

Pablo Alberto Duque Marín

Manzanares, Caldas, Colombia (remote, UTC-5) | pabdus3000@gmail.com
github.com/pabdus | linkedin.com/in/pablo-alberto-duque-marín

EDUCATION

Universidad Internacional de La Rioja (UNIR)

MSc in Artificial Intelligence

- Thesis in progress: data-driven fault identification in industrial equipment.
- Postgraduate specialization in generative AI for data analysis. Coursework in knowledge representation, fuzzy logic, NLP and research methodology.

2026 – Present
Spain (online)

Universidad Cuauhtémoc

MSc in Applied Statistics

2025 – Present
Mexico (online)

Universidad Compensar

BSc in Data Science

2023 – Present
Colombia (online)

Earlier

Data Science Bootcamp, 800+ hours (Henry, 2022–2023) · BA in Police Administration · Systems Technician

2007 – 2023

EXPERIENCE

AI Pipeline & Infrastructure Engineer (Internship)

Softart

- Own the backend of a document-to-decision system in a regulated legal domain, measured at 53 seconds from upload to generated assessment.
- Provisioned roughly 57 AWS resources with Terraform, using remote S3 state and reusable modules so the environment rebuilds from scratch.
- Built the event-driven pipeline on Lambda, DynamoDB, S3, EventBridge and SQS, decoupling stages so one step can be retried without reprocessing the document.
- Extracted fields with Amazon Textract, switching from FORMS to TABLES once FORMS collapsed the column layout, and anchoring each read on its label cell.
- Integrated LLMs through the Bedrock Converse API, grouping related documents into a single call. Ruled out a Knowledge Base: similarity retrieval cannot check a fixed list of criteria exhaustively.
- Prototyped seven interface screens ahead of frontend development. Work reviewed by a tech lead; CI/CD and frontend belong to other teammates.

2026 – Present
Remote, Colombia

Cloud Infrastructure Engineer

Independent contract

- Migrated a manually managed AWS service to Terraform, leaving the environment version-controlled and reproducible. Delivered to production and invoiced.

2026

Remote

Data Analyst

Tubos y Perfiles de Occidente S.A.S.

- Built Python ETL processes joining operational and commercial data from several internal and external sources.
- Replaced manual reporting with scheduled Python scripts and wrote the SQL behind recurring analyses used for decision-making.

Mar 2022 – Mar 2026
Dosquebradas, Colombia

Founder & General Manager · Lieutenant, Unit Commander

Narnia Producciones (Chile) · Colombian National Police

- Ran an event production company: budgeting, profitability analysis and supplier negotiation. Previously commanded operational police units across Colombia.

2011 – 2022

PROJECTS

Critical Evaluation of CTGAN for Tabular Synthetic Data

github.com/pabdus/ctgan-adult-critical-evaluation

- Evaluated three CTGAN variants on UCI Adult across fidelity (Wasserstein, Jensen-Shannon), utility (TSTR), memorization (DCR) and privacy.
- Found the baseline generating 62% logically invalid values that Wasserstein and Jensen-Shannon both scored well, since distributional metrics ignore semantic validity. A one-line clip restored validity at no cost to utility; a log transform instead induced partial mode collapse.
- Wrote a membership inference attack from scratch, which exposed an inverse relationship between utility and privacy leakage (AUC 0.852 against 0.663) and showed DCR reporting modal degeneracy as clean generalization.

Class Balancing under Extreme Imbalance: SMOTE against CTGAN

github.com/pabdus/proyecto-02-obesidad-brfss

- Benchmarked three classifiers against four balancing strategies on 289,693 CDC BRFSS records, giving 12 comparable evaluations on a minority class holding 1.68% of the data.
- Reported a negative result: CTGAN gave no real gain over SMOTE, and no strategy lifted minority F1 above 0.07. Traced it to separability rather than volume, and documented CTGAN returning 466 of 10,000 requested conditional samples.

Regional Inequality in Colombia

DANE 2018 census microdata

- Quantified urban–rural gaps over 15.3M records using Z-tests for proportions, chi-square and logistic regression; rural illiteracy odds came out 5.39 times higher. Published as a choropleth map and narrative dashboard.

WineQT Statistical Analysis and Dashboard

Deployed on GitHub and Binder

- Levene and Welch tests, multiple linear regression ($R^2 = 0.33$) and logistic regression (74% accuracy) behind an interactive Dash and Plotly interface.

TECHNICAL SKILLS

Languages: Python, SQL, HCL/Terraform, R (basic)

Cloud and infrastructure: AWS (Lambda, DynamoDB, S3, EventBridge, SQS, Bedrock, Textract), Terraform, serverless and event-driven design, infrastructure as code

Machine learning and AI: scikit-learn, XGBoost, PyTorch, CTGAN/SDV, SMOTE, LLM integration through Bedrock Converse API, model evaluation (TSTR, mode-collapse detection, membership inference)

Data and statistics: pandas, NumPy, SciPy, ETL, MySQL, PostgreSQL, MongoDB, hypothesis testing, linear and logistic regression, experimental design

Visualization and tooling: Plotly, Dash, Power BI, Matplotlib, Seaborn, Git, Jupyter

Languages spoken: Spanish (native), English (intermediate, fluent technical reading)