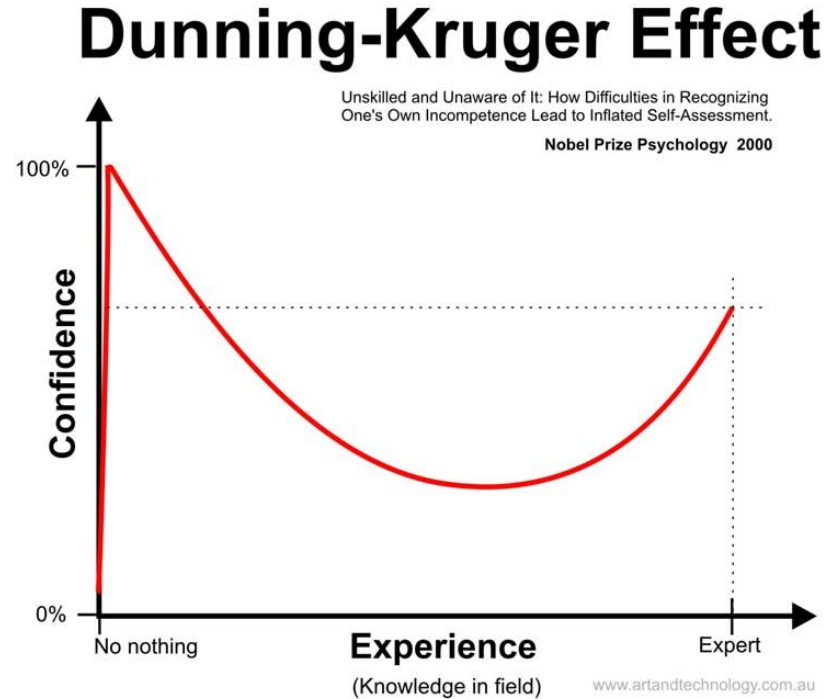


Primeros Pasos



Data Mining Economía y Finanzas - 2021 - Comisión 1

Alejandro Bolaños

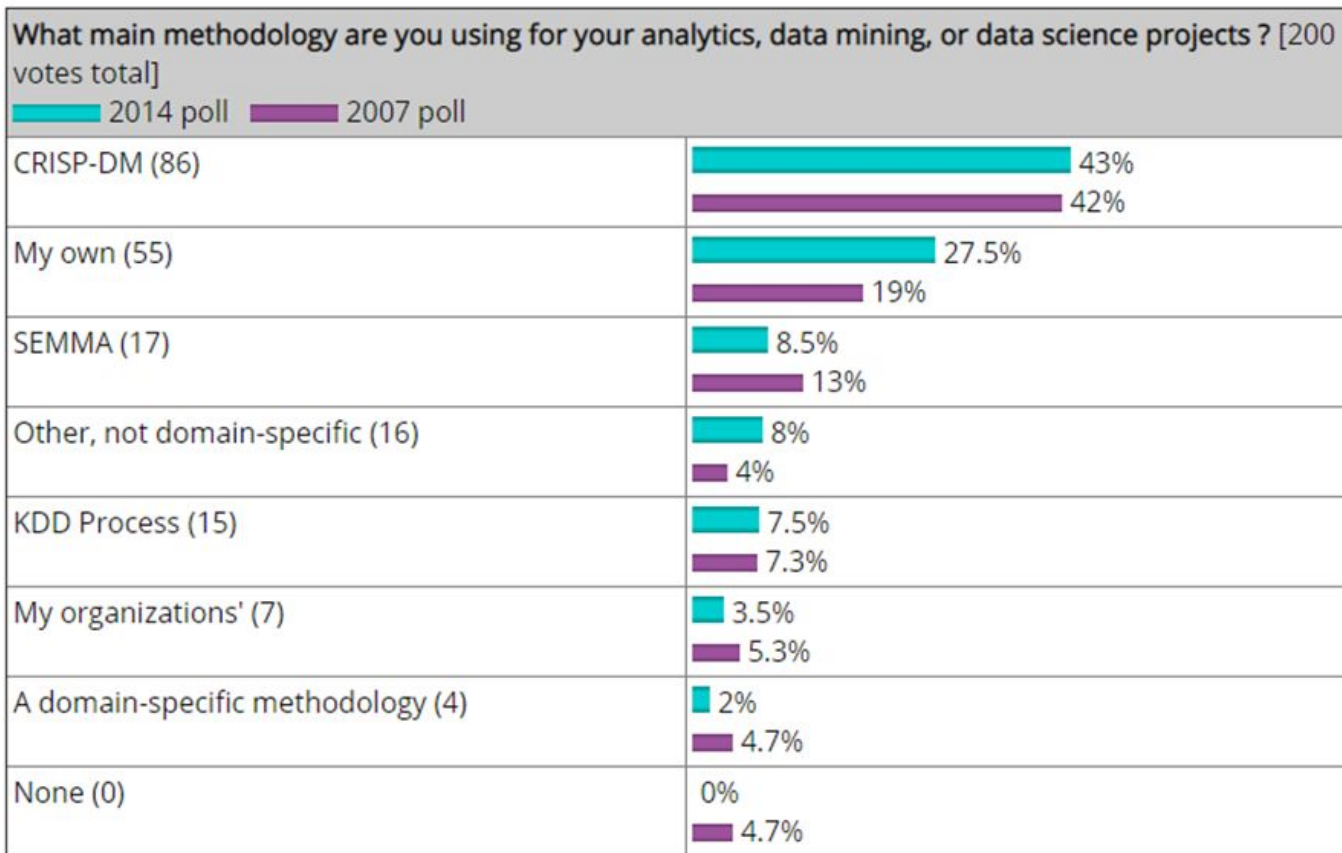
Primera Asignación > Abandono de clientes

