DIYA SAHA

Los Angeles, CA | 4084807445 | diya.s@uci.edu | https://www.linkedin.com/in/diya-saha | https://diyadotsaha.github.io

EDUCATION

University of California, Irvine   **June 2025**
Master of Science, Computer Science

University of California, Santa Cruz   **June 2023**
Bachelor of Science, Computer Science and Minor in Computer Science

WORK EXPERIENCE

Graduate Researcher - Canine Cancer Research with Dr.Pierre Baldi | Irvine, CA   **July 2024 – Oct 2024**
- Led the automation of clinical report summarization across 10,000+ canine patient records by fine-tuning LLMs (OpenAI GPT-4, LLaMA) to extract medical entities and diagnoses, eliminating the need for handwritten reports by veterinarians.
- Integrated OCR pipelines (Tesseract + OpenCV) to extract structured data from unstructured PDF clinical reports, resulting in a reduction in manual processing time.
- Utilized Named Entity Recognition to identify correlations between biomarkers, drug names, and patient diagnoses, uncovering key patterns that contributed to advancing research in canine cancer diagnostics.

Graduate Researcher - Low Resource Language Research with Dr. Baldi | Irvine, CA   **Sep 2024 – Nov 2024**
- Fine-tuned mBART on curated Nahuatl-Spanish corpus, developing a bidirectional translation system for a low-resource language.
- Engineered a parallel corpus using lexical substitution and back-translation to augment the training data, expanding the model's coverage and improving performance.
- Applied evaluation metrics like Jaccardian distance and word-based distance evaluation, validating the system's performance by achieving an 80% match with Google Translate.

Software Developer Intern - DSights.Inc | Los Angeles, CA   **June 2022 – Sep 2022**
- Developed an NLP-based tool to compare and identify similarities between 50+ public menus from web pages, building a custom corpus of ingredients, recipes, and dish names to compare dish similarity between the restaurants.
- Utilized OCR to extract menu data, performing EDA with Seaborn, Matplotlib, and Pandas to uncover key trends and patterns in food offerings.
- Identified actionable opportunities for the client to optimize their menu, such as refining dish offerings and leveraging ingredient trends for targeted marketing and business strategies.

PROJECTS

TL;DW(Too Long; Didn't Watch) | Python, Streamlit, Gemini, Whisper 🔗   **Apr 2025**
- Developed an AI-powered Streamlit app that automates video summarization, topic extraction, and transcript generation using Whisper and Gemini LLM, enabling users to quickly digest long-form content.
- Integrated interactive quiz generation with auto-grading and LLM-based explanations, allowing users to absorb key takeaways and reinforce their understanding by simply entering a single URL, making the learning process efficient.

CourseBuddy - Context-Aware Chatbot using RAG | Langchain, FAISS, HuggingFace 🔗   **Sep 2024**
- Built a custom vector database using FAISS, integrating the UCI course catalog, course prerequisites, and other academic data, and powered an interactive chatbot via Streamlit to act as a virtual academic advisor.
- Leveraged RAG to enhance the chatbot's ability to provide personalized guidance on course selection, incorporating FAQs from Reddit, Quora, and public academic sites to minimize LLM hallucinations and ensure accurate information.

Diabetic Readmission Prediction Model for Healthcare Optimization | Python, Pandas, Matplotlib, Sklearn 🔗   **Sep 2023**
- Utilized a decade-long dataset from 130 US hospitals to predict diabetic patient readmission rates, achieving 90% model accuracy with classification, ensemble, and neural network models.
- Applied oversampling techniques (SMOTE, data imputation) to address class imbalance, and optimized model performance with feature selection and engineering.
- Developed a predictive model that provided actionable insights to help hospitals reduce readmission rates, optimize resource allocation, and improve patient outcomes, potentially saving significant healthcare costs.

SKILLS

Programming Languages: Python, Git, C/C++, HTML/CSS, JavaScript, SQL, MATLAB
Developer Tools: Google Cloud Platform, AWS, Figma, Langchain, HuggingFace, Streamlit, PostgreSQL, MySQL, Docker/Kubernetes, Vercel
ML Frameworks: PyTorch, TensorFlow, Keras, Scikit-learn, GenAI, RAG, LLMs, MLOps
Certifications & Training: Certified AWS AI Practitioner, Google Certifications (Data Analytics, UI/UX Design)
Awards: Dean's Honor on multiple quarters; top 15% of my class (2023)