VATSAL PARIKH

+1 930-215-5267 | parikhvatsal7101@gmail.com | vatsalparikh.us | linkedin.com/in/vatsalparikh07 | Bloomington, IN (Open to relocate)

Education

Master of Science - Data Science | Indiana University, Bloomington — **August, 2023 – Present**
Coursework: Data Mining, Applied Machine Learning, Data Visualization, Natural Language Processing — **GPA: 3.85 / 4.0**

Bachelor of Engineering - Information Technology | Gujarat Technological University, India — **July, 2019 – June, 2023**

Technical Skills

Languages : Python, R, SQL, HTML, CSS, JavaScript, React.js, C, C++
Data Management : PostgreSQL, MySQL, MongoDB, Neo4j, Cassandra, Snowflake, SQLite
Tools : Power BI, Tableau, Microsoft Excel, PySpark, BigQuery, GCP, AWS EC2, Amazon S3, Redshift, Hadoop, MapReduce, Kafka, Git, Docker, Jenkins, Kubernetes, JIRA, Visual Studio, Jupyter, Gephi, NodeXL
Libraries : Pandas, NumPy, Matplotlib, OpenCV, SciKit-learn, Streamlit, NLTK, TensorFlow, Keras, PyTorch, ggplot2
Techniques : Machine Learning, Statistics, Deep Learning, NLP, Generative AI, LLMs, Social Media Mining
Soft Skills : Project Management, Leadership, Problem Solving, Effective Communication, Attention to detail

Work Experience

Research Assistant — **May 2024 - Aug 2024**
Kelley School of Business, Indiana University | *Python, NetworkX, Gephi, NodeXL, Topic Modeling* — **Bloomington, IN**

- Performed temporal network and topic modeling analysis on over **145,000 posts** and **623,000 comments** from the r/depression subreddit to explore community structures and engagement trends.
- Applied Latent Dirichlet Allocation (LDA) to identify **5 key themes** in mental health discussions over a three-month period revealing that **30% of user interactions** focused on existential despair, social isolation, and emotional turmoil.
- Created temporal networks to map community structures and developed a **content-content network** using Louvain method and Fruchterman-Reingold force-directed algorithm demonstrating **25% growth** in unified cluster interactions.

Data Analyst Intern — **Jan 2023 - May 2023**
IBM Corporation | *Python, Web Scraping, EDA, Machine Learning, Microsoft Azure* — **Ahmedabad, India**

- Developed a "Heart Disease Prediction System" using machine learning, analyzing a dataset of **30710 patient records** with **14 medical attribute**s achieving a **93.44% accuracy** using Support Vector Machines (SVM).
- Conducted comprehensive **exploratory data analysis (EDA**) on the web-scraped dataset by employing statistical techniques for data cleaning and preprocessing, improving data quality by **98%** & identifying **5 key predictive features.**
- Deployed an interactive web application using Streamlit and Python to predict heart disease risk based on user-input medical parameters, reducing diagnosis time by **60%**. Presented findings at the internship pitch night, securing first place among **50+ participants** and received winner certificate for innovative application of **data science in healthcare**.

Projects

StayEase : Simplified Hotel Booking Solution | *MongoDB, React.js, Vercel, GraphQL*

- Built a comprehensive hotel booking web app with React, Node.js, Express, and MongoDB, featuring advanced search, booking, and management functionalities.
- Implemented **GraphQL** for efficient data handling, optimizing query performance by **30%**, and deployed the application on **Vercel**. Utilized a Kaggle dataset with **4,000 hotels** to support dynamic hotel data management and user interaction, processing over **1,000 queries daily**.

Predictive Modelling for Personalized Diabetes Care | *Data mining, Python, EDA, Deep learning, Tensorflow, Keras*

- Spearheaded data mining techniques such as data cleaning, pre-processing, decision-making, and correlation analysis on a comprehensive dataset comprising **70,692 subjects** from the Behavioral Risk Factor Surveillance System (BRFSS).
- Optimized and fine-tuned machine learning models achieving up to **75.47% accuracy** in predicting diabetes risk using Artificial Neural Networks (ANN) and created a web-based application for real-time risk assessment through streamlit.

Waste Management and Garbage Classification Model | *Python, OpenCV, Streamlit, Deep Learning, PyTorch*

- Developed an AI-powered waste management system as part of the Luddy Hackathon securing rank 4, by utilizing **InceptionV3 transfer learning** and fine-tuning techniques, achieving accuracy rates of **89.37%** and significantly improving waste sorting efficiency and accuracy.
- Integrated **real-time video processing** capabilities into a user-friendly web application using the **Streamlit** framework, enabling efficient waste classification and providing disposal guidance to users.

PUBLICATIONS

Published a paper titled 'Analysing the Public Discourse around OpenAI's Text-to-Video Model 'Sora' using Topic Modeling' in the research category Computers and Society (cs.CY) on **arXiv:2407.13071**.