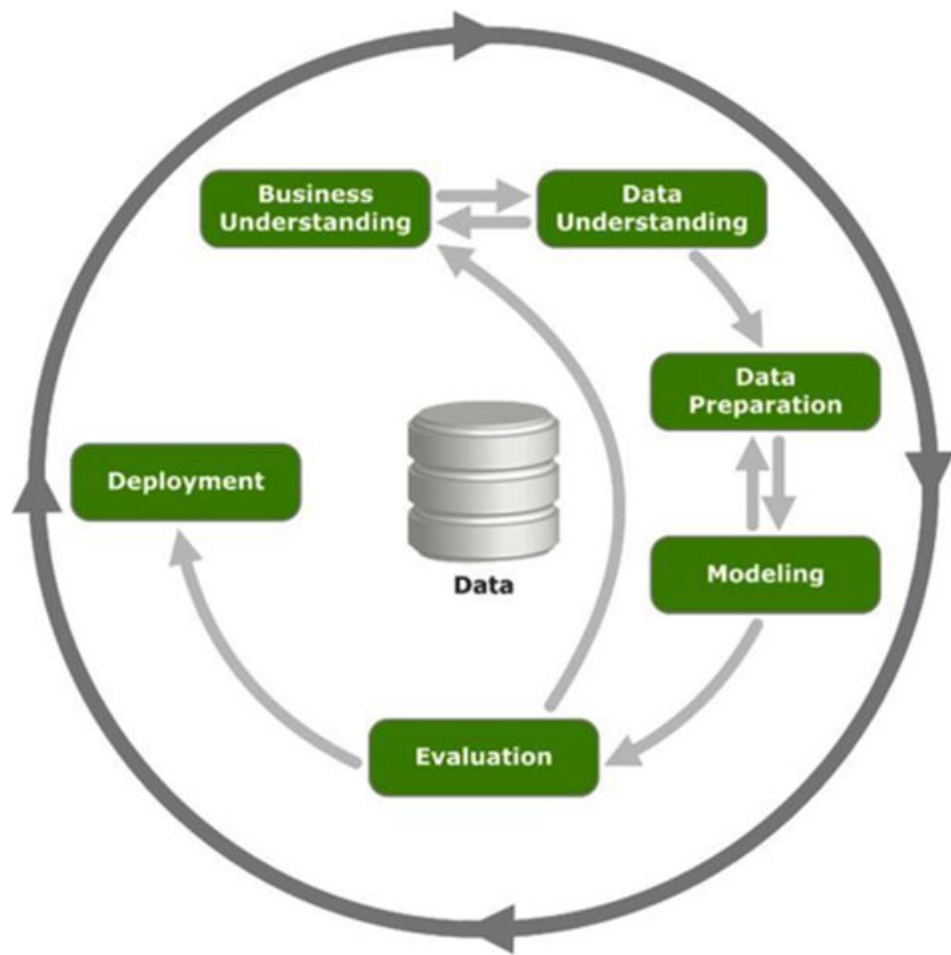
Minuta de la reunión

- Nuestra empresa tiene clientes de alto valor que son los que disponen del Paquete Premium
- Un cliente de alto valor en promedio genera a la empresa 100k pesos
- Adquirir a un cliente de alto valor es muy costoso
- Se realizó un experimento, donde sí se gastaba 1250 pesos en un estímulo para retener a un cliente premium, el 50% acepta y se queda
- Marketing quiere empezar a hacer campañas proactivas para evitar la fuga, le pide un listado de clientes a los cuales ellos deben estimular.
- Quieren la cantidad de clientes justa, les interesa maximizar la ganancia



