Arnav Hiteshbhai Desai

New York City | (631) 361-1666 | arnavdesai2025@gmail.com | linkedin.com/arnavdesai | github.com/arnavdesai

Education

Stony Brook University **Aug. 2025 – Jun. 2027**
Master's of Science in Data Science ***GPA:3.55/4.0***
Coursework: Data Analysis, Data Structures and Algorithms, Data Science Fundamentals, Statistical Computing

Gujarat Technological University **Jun. 2021 – Jun. 2025**
Bachelor of Engineering in Information and Communication Technology ***GPA:8.56/10.00***
Coursework: OS, DBMS, Software Engineering, Big Data Analytics, AI & ML, Computer Networks.

Technical Skills

Languages: Python, Java, SQL, JavaScript, C/C++, R, MATLAB, Scala, Shell scripting.
Frameworks and Libraries: TensorFlow, PyTorch, Keras, Scikit-learn, NumPy, Pandas, SciPy, Matplotlib, Plotly, Seaborn, Hugging Face, NLTK, XGBoost, OpenCV, DICOM, Apache PySpark, AWS Lambda, Librosa, Beautiful-Soup, Streamlit.
Web development: React.js, Node.js, Next.js, Vue.js, Express.js, Angular.js, TailwindCSS, FastAPI, REST API, Flask, Django.
Cloud and Tools:: Git, GitHub, Google Cloud Platform, AWS, Azure, MySQL, MongoDB, PostgreSQL, Redis, Snowflake, LangChain, Docker, Apache Kafka, Anaconda, PyCharm, JupyterLab, Jupyter Notebook, Google Colab, Tableau, Excel, PowerBI.

Professional Experience

Visual Analytics and Imaging Lab **Jan. 2026 – Present**
Graduate Research Assistant - Python, PyVis, UMAP, NetworkX, sentence-transformer, Streamlit *Stony Brook, NY*
- Modeled domain-specific **Causal Networks** for ADHD using **Python**, **NetworkX**, and causal inference methodologies, quantifying dependencies among clinical, behavioral, and socioeconomic factors affecting functional outcomes.
- Implemented a **Causal Chat** system combining **LLM-powered dialogue**, causal graph traversal and counterfactual reasoning, facilitating interpretable analysis of intervention effects and personalized outcome exploration.

Prodigy Infotech **Jul. 2025 – Aug. 2025**
Machine Learning Engineering Intern - Python, Hugging Face, NLTK, AWS EC2, React.js *Remote*
- Architected advanced resume parser leveraging **Regex** and **NLTK** tokenization, integrated **Hugging Face** NER to surpass **92%** extraction accuracy and managed data storage and retrieval with **MongoDB**, ensuring efficient candidate evaluation.
- Spearheaded **AWS EC2** to build, train and deploy robust **XGBoost** model with CI/CD pipelines automated through **GitHub Actions**, enabling seamless integration, continuous updates and better scalability for enterprise level applications.
- Facilitated **42%** improvement in accuracy and insights, developing a responsive dashboard with **JavaScript** and **React**.

Labmentix **May 2025 – Jul. 2025**
Data Science Intern - Python, Apache PySpark, Apache Airflow, AWS S3, AWS EC2, Seaborn *Remote*
- Streamlined comprehensive stock market analysis of technology and automobile companies using **PySpark**, engineered scalable ETL pipelines with **Airflow**, driving **42%** improvement in analytical accuracy and boosting reliability by **38%**.
- Validated model data on **AWS S3** and deployed with **AWS Lambda**, ensuring scalability, availability, and security.
- Advocated visualizations using **Seaborn** to analyze stock price fluctuations, extracting actionable insights from trading volumes averaging **38,000** shares/day, developing a classification model achieving **57%** precision for predictive analytics.

House Of Imaging and Radiology **Jun. 2024 – Aug. 2024**
Machine Learning Engineer Intern - Python, PyTorch, OpenCV, Flask, Docker, GCP *Gujarat, India*
- Orchestrated interactive medical dashboard using **Flask**, integrating patient information with their prescription history, deployed system on **GCP**, automating workflows, resulting in **28%** reduction of overall administrative workload.
- Supervised **OpenCV** for data cleaning, having predictive CNN model in **PyTorch** for Osteoporosis risk assessment, created endpoint inference with **Flask**, containerized by **Docker** achieving **92.3%** accuracy, deploying on **GCP Cloud Run**.

Projects

Spotflix | *Python, Pandas, Numpy, Librosa, Scikit-learn, SQL, FastAPI* **GitHub**
- Programmed music recommendation system using **Librosa** and **Beautiful Soup**, scraping music metadata enhancing content features, analyzing user behavior, achieving **91%** recommendation accuracy, boosting user engagement by **44%**.

Brain Tumor Classification | *Python, PyTorch, NumPy, OpenCV, Flask, GCP* **GitHub**
- Constructed **PyTorch** based deep learning model achieving **87.6%** accuracy predicting tumors across types of nervous system glands using **TorchVision**, **OpenCV** and **DICOM**, deploying the system on **GCP Vertex** for real-time predictions.

AskAD | Python, MySQL, Gemini, Snowflake, Docker **GitHub**
- Amplified an AI-powered natural language insights system by integrating **Snowflake** with an LLM for **SQL**-based result summarization, enabling users to query data in plain language and receive actionable insights, improving efficiency by **53%**.

Paisa Bazar Credit Score | *Python, Pandas, NumPy, Apache Airflow, Scikit-learn, Azure* **GitHub**
- Leveraged **Pandas** and **NumPy** to preprocess data, orchestrated scalable ETL pipelines with **Apache Airflow** and trained a classification model using **Scikit-learn** on an **Azure VM**, achieving **84.8%** accuracy predicting customer credit scores.

US Accidents Analysis | *Python, Pandas, Dask, Apache Kafka, SQL, Folium, Power BI* **GitHub**
- Forecasted **294,658** accident records using **Pandas** and **Dask**, unifying insights visualization with **Folium** and **Power BI**.