Gonzalo Cruz Gómez

gonza.c.gomez03@gmail.com | github.com/gonzalo-cruz | gonzalo-cruz.github.io/portfolio/

Data Science & Engineering undergraduate (3rd year, URJC) with a background in Telecommunications Engineering. Builds production-ready data systems: end-to-end ML pipelines, distributed computing, and deep learning. Founder of Asociación Datalab. Seeking an internship in data science or data engineering.

Education

B.S. Data Science & Engineering **URJC** *2024 – Present*
- Completed two academic years in one (2024–2025). Coursework: ML, Distributed Systems, Deep Learning.

B.S. Telecommunications Engineering **UC3M** *2019 – 2023*
- 120+ credits in Networking, Physics and Signal Processing. Transferred to specialize in Data Science.

Certificate of Advanced English (C2) Cambridge Assessment *2019*

Projects

TripAdvisor Restaurants Data Pipeline *Python, Apache Airflow, Kafka, Scikit-learn, Pandas*
- Orchestrated a 5-task ETL pipeline with Apache Airflow, processing **1M+ rows** in 50k-row chunks to stay memory-safe throughout — extract, clean, EDA, preprocessing, and load.
- Fitted `IncrementalPCA` and `StandardScaler` via `partial_fit` on batches; applied OHE or label encoding per column based on cardinality.
- Streamed the processed dataset to Kafka using a fault-tolerant producer (`acks=all`, sub-batch delivery) ready for downstream consumption.
- Centralised all pipeline parameters (chunk size, thresholds, Kafka config) in a single `config.toml`; all tasks read from it at runtime.

Forest Fire Spread Prediction *R, GLMs, GAMs, Statistical Modeling*
- Built GLM and GAM regression models to predict fire-burned area from meteorological covariates; GAMs outperformed GLMs on held-out data with lower residual variance.
- Resolved severe data quality issues (high skew, missing values, near-zero-variance predictors) via log transformation, imputation, and feature selection.
- Applied cross-validation to compare model fits; selected GAMs as the final model based on lower RMSE and better capture of non-linear covariate effects.

Stroke Risk Classification *Python, Scikit-learn, TensorFlow, ANN*
- Trained an MLP on an imbalanced medical dataset, applying SMOTE and optimizing the decision threshold to minimize false negatives given the clinical cost of missed strokes.
- Selected AUC-ROC and F1 as evaluation metrics over accuracy; tuned architecture and hyperparameters accordingly.
- Benchmarked the MLP against a logistic regression baseline; the non-linear model improved recall on the stroke class while maintaining acceptable precision.

Experience

Founder & President — Asociación Datalab May 2025 – Present

Established the university's first Data Science student association: defined the strategic roadmap, handled legal registration, and organized technical workshops.

Private Tutor — Mathematics & Physics Sep 2021 – Present

Mentored high-school students over 5 years, achieving a **100% pass rate** and helping 90% of students improve by at least 3 grade points through personalized plans.

Technical Skills

Programming: Python (Advanced), R, SQL, C
Data Science: Scikit-learn, PyTorch, TensorFlow, Pandas, NumPy, Hadoop, Spark, Matplotlib
Engineering: Apache Airflow, Kafka, Docker, Git/GitHub, Linux (Bash)
Languages: Spanish (Native), English (C2), German (B1), Chinese (HSK3)