

Daily life of a data scientist



Daniel García Fernández

2024-06-06



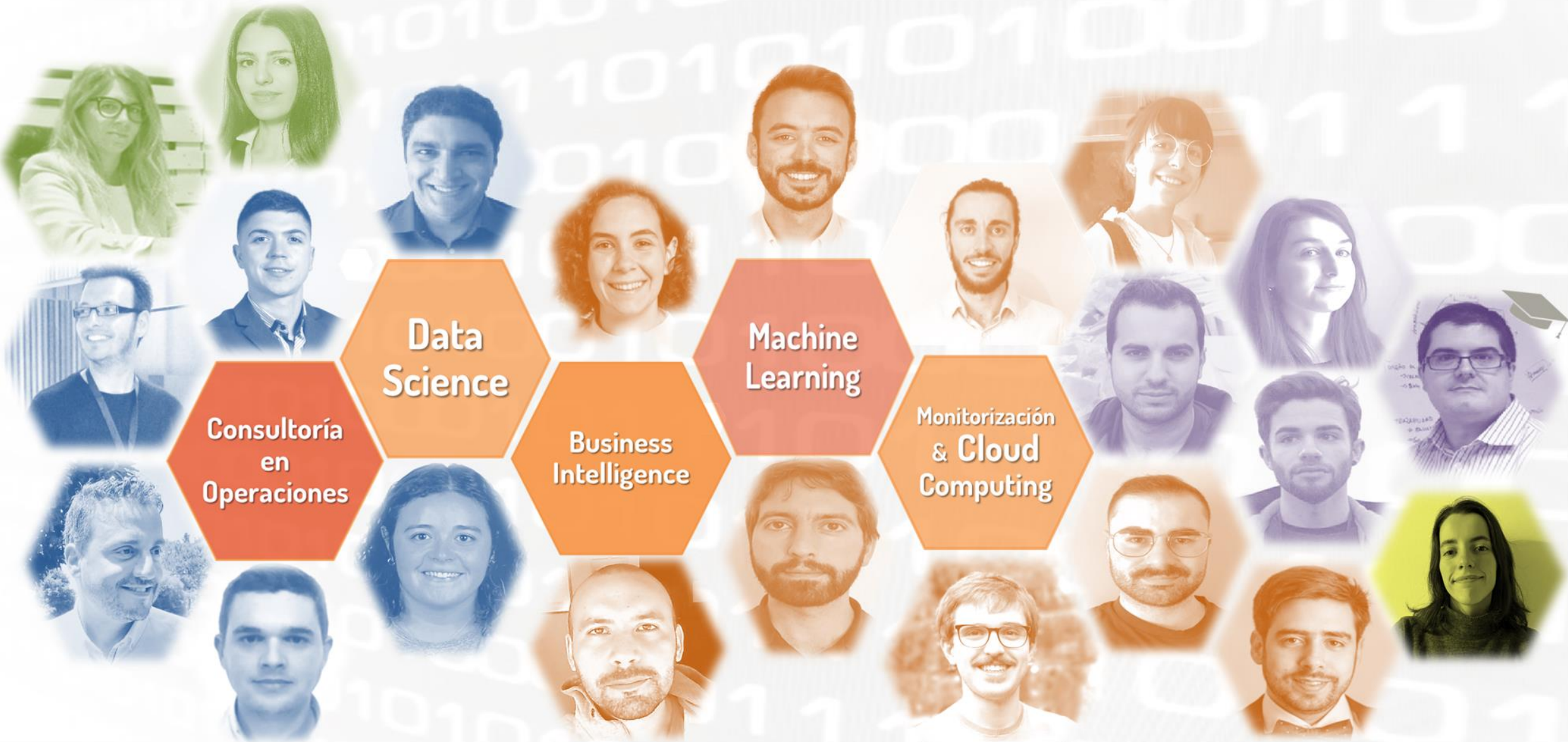
About TripleAlpha

“We help our customers during the processes of **data generation, management, and analysis** to extract the maximum **business value**”





About TripleAlpha



Consultoría
en
Operaciones

Data
Science

Business
Intelligence

Machine
Learning

Monitorización
& Cloud
Computing



About TripleAlpha

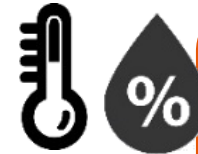
Which **process parameters** globally maximise my **profits**?



How does **logistics** affect **losses** at the production plant?



What **combination of variables** causes my defects?



How does the **raw material variability** affect the **output quality**?



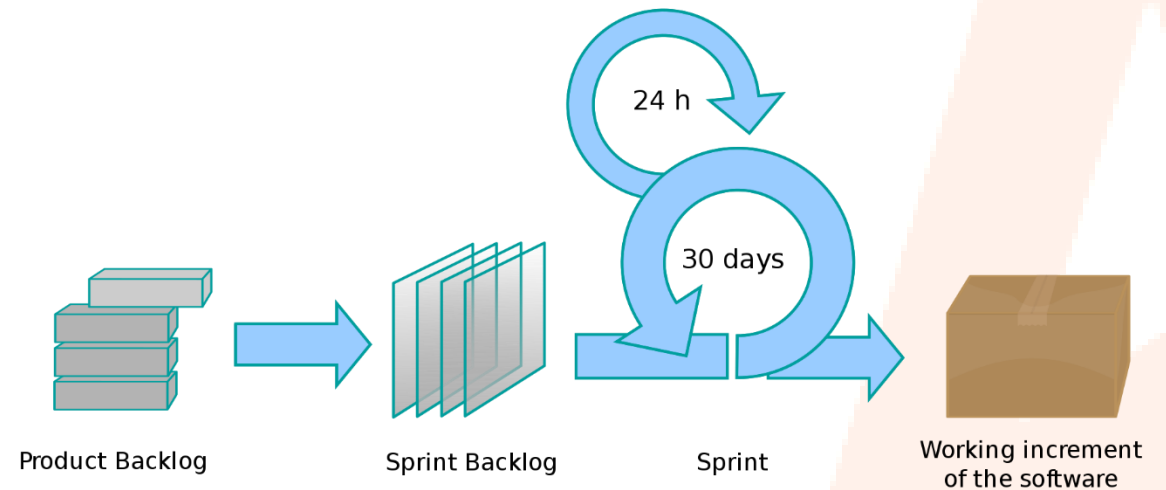
How does my **milk yield increase** if I temper the **water** in the watering?





0.1. Introduction

- Daily life of a data scientist
- We will cover ten days of an idealised *sprint* (see [Scrum, Agile](#) development)
- We will discuss data exploration, modelling, reports, presentations...



