Rhutam Mahajan

College Park, MD +1 (227) 275-1067 rhutammahajan@gmail.com
linkedin.com/in/rhutammahajan github.com/Rhutam03 rhutammahajan.com

Professional Summary

I am grad student at UMD pursuing Master's in Data Science with experience in Python, SQL, machine learning, NLP, time series forecasting, and backend deployment. Built and evaluated ML systems across 5,067 medical triage cases, 13,135 plant disease images, and 149,578 IPL records, with projects spanning multimodal and few-shot learning, forecasting, SQL analytics, and FastAPI applications.

Education

University of Maryland, College Park — Aug 2025 – May 2027
Master of Science in Data Science — College Park, MD

- GPA: 4.0/4.0
- Coursework: Probability and Statistics, Principles of Data Science, Principles of Machine Learning

Savitribai Phule Pune University — Dec 2021 – June 2025
Bachelor of Engineering in Computer Engineering — Pune, India

- GPA: 8.5/10
- Honors in Data Science

Technical Skills

Programming Languages: Python, SQL, Java, JavaScript, HTML, CSS
Machine Learning and AI: Scikit-learn, PyTorch, TensorFlow, Keras, Transformers, NLP, Neural Networks, Classification, Few Shot Learning, Multimodal Learning, Time Series Forecasting, Feature Engineering, Model Evaluation
Data Analysis and Scientific Computing: Pandas, NumPy, SciPy, Statsmodels, pmdarima, Matplotlib, Jupyter Notebook
Data Engineering and Databases: Apache Spark, PySpark, Neo4j, MySQL, Oracle SQL, SQLite, MongoDB
Backend, Deployment, and Tools: FastAPI, Flask, REST APIs, Docker, AWS, Git, Tableau, Power BI

Experience

InternPe — Feb 2024 – Mar 2024
AI/ML Intern — Remote

- Engineered machine learning pipelines for 4 prediction tasks, including cricket match outcomes, diabetes prediction, car price regression, and breast cancer classification using Python, scikit-learn, TensorFlow, and Pandas.
- Processed 149,578 ball-by-ball IPL records by cleaning missing values, engineering live match context features, and preparing train-test splits for match outcome prediction.
- Benchmarked Logistic Regression and Random Forest models for cricket prediction, improving test accuracy from 71.11% to 94.87% through feature engineering and model comparison.
- Modeled structured datasets of up to 179,078 records, achieving 77.27% diabetes SVM accuracy, 0.9210 R^2 for car price regression, and 96.49% breast cancer neural network accuracy.

Gamaka AI — Dec 2023 – Mar 2024
Junior Data Scientist — Pune, India

- Audited structured datasets with SQL and statistical checks, identifying missing values, duplicate records, inconsistent categories, and unusual entries to improve reporting reliability.
- Designed relational database schemas for financial, airline, mobile shop, and ride-sharing systems, modeling users, drivers, rides, payments, aircraft, flights, passengers, reservations, and staff.
- Built a CRUD-based expense tracker with signup, login, add, view, search, update, and delete workflows using Python, JSON storage, and date-based expense filtering.
- Wrote SQL queries on financial records across 7 companies from 2010 to 2012, analyzing revenue, expenses, profit, earnings per share, yearly totals, and company-level trends.

Projects

Multimodal Medical Triage — Dec 2025 – Mar 2026

- Architected an end-to-end triage system combining lesion images and clinical text with a ResNet-18 image encoder, 128-dimensional text encoder, and fused classifier for Low, Medium, and High Risk prediction.
- Trained the 3-class model using an 80/20 stratified split, class-weighted cross-entropy loss, AdamW optimization, batch size 16, max text length 48, and 12 epochs.
- Validated performance on 5,067 cases, achieving 80.13% accuracy, 77.32% macro F1 score, 77.48% balanced accuracy, and 80.27% weighted F1 score.
- Prioritized high-risk case detection with 78.33% recall and 74.71% F1 score, correctly identifying 1,229 of 1,569 high-risk validation cases.
- Implemented a FastAPI and React TypeScript application with 4 prediction routes, image upload, confidence scores, class probabilities, 3 clinical note templates, and recent case history.

Plant Leaf Disease Classification using Few Shot Learning — Aug 2024 – May 2025

- Designed a multilevel few-shot plant disease classifier using EfficientNet-B0 and prototypical networks, cleaning 38 source classes down to 12 high-consistency disease classes.
- Curated few-shot train and validation splits with 240 total images, including 120 training images and 120 validation images, using 10 images per class across 12 plant disease categories.
- Trained a 384-dimensional EfficientProtoNet with normalized embeddings, achieving 95.73% best combined validation accuracy, 96.13% plant-level accuracy, and 95.33% disease-level accuracy.
- Assessed the saved model on 13,135 cleaned PlantVillage images, reaching 97.77% accuracy, 95.46% macro precision, 96.40% macro recall, and 95.83% macro F1 score.
- Deployed a Flask web application with image upload, SQLite login and signup, 12-class disease prediction, and non-leaf rejection using a 0.6 similarity threshold.

Spark Hospital Readmission and Graph Data Engineering Workflow — Apr 2026

- Built a PySpark data cleaning and modeling workflow on 12,465 hospital readmission records, checking duplicates, missing values, type inconsistencies, placeholder values, and numeric outliers.
- Engineered readmission features from 8 numeric and 8 categorical variables, applying StringIndexer, OneHotEncoder, VectorAssembler, and Spark ML pipelines for supervised classification.
- Trained a Spark MLlib Random Forest classifier with 100 trees and max depth 8 using an 80/20 train-test split of 10,025 training rows and 2,440 testing rows.
- Evaluated the readmission model with 60.57% accuracy, 58.29% F1 score, 61.60% precision, 60.57% recall, and 0.636 AUC on unseen test data.
- Analyzed readmission patterns using Spark SQL, showing readmitted patients averaged 4.63 hospital days and 0.70 inpatient visits compared with 4.33 days and 0.34 inpatient visits for non-readmitted patients.
- Connected Spark to Neo4j using the Neo4j Spark Connector and imported a PlantVitals graph as vertex and edge DataFrames for graph-based data analysis.

Time Series Forecasting on Air Passenger and Alcohol Sales Data — Feb 2024 – Mar 2024

- Built two reproducible forecasting workflows using Python, Pandas, NumPy, Matplotlib, statsmodels, pmdarima, and SARIMAX on airline passenger and alcohol sales datasets.
- Processed 144 airline passenger records from 1949 to 1960 and 325 alcohol sales records from 1992 to 2019 for trend, seasonality, and forecast analysis.
- Applied log transformation, differencing, ACF/PACF analysis, and seasonal decomposition to diagnose non-stationarity, yearly seasonality, and long-term demand patterns.
- Selected SARIMAX configurations with auto_arima, using ARIMA order $(2, 0, 0)$ with seasonal order $(0, 1, 1, 12)$ for passengers and ARIMA order $(2, 1, 1)$ with seasonal order $(1, 0, 2, 12)$ for alcohol sales.
- Generated 12-month forecasts, predicting 421.78 to 660.62 passengers for 1961 and monthly alcohol sales from $11,325.15 to $16,870.14 from Feb 2019 to Jan 2020.

Publications

- Plant Leaf Disease Multilevel Classification Using Few-Shot Learning, International Journal of Research and Analytical Reviews — May 2025