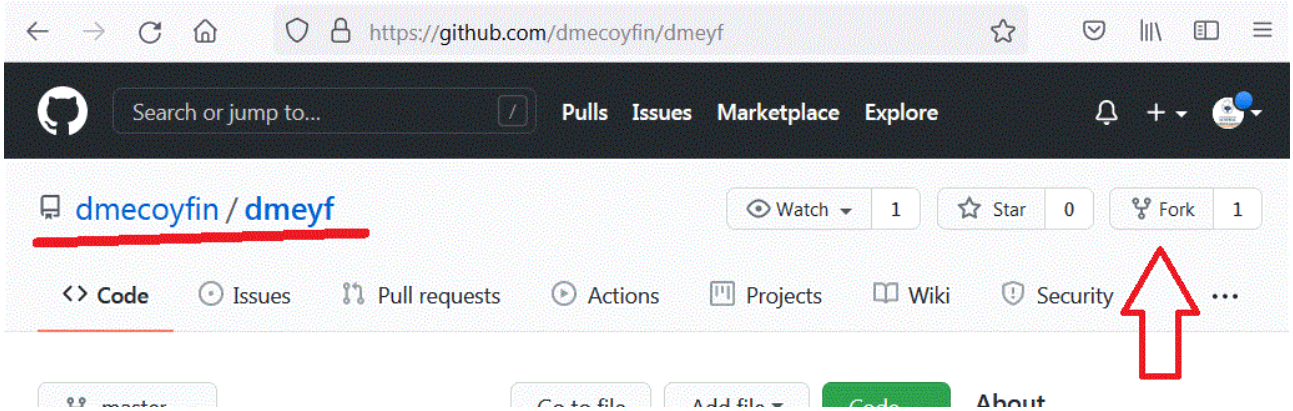
Instalar Git en su computadora local siguiendo estas instrucciones http://git-scm.com/downloads

2.11 Configuración GitHub

Para poder compartir nuestros proyectos el repositorio estará en la web. Para ello crearemos una cuenta en los servicios de Github

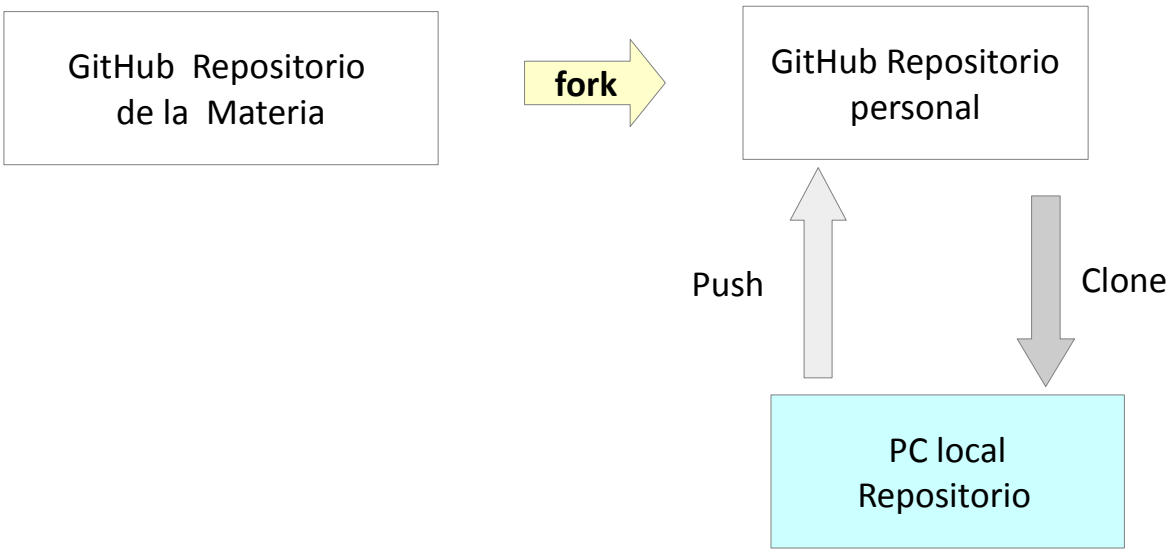
Registrarse en https://github.com/signup?ref_cta=Sign+up

Una vez hecho el login a su GitHub usted no tendrá ningun repositorio creado inicialmente, ir al repositorio GitHub de la materia (indicado en la sección 1.1 de este documento) y hacer un fork del repositorio presionando el botón que está arriba a la derecha de la página de GitHub .



De esta forma usted pasará a tener una copia del repositorio de la materia en su propia cuenta de GitHub en la nube.

En la siguiente sección hará una copia del mismo en su PC local que es donde trabajará y ejecutará los scripts.



