Manuel Alejandro Diaz Rubiano

+57 3196371560 - alejandromadr@gmail.com - www.linkedin.com/in/manuel-díaz-96b821131/
Webpage: https://manueldiazai.com/

EDUCATION

Universidad Santo Tomas Bogotá, CO.
April - 2022

Degree in Statistics and Math
Thesis: Topic analysis using Twitter: an application of the LDA model to the Colombian case. Relevant Coursework: Probability and Bayesian statistics.

Universidad de la Sabana Bogotá, CO
Diploma in Machine Learning. September - 2023
Subject: LLM models apply to business case, to develop Chatbots with Llama2 and GPT. Understanding the LLM structures, such as their transformers architecture and Self-Attention functions.

Universidad del Rosario Bogotá, CO
Diploma in Big Data and Cloud Computing with AWS. September - 2022
Subject: Use Python and R Language to create Machine Learning Models, Apache Spark and Apache Hadoop for Big Data processing and analytics. Using AWS tools to train big data models, using SageMaker and S3.

EXPERIENCE

H&Co Latam: Bogota, CO (Sept 2023 – Now)

AI Engineer

- Led end-to-end GenAIOps pipelines using AWS Lambda, AWS Bedrock, and OpenAI API, deploying production-grade LLM applications with structured outputs via Pydantic and Instructor for reliable schema-adherent responses from Claude and GPT models.
- Engineered graph-based AI systems using LangGraph and LangChain for multi-agent orchestration, enabling complex workflow automation through state management, conditional routing, and parallel tool execution across distributed AI agents.
- Built production APIs leveraging OpenAI's function calling and structured outputs (strict mode with JSON Schema), achieving 100% schema adherence for critical business logic including data extraction, classification, and multi-step reasoning tasks.
- Deployed RAGOps architectures combining vector databases (mainly OpenSearch but also testing with Chroma and Pinecone), Redis for session state and fast key-value caching and LangChain/LangGraph orchestration, and LLMs for context-aware intelligent assistants with semantic search and retrieval-augmented generation capabilities.
- Established MLOps CI/CD pipelines using GitHub Actions, AWS CDK, and Git-based versioning for infrastructure-as-code deployment, ensuring reproducible model training, automated testing, and seamless production releases with rollback capabilities.
- Integrated monitoring and observability using AWS CloudWatch, MLFlow, and custom logging to track model performance, AI API Inference usage, data drift, and system reliability across generative AI applications.

Manuel Alejandro Diaz Rubiano

Finanzauto: Bogota, CO (Jan 2022 – Sept 2023)

Machine Learning Specialist

- Developed time-series forecasting models using Python (Prophet, ARIMA, LSTM) to predict portfolio recovery rates through cohort analysis, enabling proactive risk management and improving collection strategy accuracy by analyzing default patterns and payment behaviors.
- Built predictive analytics pipelines for credit risk assessment and portfolio depreciation forecasting, leveraging machine learning techniques to model payment default probability and optimize loan portfolio performance metrics.
- Optimized model serving infrastructure for real-time inference using TorchServe and AWS Lambda, implementing model quantization and batching strategies to reduce latency by 45% while maintaining prediction accuracy for high-throughput credit scoring endpoints.
- Automated model training, validation, and deployment processes with MLOps frameworks, ensuring reproducibility and scalability for BERT, GPT, and T5 models.
- Implemented fine-tuning pipelines for domain-specific BERT and GPT models on financial text data, leveraging PyTorch and HuggingFace Transformers to improve credit risk classification accuracy by 18% through transfer learning and custom tokenization strategies.
- Architected end-to-end MLOps CI/CD pipelines using GitHub Actions, AWS SageMaker, and Docker for automated model training, validation, and deployment, with DVC for data versioning and A/B testing frameworks to evaluate model performance in production with statistical significance testing.

Concentrix: Remote, USA (Feb 2020 – Jan 2022)

Sr. Specialist Global Analytics Insights.

- Developed comprehensive KPI dashboards in PowerBI and Looker Studio, providing real-time business insights that led to a 15% improvement in strategic decision-making.
- Enhanced data governance and quality frameworks using Python, Pytorch and Scikit Learn, ensuring compliance with industry standards and reducing data-related errors.
- Conducted advanced statistical analysis to identify key performance drivers, translating complex data sets into actionable insights that improved operational efficiency.
- Devised and implemented NLP-driven improvement strategies for call center agents, optimizing performance and customer satisfaction through advanced analytics, using Hugging Face models like BERT, T5 and tools such as Pytorch, Tensorflow.
- Optimized data extraction and integration processes using SQL and ETL techniques, reducing data retrieval times by 50% and increasing the reliability of data feeds.

FREELANCE EXPERIENCE AND INDEPENDENT PROJECTS

Outlier AI (Nov 2023 – Jan 2025)

Data Annotator and Specialized RLHF.

- Performed precise data annotation on Outlier AI, meticulously labeling datasets to enhance the accuracy and performance of machine learning models during their training phases.
- Specialized in Reinforcement Learning from Human Feedback (RLHF), providing targeted evaluations for programming and mathematical outputs to train models that align closely with human preferences.

- Utilized advanced AI tools and methodologies to streamline data labeling processes across various projects, ensuring the efficient preparation of training data for diverse AI applications.

BlueBoost AI (Nov 2024 – Dec 2025)

Machine Learning Engineer.

- Architected end-to-end workflow automation pipelines using N8N to orchestrate complex AI-powered business processes, integrating custom APIs with OpenAI API for automated lead generation, customer communication, and data processing workflows.
- Designed and implemented RAGOps pipelines by developing scalable vector databases in OpenSearch and integrating them with LangChain for efficient semantic search and context retrieval in production chatbot applications.
- Built production-grade API orchestration layers using N8N and FastAPI to connect proprietary business logic with LLM providers (OpenAI, Anthropic), enabling automated multi-step workflows with conditional routing, error handling, and real-time monitoring.
- Established GenAIOps workflows for managing LLM-based agents with automated deployment using AWS Lambda and API Gateway, implementing model versioning, A/B testing, and real-time performance tracking for continuous optimization.
- Implemented secure, scalable infrastructure integrating AWS WAF for API security, Docker containerization for reproducible deployments, and CloudWatch monitoring for production-ready AI solutions.

Skills and Interests

Soft Skills

- Leadership and team management
- Strategic thinking and innovation
- Strong communication and collaboration
- Analytical and problem-solving skills
- Adaptability and continuous learning
- Stakeholder engagement

Tools & Technologies

- **Programming:** Python, R, SQL, TypeScript. Git
- **ML/DL Frameworks:** TensorFlow, PyTorch, Scikit-learn, Jax
- **NLP/GenAI:** Transformers, LangChain, LangGraph, LlamaIndex, OpenAI, Claude, HuggingFace,
- **Vector Databases:** FAISS, Chroma, OpenSearch, Pinecone
- **MLOps/GenAIOps:** AWS Sagemaker, AWS Bedrock, AWS Lambda, AWS CDK, Terraform, FastAPI, Docker, Kubernetes, GitHub Actions
- **CI/CD & Monitoring:** Grafana, AWS CloudWatch, Prometheus, Phoenix Arize, MLFlow
- **Data Visualization:** Power BI, Plotly, Looker Studio, Streamlit, Shiny
- **Cloud Platforms:** AWS, GCP, Azure
- **ETL/Data Pipelines:** Apache Spark, Hadoop, SQL

Languages: English, Spanish