

Data Science Project III

14021 – PRYIII – "Proyecto III, análisis de datos"

Project Offers 2025

In the following pages, there's a list of project offers. They're meant for a team of around six students each.

Note that there are four class groups:

- 3CDA1 Mon 9:00-11:00 LAB 1E 1.0 (von Neumann), Fri 11:30-13:30: LAB DSIC 9
- 3CDA2 Mon 12:30-14:30 LAB 1E 1.0 (von Neumann), Fri 9:30-11:30: LAB DSIC 9
- 3CDB1 Mon 16:00-18:00 LAB 1E 2.1 (Boole), Fri 17-19 - LAB 1E 1.0 (von Neumann)
- 3CDB2 Mon 18:00-20:00 LAB 1E 1.0 (von Neumann), Fri 15-17 - LAB 1E 1.0 (von Neumann)

Projects offered by Joaquín are preferably for group 3CDA1, those offered by Alberto preferably for group 3CDA2 and those for group 3CDB1/3CDB2 for Jose. Projects about challenges, those in coordination with courses and digital twins could be an option for all groups, and are listed at the end.

Teams with members of different groups are **not** allowed.

When choosing your preferences in the questionnaire, use the shortcut of the project, e.g., "MEDIMG".

Project Offers – Joaquín - 3CDA1

CLIMACT: Predicting the Future of Biodiversity under Climate Change Scenarios

Data description: The world is changing, and species must adapt or face extinction. How can we predict their survival under future climates? Here a rich geospatial dataset of 301 populations of *Arabidopsis thaliana* across the Iberian Peninsula. The data combines genetic diversity (cluster memberships) with high-resolution bioclimatic variables, such as temperature and precipitation patterns, all georeferenced to specific locations.

Main source of data: Available at Zenodo (<https://zenodo.org/records/2552025>)

Possible value: Climate change threatens ecosystems worldwide. Predicting how species will respond can guide conservation efforts and policy-making. In this project, you will model range shifts under future climate scenarios (e.g., RCP 2.6 and RCP 8.5 for 2070), helping identify areas at risk and strategies for biodiversity preservation.

Innovation: Explore the use of different modeling approaches to predict species distribution under climate change. Whether it's Bayesian models, machine learning, or other statistical tools, students can choose their preferred methods to analyze genetic, climatic, and geospatial data and evaluate their performance.

Related data to add value: Incorporate additional data from climate projections (<https://www.worldclim.org/data/index.html>) and ecological insights to enhance predictions. Explore how global trends and local factors interact to shape species survival in a rapidly warming world.

TRACKBIO: Transforming Species Distribution Models with GPS Movement Data

(Collaboration with the Instituto de Ciencias del Mar)

Data description: Traditional species distribution models (SDMs) rely on static presence-absence data, but GPS tracking provides a dynamic perspective on habitat use. This project explores movement data from birds and mammals in Antarctica, recorded with multiple tracking technologies, to investigate how species interact with their environment.

Main source of data: Public tracking data from various Antarctic species (<https://doi.org/10.15468/uzmdpm>), with detailed dataset descriptions and case studies available (<https://doi.org/10.1038/s41597-020-0406-x>, <https://doi.org/10.1038/s41586-020-2126-y>).

Possible value: Can movement data improve species distribution models? Unlike traditional SDMs, tracking data introduces spatial and temporal dependencies that must be carefully handled. The main challenge here is to rethink how movement patterns can refine habitat suitability models and conservation strategies.

Innovation: The integration of tracking data into SDMs is an emerging and largely unexplored field. Students will tackle complex challenges in ecological modeling, testing novel ways to account for movement and temporal bias in species predictions. This work could lead to more accurate, data-driven conservation insights that go beyond standard ecological modeling.

Related data to add value: Exploring different species and tracking technologies can provide insights into habitat use. Combining spatial modeling techniques with machine learning approaches like Random Forests or Neural Networks can enhance species distribution models by capturing both environmental drivers and movement dynamics from GPS tracking data.

PREGNANCY: Do Prenatal Pollutants Shape Pubertal Development?

(Collaboration with the Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunitat Valenciana (FISABIO))

Data description: Could exposure to environmental pollutants during pregnancy influence when puberty begins? This project explores the impact of prenatal exposure to perfluoroalkyl substances and organochlorines on pubertal milestones, such as age at menarche and Tanner stage. Instead of analyzing pollutants in isolation, students will investigate whether they amplify, counteract, or interact unpredictably to shape hormonal development.

