

DYLAN DSOUZA

San Diego, CA

✉ dydsouza@ucsd.edu  dylandsouza.com  linkedin.com/in/dsouza-dylan  github.com/dsouza-dylan

EDUCATION

University of California San Diego

Bachelor of Science in Data Science – GPA: 3.95 (Provost Honors)

San Diego, CA

Expected March 2027

EXPERIENCE

Machine Learning Researcher

Oct 2025 – Present

Engineers for Exploration

San Diego, CA

- Engineered a Google Earth Engine pipeline to stage 69-band Sentinel-2 and satellite embedding training tiles across Madagascar and Mozambique, gating downloads to mangrove regions using ESA WorldCover land-cover classification.
- Integrated 7 geospatial datasets to curate a structured training dataset for pixel-level mangrove mapping, processing 21,000+ GeoTIFF image-mask pairs across 4 land-cover classes for downstream model training.

Research Data Analyst

May 2025 – Present

Scripps Institution of Oceanography

San Diego, CA

- Automated 11 years of coral reef data processing (60% cut in manual processing time), building a Python ETL pipeline that standardized 40,000+ benthic annotations into clean datasets for downstream analysis across 15 reef sites.
- Analyzed treatment-differentiated recovery from a 2023 mass bleaching event using PERMANOVA ($R^2=0.74$, $p=0.002$), NMDS ordination, and piecewise mixed-effects models across 5 experimental groups and 16 survey time points.

Market Research Intern

Jul 2023 – Aug 2023

Bennett Coleman & Co. Ltd. (The Times Group)

Mumbai, India

- Built a multiclass NLP classifier achieving 85% accuracy across 5 market sectors, processing 56,000+ newspaper headlines using TF-IDF and BERT embeddings to automate market sector categorization at scale.
- Produced an internal research brief on data infrastructure modernization – covering data lakes, CDPs, and CRM optimization – to support a 5-person team's evaluation of editorial and business workflow improvements.

LEADERSHIP & ACTIVITIES

AI Research & Development Analyst

Oct 2025 – Present

Computer Science & Engineering Society

San Diego, CA

- Implemented a hybrid RAG retrieval pipeline for an agentic tutoring platform, fusing MiniLM vector search and BM25 retrieval via Reciprocal Rank Fusion with graph-aware context expansion across Neo4j NEXT-chain relationships.
- Built a knowledge graph pipeline that extracts LLM-generated concepts from research documents, mapping relationships between them using contextual proximity scoring and Girvan-Newman community detection across 8+ entity categories.

Quantitative Analyst

Oct 2025 – Dec 2025


Sustainable Investment Group

San Diego, CA

- Built an automated Python pipeline using FinBERT and Yahoo Finance API to quantify lagged correlations between 25 weeks of news sentiment and 5+ market KPIs, generating actionable trading signals across returns, volatility, and volume.
- Co-led quantitative investment thesis for Under Armour integrating ARIMA forecasting, FinBERT sentiment analysis, and DCF modeling to deliver a sell recommendation in a competitive 5-team stock pitch.


PROJECTS

ENSOcast: Interactive Climate Analytics Dashboard | Conference Project (SDUTC 2025)

[View Project](#) |  [GitHub](#)

- Engineered a Streamlit climate dashboard on 40+ years of NOAA data; benchmarked Random Forest, XGBoost, 1D CNN, LSTM, and Ensemble models achieving 82% accuracy in 3-class ENSO phase prediction.

DegreeDelta: Quantifying the Income Advantage of College Degrees | Personal Project

[View Project](#) |  [GitHub](#)


- Quantified the college income premium across all 50 U.S. states using weighted ACS PUMS microdata and state-level aggregation, revealing geographic variation in the financial return to higher education.

Sweet but Healthy? Evaluating Honey vs. Sugar in Recipes | Personal Project

[View Project](#) |  [GitHub](#)


- Applied permutation tests to 80,000+ ingredient-filtered Food.com recipes to assess whether honey-based desserts are nutritionally distinct from sugar-based ones across calories, sugar content, and total fat.

America on Fire: Mapping U.S. Wildfire Distribution | Course Project (Data Visualization)

[View Project](#) |  [GitHub](#)

- Visualized 2024 U.S. wildfire distribution through an interactive D3.js map using NASA MODIS satellite data, surfacing spatial burn patterns, seasonal trends, and unexpected regional hotspots.

NextMove: Predicting Next Workout Activity | Course Project (Recommender Systems)

[View Project](#) |  [GitHub](#)

- Predicted next workout type from 20,000+ Endomondo records as a multiclass problem, engineering temporal and physiological features to achieve 87% accuracy with Random Forest over 3 heuristic baselines.

SKILLS

Languages: Python, SQL (PostgreSQL), R, Java, HTML, CSS, JavaScript

Libraries/Frameworks: pandas, NumPy, scikit-learn, statsmodels, Matplotlib, Plotly, Neo4j, Git, PyTorch, TensorFlow, NLTK, LangChain

Core Competencies: Data Cleaning/Visualization, EDA, ETL Pipelines, A/B Testing, Financial Modeling, Time Series Analysis, NLP