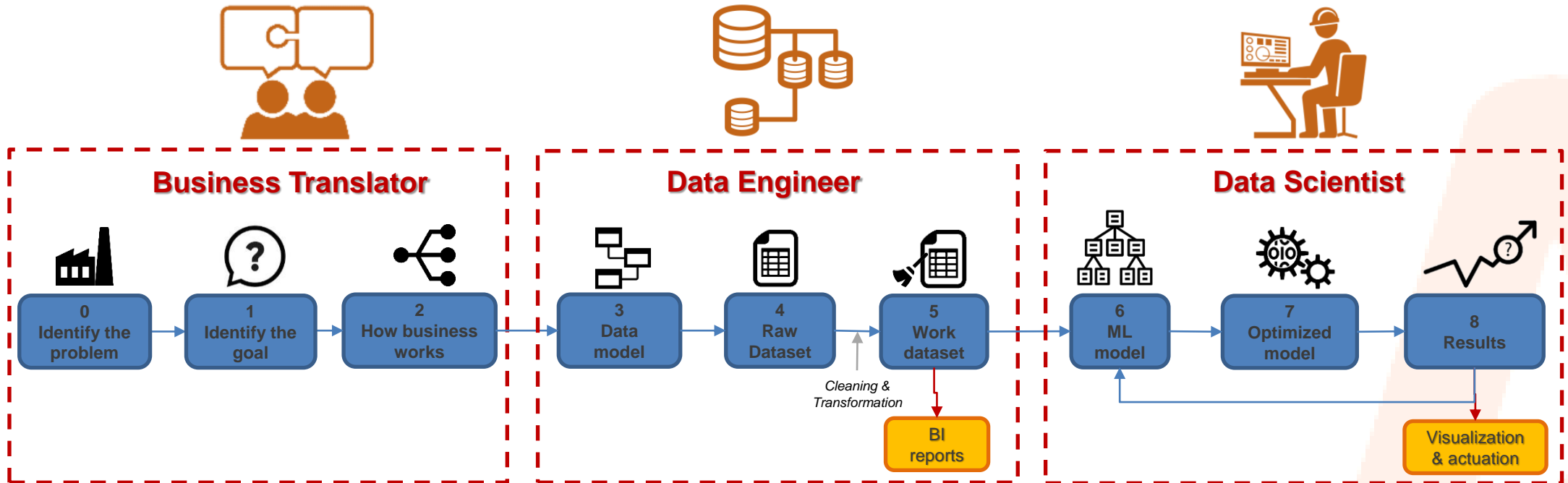


0.2. Previous requirements

- Let us assume you are a data scientist working for a company.
- Previous knowledge:
 - Programming:
 - Python, R (more statistical background), Julia, perhaps C...
 - Libraries: sklearn, pandas, matplotlib, scipy, numpy...
 - Statistics and mathematical modelling. The more, the better.
 - Linear models, Support Vector Machine, random forests, perhaps neural networks...
 - Databases. SQL, MongoDB.
 - Cloud computing basics (Azure, AWS...)
 - Important – either experience or will to gain experience facing real-world data.



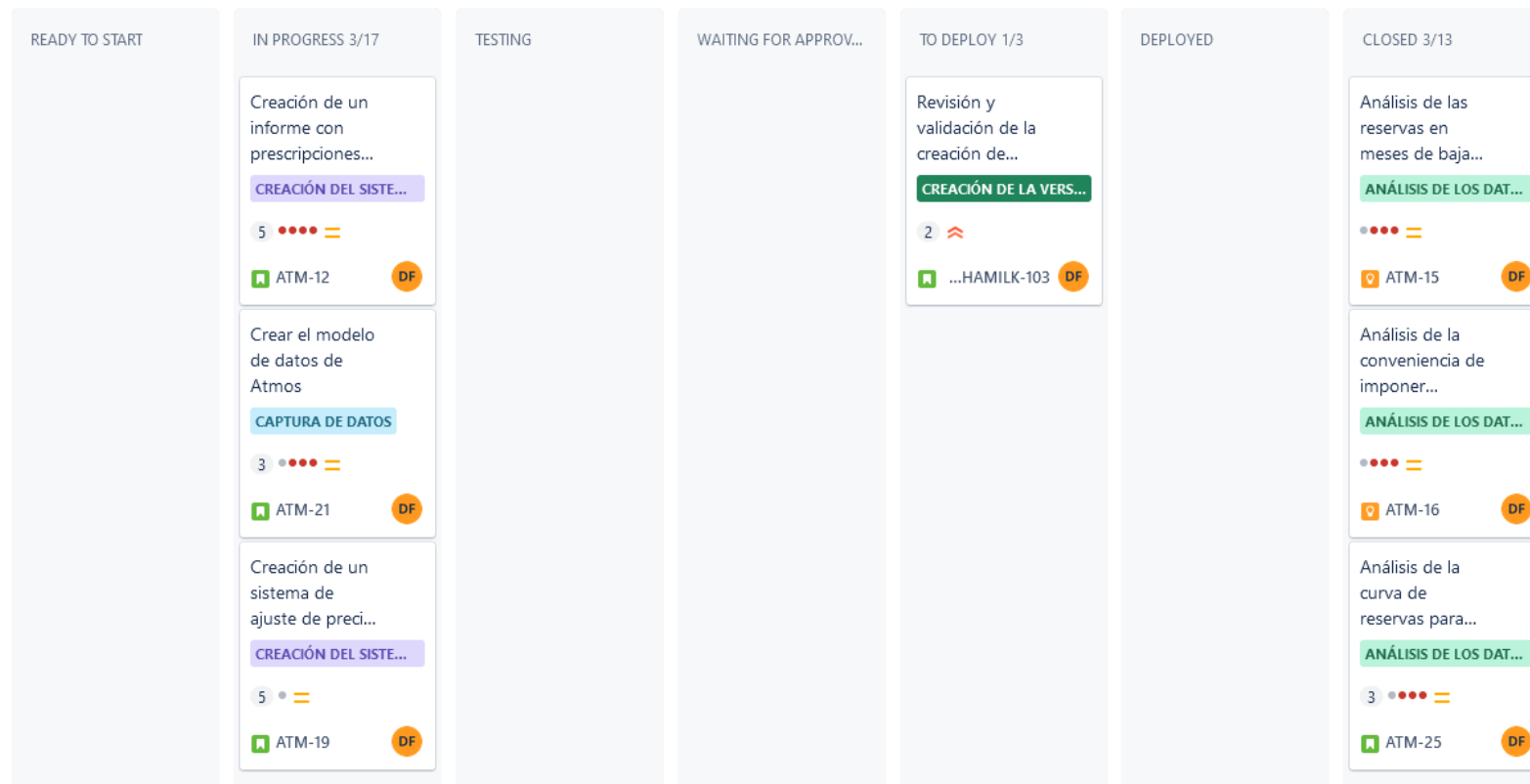
0.3. Our place





0.4. Task choosing

- From our task software (e.g. Jira) we pick the sprint tasks that have been created for US.
- We work on them, moving them through the stages until they are closed.





Day 1. EDA

- EDA: Exploratory Data Analysis.
- Crucial part of the analysis, despite its usual treatment.
- Task: to study the dataset that the data engineer has passed us.
 - Pandas for handling the data.
 - Plots: matplotlib, pandas profiling, seaborn, ggplot.
- We need to build insight on what is inside the dataset.

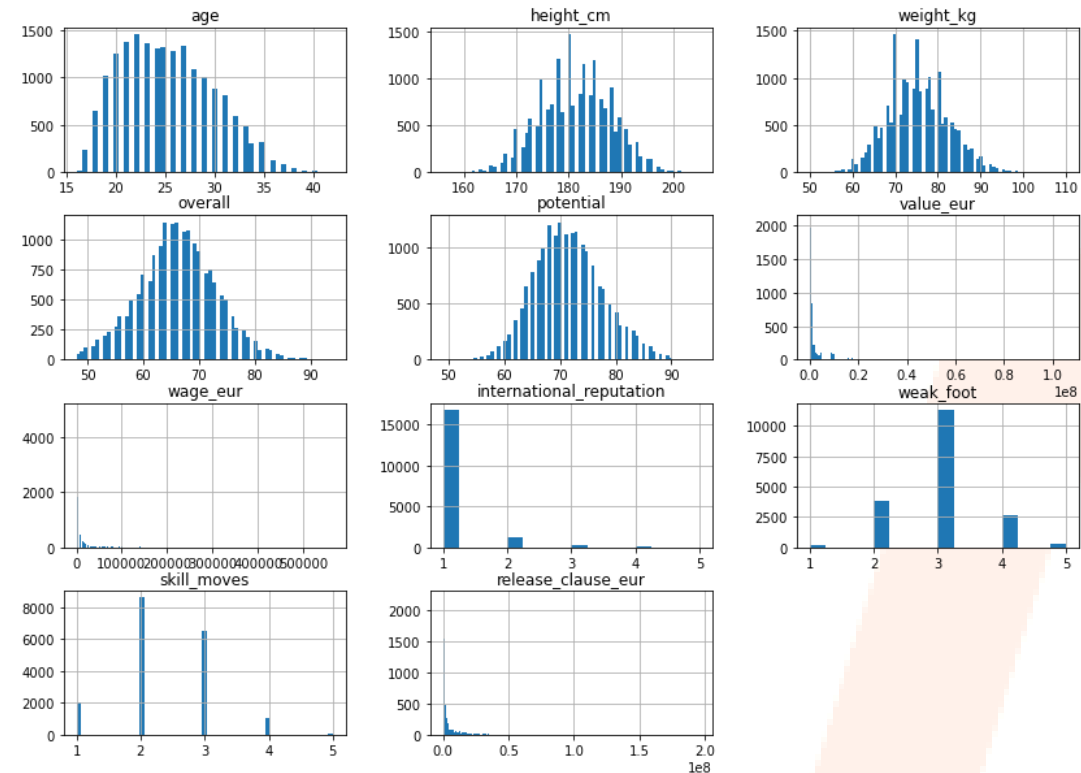
	sofifa_id	player_url	short_name	long_name	age	dob	height_cm	weight_kg	nationality	club	...	lv
0	158023	https://sofifa.com/player/158023/lionel-messi/...	L. Messi	Lionel Andrés Messi Cuccittini	32	1987-06-24	170	72	Argentina	FC Barcelona	...	6
1	20801	https://sofifa.com/player/20801/cristiano-ronaldo-dos-...	Cristiano Ronaldo	Cristiano Ronaldo dos Santos Aveiro	34	1985-02-05	187	83	Portugal	Juventus	...	6
2	190871	https://sofifa.com/player/190871/neymar-da-sil-...	Neymar Jr	Neymar da Silva Santos Junior	27	1992-02-05	175	68	Brazil	Paris Saint-Germain	...	6
3	200389	https://sofifa.com/player/200389/jan-oblak/20/...	J. Oblak	Jan Oblak	26	1993-01-07	188	87	Slovenia	Atlético Madrid	...	N
4	183277	https://sofifa.com/player/183277/eden-hazard/2...	E. Hazard	Eden Hazard	28	1991-01-07	175	74	Belgium	Real Madrid	...	6