Main source of data: A dataset measuring prenatal exposure to exposure to perfluoroalkyl and organochlorines, alongside key maternal and child variables (BMI, smoking, preterm birth, child's sex, maternal age at menarche, socioeconomic status, and more). Organochlorines must be adjusted for lipid levels due to their chemical properties.

Possible value: Are we underestimating the effects of pollutants by studying them separately? This project challenges students to apply multivariate techniques such as Principal Component Analysis (PCA), Principal Component Regression (PCR), or Partial Least Squares (PLS) to uncover hidden patterns in exposure data. By analyzing all pollutants together, they will assess whether contaminants reinforce or cancel each other out, providing a clearer picture of environmental risks affecting pubertal development.

Innovation: to be found!

Related data to add value: Incorporating additional environmental exposure data can refine predictions. Analyzing **individual vs. combined pollutants** will determine if interactions matter more than previously thought.

AIRPOLLUTION: Predicting Pollution in the Valencian Community

Data description: Air pollution varies across time and space, affecting environmental and public health. This challenge focuses on modelling and predicting pollutant concentrations in the Valencian Community using spatiotemporal techniques. The dataset includes historical measurements of NO₂, PM_{2.5}, PM₁₀, and O₃, along with meteorological and geographic data.

Main source of data: Air quality datasets for the Valencian Community, available for download ([Link to the data](#)), containing pollutant levels recorded at multiple monitoring stations with timestamps and spatial coordinates.

Possible value: Can we accurately predict pollution levels across different areas and time periods? This project explores spatial and temporal dependencies, offering insights for forecasting air quality and guiding environmental policies.

Innovation: You could explore spatiotemporal modeling techniques to predict pollution levels, applying geostatistical methods (Kriging), time series forecasting (ARIMA, Prophet), and machine learning approaches (Random Forest, Gradient Boosting). Evaluating the strengths of these techniques will help determine the most effective approach for air quality prediction.

Related data to add value:

<https://doi.org/10.1016/j.atmosenv.2021.118192>,

<https://doi.org/10.1002/env.2723>.

EMERGENCY: Can Data-Driven Strategies Save Lives?

(Collaboration with SOA group ITI-UPV)

Data description: In emergency care, every second counts. This challenge focuses on analyzing ambulance response times using geospatial data from 77 areas in Valencia, including ambulance bases, hospitals, emergency locations, travel times, etc.. Emergency calls are synthetically generated with Poisson-distributed timestamps, incorporating dynamic demand, travel constraints, and resource allocation trade-offs. Various emergency management strategies have been tested, considering different starting points for ambulances and the impact of traffic restrictions on response efficiency.

Main source of data: Synthetic data were generated, including ambulance and hospital locations, travel distances and times, and 600 emergency scenarios testing different emergency response strategies.

Possible value: Are ambulances reaching emergencies as fast as possible? The main challenge is to analyze how different management strategies affect response times. Does starting from different ambulance bases improve arrival times? How do on-site care duration and hospital delays impact overall efficiency? Can a data-driven approach reveal inefficiencies and suggest improvements?

Innovation: Exploring machine learning for response time prediction or measurement agreement techniques of different emergency strategies could provide new insights into improving response times.

Related data to add value: Check papers derived from the original project:

<https://doi.org/10.1016/j.eswa.2022.118773>

<https://riunet.upv.es/handle/10251/207661>

MARKETDYNAMICS: Understanding Brand Competition through Market Shares

Data description: Market shares are key for companies like Amazon and Google, as well as major retail and food brands, to analyze competition and optimize strategies. Explore weekly market share data from 52 brands across four supermarkets in India, covering sales, product variants, packaging sizes, and distribution metrics.

Main source of data: Real-world retail data containing the variables mentioned above, providing a realistic scenario for analysis. The dataset includes 52 brands sold across four major supermarkets, with variables such as: Date (Weekly observations spanning three years), Brand (Identification of 52 competing brands), Variant (Different product types), Pack size: Product packaging sizes (e.g., 0-350g, 351-500g), Value.sales, etc.

Possible value: How do pricing strategies, distribution networks, or product variations impact a brand's market share? What happens when a leading brand raises its prices, or a new competitor enters the market?

Innovation: The combination of compositional data analysis with time series modeling is an emerging field. Students can explore ARIMA, ARIMAX, or Prophet for trend detection and apply machine learning techniques to analyze brand interactions and predict the impact of pricing, distribution, or promotions.