Methodology





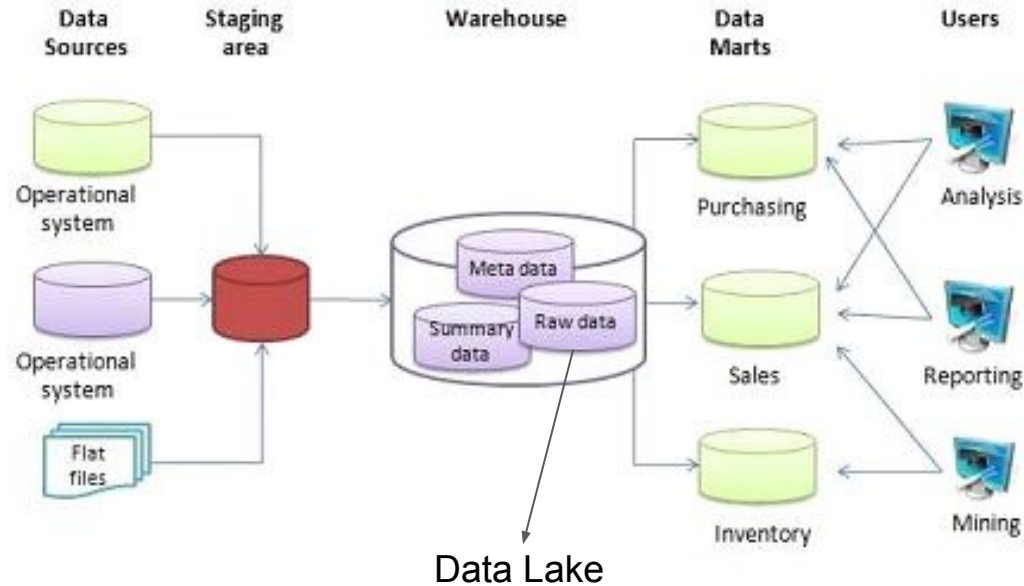
CRISP – DM

Cross-industry standard
process for data mining

Phases and Tasks

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives <i>Background</i> <i>Business Objectives</i> <i>Business Success Criteria</i>	Collect Initial Data <i>Initial Data Collection Report</i>	<i>Data Set</i> <i>Data Set Description</i>	Select Modeling Technique <i>Modeling Technique</i> <i>Modeling Assumptions</i>	Evaluate Results <i>Assessment of Data Mining Results w.r.t. Business Success Criteria</i> <i>Approved Models</i>	Plan Deployment <i>Deployment Plan</i>
Situation Assessment <i>Inventory of Resources</i> <i>Requirements, Assumptions, and Constraints</i> <i>Risks and Contingencies</i> <i>Terminology</i> <i>Costs and Benefits</i>	Describe Data <i>Data Description Report</i>	Select Data <i>Rationale for Inclusion / Exclusion</i>	Generate Test Design <i>Test Design</i>	Review Process <i>Review of Process</i>	Plan Monitoring and Maintenance <i>Monitoring and Maintenance Plan</i>
