Arthi Murali

9940627763 · meetarthi@gmail.com · meetarthi.com
linkedin.com/in/meetarthi · github.com/meetarthi · medium.com/@meetarthi · kaggle.com/arthimurali

---

I'm a Data Scientist with expertise in Large Language Models (LLMs), OCR, and ML Models. I have experience with OpenAI API, Tesseract, SpaCy, and audio-to-text models. I build data pipelines using SQL, MongoDB, Apache Airflow, and AWS, and created user-friendly UIs with Streamlit and Chainlit. My skills include data transformation with Pandas, chatbot development, and server monitoring. I thrive on new challenges and enjoy contributing to organizational success through AI-based automation and data-driven projects. I'm eager to apply my skills to drive innovation and deliver high-impact results.

WORK

Heyram infrastructure

ML engineer Intern
During my internship, I developed an interactive learning application using Streamlit, hosted on Hugging-Face Spaces, for middle school children. It featured entity recognition, parts of speech tagging, image-to-text, and audio-to-text conversion powered Spacy, Tesseract, and OpenAI APIs

iMorph

Product Development Intern
During my product internship tenure at iMorph, I was responsible for prototyping and building micro products. I led the initiative for python assessment bot which is based on openai assistant used to screen candidates and tech news aggregator tool that is used for building datasets.

EXPERIENCE/PROJECTS

1. Innovative AI Playground

AI leveraged Application for enhancing the learning journey of middle school children through interactive learning developed using streamlit framework and hosted on huggingface spaces. Application features four categories: entity recognition, parts of speech, image to text, and audio to text recognition. Employed Spacy and NER for entity recognition and POS tagging, CSS for providing a visual cue for different grammatical categories, tesseract for OCR allowing the app to extract text from uploaded images, OpenAi API features for extracting information for user input (search term) and Audio to text conversion.
Technology - Streamlit, Openai key, spacy ,tesseract, Css, regex

2. Generating Summary of Startup Based on Company Profile for Investors using Language Models

Developed a startup application summarisation tool to automate the summarisation process and provide the investor with the overview of company, financials, team, and products. Extracted text from startup application PDF, used precise prompts, and stored generated summaries in a text file. Initially used Falcon-7b, llama-2-7b, and llama-2-13b models. The OpenAPI model outperformed others by giving a more concise and accurate summary of the company.
Technology - Openai key, langchain, Language Models(Falcon-7b, Llama-2-7b and Llama-2-13b), PyPDF2, File handling

3. Chest X-ray Classification Using Deep Learning for Automated Detection of Tuberculosis

Developed an automatic tuberculosis (TB) detection system using deep learning models, specifically convolutional neural networks (CNNs). Implemented data augmentation techniques and preprocessing steps to improve model performance. Worked on feature engineering to identify important attributes contributing to classification. Evaluated performance using various metrics like accuracy, sensitivity, specificity, precision, AUC, and F1 score. Achieved a 97.62% accuracy on the test set.
Technology : HTML, CSS, Python, JavaScript, Kaggle image datasets, Keras, TensorFlow, Wamp Server, Google Collab, Visual Studio, pickle, and Pandas.

4. Python Assessment Bot

Developed a Python assessment bot using Chainlit to evaluate individual's proficiency in Python. Leveraged Chainlit to create a question and answer platform, prompts from the assistant ID will be interpreted to generate questions and format the feedback. Used Openai to interpret the answer of the user and give relevant feedback and score. Implemented a scoring system for user response, stores scores in a database and upon completion the Final score would be displayed . Customised the user interface for better a presentation.

Technology - Chainlit, Openai key, Assistant ID, Css, sqlite3, Visual Studio

---

MINOR PROJECTS

1. Forecasting whether the patient has Diabetes based on symptoms from Pima Indian Diabetes dataset, using comparative analysis of machine learning algorithms - **Decision Tree, Random Forests, SVC, XGBoost, Naive Bayes,** and **AutoSklearn**.
2. Amazon privileges access prediction based on the employer role using **H2O AutoML**.
3. Tweet's scrapping app to extract tweet data from User-Specified Details.
4. PhonePe Data Visualisation Tool - Insights into Digital Payments Using **Plotly**, and **Geopandas**
5. YouTube Data Harvesting and Warehousing using **MongoDB** and **SQL** to analyse YouTube channel information.
6. AWS Lambda leveraged Server Monitoring and Slack Notification System using **AWS Lambda, CloudWatch, Slack API**
7. Backend API for E-commerce Website with User Registration, Product Rating, Sorting using **Flask**.
8. Predicting the need of specific target drug for a patient from Longitudinal Record using **Xgboost**.
9. Tech News Aggregator - CLI-based RSS Scraper for Technology News Using **Requests, SQLite3,** and **Feedparser**

SKILLS

Data Engineering

- Pandas - Data manipulation, analysis, and handling
- Cloud Computing (AWS) - S3, EC2, Lambda, Glue, Athena, CloudWatch, IAM
- Databases - SQL, MongoDB (Compass), AWS(DynamoDB, RDS)
- Big Data & Data Processing Tools - Hadoop, Airflow, PySpark, Prometheus, Grafana
- DevOps Tools - Docker, Cron Jobs, Git

Data Science

- Language models - Falcon, llama, OpenAi
- Statistics - Standard Deviation, Covariance, Correlation, A/B Testing
- Data Visualisation and analysis- Numpy, Matplotlib, Seaborn, Plotly
- Machine Learning - Bayesian method, Decision Trees, Random Trees, XGBoost, Support Vector Machine(SVM), KNN, Principal Component Analysis (PCA), Clustering, NLP.
- Deep learning - Convolutional Neural Network (CNN)
- Recommender Systems - User-Based Filtering, Item-Based Filtering
- AutoML - H2O, TPOT
- Platform and Tools - HuggingFace, Pickle, Gradio, Pandas Profiling

---

EDUCATION

Master Data Engineering
GUVI, Chennai

Bachelor of Science Honors
Biomedical Sciences
Sri Ramachandra Institute of Higher Education and Research, Chennai

Higher Secondary Education
Kumararani Meena Muthiah Matric Higher Secondary School, Chennai

Secondary Education
Kendriya Vidyalaya CLRI, Chennai

CERTIFICATION

- Machine Learning, Data Science and Deep Learning with Python
- The Complete Data Structures and Algorithms Course in Python

LANGUAGES

- English - Full Professional Proficiency
- Hindi - Limited Working Proficiency
- Tamil - Native / Bilingual