SNEH PILLAI

https://www.linkedin.com/in/snehpillai/ | https://github.com/snehsuresh
https://snehpillai.vercel.app | snehpillai02@gmail.com | Boston, MA | 774-704-0786

Education

University of Massachusetts, Dartmouth *Dartmouth, MA* **Expected May 2025**

- Master of Science in **Data Science** (Currently Enrolled, **3.96 GPA**)

Relevant Courses: Advanced Mathematical Statistics | Database Design | Applied Business Analytics and Information Visualization | Business Intelligence and Knowledge Management | Big Data Analytics | Advanced Data Mining | Data Visualization | High Performance Scientific Computing

APJ Abdul Kalam Technological University *Kerala, India* **05/2015 - 07/2019**

- Bachelor of Technology, **Computer Science and Engineering**

Professional Experience

Research and Industry Support (Data Science) **University of Massachusetts, Dartmouth** *Dartmouth, MA* **10/2024 - Present**

- **Developed a Bayesian forecasting model** utilizing large language models (Mistral 7B, gpt-j-6B, LLaMa2-8B), processing over 1 million data points and achieving a 95% accuracy in demand projections.
- **Conducted an econometric time series analysis** on 12 macroeconomic indicators and 5 years of historical sales data, identifying key drivers of consumer behavior and enhancing forecast accuracy by evaluating various KPIs.

Data Scientist I **Playerzpot** *Navi Mumbai, MH, India* **08/2022 - 08/2023**

- **Designed a BERT-based model** that enhanced transactional data analysis, surpassing baseline models through masked language modeling and random token detection, contributing to a **76% F1 score** in player retention classification.
- **Analyzed player conversations** with unsupervised clustering techniques and **LLaMA2-13b**, revealing critical factors driving player churn and improving retention strategies.
- **Increased transaction volume by ₹5 million** through the development of recommendation models that optimized player engagement strategies.

Associate Data Scientist **Medfanum Pharmaceuticals** *Remote, India* **10/2021 - 7/2022**

- **Developed a predictive model**, **reducing inventory holding costs by 25%** through accurate demand forecasting for seasonal medications.
- Improved sales **prediction accuracy by 15%** using ARIMA time series analysis in R, enabling better inventory management for critical drugs.

Software Engineer **Organiza Tech Pvt. Ltd.** *Navi Mumbai, MH, India* **10/2019 - 08/2021**

- Enhanced XBRL mapping accuracy from 75% to **98%** by creating an auto-mapping feature in Unifeye using Python and Node.js.
- Reduced **server response times by 35%** for high-concurrency tasks by optimizing server performance with Golang and Redis.
- Led the migration of the KYC system to a Node.js microservices architecture, improving identity verification speed and code maintainability.

Academic Projects

- **Smart City Real-Time Data Streaming Pipeline:** Developed an end-to-end data pipeline using IoT devices, Apache Kafka, and Apache Spark, enhancing urban mobility and decision-making through actionable insights.
- **Real-Time Facial Emotion Detection and Audio Feedback System for the Visually Impaired:** Created a computer vision system with YOLOv5 to detect facial emotions, providing real-time audio feedback for visually impaired users. Link
- **dbLinkPro:** A Python package with Docker integration for MySQL, PostgreSQL, MongoDB, and Cassandra, for CRUD interactions. Link
- **British Airways Review Dashboard:** Designed a dashboard for visualizing customer feedback, aggregating reviews to highlight key metrics such as monthly ratings and comfort scores.
- **EigenFace Detection**: Comparing PCA, 2D PCA, and 2D Square PCA efficiency for face detection. A Data Mining Project. Link
- **Style Transfer for Image Transformation:** Implemented neural style transfer techniques with TensorFlow and PyTorch, allowing users to transform photos into artistic styles.
- **Helmet Detection:** Computer Vision End to end pipeline using PyTorch. Roboflow v5.
- **Brain Tumor Detection**: Uses MATLAB and App Designer for brain tumor detection. Link
- **Assembled a large language model using Python and PyTorch, focusing on tokenization and data compression for various NLP tasks.** Link
- **Query Optimization with Generative AI and TPC-H Data:** Analyze, rewrite, and suggest improvements to queries using React for frontend, Node for backend, Postgre, Gemini Link
- **Adaptive Distillation Pipeline Based on Hardware Constraints:** Innovative technique for compressing large teacher models with HuggingFace, improving model efficiency through adaptive knowledge distillation. Link
- **Reddit Data Pipeline Engineering**: Integration of Reddit, Airflow, Celery, Postgres, S3, AWS Glue, Athena, and Redshift to create a seamless ETL process. Link