Determine Data Mining Goal <i>Data Mining Goals</i> <i>Data Mining Success Criteria</i>	Explore Data <i>Data Exploration Report</i>	Clean Data <i>Data Cleaning Report</i>	Build Model <i>Parameter Settings</i> <i>Models</i> <i>Model Description</i>	Determine Next Steps <i>List of Possible Actions</i> <i>Decision</i>	Produce Final Report <i>Final Report</i> <i>Final Presentation</i>
Produce Project Plan <i>Project Plan</i> <i>Initial Assessment of Tools and Techniques</i>	Verify Data Quality <i>Data Quality Report</i>	Construct Data <i>Derived Attributes</i> <i>Generated Records</i>	Assess Model <i>Model Assessment</i> <i>Revised Parameter Settings</i>		Review Project Experience <i>Documentation</i>
		Integrate Data <i>Merged Data</i>			
		Format Data <i>Reformatted Data</i>			

Datos > Un Mundo Ideal ...



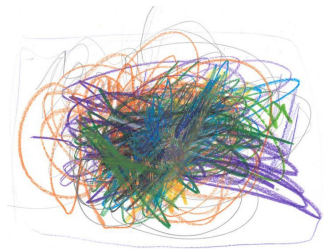
https://en.wikipedia.org/wiki/Data_warehouse#/media/File:Data_warehouse_architecture.jpg

Datos > Sucios como Berry en un día de Lluvia



Datos > Ejercicio de Construcción

Tenemos que pasar de



a \mathbb{R}^N

(hablando con licencia poética)

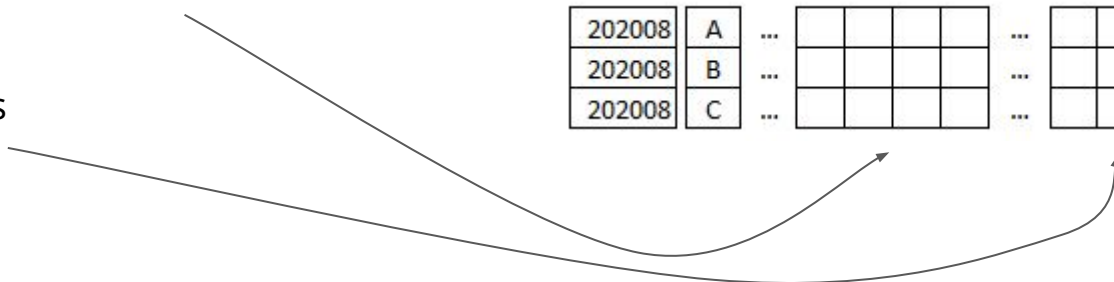
Pregunta: ¿Cómo pasamos info de ... a nuestro Dataset?

- Tarjeta de Crédito
- Payroll
- Movimiento entre cuentas
- Home Banking
- Atención al cliente
- Presencia en Sucursales

