Yuvraj Singh

yuvraj.singh212@gmail.com 7354460006 LinkedIn GitHub

Education

- **B.tech in Computer Science** *(e-commerce technology)*, VIT, Bhopal (May, 2026) **8.02 CGPA**

Technical Skills

- **Languages & Databases:** Python, PostgreSQL, Bash, Azure PostgreSQL
- **Data Libraries & Tools:** Pandas, PySpark, NumPy, Streamlit, RegEx, Apache Airflow, Docker, Git
- **Technical Proficiencies:** Web Crawling, Web Scraping, ETL Pipelines, API Integration, Automation Scripts

Projects

Shopinion

[Python, Apache Airflow, Docker, Selenium-Stealth, Pandas, Azure PostgreSQL, ETL]

- **Situation:** A project to train a **context-aware BERT sentiment model** required a **large-scale dataset** of over 100,000+ product reviews, but manual collection was unfeasible due to Flipkart's robust **anti-scraping measures**.
- **Task:** Engineer a fully automated, **end-to-end ETL pipeline** to autonomously handle the entire data lifecycle—from stealthy web scraping against a protected site to the structured storage of only unique entries and provide a simple interface for others to use.
- **Action:** Orchestrated the entire workflow using **Apache Airflow** and **Docker**, designing DAGs for both sequential and parallel execution. Engineered a resilient scraper with **Selenium-Stealth** to bypass anti-bot measures, **automatically extracting over 100,000+ clean, unique reviews in under 8 hours**. Implemented an intelligent data loading module in **Pandas** that performs a **pre-emptive check** against **Azure PostgreSQL** to ensure idempotent writes and prevent data duplication.
- **Result:** Deployed a robust, **self-service data collection platform** that fully automates the complex scraping and cleaning process. The parallel processing mode, running in controlled batches of three, demonstrated a **20-22% improvement in execution time** over sequential scraping. The system now enables others to self-serve the creation of large-scale, clean datasets for their own analysis projects.

Outbreak Tracker

[Random-Forest, Python, Pandas, Streamlit, RDBMS, SQL]

- **Situation:** Motivated by a personal experience with a localized jaundice outbreak, I identified that standard symptom-checkers often fail to distinguish between diseases with similar symptoms without crucial **regional context**.
- **Task:** Design and deploy a **proof-of-concept dashboard** to predict diseases with higher accuracy by integrating location-specific, simulating realistic case counts from an **SQL database** into its diagnostic logic.
- **Action:** Engineered a synthetic dataset of **100,000+ records** to model 8 diseases across 8 Indian states. Trained a **Random Forest Classifier** to a baseline accuracy of **92%** and developed a function to dynamically adjust prediction confidence based on active regional case counts.
- **Result:** The deployed dashboard increased diagnostic accuracy by **5-10%** for diseases with overlapping symptoms compared to the context-free model. This approach led to a **22% reduction** in potential misclassifications for critical diseases in high-prevalence zones.

Advanced Data Cleaning and Feature Engineering [GitHub]

[Python, Pandas, RegEx, Data Cleaning, ETL, Feature Engineering]

- **Transformed** a raw **9,999-entry dataset** into a validated set of **6,816 unique records** by scripting the removal of 3,183 duplicates and parsing inconsistent text into structured **start/end date columns** using Pandas and RegEx.
- **Engineered** a new, high-fidelity **Type** column by developing a **logical pipeline** that first classified content based on the presence of a **Director** tag, then refined categories using runtime data (runtime $\leq$ 40 min → **Short-movie**) and genre-based **keyword matching** (**Animation**, **Documentary**).
- **Eliminated** data gaps by imputing over 1,500 missing **Runtime** values (˜22% of data) and over 1,000 missing **Rating** values (˜13% of data) using **context-aware group means**—a **mathematically more robust** approach than using a single global average.

Extracurricular Activities

- **SGFI National-level Basketball Player**; represented at multiple regional and inter-school tournaments
- Member of **university sports council**, contributing to **planning and execution** of intra-college leagues

Certifications

- **Data Engineer Associate** (Data Camp, Feb 2025)
  Learned ETL workflows, SQL for data modeling, and database design using PostgreSQL and Snowflake.
- **Financial Modeling And Valuation** (Internshala, Sep 2021)
  Learned corporate finance fundamentals including DCF modeling, ratio analysis, valuation techniques, and financial forecasting using Excel.