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: 8.35/10

Technical Skills

Programming: Python, R, SQL.
ML Frameworks: TensorFlow, Scikit-learn, NLTK, OpenCV .
Data Tools: Pandas, NumPy, Git, Tableau, Power BI, LabelImg, LabelStudio.
Domains: Machine Learning, Computer Vision, NLP, 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

Automated Resume Classification System — June 2024 – Aug 2024

- Developed and optimized 5 ML models (Random Forest, SVM, Naïve Bayes, Logistic Regression, KNN) for automated resume classification, with top models achieving 99.5 percent accuracy in candidate qualification.
- Designed and implemented a robust text processing pipeline (tokenization,lemmatization,TF-IDF) that became the foundation for all model's high accuracy (98.4-99.5 percent).
- Tools Used: Python, Scikit-learn, NLTK, Pandas, Numpy, Matplotlib, GitHub.

IoT-Based Air Quality Prediction Model — Oct 2023 – Dec 2023

- Developed and optimized Random Forest/Gradient Boosting models for real-time CO prediction from IoT sensors (R^2: 0.85-0.9, RMSE: 0.7), employing feature engineering and time-series analysis to enhance accuracy and ensure production-ready robustness.
- Built end-to-end data pipelines for processing streaming sensor data and generating actionable air quality insights, enabling continuous environmental monitoring.
- Tools Used: Python, Scikit learn, Pandas, numpy, matplotlib, Seaborn, GitHub.

Early-Stage Lung Cancer Detection Classifier — Aug 2022 – Oct 2022

- Developed and evaluated predictive models (Logistic Regression, KNN, Decision Tree, SVM, Naïve Bayes) for early lung cancer detection, with Logistic Regression achieving 91.3 percent accuracy.
- Conducted end-to-end data preprocessing and model evaluation using confusion matrices, accuracy, and F1-scores for clinical reliability.
- 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 100+ DSA questions on GeeksForGeeks, Institution Rank- 325.
Club: Member of Data Science Club at VIT Bhopal (Nov 2021 – Oct 2022).