<https://www.kaggle.com/code/swamita/exploratory-data-analysis-on-fifa-20-dataset>



Day 1. EDA

- Data cleaning.
 - We need to discover which ones are relevant for our goal.
 - It is possible that some of them are NaN or meaningless. We need to impute them.
- We plot the data to better understand the dataset.

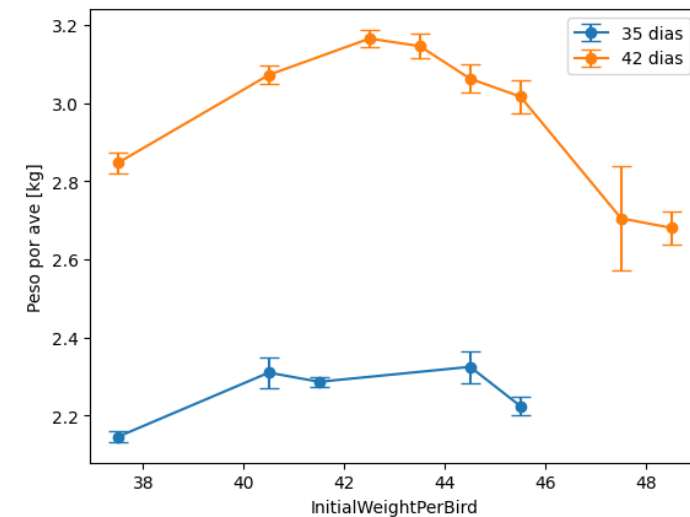
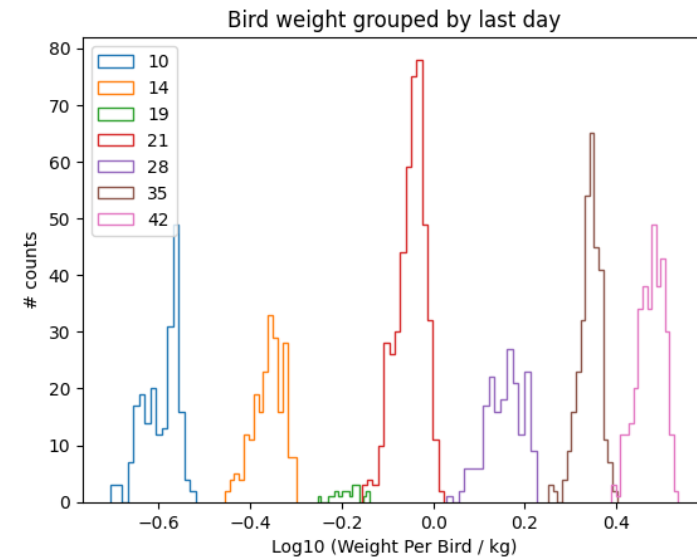
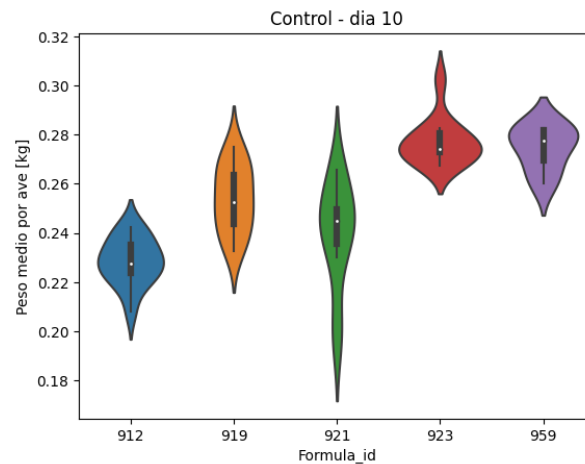


<https://www.kaggle.com/code/swamita/exploratory-data-analysis-on-fifa-20-dataset>



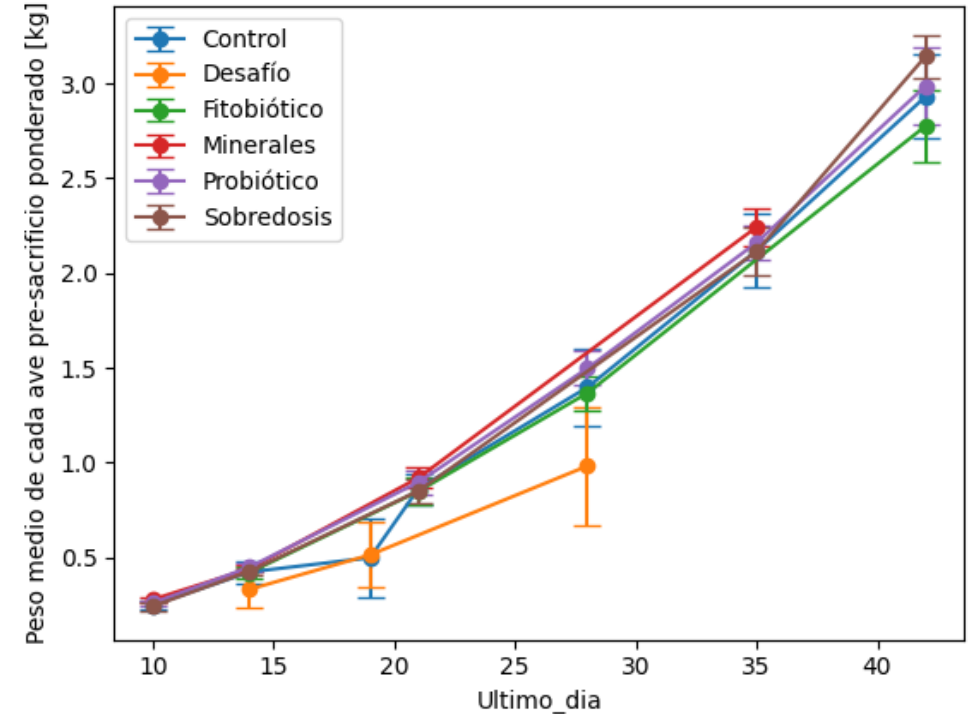
Day 1. EDA

- Comparing variables and extract correlations.
- Performing operations on data to reach meaningful variables.
- Identifying key target variables and calculating them.
- Time, environmental variables...