SNEH PILLAI

https://www.linkedin.com/in/snehpillai/ | https://github.com/snehsuresh
https://snehpillai.vercel.app | snehpillai02@gmail.com | Boston, MA | 774-704-0786

- **Personalized Recipe Recommendation System:** Selenium for data extraction, BERT for semantic analysis and EnsTM for advanced topic modeling. Link
- **RAG QA**: Demonstrates deploying a Retrieval-Augmented Generation (RAG) application using AWS services such as ECR, and Langchain, Hugging Face, and Docker. Link
- **CDC Data Analysis:** Conducted exploratory data analysis on data from the Centers for Disease Control and Prevention, focusing on US county rates of diabetes, obesity, and lack of physical activity. Applied techniques such as regression and clustering for effective data curation, querying, and aggregation to derive actionable insights. Link
- **Fatal Police Shooting Analysis**: Analyzed data from the Washington Post data repository on fatal police shootings in the United States. Dataset: Utilized data on fatal police shootings from the Washington Post.
- **Economic Indicators Analysis**: Analyzed economic data from Analyze Boston open data hub by applying stochastic processes such as ARIMA, SARIMAX models, and STL decomposition for trend, seasonality, and residual analysis. Time series analysis Link

Research Experience

- **(Ongoing) Enhancing Demand and Risk Forecasts with Macroeconomic Data:** Developing a Bayesian forecasting model combining macroeconomic indicators with historical sales data to improve forecasting accuracy.
- **(Ongoing) Optimizing Knowledge Distillation for Balanced Cloud and On-Device Inference Workloads:** Creating a knowledge distillation pipeline to balance computational loads, optimizing resource usage and model performance.
- **(Ongoing) Keyword Masking for LLM Privacy Protection:** Innovating keyword masking techniques for large language models to prevent inference of sensitive personal attributes.
- **Enhancing Supply Chain Management through Business Intelligence and Knowledge Management:** Integrating BI with SCM and KM practices to enhance transparency and optimize costs in supply chain processes.

Skills

- **Language**: Python | JavaScript | Typescript | R Language | C/C++
- **Artificial Intelligence / Machine Learning :** Machine Learning | Deep Learning | PyTorch | GPT | LangChain | Few-shot classification | Retrieval Augmented Generation and Fine Tuning | Transformers | YOLO Object Detection | Topic Modelling (LDA, NMF, BERTopic) | Autoregressive Models | Diffusion Models | Generative Adversarial Networks (GANs) | Prompt Engineering
- **Cloud Computing and Big Data Technologies:** AWS | Kubernetes | Terraform | Celery | Azure | GCP
- **Databases / DBMS:** MySQL | PostgreSql | MongoDB | Cassandra | CosmosDB | Redis
- **Frameworks and Libraries:** Node | Express | React | Numpy | Pandas | Firebase
- **DevOps and Automation:** Docker | Airflow | CI/CD | MLFlow | DVC | Github Actions | Jenkins
- **Data Visualization:** Tableau | Amazon Quicksight | D3.js
- Stochastic Processes and Forecasting | CUDA | OpenMP | Linear Algebra | Data Structures & Algorithms | OOP | Full-stack Development | BI Tools | Knowledge Management | Shell Scripting

Certifications

- **Inferential Statistics,** Coursera
- **Machine Learning Course,** Udemy
- **Introduction to Probability and Data with R,** Coursera
- **Python Mega course,** Udemy