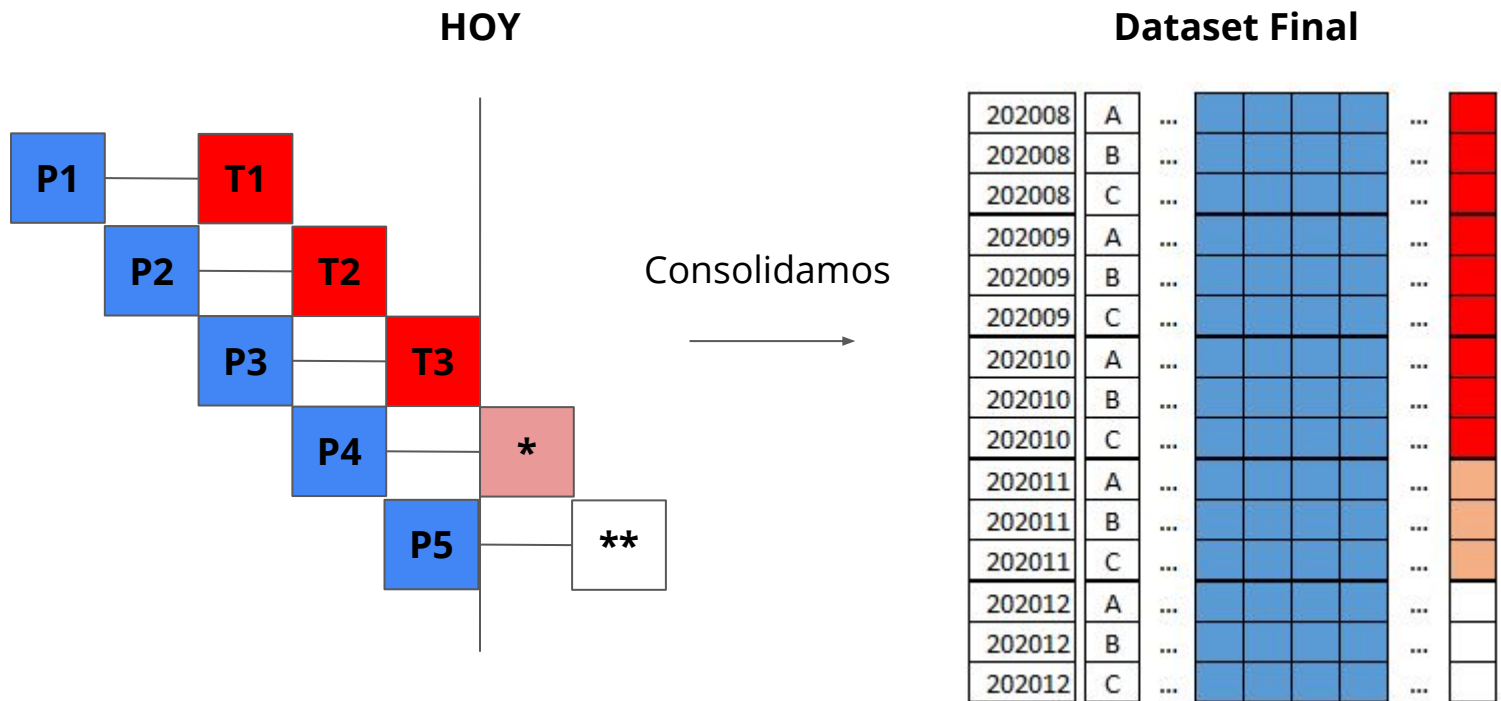
Identificador del intervalo de tiempo (Foto).

Identificador del sujeto analítico.

202008	A			
202008	B			
202008	C			



Datos > Construcción de Target



¿Cuál es el escenario actual de Kaggle?

Manos a la obra



Herramientas > Git

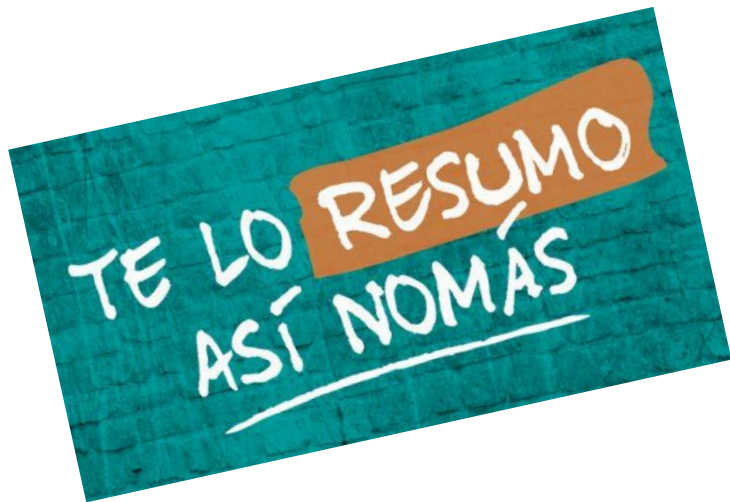


- ¿Qué es el código fuente?
- ¿Cuando un Data Scientist usa códigos fuente?
- ¿Qué es un repositorio de código?
- ¿Qué es Git?
- ¿Qué es Github?