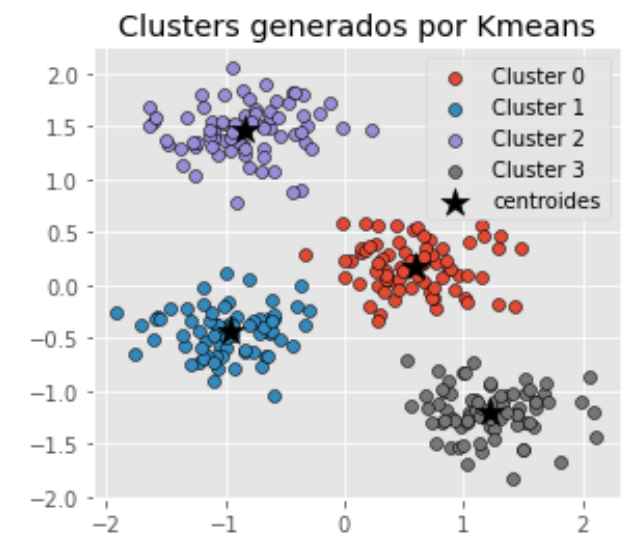
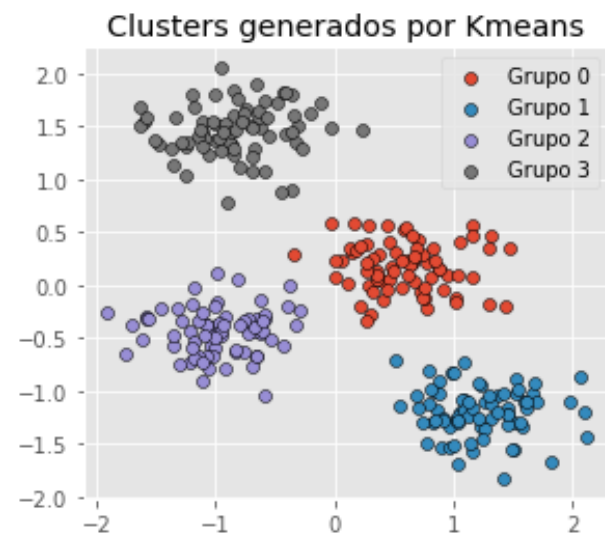
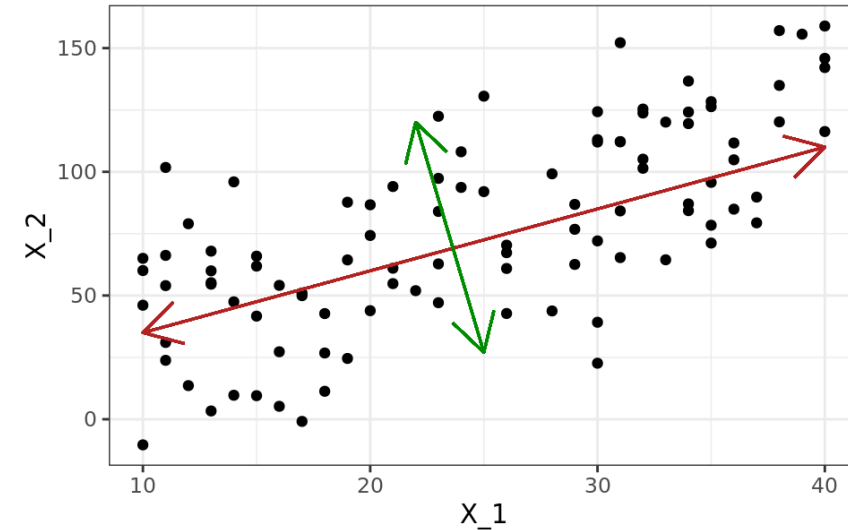


- Comparing variables and extract correlations.
- Performing operations on data to reach meaningful variables.
- Identifying key target variables and calculating them.
- Time, environmental variables...





- Advanced techniques:
 - Principal Component Analysis (PCA)
 - Clustering.
- Useful for increasing the interpretability of data.
- EDA is a fundamental part, if often overlooked, of the data analysis.

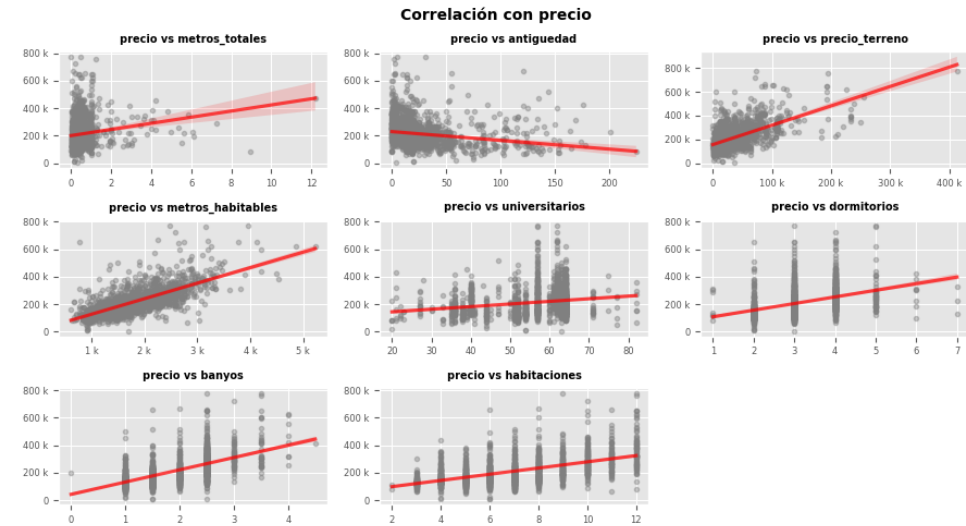




Day 2. Regression models

- We need to predict a numerical variable.
 - Sales of an article during a period, weight of a chicken after N days of breeding, price of a house...
- First step – choose and create the right variables.

https://cienciadedatos.net/documentos/py06_machine_learning_python_scikitlearn

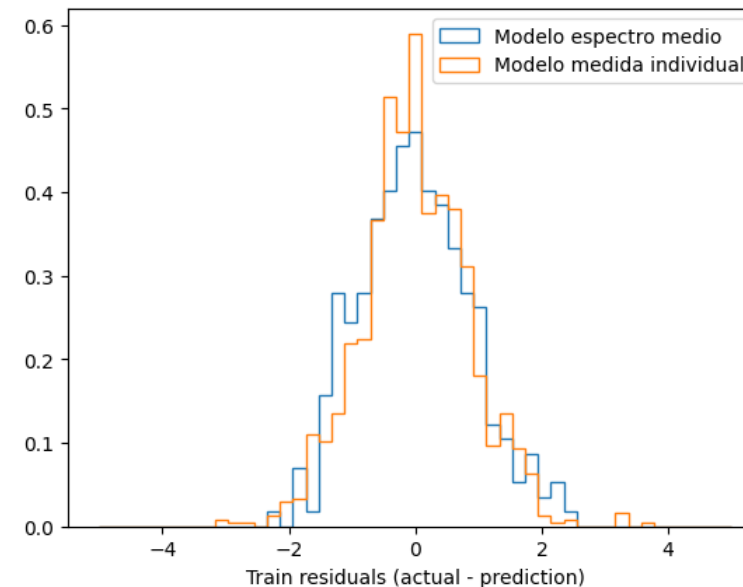
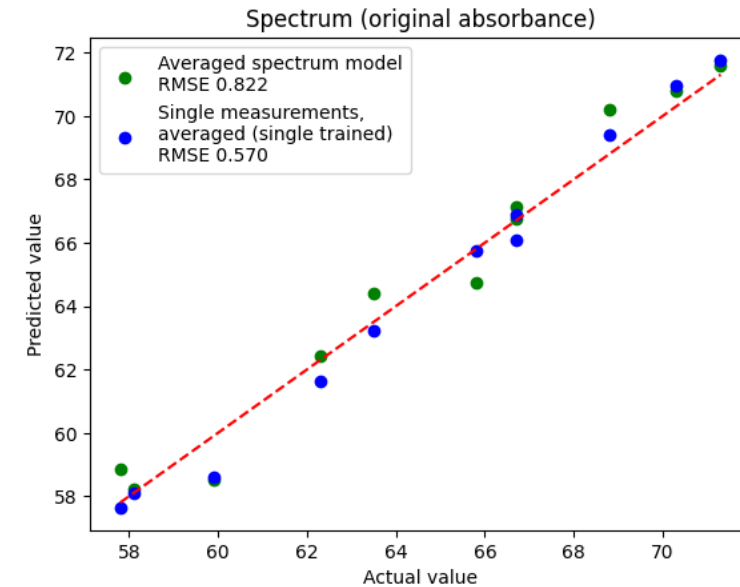


precio	1	0.16	-0.19	0.58	0.71	0.2	0.4	0.6	0.53
metros_totales	0.16	1	-0.016	0.059	0.16	-0.033	0.11	0.085	0.14
antiguedad	-0.19	-0.016	1	-0.022	-0.17	-0.038	0.027	-0.36	-0.082
precio_terreno	0.58	0.059	-0.022	1	0.42	0.23	0.2	0.3	0.3
metros_habitables	0.71	0.16	-0.17	0.42	1	0.21	0.66	0.72	0.73
universitarios	0.2	-0.033	-0.038	0.23	0.21	1	0.16	0.18	0.16
dormitorios	0.4	0.11	0.027	0.2	0.66	0.16	1	0.46	0.67
banyos	0.6	0.085	-0.36	0.3	0.72	0.18	0.46	1	0.52
habitaciones	0.53	0.14	-0.082	0.3	0.73	0.16	0.67	0.52	1



