# KISHORE KUMAR B

**Data Scientist**

Madurai South, TN | +91 8428500923 | Portfolio | Github | Linkedin | kishore.kumar4907@gmail.com


## OBJECTIVE

Aspiring Data Scientist with a strong foundation in Python, SQL, and machine learning, experienced in developing predictive models, deploying interactive data apps, and working with large datasets using PySpark and Scikit-learn. Passionate about extracting actionable insights from data and building AI-powered solutions that drive business impact.

## EDUCATION

**PSNA College of Engineering and Technology - Dindigul** **2020 - 2024**

***Bachelor of Technology, Information Technology*** **CGPA: 8.71**

## SKILLS

- **Programming & Query Language**: Python, SQL, Java, HTML, CSS
- **Data Analysis & Visualization**: Power BI, Pandas, Numpy, Matplotlib, Seaborn, Plotly, Excel, Google Sheets, Tableau
- **Machine Learning & Frameworks**: Scikit-learn, TensorFlow, PyTorch, OpenCV, Pillow, Mediapipe, NLTK, PySpark, LangChain
- **Deployment & Tools:** Streamlit, Flask, FastAPI, Docker, Git, GitHub, Google Colab, Cursor
- **Cloud & Platforms:** AWS, Microsoft Office
- **Languages**: Tamil, English, Hindi

## WORK EXPERIENCE

### Data Scientist | Code Clause (Remote) June 2024 – July 2024

- Developed a movie genre prediction model using Python, **NLP**, and machine learning techniques, leveraging **TF-IDF** vectorization and a **Multinomial Naive Bayes** classifier for classifying movie synopses. Achieved **89%** accuracy in genre prediction model on test data.
- Implemented robust data preprocessing, feature extraction, hyperparameter tuning, and model evaluation using Pandas, **NLTK**, **Scikit-learn**, and **GridSearchCV** to optimize performance.
- Developed a heart disease risk prediction web app with Python, Streamlit, and Scikit-learn, incorporating dynamic dashboard elements to provide users with personalized risk assessments and lifestyle recommendations.

### Data Analyst | Turing (Remote) July 2024 - Present

- Created advanced data visualizations such as swarm plots, heatmaps, waffle plots, Sankey diagrams, 3D charts, and more to support the training of large language models (**LLMs**). Used various libraries like Matplotlib, Seaborn, Plotly, Altair, Bokeh, NumPy, Pandas.
- Created diverse image manipulation datasets to evaluate Gemini's understanding of deterministic image transformations (resize, skew, brightness, contrast, vibrance, posterize, etc.). Generated input images using prompts and crafted both specific and ambiguous queries per task. Used Python (Pillow/OpenCV) for image operations.

## ACADEMIC PROJECTS

### Heart Disease Risk Prediction Jun 2023 - Jul 2023

- Implemented a predictive web application using **Python and Streamlit** to assess heart disease risk based on user-inputted health parameters like cholesterol, blood pressure, and BMI. Deployed a Streamlit web app used by **100+ users** for health risk assessment.
- Trained a classification model using **Scikit-learn (Logistic Regression)** and processed medical data with **Pandas**, including handling missing values, feature scaling, and input validation.

### Resume Analyzer – Resume Scoring & Optimization Tool Dec 2023 – Jan 2024

- Built an **AI-powered resume analyzer** using **Streamlit**, integrating **Gemini API** to enhance resume evaluation by generating context-aware feedback and improvement suggestions.
- Implemented **NLP**-based **text extraction** and **analysis** using **spaCy, NLTK**, and **PyPDF2** to assess resume content for **skills, experience, education, achievements**, and formatting.
- **Provided real-time scoring and enhancement recommendations** across various input formats (PDF, DOCX, TXT), helping users optimize resumes for targeted job roles.

### Churn Prediction Using PySpark Feb 2024 – Mar 2024

- Developed a **PySpark**-based churn prediction pipeline to analyze large-scale telecom customer data and identify churn risk using distributed data processing. Optimized data processing time by **30%** using efficient PySpark transformations and caching strategies.
- Trained a **Gradient Boosted Tree (GBT) model** with **Spark MLlib** and evaluated performance using accuracy, **ROC-AUC**, and **confusion matrix** metrics on test data.
- Demonstrated expertise in PySpark, MLlib, and scalable machine learning model development for predictive analytics.

### Retail Sales Dashboard

- Designed and deployed an interactive **Power BI dashboard** to analyze retail sales data across **item types, outlet sizes, establishment years, and fat content categories**, enhancing business decision-making.
- Performed end-to-end **data cleaning and transformation (ETL)** using Excel and Power Query, handled missing values, corrected data types, created calculated **m**etrics and normalized data for consistent visualization.
- Developed **dynamic KPIs and visuals,** implemented **slicer filters**, and applied a custom yellow-black Blinkit-style theme for user experience.

## CERTIFICATIONS

- **Fundamentals of Agents**: Hugging Face - February 2025: Acquired expertise in AI agent fundamentals, automation workflows, and model integration.
- **LangChain for LLM Application Development –** DeepLearning.AI - July 2025: Gained hands-on experience in building LLM-powered apps using chains, agents, memories, and custom data.