Ahora pasaremos a crear el token, que es el password de GitHub

1. En Github ir al icono del usario en la parte superior derecha de la pantalla
2. `Settings` (última opción)
3. `Developer Settings` , en el menú de la izquierda
4. `Personal access tokens`
5. `Generate New Token`, copiar en un lugar seguro el token

2.12 Clonando el repositorio desde Rstudio Desktop

Ingrese a Rstudio Desktop en su PC local

Siga estas instrucciones

1. `File` en el menú principal, a la izquierda
2. `New Project` (la segunda opción)
3. `Version Control` (la tercer opción)
4. `Git "Clone a project from a Git repository"`
    1. `Repository URL`: https://github.com/dmecoyfin/dmeyf
    2. `Project Directory Name`: dmeyf
    3. `Create Project as subdirectory of`: buscar la carpeta creada en el punto 2.9

Se habrá creado una carpeta de nombre `dmeyf` dentro de su carpeta de la asignatura en su PC local, con la siguiente estructura inicial

- `dmeyf`
    - `dic`
    - `src`
        - `rpart`
            - `101_PrimerModelo.R`
            - `101_PrimerModelo.ipynb`

Ahora usted tiene en su PC local una foto del repositorio de la materia.

2.13 Prueba de entorno RStudio y Kaggle

Ingresar al RStudio y abrir el archivo `dmeyf/src/rpart/101_PrimerModelo.R`

Ejecutar el script línea a línea.

Finalmente deberá generar en la carpeta kaggle al archivo `K101_001.csv`

Ingresar a Kaggle a la página de la competencia, y presionar el boton negro con letras blancas en el menú superior izquierdo que dice "Submit Predictions", una vez allí ir al Step 1 que consiste en hacer el upload del archivo `K101_001.csv` , luego el Step 2 ponele un nombre significativo al submit y finalmente presionar el boton que esta en el cento inferior de la pantalla y dice "Make Submission".

Verificar que ha sido exitosa.

Felicitaciones, usted ha corrrido RStudio y subido su primer predicción a Kaggle; si visita el Leaderboard verá que está obteniendo una ganancia de 6.84514 o sea 6.8 millones de pesos !

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

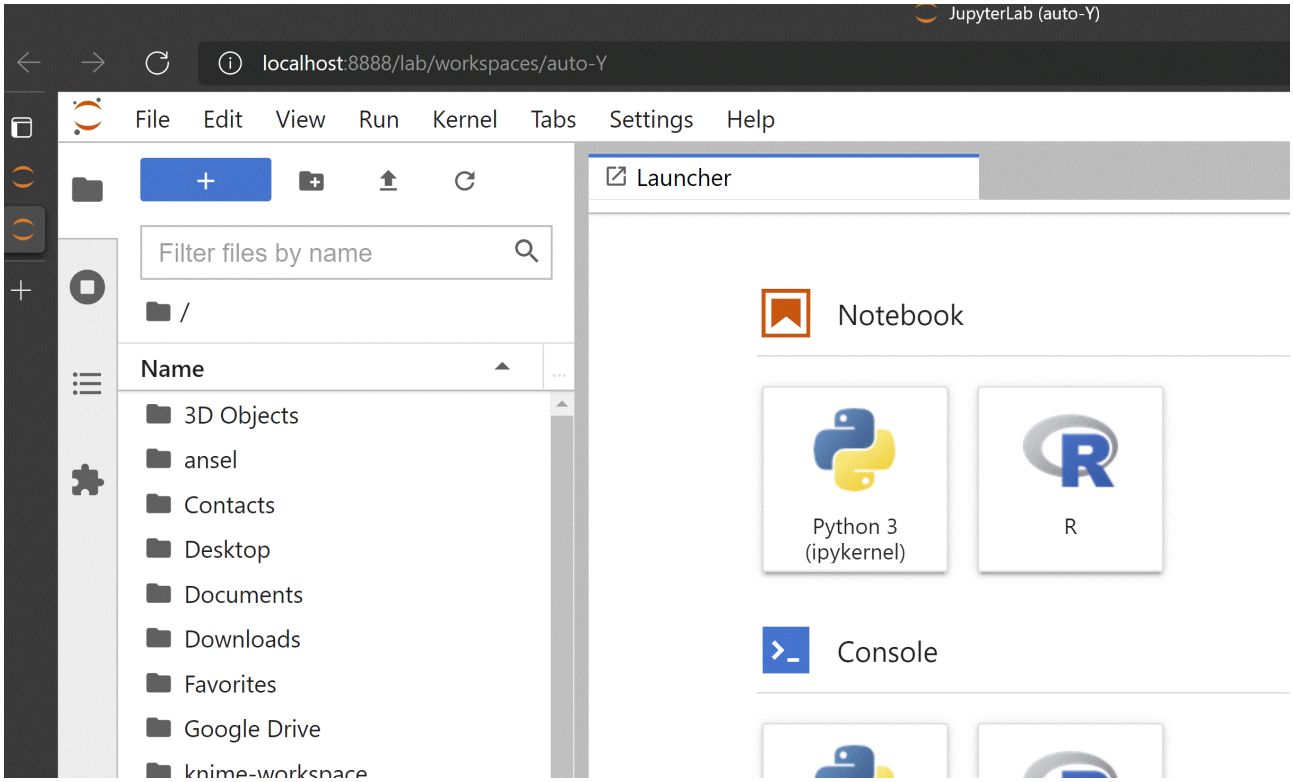
2.14 Prueba de entorno Jupyter Lab y Kaggle

Por la línea de comandos ir a la carpeta `dmeyf` dentro de la carpeta de materia.