Day 2. Regression models

- Choose the model to apply
 - Linear (Ridge, Lasso)
 - Support Vector Machine
 - Random forest (xgboost, lightgbm)
 - Neural networks (careful!)
 - Craft model (e.g. with scipy)
- Creation of train and test datasets.
- Cross validation (CV)
- Errors for the chosen models (RMSE, SMAPE)

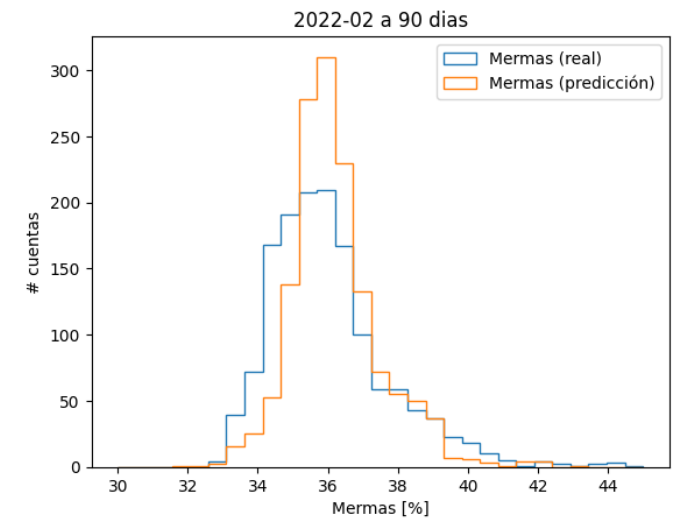
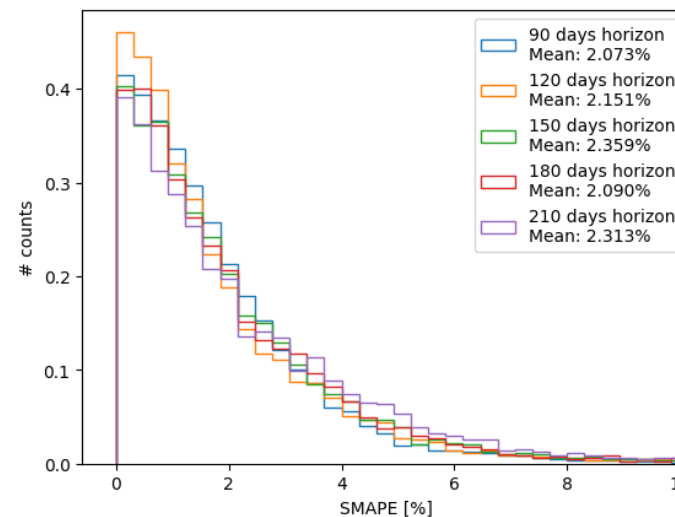
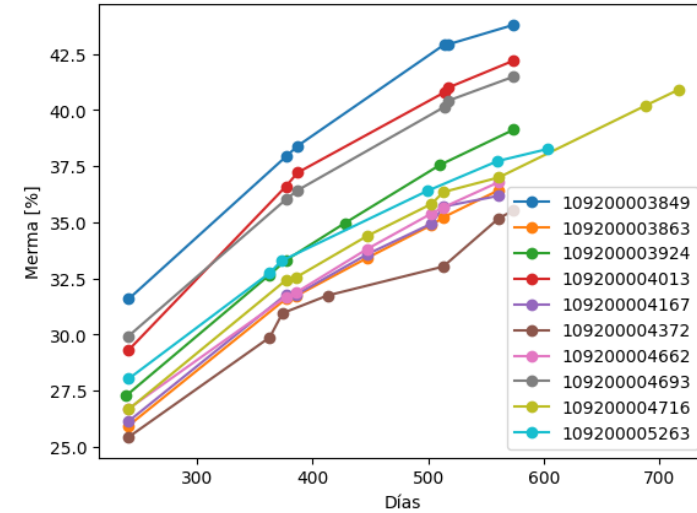




Day 2. Regression models

- Linear model for weight loss in cured hams
- Ideal case is exponential fit. Not feasible.
- Linear model – good accuracy. Compromise solution.

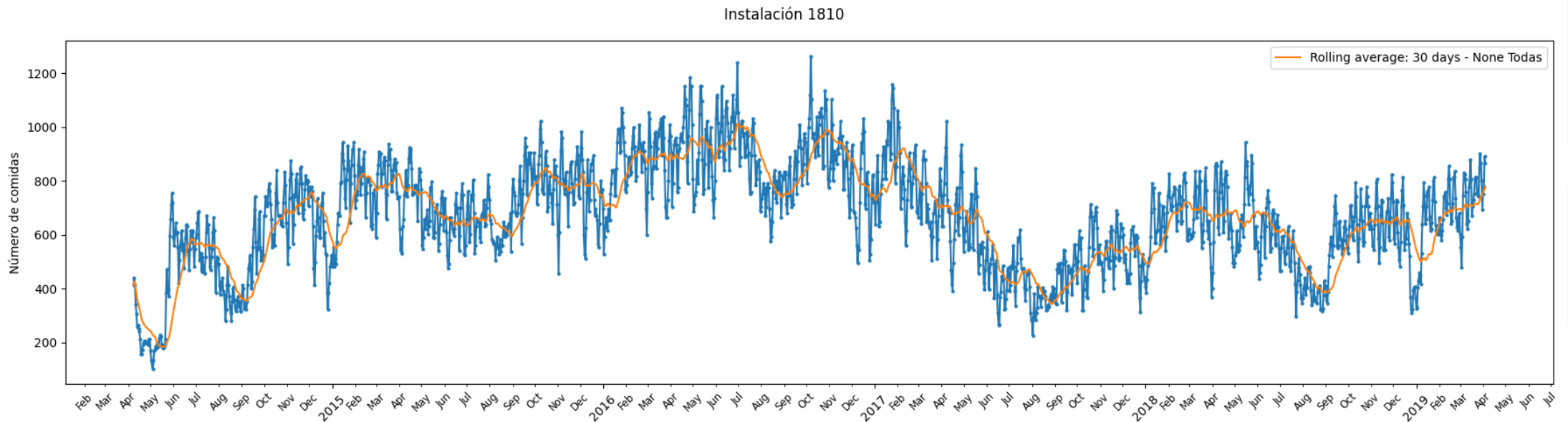
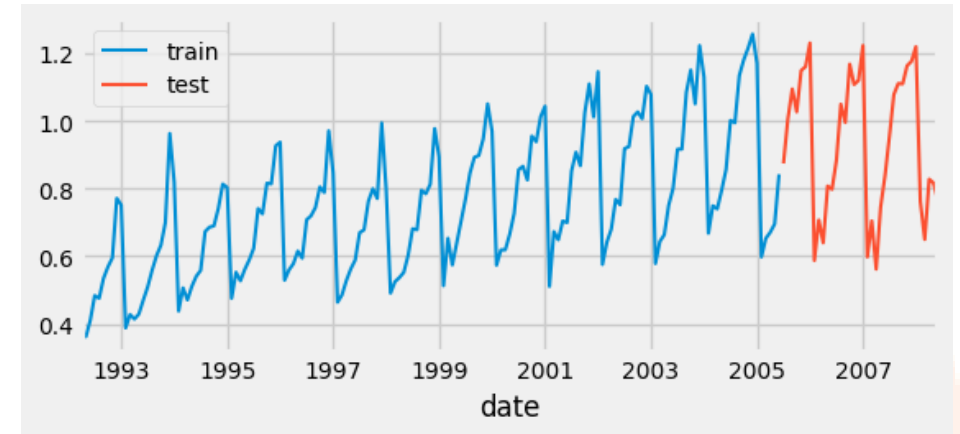
$$\mu(t) = \mu_0 + \gamma t + \sum(\alpha_i \mu_i + \beta_i t_i)$$