Related data to add value: time series techniques, multivariate analysis and Compositional data analysis.

Project Offers – Alberto - 3CDAB

UPV-EARTH: Contribution of UPV Scientific publications in connection to Planetary Boundaries.

Data description: A dataset consisting of **50,000 abstracts** of scientific contributions from the university (Universitat Politècnica de València)

Main source of data:

This data has been downloaded from Scopus and OpenAlex. We would like to estimate how we can identify the scientific contribution of our university to the achievement of the Planetary Boundaries

Possible value: Planetary boundaries is a framework outlines nine key processes that regulate the stability and resilience of the Earth system. The situation is dramatic and many of them have been surpassed. Beyond the contribution to the Sustainable Development Goals, what can we say about the impact of the scientific contribution of the UPV to this agenda. A similar work on Sustainable Development Goals can be found here <https://pubs.acs.org/doi/10.1021/acssusresmgmt.4c00074>

Innovation: Use of a NLP approach in order to examine the texts, including LDA, BERT, Top2VEC and the use of LLM's

Related data to add value: Incorporating information from World Bank at country level can also improve the impact of the analysis.

MATH-VISION: Measuring Multimodal Mathematical Reasoning with the MATH-Vision Dataset

Data description: A dataset consisting of questions from the Kangaroo contest (Proves CANGUR a la Comunitat Valenciana).

Main source of data: Questions posed in the context in the last years. <https://huggingface.co/datasets/MathLLMs/MathVision> and questions from the Cangur Matemàtic.

Possible value: Generative models are resembling the capabilities of humans for reasoning. LLMs can proceed with text, but when images appear the situation is more complex. Some mathematical problems can also be solved with text, but in other situations the use of images is required to completely explain the project.

Innovation: Use multimodal Generative AI models to understand its performance and its limitations.

Related data to add value: It can help to promote the participation of students in the Proves Cangur. <https://neurips.cc/virtual/2024/poster/97639>

STOCKMARKET: Volatility of Investment Portfolios

Data description: Deviations from Brownian motion leading to anomalous diffusion are found in transport dynamics from quantum physics to life sciences. The characterization of anomalous diffusion from the measurement of an individual trajectory is a challenging task, which traditionally relies on calculating the trajectory mean squared displacement. We already know that volatility can be inferred from synthetic data and extracted through the use of machine learning methods

Main source of data: Data will be provided. It is described in [Stock volatility as an anomalous diffusion process](#)

Possible value: Set a new approach for controlling the volatility of an investment portfolio.

Innovation: Propose an alternative method for designing investment portfolios through machine learning.

Related data to add value: Use information of several stock markets such as Nasdaq, S&P 500, or Dow Jones.

SOCCKER-DIFFUSION: Volatility of Investment Portfolios

Data description: Changes in a player's movement are closely tied to the strategies and tactics within a soccer match, where encounters and interactions between players play a pivotal role. Examples include goal-scoring opportunities, defensive positioning, and precise passing to specific locations on the field.

Using real-time player tracking combined with motion analysis is a powerful approach to study these movement patterns. Recent advancements in tracking technology and data analysis have significantly improved methods for analyzing individual players' movements during a game.

This progress has led to the development of various techniques for accurately detecting changes in movement. These changes act as valuable indicators for key moments, such as player interactions, formations, and strategic adjustments. We want to develop a combination of physics and machine learning models in the line of the AnDi Challenge 2 <https://codalab.lisn.upsaclay.fr/competitions/16618> to analyze how can be described how soccer teams play through the use of diffusion measures.

Main source of data: Statsbomb data <https://github.com/statsbomb/open-data> Synthetic data from <https://codalab.lisn.upsaclay.fr/competitions/16618>

Possible value: Use this information for understanding the global performance of a team.

Related data to add value: All you can add after watching hundreds of hours of soccer in your life.

Project Offers – Jose - 3CDB1/3CDB2

CHARGING: Explaining Charging Curves Classifiers

Data description: Information about the charging curves of electric vehicles.

Main source of data: This dataset, <https://www.kaggle.com/datasets/michaelbryantds/electric-vehicle-charging-dataset>, sources from here <https://www.mdpi.com/1996-1073/14/8/2233> or elsewhere.

Possible value: many charging sessions fail or are suboptimal. Looking at the charging curves we can build models that anticipate these problems. We're interested in explaining these models to users, so they can understand why and when their car is not charging well.