Ejecutar el comando `jupyter lab`

Se abrirá el navegador y tendrá disponible los lenguajes Python 3 y R

deberá ver una pantalla como esta



Navegar hasta `src/rpart` y abir el notebook `101_PrimerModelo.ipynb`

Ejecutar linea a linea el notebook.

Al final de todo, debería haber quedado en la carpeta kaggle el archivo `K101_002.csv` , ingresar a Kaggle y subir dicha predicción.

Comparar la diferencia de ganancias en el Public Leaderboard

Si se le ha presentado algún error que no pudo solucionar, espere a ingresar a Zulip y consúltelo.

2.15 Hypothes.is herramienta de anotación colaborativa

Esta debe ser de todas las herramientas la más novedosa que utilizaremos en la asignatura.

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.

Hypothes.is es una herramienta de anotación colaborativa que utilizaremos para anotar y debatir notas de interés, artículos especializados pero principalmente la innovadora anotación de scripts en formato html y Jupyter Notebooks en formato html.

También será posible anotar los slides de la materia.

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>

Luego de instalar la extensión para Google Chrome, el mundo va a ser un lugar más feliz para la anotación colaborativa si utiliza Chrome, en caso contrario igualmente en esta materia le facilitaremos la vida con las anotaciones obligatorias.

No haremos públicas las anotaciones en la materia, sino que las acotaremos a tres grupos { dmeyf, dmeyf-art , dmeyf-code } , para sumarse a estos grupos estando logueado a Hypothes.is en su browser y luego seguir estos links

https://hypothes.is/groups/oZ5iroYw/dmeyf	anotaciones en general
https://hypothes.is/groups/X3LveG7X/dmeyf-art	anotaciones de artículos
https://hypothes.is/groups/dRALPikb/dmeyf-code	anotaciones en scripts y notebooks

Finalmente pondremos en practica la anotación colaborativa, para lo cual ya están disponibles dos artículos para anotar colaborativamente en el grupo `dmeyf-art`

https://storage.googleapis.com/dmeyf/annotation/general/why_so_many_data_scientists_are_leaving_their_jobs.html

https://storage.googleapis.com/dmeyf/annotation/general/why_business_fail_at_machine_learning.html

Y el script en R y el Jupyter Notebook en R, que deben ser comentados en el grupo `dmeyf-code`

https://storage.googleapis.com/dmeyf/annotation/code/101_PrimerModelo.R.html

https://storage.googleapis.com/dmeyf/annotation/code/101_PrimerModelo.html

2.16 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
8. Educación previa en ciencia de datos (Digital House, Coursera, edX, etc) detalle
9. ¿A que se dedica en su trabajo actual? ¿Trabaja en ciencia de datos o en algo relacionado? En caso afirmativo detalle.
10. ¿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 Fisica Cuántica, considerado el mejor profesor del siglo XX https://www.youtube.com/watch?v=BY6VntTmtIo
11. 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 de programación

    1. R
    2. Python
    3. Julia
    4. SQL
    5. GitHub
    6. Sistema operativo Linux
    7. Cloud Computing (Google Cloud, Azure, AWS, ...)
    8. Arboles de Decisión
    9. Gradient Boosting (XGBoost/LightGBM)
12. Utiliza algun entorno en particular de R (Rstudio, R Base, etc)
13. Utiliza algun entorno en particular de Python (VS Code, PyCharm)
14. Detalle experiencia en Sofware de Ciencia de Datos del estilo SAS, BeSmart, Knime, Weka, etc
15. Envíe CINCO números primos de 6 dígitos (del 100003 al 999983) intentando que sus números no se solapen con los que elijan sus compañeros de curso. Serán utilizados como semillas de generadores números pseudoaleatorios a lo largo de la materia.

2.17 Lenguaje de programación Julia

Se disponibilizará esta sección el lunes 18 de octubre.

2.18 Librerias de Julia

Se disponibilizará esta sección el lunes 18 de octubre.

2.19 Acceso de Jupyter Lab a Julia

Esta sección será disponibilizada el lunes 18-octubre

2.20 Entorno Google Cloud

A partir de la cuarta clase trabajaremos en Google Cloud donde crearemos máquinas virtuales poderosas (16vCPU, 256 GB de memoria RAM) . Recién procederemos a la instalación del entorno como Tarea para el Hogar entre la tercer y cuarta clase.

3 Sendero EDA

4 Sendero Arboles

5 Montecarlo Estimation y Cross Validation

6 Sendero Optimización de Hiperparámetros

7 Sendero Feature Engineering local

8 Sendero Random Forest

9 Sendero Gradient Boosting

10 Sendero Feature Engineering histórico

11 Sendero Shapley Values

12 Sendero Machine Learning Operations

13 Anexo Google Cloud