Day 3. Forecasting models

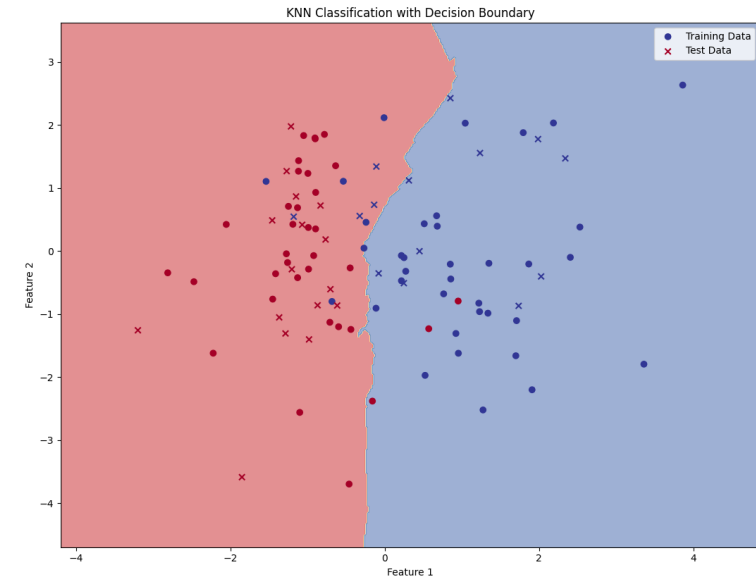
- In data science, *forecasting* usually refers to time series forecasting.
- Similar to regression models, but for long enough time series and if prediction within a horizon is desirable.
- Models: Prophet, ARIMA...





Day 4. Classification models

- Let us assume we have spectral NIR data for olive samples. We want to know if pesticides are present (1) or not (0). Independent lab analysis measuring the pesticide concentration are available that can be compared to the NIR spectra.
- A classification model can be used:
 - SVM, k-nearest neighbours, logistic regression, neural network...



		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

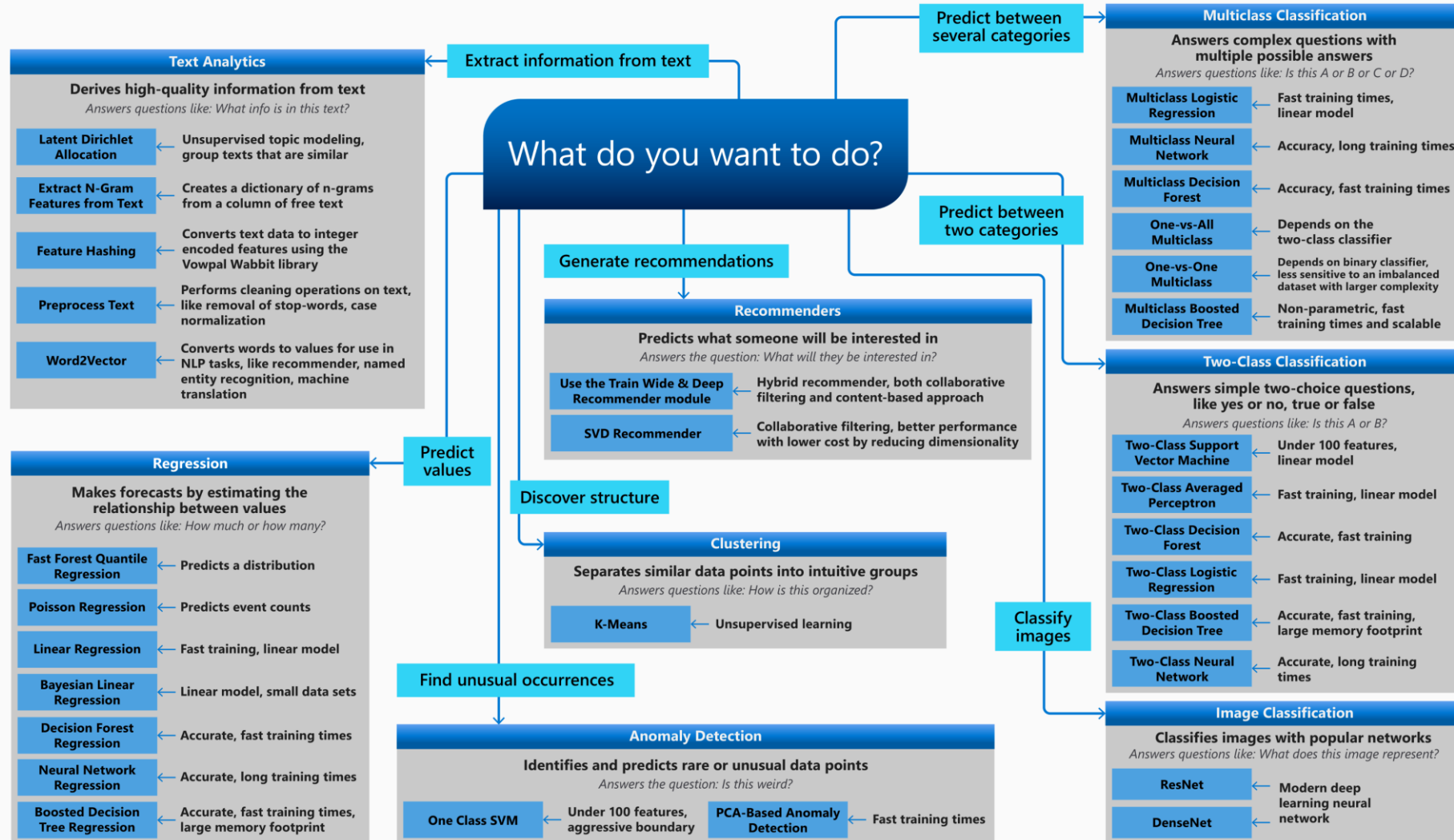


Cheat sheet



Machine Learning Algorithm Cheat Sheet

This cheat sheet helps you choose the best machine learning algorithm for your predictive analytics solution. Your decision is driven by both the nature of your data and the goal you want to achieve with your data.





Day 5. Reports

- Communicating your results is imperative.
- Generate a Jira report.
- Create notebooks, plots, documents useful for understanding the analysis.
 - Just like research.
- Comment and document every code that goes into production.

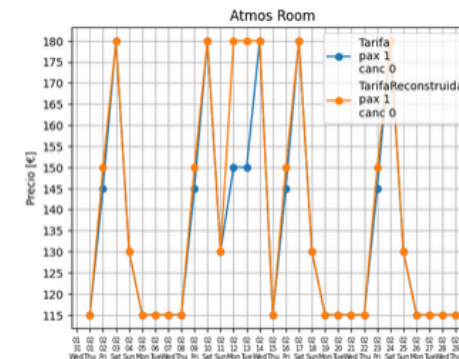
🔍 ATM-9 / 📄 ATM-12

DF

Add a comment...

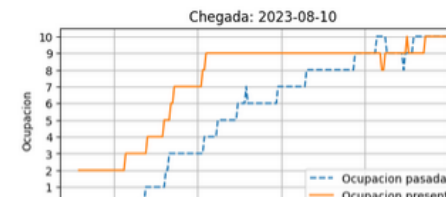
Pro tip: press **M** to comment

Con estas reglas podemos reconstruir el precio. No de manera exacta, ya que no es el objetivo, pero sí lo suficientemente bien como para poder saber cómo van a cambiar los precios en función de la ocupación.



