Nathaniel del Rosario

natdosana@gmail.com | linkedin.com/in/natdosan | natdosan.github.io

Education

University of California, San Diego — La Jolla, CA
B.S. Data Science — *2025*
- Graduate (cross enrollment): Computer Vision, Recommender Systems, Deep Learning
- Principles & Techniques of Data Science, Statistics, Relational Databases, Operating Systems, Cloud Computing, Scalable ML

University of California, Berkeley — Berkeley, CA
Computer Science, Cross Enrollment — *2023*
- Artificial Intelligence, Machine Learning, Reinforcement Learning, Search Algorithms, Probabilistic Modeling & MDPs

Skills

Python, Pandas, NumPy, PyTorch, Sci-Kit Learn, HuggingFace, OpenCV, Tensorflow, Dash/Plotly, Cuda, Dask, Streamlit
AWS, Azure, Google Cloud, Docker, Kubernetes, SSH, Snowflake, Databricks, Spark, Hadoop, PostgreSQL, Github, Jira, ArcGIS

Experience

University of California, San Diego — April 2024 - Present
Machine Learning Researcher — *La Jolla*
- Designed metrics and regression model to predict public transportation accessibility in New York City (RMSE of .1785)
- Utilizing machine learning to identify and predict crime hotspots in cities supervised by Prof. Zaslavsky
- Implementing Multi-Modal datasets into LLM-Reasoners Library to improve Process Reward Model (PRM) capabilities of smaller LLM models. Advised by Prof. Zhiting Hu, Kun Zhou (PostDoc), Shibo Hao (PhD)

Bio-Rad — June 2024 - August 2024
Data Science Intern - Clinical Diagnostics Group — *Pleasanton*
- Built ETL pipeline w/ Pandas, RestAPI to ensure 100% data integrity & improved consistency from 91% to 99.9%
- Utilized AWS EC2 to deploy web-app & unit tests utilizing Dask to achieve 5.1x / 80.1% speedup on data validation
- Leveraged AWS RDS, Docker, PostgreSQL to deploy database, reduced storage usage by 35% through schema optimization
- Ensured fault tolerance through distributing across multiple availability zones & heuristics achieving persistent Database I/O

San Diego Supercomputer Center — June 2023 - September 2023
Machine Learning Engineer Intern — *Remote*
- Designed Content-Based Filtering Recommender System utilizing Cosine and Jaccard similarity for baseline output
- Trained an RL agent using Stable Baselines and Q-Learning to improve recommendation quality after 100 iterations
- Utilized AWS S3, PostgreSQL for database queries & vectorized code to achieve 1.7x runtime speedup in feature engineering
- Deployed Recommender System on AWS EC2, Lambda, achieving a design that scaled to process 200,000+ points

Deloitte — February 2023 - June 2023
Data Science Fellow — *Remote*
- Cleaned data w/ 3000+ features, 1 billion observations using Dask, vectorized Pandas to decrease cleaning runtime by 20%
- Leveraged XGBoost, Lasso to identify 850 significant features, predict drug use in young adults with 81% accuracy
- Tuned Hyperparameters, class weighting to improve F1 score from .35 to .70 and identify 10 highest risk demographics

Chan Zuckerburg Biohub — June 2022 - January 2023
Data Science Intern - Infectious Disease Group — *San Francisco*
- Built 9 interactive visualizations of CRISPR screen comparisons between 20000 features using Pandas/Dash/Plotly
- Improved data processing of a Nextflow data pipeline (16,000,000 data points) to minimize runtime by 10%
- Designed algorithms to compare across 30+ virus screens to yield insights in virus-host interactions using vectorized code
- Wrote documentation for 23 functions from scratch and improved 3K+ codebase readability using Readthedocs

Projects & Leadership

University of California, San Diego — September 2023 - Present
Instructional Assistant — *La Jolla*
- Beta Testing assignment and exam questions, hosting Office Hours for a data science course of over 500 students
- Updated deployment of course website using github pages & Docker supervised by under Suraj and Tiefenbruck
- Grading and hosting Office Hours for upper division data science course of over 700 students under Shannon Ellis, Sam Lau

Exploring CNN Architecture for Semantic Segmentation — PyTorch, OpenCV, HuggingFace — February 2024
- Implemented different UNet architectures with AdamW Optimization, Data Augmentation, weighted cross entropy loss, learning rate scheduling to improve IoU score from .055 to .071 and pixel accuracy from 73.4% to 75.1%
- Utilized FCN ResNet-101 for transfer learning further improving IoU score to .33 and validation accuracy to 87.3%

Spotify User Persona Clustering — SpotiPy, Scikit-Learn — June 2023
- Wrote an automated pipeline using SpotiPy, Spotify API to scrape, preprocess, feature engineer data (200+ unique songs)
- Performed PCA and K-Means to identify 6 unique listening personas for identifying target audiences