Soham Pahari

Github: github.com/suhanpahari
Portfolio: SuhanXD.github.io
Leetcode: leetcode.com/suhanpahari

Email: paharisoham@gmail.com
Mobile: +91-9064520673
LinkedIn: linkedin.com/in/sohampahari

Education

University of Petroleum and Energy Studies Dehradun, India
Bachelor of Technology - Computer Science & Engineering; *2022 – 2026*
Contai High School Contai, India
Higher Secondary Education - (Mathematics, Statistics); Marks: 88.4% *2020 – 2022*

Skills Summary

Programming Languages:	Python, Java
Python Libraries:	Pandas, NumPy, Matplotlib, Scikit-learn, NLTK, Flask, Hugging Face, TensorFlow
Visualization Tools:	Power BI, Tableau, Talend Open Studio
Data Engineering Skills/Tools:	SQL, ETL, Data Streaming, AWS, Snowflake, Apache Spark, Apache Airflow
Soft Skills:	Communication, Problem Solving, Active Learner, Critical Thinking, Creativity

Internship Experience

Bahas Pvt Ltd Remote
ML Development Intern *May 2024 – August 2024*

- **Developed Multi-Model Classification System**: Built and deployed multiple models including Fine-Tuned BERT, Random Forest, SVM, Logistic Regression, and Naive Bayes for emotion classification tasks.
- **Implemented BERT-Based Model**: Utilized the bert-base-multilingual-uncased model with mBERT for advanced text tokenization and embeddings, significantly enhancing accuracy in emotion predictions.
- **Designed and Deployed Streamlit Application**: Created an interactive application enabling real-time model selection and emotion classification for user inputs, leading to improved user engagement.
- **Model Optimization**: Fine-tuned, optimized, and executed advanced hyperparameter tuning techniques, resulting in enhanced performance and accuracy across models.
- **Technologies Used**: Python, TensorFlow, Hugging Face Transformers, Scikit-Learn, Streamlit, mBERT, TF-IDF, Numpy, Pandas.

Projects

- **Emotion Classification System**:
  - Developed an advanced emotion classification system for Bengali text using multiple machine learning models, including Fine-Tuned BERT, Custom BERT, Random Forest, SVM, Logistic Regression, and Naive Bayes.
  - Improved model efficiency by 16.23% through optimization techniques, resulting in a system with 87.4% accuracy in emotion prediction tasks.
  - Integrated the models into a user-friendly Streamlit application that enables real-time model selection and emotion classification based on user inputs.
  - Utilized mBERT for tokenization, significantly enhancing the accuracy of emotion detection by capturing contextual nuances in the Bengali language.
  - Technologies used: Python, TensorFlow, scikit-learn, Hugging Face Transformers, Streamlit — GitHubLink

- **Delhi Pollution Prediction in Time Series with Sequential Models**:
  - Developed a model to tackle Delhi's pollution issues by leveraging ARIMA, LSTM, and custom hybrid metaheuristic algorithm (Dung Beetle Algorithm, Quantum Swarm Algorithm, Hybrid Genetic Algorithm, Red Deer Algorithm, and Gravitational Algorithm).
  - Increased model efficiency by 15%, while improving prediction accuracy by 9-10% through advanced optimization techniques.
  - Focused on minimizing error and optimizing model performance, ensuring long-term, reliable pollution forecasting.
  - The project is in its final phase, promising scalable solutions for future environmental challenges.
  - Technologies used: TensorFlow, Keras, scikit-learn, pmdarima (for ARIMA, SARIMAX). — GitHubLink

- **Differential Gene Expression Analysis on GEO Datasets**:
  - Conducted differential gene expression analysis on GEO datasets, focusing on GSE199135 to uncover biological insights.
  - Focused on minimizing error and optimizing model performance, ensuring long-term, reliable pollution forecasting.
  - Applied statistical testing to identify significantly differentially expressed genes between sample groups.
  - Conducted functional enrichment analysis to link differentially expressed genes to biological pathways.
  - Visualized results using heatmaps, volcano plots, and enriched pathway diagrams for clear interpretation.
  - Tools used: R, limma, ggplot2, GEOquery, Bioconductor packages. — GitHubLink