Construyamos ahora las curvas de ocupación proyectada para una fecha concreta en función de la fecha de corte. La diferencia entre la ocupación proyectada para la fecha de llegada y la ocupación proyectada pasada para la fecha equivalente nos da una guía para subir o bajar el precio. Supongamos que por cada habitación de diferencia subimos o bajamos el precio un 5% y pongamos la variación de precio en el gráfico:

📄 Gráfico de ocupaciones proyectadas vs factor corrector de precio





Day 6. Database

- Monday, again.
- The data engineer is on vacation and I don't have the new dataset. Let us pull it from the database.
 - SQL, MongoDB.
- We need to know how to extract the data and cast them into a useful file format (e.g. CSV, parquet, HDF5)
- Good command of queries. Select, join tables, filter data, etc.

The screenshot shows a database query interface. At the top, there are tabs for "Query" and "Query History". The "Query" tab is active, displaying a SQL query:

```
1 SELECT film_id, title, release_year, rating FROM film
2 WHERE title ILIKE 'A%'
```

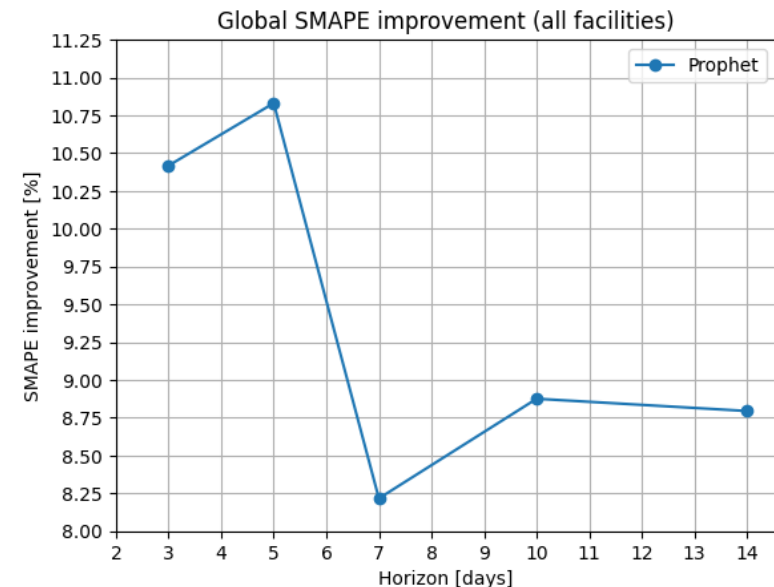
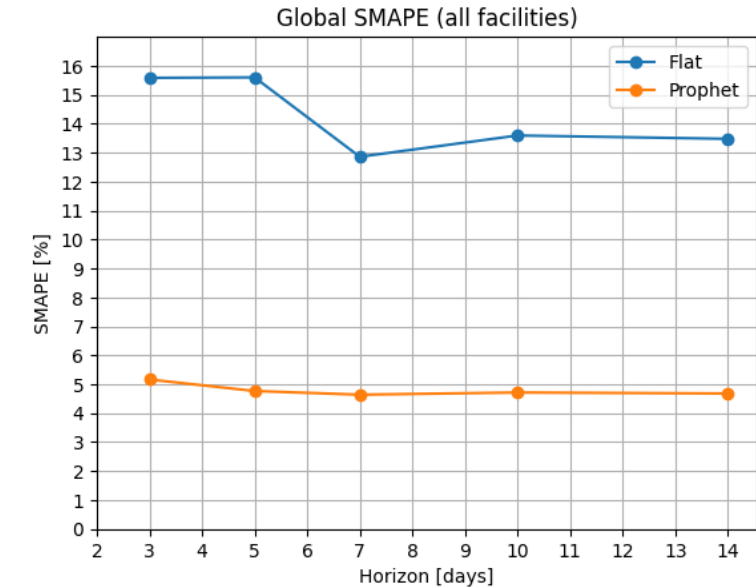
Below the query, there are tabs for "Data Output", "Messages", and "Notifications". The "Data Output" tab is active, showing a table of results. The table has five columns: "film_id", "title", "release_year", and "rating". The "film_id" column is marked as a primary key (PK) and integer. The "title" column is marked as character varying (255). The "release_year" column is marked as integer. The "rating" column is marked as mpaa_rating. The table contains 13 rows of data, with the first row being "Airport Pollock" (2006, R) and the last row being "Ali Forever" (2006, PG).

	film_id [PK] integer	title character varying (255)	release_year integer	rating mpaa_rating
1	8	Airport Pollock	2006	R
2	1	Academy Dinosaur	2006	PG
3	2	Ace Goldfinger	2006	G
4	3	Adaptation Holes	2006	NC-17
5	4	Affair Prejudice	2006	G
6	5	African Egg	2006	G
7	6	Agent Truman	2006	PG
8	7	Airplane Sierra	2006	PG-13
9	9	Alabama Devil	2006	PG-13
10	10	Aladdin Calendar	2006	NC-17
11	11	Alamo Videotape	2006	G
12	12	Alaska Phantom	2006	PG
13	13	Ali Forever	2006	PG



Day 7. Presentation with the customer

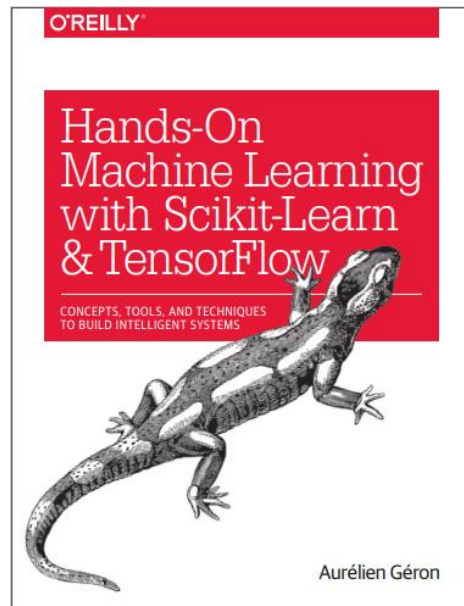
- We need to present the customer our results.
- Try to speak always with vocabulary close to the customer's.
- Keep in mind the customer's goal and organise your presentation accordingly.
 - Don't lose the 'commercial' mindset but don't sacrifice rigour.
- Try to show insightful figures of merit.
- Allow the customer to interrupt and ask questions. Jot their suggestions down for new tasks.













Day 8. Formation

- Keeping up to date is important. New libraries, new technologies.
- A problem is going to require the use of tools that you do not master yet. Almost guaranteed to happen.
- [Udemy](#), [Coursera](#), books ([O'Reilly](#))
- Not only data science or data engineering. Frontend, backend, html, css, Azure, AWS, Kubernetes, DevOps...



 <p>The Full Stack Web Development Eduonix Learning Solutions, 1+ Million Students Worldwide 200... 93% Complete ★★★★★ Leave a rating</p>	 <p>Complete FrontEnd Web Development and Design HTML CSS JS Laurence Svekis, Instructor, GDE, Application Developer 100% Complete ★★★★★ Leave a rating</p>	 <p>SQL - MySQL for Data Analytics and Business Intelligence 365 Careers, Creating opportunities for Business &... 80% Complete ★★★★★ Leave a rating</p>	 <p>Scrum Certification 2022 + Scrum Master+ Agile Scrum... Paul Ashun, CSM, Author-Scrum Productivity Books, CEO Pashun... 22% Complete ★★★★★ Leave a rating</p>
 <p>Advanced CSS and Sass: Flexbox, Grid, Animations and... Jonas Schmedtmann, Web Developer, Designer, and Teacher Complete ★★★★★ Leave a rating</p>	 <p>R Programming A-Z™: R For Data Science With Real... Kirill Eremenko, Data Scientist 2% Complete ★★★★★ Leave a rating</p>	 <p>Tableau for Beginners: Get DA Certified, Grow Your Career Lukas Halim, Analytics Professional 44% Complete ★★★★★ Leave a rating</p>	 <p>Tableau 20 Advanced Training: Master Tableau in Data Science Kirill Eremenko, Data Scientist 28% Complete ★★★★★ Leave a rating</p>



Day 9. Research and refinement

- Research. Well, what can I tell you?
- Looking for academic papers that help us understand the problem at hand. Grasping important variables, learning relevant models, etc.
- Examples: pectine esterification degree. Temperature in a cereal silo.

