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, R, SQL.
ML Frameworks: TensorFlow, Scikit-learn, HuggingFace Transformers, NLTK, OpenCV.
Data Tools: Pandas, NumPy, Git, Tableau, Power BI, LabelImg, LabelStudio.
Domains: Machine Learning, Deep Learning, NLP, Computer Vision, Generative AI, OCR, Data Annotation.

Experience

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

- Annotated 5000+ images using LabelStudio and LabelImg for computer vision model training extraction efficiency.
- Developed OCR pipelines with PyTesseract, PaddleOCR and PyOCR, and automated preprocessing using OpenCV and Pillow.
- Collected and preprocessed 10,000+ image datasets to enhance model training quality.
- Tools: Python, OpenCV, Pillow, PyTesseract, PaddleOCR, PyOCR, LabelStudio, LabelImg.

Projects

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

- Developed a resume classification system leveraging DistilBERT (transformer-based deep learning model), achieving 78% accuracy in categorizing resumes across diverse job roles.
- Collected and preprocessed resumes from multiple online sources, building a robust dataset pipeline to support effective model training and evaluation.
- Tools Used: Python, Transformers, DistilBERT, Pandas, Numpy, Scikit-learn, Google Gemini API, GitHub API.

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

- Built an interactive study assistant using Google GenAI API to generate summaries, quizzes, and resolve doubts from uploaded PDFs.
- Integrated Streamlit frontend with modular backend for PDF parsing, context-aware QnA, and dynamic quiz generation.
- Tools Used: Python, Streamlit, Google Gemini API, PyMuPDF, dotenv, GitHub.

Lung Cancer Risk Prediction System [GitHub-link] — May 2025 – June 2025

- Engineered a binary classification model to predict patient lung cancer risk (Yes/No) by analyzing 15 clinical and lifestyle features, including age, smoking status, and respiratory symptoms.
- Optimized a Decision Tree Classifier to achieve 85.5% prediction accuracy and a 0.91 weighted avg F1-score, effectively identifying high-risk individuals for early medical intervention..
- Tools Used: Python, Pandas, Numpy, Scikit-learn, Matplotlib, GitHub.

Certifications and Trainings

Coursera: Applied Machine Learning in Python – University of Michigan. — Jan 2023
Coursera: The Data Scientist Toolbox – Johns Hopkins University. — Dec 2022
Udemy: Statistics for Data Science and Business Analysis. — Aug 2023

Extracurricular Activites

HackerRank: Five-star rank in HackerRank for SQL.
DSA: Solved 200+ DSA questions on GeeksForGeeks, Institution Rank- 325.
Club: Member of Data Science Club at VIT Bhopal (Nov 2021 – Oct 2022).