Domenico **Lacavalla**

DATA SCIENTIST · ML/AI ENGINEER

Bari, Italy

✉ domenicolacavalla8@gmail.com | d0men1c0.github.io | D0men1c0 | D0men1c0 | domenico-lacavalla | domenicolacavalla8

Education

UNIBA (Università Degli Studi di Bari "Aldo Moro") *Bari, Italy*

MSC IN DATA SCIENCE *Sept. 2023 - Apr. 2026*

Major courses: machine learning, data mining, statistics, linear algebra, deep learning, numerical methods

UNIBA (Università Degli Studi di Bari "Aldo Moro") *Bari, Italy*

BSC IN COMPUTER SCIENCE, GRADUATION WITH 110/110 WITH HONOURS *Oct. 2020 - July 2023*

Experimental thesis on reproducibility in recommendation systems, applied to the ClayRS codebase with data pre-filtering and metrics.

Work Experience

IBM *Bari, Italy*

DATA SCIENTIST AI ASSOCIATE - PYTHON, TRANSFORMERS, OPENCV, SCIKIT-LEARN, BERT, PYTORCH, SQL, IBM CLOUD *Oct. 2023 - Present*

- Utilized generative AI and fine-tuned BERT to analyze 4M conversations, identify trends, apply clustering, and present results to stakeholders.
- Built and deployed customer segmentation models handling 6M data points, integrating a full training pipeline with automated retraining.
- Engineered production AI pipeline leveraging OCR/OMR, NER, LLMs & Speech-to-Text for 10M+ assets, optimizing digitization to minute-scale.
- Applied ResNet (86% score) & CLIP (93% score) to evaluate fidelity of OMR output (Audiveris) against original digitized scores.
- Led temporal data extraction (OCR/Tesseract); initiated HPC (H100 GPUs) parallelization using SLURM aiming to halve processing times.
- Optimized complex ETL query execution from 20 hours to 1 hour by refactoring SQL and leveraging Spark on IBM Cloud with multithreading.

Google Summer of Code (GSoC) 2024 HumanAI *United States, Remote*

OPEN SOURCE CONTRIBUTOR - PYTHON, JUPYTER NOTEBOOK, BERTOPIC, DATA ANALYSIS, SQL, CLIP, VIT *May 2024 - Sept. 2024*

- Used NLP models to analyze 500k Dark Web discussion points, identifying key topics and establishing 5 baseline categories.
- Enhanced the model to interpret both images and text using BERT (170 topics) and CLIP/Vision Transformer (3 topics).
- Validated clustering results with Machine Learning algorithms and LSTM, examining topic evolution and sentiment analysis over time.
- Deployed 8 predictive models on Hugging Face to forecast trends and topics identified in the analysis.
- Read more in this blog post and explore the project on GitHub Repo.

Personal Projects

Gemma Model Benchmark Suite *GitHub Repo*

PYTHON, PYTORCH, TRANSFOMERS, SCIKIT-LEARN, TENSORFLOW, HUGGINGFACE *Mar. 2025 - May. 2025*

- Authored Medium blog post (10min+ read) detailing framework architecture, customization & performance.
- Engineered customizable LLM suite: 4+ LLM families, 5+ tasks (MMLU+), 15+ metrics (ROUGE+), full custom script integration.
- Optimized for Colab T4: 4 min/500-sample (Gemma 2B 4-bit); enabled broad LLM experimentation on accessible hardware.
- Ensured robustness (73% Pytest coverage, Pydantic-validated YAML); delivered insights via Streamlit & 3+ report formats.

Portuguese public procurement Analysis *GitHub Repo*

PYTHON, JUPYTER NOTEBOOK, PANDAS, CLUSTERING, SCIKIT-LEARN, MLXTEND *Jan. 2025 - Feb. 2025*

- Analyzed 5,214 contracts, revealing 3 contract profiles and key award criteria impacts with high-confidence rules (lift > 9).

Smart Traffic Lights - Team Project *GitHub Repo*

PYTHON, PROLOG, PANDAS, NUMPY, MATPLOTLIB, SCIKIT-LEARN *Nov. 2022 - Feb. 2023*

- Optimized A-to-B travel time using Prolog (OpenStreetMap KB), A* search, and HMMs, achieving 85% traffic prediction accuracy.

Extracurricular Activity

Mentee Superhero Valley 2025: Selected among top Italian students for an exclusive mentorship with Big Tech leaders.
LauzHack 2024: Partecipated in the LauzHack hackaton at EPFL with a vision AI assistant.
Samsung Innovation Campus 2022-2023 Edition: Top 25 STEM student to partecipate in the program.

Skills

Technology Stack	Python, Java, MySQL, PostgreSQL, PostGis, MongoDB, IBM Cloud, AWS, OpenShift, Docker
Python Library	Pandas, Numpy, Scikit-learn, TensorFlow, PyTorch, Keras, OpenCV, NLTK, Spacy, Gensim, Transformers, Hugging Face
Languages	Italian (Native), English (Proficiency)