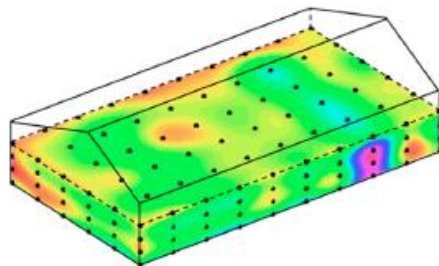


Figure 1. A 3D grain temperature field in a granary. (Note: the cloud is the grain temperature field, and the black dots are the sensor locations).

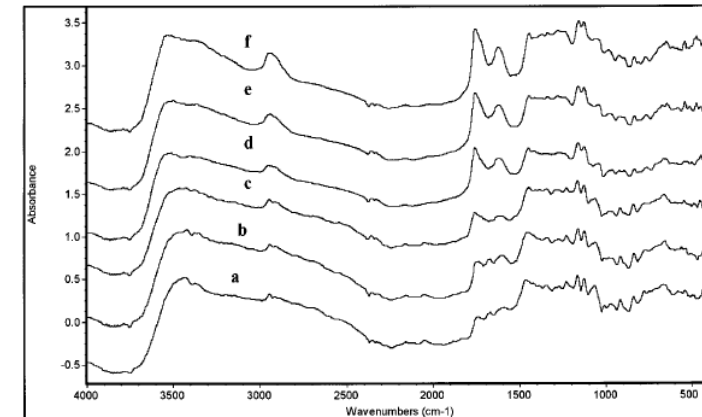
Improved Method for Determination of Pectin Degree of Esterification by Diffuse Reflectance Fourier Transform Infrared Spectroscopy

M. A. Monsoor, U. Kalapathy, and A. Proctor*

Department of Food Science, University of Arkansas, 2650 North Young Avenue, Fayetteville, Arkansas 72704

An improved method for the determination of pectin degree of esterification (DE) by diffuse reflectance Fourier transform infrared spectroscopy (DRIFTS) was developed. Pectin samples with a range of DE as determined by gas chromatography were used for developing a calibration curve by DRIFTS. A linear relationship between the DE of pectin standards and FTIR peak ratio for ester carboxyl peak area to total carboxyl peak area was found ($R^2 = 0.97$). Pectin DE of various samples was calculated from the linear fit equation developed by DRIFTS. Accuracy of the DRIFTS method was determined by comparing the DE values of four commercial pectins obtained by DRIFTS methods to the values obtained by the gas chromatography method. Greater precision was obtained for the FTIR measurement of test pectin samples when the ester peak ratio was used relative to the ester peak area.

Keywords: Pectin; degree of esterification; DRIFTS; FTIR



Article

Modeling of a 3D Temperature Field by Integrating a Physics-Specific Model and Spatiotemporal Stochastic Processes

Di Wang and Xi Zhang *^{id}



Day 9. Research and refinement

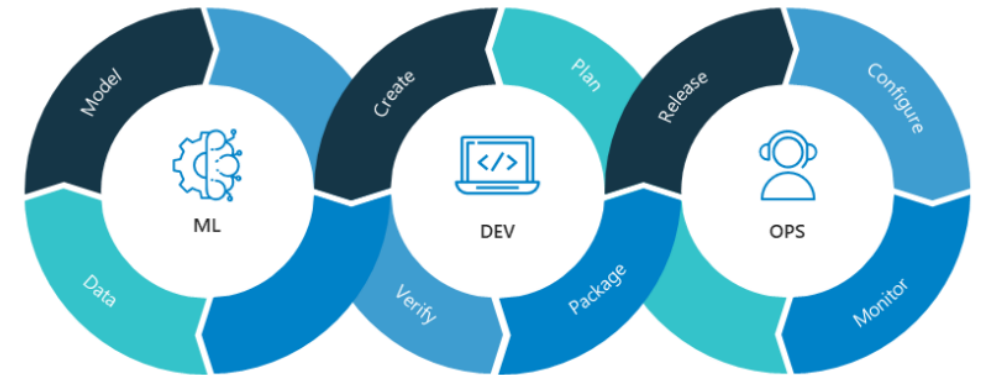
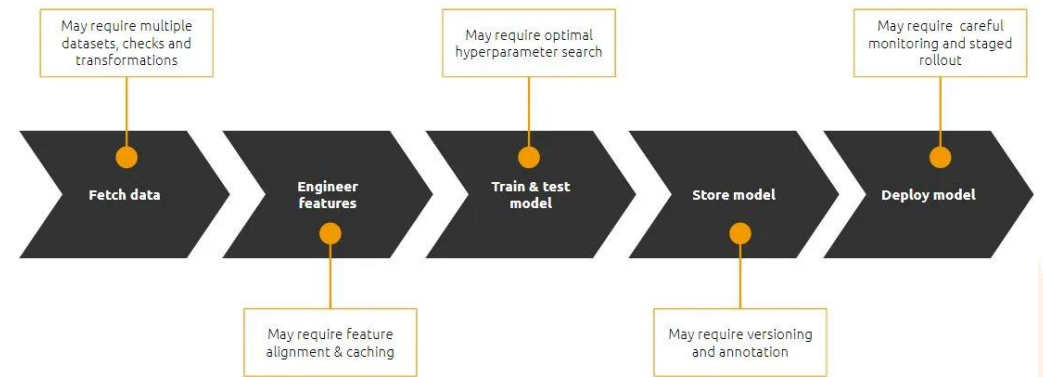
- The product owner or CTO creates tasks.
- It is necessary to clearly define the means and objectives for these tasks.
- The workload is also estimated.
- Refinement meeting.
- Tasks are organised into milestones, epics, stories (Jira nomenclature).

<input checked="" type="checkbox"/> ALPHAMILK-76 Revisión de la recopilación de datos de Ligal y CEGACOL en Granxa Pazos	AÑADIR FUENTES AUTOM	IN PROGRESS
<input checked="" type="checkbox"/> ALPHAMILK-82 Instalación de los tests automáticos de los Excel económicos en un contenedor	AÑADIR DATOS ECONÓMI	IN PROGRESS
<input checked="" type="checkbox"/> ALPHAMILK-80 Descarga automática de los datos del OVGAN	AÑADIR FUENTES AUTOM	IN PROGRESS
<input type="checkbox"/> <input checked="" type="checkbox"/> ALPHAMILK-83 Instalación del código de transferencia de datos económicos a la aplicación en un contenedor	AÑADIR DATOS ECONÓMI	IN PROGRESS
<input checked="" type="checkbox"/> ALPHAMILK-103 Revisión y validación de la creación de todos los KPIs	CREACIÓN DE LA VERSIÓN	TO DEPLOY
<input checked="" type="checkbox"/> ALPHAMILK-96 Análisis de la integración de los datos de Ligal, CEGACOL y OVGAN en la aplicación	AÑADIR FUENTES AUTOM	READY TO START



Day 10. Deploy, MLOps

- Not usually the job of a data scientist, however:
- Once the model is created and validated by the customer, it needs to be deployed so that the customer can use it.
- Deploying consists in integrating the new code into an app, program, webapp, database, or similar, depending on the problem and the customer's needs.
- MLOps – techniques for deploying and maintaining machine learning models in a consistent and robust way.
 - Reproducibility, version control, retraining, metrics history, unit tests...
- The more versatile you are, the better for your career.
- Sprint review and end!





Why data science as a career path?

- On high demand right now.
- Really advantageous working conditions. Remote working, schedule flexibility (except for meetings).
- People with experience handling data are needed.
- People having academic backgrounds are really appreciated. They know how to wrestle complicated data, after all.
- Entry salaries for doctors around 30k (before tax) in Spain. Earning much more is possible.
- Careful with hype and misdirections. Remember that a model is only as good as the data allows.
- Personal advice: a larger company does not mean a better company (for **your** needs).





- Didactic introductory tutorials: cienciadedatos.net (in Spanish), [towardsdatascience](http://towardsdatascience.com).
- Online courses. Data science, machine learning, or any other IT topic (paying).
 - [Udemy](https://www.udemy.com/)
 - [Coursera](https://www.coursera.com/)
- Contest page to extract ideas from: [Kaggle](https://www.kaggle.com/)
- [OpenML](https://www.openml.org/), datasets ready to be used for practice.
- Practice, practice and practice.





TRIPLE**ALPHA**
innovation