Innovation: We suggest to use rule extraction techniques or fuzzy rules/logic, or the use of inference engines, seen in the course RCR, or instance-based explanation (seen in EDM).

Related data to add value: Related data: calendar, weather, prices, ... 8

RESC-AI-LING: New scaling laws for AI “reasoning” and distilled models

Data description: Analyse the data of progress of large language models (LLMs) according to some other parameters that are not only #parameter, compute and data size, especially reasoning LLMs, including time to think, distillation, etc. We can see how benchmark results are compared as in <https://github.com/deepseek-ai/DeepSeek-R1>, <https://openai.com/index/openai-o1-mini-advancing-cost-efficient-reasoning/> and many other papers and comparisons.

Main source of data: [https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard#/,](https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard#/) and many other papers about performance, e.g., epoch datasets: <https://epoch.ai/data>

Possible value: anticipate the evolution of LLMs with some other features beyond #parameters, especially seeing how very small models (<https://arxiv.org/pdf/2501.05465>) and distillation (<https://github.com/deepseek-ai/DeepSeek-R1>) are achieving spectacular results with much less compute.

Innovation: to be found!

Related data to add value: Papers about scaling laws. 8

LASTEXAM: Analysing the particularities of the tasks in Humanity's Last Exam

Data description: Humanity's Last Exam (<https://agi.safe.ai/>) is a collection of very challenging tasks to evaluate the AI of today and in the future.

Main source of data: the benchmark itself

Possible value: what can we say about the questions in that dataset in terms of diversity, complexity of the language, similar questions found online, etc., beyond the statistical analysis of domain provided?

Innovation: to be found! But this is a rapidly changing area, and not many people have analysed the characteristics of the dataset.

Related data to add value: Other benchmarks used for AI, such as MMLU-Pro or GPQA or MATH-500.

CARSHARE: Adaptive tariffs

Data description: Data from car sharing companies, especially if only or mostly about electric cars. For instance, one example of this is <https://www.alterna.coop/>, having a fleet of Renault cars (Zoe) and vans. There are other open sources, some about simulated data (<https://www.kaggle.com/datasets/mexwell/car-sharing-simulation>) but we could try to put the team in contact with alterna.coop.

Possible value: Make discounts when the cars are rented in low-consumption periods, and also given back when electricity is cheaper. Also consider users that can charge them in places with free charge, and can be compensated if they return it with more battery level.

Innovation: Possibly the combination of use patterns and electricity prices for charging is an underexplored combination, also in car sharing services with relatively low demand such as this one..

Related data to add value: data from some other sources or synthetic data.

6

TURINTGTEST: Imitating Diversity

(this project has strong ethical factors, perhaps collaborating with <https://www.lcfi.ac.uk/>)

Data description: The imitation game, introduced by Alan Turing in 1950, was originally inspired by a Victorian game to tell men from women, with a woman pretending to be a man, or otherwise. Then this was replaced by human vs machine. From the context of Alan Turing's being homosexual and beyond the obsolete perspective of "binarism" today, this project will collect and create data about people (whatever their sexual orientation, gender, sex, etc.), of actual conversations and pretended conversations, jointly with "personas" and "impersonations" using large language models.

Possible value: In the 75th anniversary of the Imitation Game, explore gender roles by language models, and the very concept of what means to be human in the 21st century.

Innovation: Can we update the imitation game to the 21st century? Can we explore diversity in the context of the Turing test using data analysis of conversations from humans and machines, and analysed by humans and machines? Can we identify (false) stereotypes of the most masculine, most feminine, most human, most agreeable, smartest, etc., chatbot?

Related data to add value: In the early stages of the project the team can devise a Turing Test extended for gender diversity and large language models. Datasets about the Turing test, Loebner's prize, etc. Turing's 1950 paper: <https://courses.cs.umbc.edu/471/papers/turing.pdf> and general info about the Turing Test: <https://plato.stanford.edu/entries/turing-test/>


2

Project Offers – All groups

RGB: True Colours

(provided by Samuel Morillas)

Data description: Characterising displays of technologies LCD, OLED, and QLED to achieve accurate color reproduction can be based on input (device-dependent RGB data) and output (device-independent XYZ coordinates) data obtained from three different displays. Training and test datasets are built using \$RGB\$ data measured directly from the displays and corresponding \$XYZ\$ coordinates measured with a high-precision colorimeter.

