Hitesh Matharu

 github.com/HiteshM30 linkedin.com/in/hitesh-matharu hitesh04matharu@gmail.com +91-7888933406

EDUCATION

University of Petroleum and Energy Studies   Dehradun, India
B.Tech (Hons.) Computer Science (AI/ML)   *Aug 2022 – Current — Current GPA: 7.6*

EXPERIENCE

Data Analyst Intern | *Cognifyz Technologies — Remote*   Mar 2025 – Apr 2025
- Cleaned, transformed, and standardized large datasets containing 10,000+ records using Python and Excel, eliminating duplicates and formatting inconsistencies, which improved overall data quality and reliability by 25%.
- Performed comprehensive exploratory data analysis (EDA) to uncover hidden patterns, correlations, and trends within complex datasets, resulting in actionable insights that informed strategic decisions for over three business use cases.
- Designed and developed visually engaging dashboards and charts using tools like Matplotlib and Excel, presenting key findings to non-technical stakeholders in a clear, concise manner and facilitating better understanding of data-driven outcomes.

PROJECTS

Next-Word Prediction Model | *Python, PyTorch, TensorFlow*   Nov 2024
- Constructed an **LSTM-based machine learning model** achieving **90% accuracy** in predicting the next word.
- Handled over **2 million text entries** from diverse sources, including **Wikipedia** and **OpenAI GPT** datasets, optimizing training time by **25%**.
- Implemented text preprocessing techniques (**tokenization**, **stemming**, and **lemmatization**) to reduce dataset noise, improving model accuracy by **5%**.
- Reduced model training time by **30%** by leveraging **GPU acceleration** and optimizing batch sizes.

Depression Diagnosis Prediction Model | *Python, XGBoost, RandomForest, Streamlit, scikit-learn, Pandas* Jan 2025
- Engineered a predictive model using **XGBoost**, achieving **precision: 0.88**, **recall: 0.85**, and **F1 score: 0.86** by analyzing **1,000+ mental health survey records**.
- Deployed **XGBoost** as the primary classification model after rigorous testing against **RandomForest**, resulting in a **15%** decrease in **misclassification** rates and enhancing lead conversion efficiency across marketing campaigns.
- Enhanced preprocessing efficiency by automating data cleaning and feature engineering with **SimpleImputer** and **LabelEncoder**, reducing manual intervention by **30%**.
- Developed a user-friendly **Streamlit** web application for real-time predictions, boosting accessibility and deployment speed.
- Utilized **pickle** for efficient model serialization, improving model loading and inference time by **25%**.

PDF Summarizer from Scratch | *Python, NLTK*   Nov 2024
- Designed an **extractive summarizer** using **TF-IDF**, achieving **85% relevance accuracy** in summarizing text inputs, reducing document size by an average of **60%** while retaining key information.
- Dealt with over **10,000 text entries**, achieving a **25%** reduction in preprocessing time by implementing optimized **tokenization**, **stop word removal**, and **lemmatization** pipelines.
- Evaluated the summarizer on a dataset of **5,000+ documents**, achieving a **20%** improvement in processing efficiency compared to baseline models, and reducing the average runtime to under **2 seconds** per document.

SKILLS

Programming Languages: Python, Java, C, JavaScript, SQL
Tools/Platforms: Git/GitHub, VS Code, IntelliJ IDEA, Jupyter Notebook
Libraries/Frameworks: Pandas, NumPy, Matplotlib, Scikit-learn, TensorFlow

ACADEMIC/EXTRA-CURRICULAR ACHIEVEMENTS

Cloud Security Alliance Student Chapter, UPES | *Public Relations Head*   Sept 2022 – Present
- Conducted a series of 15 public speaking competitions for students, fostering a competitive but supportive environment; encouraged over 75% of participants to pursue further training in public speaking and leadership development.

CERTIFICATIONS

Accenture North America Data Analytics and Visualization Job Simulation on Forage — Nov 2024
- Participated in a social media advisory simulation as a Data Analyst, delivering insights to a hypothetical client.
- Processed and analyzed 7 datasets, identifying key content trends that informed 3+ strategic recommendations for improved engagement and reach.