Sooraj Sreekumar

sooraj12sreekumar@gmail.com | 8122152267 | LinkedIn | GitHub

Education

Vellore Institute of Technology, Bhopal- M Tech Integrated In Computer Science and Engineering specialized in Computational and Data Science — Sept 2021 – May 2026

- GPA: 8.37/10

Technical Skills

Programming: Python, SQL.
ML / AI: Scikit-learn, TensorFlow, HuggingFace Transformers.
NLP & CV: NLTK, OpenCV, OCR (Tesseract, PaddleOCR).
Data & Visualization: Pandas, NumPy, Tableau, Power BI.
Tools: Git, Streamlit, LabelStudio, LabelImg.

Experience

AI Intern, Kumaran Systems, Chennai — Sept 2024 – Jan 2025

- Built high-quality computer vision training datasets by annotating 5,000+ images, directly contributing to improved OCR accuracy and model robustness. extraction efficiency.
- Designed and implemented end-to-end OCR pipelines using PyTesseract, PaddleOCR, and PyOCR with OpenCV-based preprocessing to automate text extraction from noisy, real-world images.
- Collected, cleaned, and standardized 10,000+ images, establishing a scalable data preprocessing workflow to enhance downstream model training quality.
- Tech Stack: Python, OpenCV, Pillow, PyTesseract, PaddleOCR, PyOCR, LabelStudio, LabelImg.

Projects

Resume Classification System Using BERT [GitHub-link] — June 2025 – Aug 2025

- Built an automated resume classification system using DistilBERT, trained on 14,000+ resumes, to categorize resumes across multiple job roles with 78% accuracy on unseen data.
- Implemented a modular training and evaluation workflow with saved tokenizers, label encoders, and model checkpoints, enabling reproducible and scalable inference.
- Tech Stack: Python, HuggingFace Transformers, DistilBERT, Pandas, Numpy, Scikit-learn, Git.

AI-Powered Study Assistant [GitHub-link] — July 2025 – Aug 2025

- Built a document-aware AI study assistant using Google GenAI, enabling users to upload PDFs and generate summaries, quizzes, and context-aware Q&A.
- Designed a modular backend architecture for PDF parsing, prompt-driven LLM interaction, and dynamic quiz generation, integrated with a Streamlit frontend.
- Tech Stack: Python, Streamlit, Google Gemini API, PyMuPDF, dotenv, GitHub.

Multi-Sensor Carbon Monoxide Prognosis System [GitHub-link] — May 2025 – June 2025

- Developed and validated a Carbon Monoxide prediction system using multi-sensor IoT data, achieving R^2=0.9963 for accurate environmental risk forecasting.
- Deployed a Streamlit-based web application for real-time CO concentration prediction, featuring user-driven inputs and clear risk-level interpretations for actionable safety decisions.
- Tech Stack: Python, Pandas, NumPy, Scikit-learn, Streamlit, Matplotlib, Seaborn.

Certifications and Trainings

Finlatics: Generative AI Experience Certification — July 2025 - Sept 2025
Coursera: The Data Scientist Toolbox – Johns Hopkins University. — Nov 2022 - Dec 2022
Udemy: Statistics for Data Science and Business Analysis. — July 2023 - Aug 2023

Technical Achievements

- HackerRank: Achieved five-star rating in SQL.
- Solved 100+ Data Structures and Algorithms problems on GeeksForGeeks; Institution Rank: 325.