Main source of data: See the following article for a better description of problem and data: 

[Extended IbPRIA 2023 paper Displays .pdf](#)

Possible value: A key aspect of this research is to achieve good performance in color reproduction, but also providing physical insights into the relationships between the \$RGB\$ inputs and the resulting \$XYZ\$ outputs. This interpretability allows for a deeper understanding of the display's behavior.

Innovation: the approach could reuse ideas from the above paper but should try to add some innovative approaches in the transformation or modelling of the data.

Related data to add value: data from some other sources or synthetic data.

2

ENERGY: Household Energy Profiles

Data description: The behaviour of a household is captured by its energy use profile.

Main source of data: Consumption in a household and activity recognition. Several sources. Possible data from https://data.open-power-system-data.org/household_data/ (or many other sources with similar data) or the option of local data from some Alternacoop's customers (we will make the contact with the company).

Possible value: consumption estimate in a household and activity recognition, recommend when to switch on/off appliances.

Innovation: the tricky part is activity recognition, as in many houses we don't have labelled information about the appliances that are on, but this can be inferred from the electricity consumption. Of course this has been done to an extent, so an innovative angle has to be found!

Related data to add value: Sphere project. - smart meter data. Related data: calendar, weather, prices, ...

6

MBA: Financial Times MBA Rankings

(provided by ESADE)

Data description: ESADE Business School's Full-Time MBA has established itself as one of the top global options, reaching 17th place worldwide and 6th in Europe according to 2024 Financial Times ranking. However, the exact methodology used by Financial Times to rank business schools is not entirely transparent and is based on criteria that may evolve over time, leading to fluctuations, sometimes abrupt, in the positioning of the different institutions.

Main source of data: Data collection and modeling from Financial Times, ESADE, and its main competitors, using public information, academic databases, and scientific outputs.

Possible value: This project aims to achieve a deep and quantifiable understanding of the factors that determine ESADE's position and that of other business schools in this ranking. To this end, a team of data science researchers will analyze the historical evolution of the key variables impacting rankings, identifying which have been most significant in the past and which are expected to gain relevance in the coming years. Additionally, the study will explore the potential emergence of new evaluation criteria, as anticipating such variables would provide ESADE with a significant strategic advantage. Data-driven strategic recommendations to ensure ESADE remains in the top 20 globally, maximize the chances of rising to top 15 by 2026, and lay the groundwork for reaching top 10 before 2030.

Innovation: Identify patterns in the evolution of ranking factors, and the relative importance of each variable in the coming years, anticipating potential changes in the Financial Times ranking methodology.

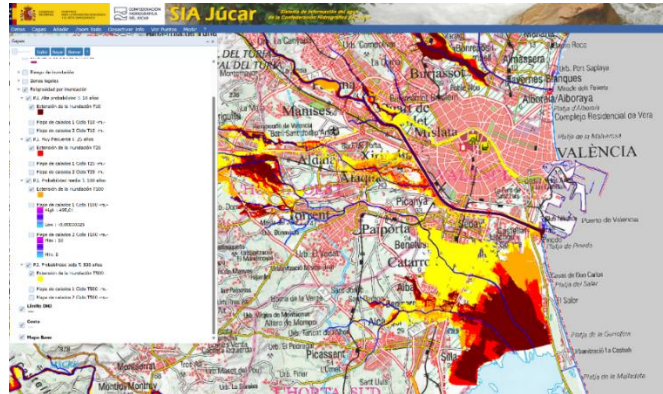
Related data to add value: data from some other sources or synthetic data.

Note: The selected research team will be invited to ESADE Barcelona to present their findings and strategic recommendations to the ESADE Dean and the MBA leadership team, offering a unique opportunity to directly influence the strategy of one of the world's top business schools.

SAVE-U-CAR: Flood-aware safe route and parking

(possible collaboration with the Aula Renault – AI and mobility)

Data description: we have access to floodable areas and <https://www.openstreetmap.org/> for roads and the SIG from the Confederación Hidrográfica del Júcar (<https://aps.chj.es/siajucar/>) showing floodable areas (“riesgo de inundación fluvial por avenidas”) with return periods of 10, 100 and 500 years: “Añadir” capa > Buscar “Estudio Cartográfico de zonas inundables (SNCZI) > Peligrosidad por inundación”, jointly with forecast information (<https://open-meteo.com/>, or any other free weather forecast source for non-commercial purposes).