Herramientas > GitHub Desktop (testing)



Herramientas > Notebook



• Literate programming - Donald Knuth

• Mathematica - Stephen Wolfram

• IPython - Fernando Pérez - Interactive Python

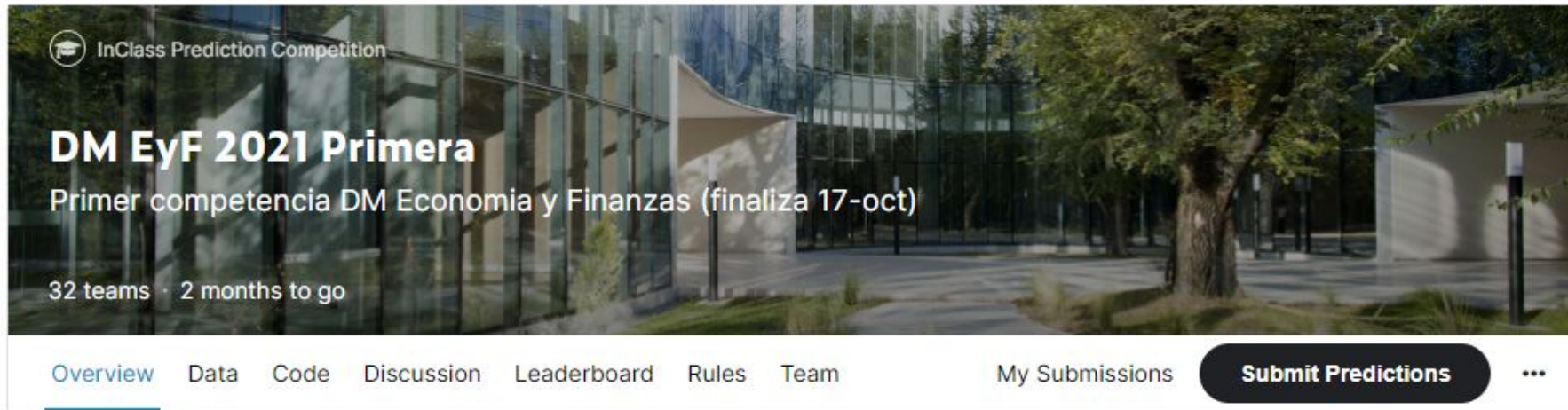
• IPython Notebook rename a Jupyter (Julia - Python - R)

• Alternativas

- RMarkdown
- Pluto.jl
- etc.

<https://yihui.org/en/2018/09/notebook-war/>

Kaggle > Un poco de Todo



The screenshot shows the top section of a Kaggle competition page. The background is a photograph of a modern glass-walled building with trees in front. In the top left corner, there is a small icon of a graduation cap and the text 'InClass Prediction Competition'. The main title 'DM EyF 2021 Primera' is in large, bold white letters. Below it, the subtitle 'Primer competencia DM Economía y Finanzas (finaliza 17-oct)' is in a smaller white font. Further down, it says '32 teams · 2 months to go'. At the bottom, there is a navigation bar with links: 'Overview' (underlined), 'Data', 'Code', 'Discussion', 'Leaderboard', 'Rules', 'Team', 'My Submissions', and a dark button labeled 'Submit Predictions' followed by three dots.

InClass Prediction Competition

DM EyF 2021 Primera

Primer competencia DM Economía y Finanzas (finaliza 17-oct)

32 teams · 2 months to go

[Overview](#) [Data](#) [Code](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Team](#) [My Submissions](#) [Submit Predictions](#) ...

- Sobre Kaggle
- Primera competencia - Entendiendo el Overfitting
- Diferencias sobre el leaderboard Público / Privado
- Subiendo un conjunto de datos

Pa' clase que viene > Repaso de árboles

Les recomiendo mirar los siguientes links donde muestra cómo funcionan los árboles de decisión.

Si usted tiene muy fresco estos conocimientos, lo invito igualmente a que vea el despilfarro de animaciones para explicar su funcionamiento:

<http://www.r2d3.us/visual-intro-to-machine-learning-part-1/>

<http://www.r2d3.us/visual-intro-to-machine-learning-part-2/>