Possible value: Calculate and evaluate the risk of different routes from departure to destination given a time of the day and the weather forecast. Ideally, suggest alternative routes (at least broadly). Also consider how long the car is going to be parked at the destination, and ensure that the car can be picked up after that time.

Innovation: The use of weather and routes is used by many driving apps, but not the information about floodable areas. Also, many do not care what is going to happen hours after the trip, when the car is parked.

Related data to add value: data from some other sources or synthetic data. Possibly this: <https://dana2024.upvusig.car.upv.es/>

4

CARALARM: Anticipatory alarms for abandoning a vehicle

(possible collaboration with the Aula Renault – AI and mobility)

Data description: this requires exploration to find datasets with recorded data of the car, including sensors, GPS, LIDAR, RADAR, cameras outside or inside the car, etc. For instance, this is one for trucks: <https://www.man.eu/truckscenes>, and then combine it with information of floodable areas and weather forecast (see description in the proposal of the save-u-car project). Other platforms that collect this data are: <https://openxcplatform.com/>, <https://www.openvehicles.com/>, etc., and there may be personal data available for analysis.

Possible value: Alert the owners/users not to come to rescue the car, when the sensors, GPS, etc, has some information that the car may be in a flooded area, surrounded by water or with risk of becoming so. Similarly, there should be another alert to recommend the users leave the car, with several levels: go back, find a place to park and leave the car, stop and leave the car immediately, etc.

Innovation: There are many coded alerts in assisted driving components of cars and apps, but none specifically to deal with floodings.

Related data to add value: data from some other sources or synthetic data. Possibly this: <https://dana2024.upvusig.car.upv.es/>

5

DELIVETTER: Deliver better by using simulated data

(possible collaboration with the Aula Renault – AI and mobility)

Data description: there is simulated data about vehicle trip data in <https://www.nature.com/articles/s41597-023-01997-4>. This data is simulated and it can be used as is, or modified according to the space.

Possible value: Analyse possible interventions and hybrid designs in last-mile delivery (e.g., a van with small robots inside) that could be simulated to see if there are more effective delivery logistics.

Innovation: this is a very open project in terms of the source of data, use of simulation and possible ideas to improve delivery. There is a huge amount of literature and existing ideas, but we suggest to compare with some simulation approaches for more sustainability in city delivery (e.g., <https://www.sciencedirect.com/science/article/pii/S2192437624000050>), and propose some innovative solutions.

Related data to add value: data from some other sources or synthetic data.

5

SEXISM: Detecting sexism in social channels

(provided by and in collaboration UNICC: <https://www.unicc.org/who-we-are/history/>).

The project continues and extends previous capstone projects aiming to identify, prevent and better respond to sexual violence against women and girls in the digital space.

Develop approaches to build safe digital spaces for women and girls

Data description: UNICC is an IT Service Provider to the UN and provides many different services to different organizations and agencies. There is data from research, interview and highlight key commitments from the private sector, universities, students and social work, local governments, and women's rights organizations recognising the problem.

Main source of data:

- There is data data collected and used in previous project.

Possible value:

Currently there are ways/features in social media platforms to hide sensitive content with a focus on hate speech and fake news, but content that promotes violence against women is not flagged or identified, especially in Spanish, and considering all its varieties (e.g., Mexican Spanish).

Innovation: It will use modelling and ML/NLP techniques and possibly language models to propose innovative AI driven approaches to flag content, to identify harassment and other forms of sexual violence in social media platforms experienced by women and girls. From the extracted knowledge or the built models, include actions that social media platforms can take that may contribute to the fight against violence in the digital space.

Related data to add value: <https://research.ibm.com/blog/new-spanish-llm-ai>, <https://huggingface.co/spaces/mteb/leaderboard>, Data from other domains might be useful. The use of language models is more than likely. Additionally, for training models the students can use external sources such as SemEval 2023 shared task dataset [CodaLab - Competition \(upsaclay.fr\)](https://codalab.lisn.fr/competitions/semeval2023)

AUTOFORECAST: Intelligent Forecasting with Autoencoders

(provided by and in collaboration with Inditex)

Data description: public dataset: M5 Forecasting Dataset, which contains historical sales data at the store and product level (*M5 Forecasting Dataset: <https://www.kaggle.com/c/m5-forecasting-accuracy/data>*).

Possible value: The model will use historical data to identify complex patterns, detect anomalies, and minimize forecasting errors. The system should quickly identify trends and variations in demand, significantly improving the quality of predictions.

Innovation: it is suggested to develop a model based on autoencoders to predict product demand in a supply chain using the dataset, and compare with traditional models, with WMAPE being the main accuracy metric but also the KPIs established as a baseline. Because it is Inditex, it is of special interest to focus on clothes. One innovative part is to analyse whether the use of autoencoders significantly improves performance in time series prediction compared to traditional models used in dynamic environments.

Related data to add value: calendars, weather, macro-economic data, etc.

6.5

SOCIAL-LISTENING: What people really think about a company or a product?

(provided by Ernst and Young)

Data description: Social listening relies on diverse data sources to track conversations, analyze sentiment, and extract insights. Key sources include social media platforms like Twitter, Facebook, and Instagram, where real-time discussions occur. Blogs, news sites, and online forums such as Reddit and Quora provide in-depth opinions. E-commerce and review platforms like Amazon, Yelp, and Trustpilot offer direct customer feedback. Search engines and SEO tools track trending topics, while media monitoring covers brand mentions in the news. Privacy compliance and data quality are essential for effective analysis. Using APIs and advanced tools, businesses can leverage these sources to enhance strategy and decision-making.

Main source of data: Data will be extracted from different platforms and webs from internet.

Possible value: Social listening provides companies with real-time insights into customer opinions, industry trends, and competitor strategies. It enables brands to proactively manage their reputation, improve customer relationships, and drive business growth.

Innovation: It can transform a company's approach by using AI and machine learning to analyze vast data in real-time, uncovering emerging trends, automating sentiment analysis, and predicting customer behavior. This enables faster, more accurate decision-making, personalized engagement, and proactive problem-solving, driving business growth and competitive advantage.

Related data to add value: data from the company to be studied.

5

MEDIMG: A challenge on medical image analysis

Data description: Are you fan of challenges? You can register to one of the medical image analysis challenges currently active and use the data to build your project for this subject, but also to win a prize!!

Main source of data: <https://grand-challenge.org/challenges/>

Innovation: Develop a predictive model to solve the problem posed by the challenge in a way that hasn't been tried before, e.g., crossing with some other data.

Related data to add value: Similar studies available in the Internet.

4

TWINS: Smart cities – digital twins - simulation

Data description: The smart cities revolution has now incorporated to the concept of digital twin, an accurate model of the whole city (the digital twin) in which policy-makers can do simulations about changes or events at the city and a range of interventions.

Main source of data: smart cities (Valencia, Alcoi, any other), and simulated data: use models to generate data and simulate with the model and compare results. Cèsar Ferri has been collected data from Valencia Open Data: <https://github.com/ceferra/citybikes>, https://github.com/ceferra/estat_trafic_VLC, <https://github.com/ceferra/valenbici>, https://github.com/ceferra/espires_VLC, https://github.com/ceferra/espires_bici_VLC, https://github.com/ceferra/citybikes_prev (this has two years of data). Format is very similar as <https://github.com/ceferra/valenbici>, where there's a readme explaining it.

Possible value: Create a simulation of a city or another environment, evaluate possible policies (change of traffic regulations, opening times, etc.)

Innovation: to be found, but this is one of the few projects offers that includes simulation, which is still less common in data-driven projects today!

Related data to add value: urban mobility

7

SDG-TRADEOFFS: What is the connection between the progress of different Sustainable Development Goals Targets.

Data description: To analyze the progress towards the achievement of the 17 Sustainable Development Goals, there are targets and indicators. Targets are specific, actionable objectives designed to achieve the broader ambitions of each SDG. They describe the "what" that needs to be accomplished. There are 169 targets across the 17 SDGs. Indicators are measurable metrics used to track and evaluate progress toward each target. They answer the "how much" or "to what extent" progress has been made. There are **232 unique indicators** defined by the UN Statistical Commission.

Main source of data: SDG Tracker and World Bank from Our World in Data.

Possible value: Analyze potential synergies and trade-offs between different SDG's through the analysis of the progress of indicators and trade-offs.

Innovation: Use of statistical models to study correlations between different indicators.

Related data to add value: Existing documentation from the United Nations as the Voluntary Reviews or the Nationally Determined Contributions.

3

PARKINSONWATCH: Predicting Parkinson's Disease Using Smartwatch Data

Data description: This project analyzes smartwatch data from Parkinson's Disease (PD) patients, individuals with other movement disorders, and healthy controls. The dataset includes sensor-based movement data, clinical symptom reports, and patient metadata, collected in a supervised hospital setting. More details in the Parkinson's Disease Smartwatch (PADS) study: <https://www.nature.com/articles/s41531-023-00625-7>.

Main source of data: The **PADS (Parkinson's Disease Smartwatch) dataset**, collected over 3 years with more than 500 participants in a hospital setting. Dataset available ([PADS - Parkinsons Disease Smartwatch dataset](#)).

Possible value: Can machine learning models detect Parkinson's Disease using only smartwatch data? The findings could contribute to developing **remote health monitoring solutions** and **early intervention strategies**.

Innovation: to be found!

Related data to add value: it can be enhanced by exploring alternative modeling approaches for capturing temporal dependencies, or Bayesian models to estimate uncertainty in predictions.

2

PNEUMO: Can We Detect Pneumothorax from X-Rays?

Data description: Medical imaging is a crucial tool for diagnosing lung conditions. We focus on classifying chest X-ray images into healthy vs. pneumothorax cases, using a dataset that has been preprocessed to grayscale and standardized for consistency in analysis.

Main source of data: A dataset of chest X-ray images, preprocessed and categorized into healthy and pneumothorax cases. The dataset is publicly available and derived from a larger medical imaging repository (<https://nihcc.app.box.com/v/ChestXray-NIHCC/folder/36938765345>).

Possible value: Can medical images be classified efficiently using structured approaches? This project explores how feature extraction techniques (e.g., texture analysis, edge detection, statistical descriptors) combined with classification algorithms can distinguish between healthy and diseased lungs.

Innovation: To be found!

Related data to add value: Check other approximations ([Link here](#))

2

HOUSEPRICE: Can We Predict Housing Prices in Spain's Major Cities?

Data description: Housing prices depend on property characteristics, location, and proximity to urban amenities. This project challenges students to analyze geo-referenced real estate listings from Madrid, Barcelona, and Valencia to build a predictive model for housing prices. The dataset, available at Environment and Planning B: Urban Analytics and City Science (<https://journals.sagepub.com/doi/10.1177/23998083241242844>), includes property details (size, bedrooms, bathrooms, construction year), location attributes (distance to metro, city center, roads), and neighborhood data (housing density, cadastral quality).

Main source of data: A large-scale, geo-referenced microdata set from real estate listings enriched with Spanish cadastre data. The dataset is accessible at GitHub (<https://github.com/paezha/idealista18>).

Possible value: The main idea is to explore how location and property attributes influence pricing, offering insights for buyers, investors, and urban planners.

Innovation: to be found!

Related data to add value: Adding socioeconomic indicators, testing different predictive models, or spatially analyzing price trends could refine predictions and offer deeper insights into housing market dynamics.

HIT-OR-FLOP: Can We Predict the Success of Songs?

Data description: Music streaming platforms generate vast amounts of data on song characteristics and listener preferences. This project focuses on building a classification model to predict whether a song will be a hit or a flop based on its audio features. The dataset includes attributes such as energy, danceability, acousticness, instrumentalness, and tempo, extracted from Spotify's API.

Main source of data: A dataset containing over 160,000 songs from 1921 to 2020, including audio features and popularity scores. Available at GitHub (<https://github.com/yashrajakkad/song-popularity-prediction>).

Possible value: What makes a song popular? This project explores whether machine learning models can identify patterns in musical attributes that correlate with success. The findings could provide insights for music producers, streaming services, and record labels.

Innovation: Classification models such as Logistic Regression, Random Forest, XGBoost, and Neural Networks can be applied to predict song popularity. Feature selection techniques and dimensionality reduction (PCA, PLS) can enhance model interpretability and performance. You could detect if it has changed along time.

Related data to add value: Incorporating lyrical sentiment analysis, artist popularity, or social media engagement metrics could refine predictions and reveal deeper patterns in hit song success.

5

MISC – Miscellaneous Challenges

Data description: Are you fan of challenges? You can register to one of the challenges currently active and use the data to build your project for this subject, but also to win a prize!!

Main source of data: <https://dreamchallenges.org/open-challenges/>, <https://zindi.africa/competitions/>, <https://mlcontests.com/>, or <https://challengedata.ens.fr/challenges/year/2024>

Innovation: Develop a predictive model to solve the problem posed by the challenge. There are many topics to choose: take a look!

Related data to add value: Similar studies available on the internet.

3