

Harshal Kakaiya

Guelph, ON, N1E 2M4 | (437) 661-4803 | harshalpkakaiya@gmail.com
linkedin.com/in/harshal-kakaiya | github.com/harshalk612 | harshalk.netlify.app

Data Scientist with expertise in ML, deep learning, and data analysis. Proficient in Python, R, and SQL, with a track record of workflow optimization (50% faster) and achieving top 15% in Kaggle competitions.

WORK EXPERIENCE

ML Research Assistant

University of Guelph

May 2025 - August 2025

Guelph, ON, Canada

- Applied data cleaning and data analysis to **18k+** records spanning **10 years** of chemical compound concentrations across **4** transboundary rivers.
- Developed reproducible Python/R workflows integrating climate, hydrology, and water quality datasets, reducing manual processing time by **50%**.
- Delivered insights to support **long-term** monitoring data frameworks for environmental government agencies.

Machine Learning Intern

Tops Technologies

January 2023 - June 2023

Surat, GJ, India

- Designed and implemented robust preprocessing and feature selection workflows, reducing model training time by approximately **30%** and improving F1-score by **~10%** simultaneously.
- Collaborated with cross-functional teams to deploy ML models into production, enabling real-time insights for business and clinical stakeholders.
- Interpreted and visualized large-scale healthcare datasets using Python and its libraries, uncovering patterns that boosted precision and recall by **8%**.

TECHNICAL SKILLS

Programming Languages: Python, R, SQL

Packages & Frameworks: Numpy, Pandas, Matplotlib, Seaborn, Scikit-learn, Tensorflow, Keras, Tidyverse, RShiny

Tools & Platforms: Linux, Git/Github, MySQL, MongoDB, Azure, AWS, GCP

Core Skills: Machine Learning, Deep Learning, NLP, Data Visualization, Data Manipulation

EDUCATION

Master of Data Science

University of Guelph | GPA: 4.0/4.0

September 2025

Guelph, ON, Canada

- Relevant Coursework:** Introduction to Data Science, Data Manipulation & Visualization, Analysis of Big Data, Machine Learning for Sequences, Analysis of Spatial-Temporal Data, **Major Research Project**.
- Student Representative** for 2024-2025, **Treasurer** in Graduate Math & Stats Club

Bachelor of Engineering in Information Technology

Sarvajanik College of Engineering & Technology | GPA: 3.7/4.0

July 2023

Surat, GJ, India

- Relevant Coursework:** Probability & Statistics, Data Structures, Database Management Systems, Analysis and Design of Algorithms, Artificial Intelligence, Pattern Recognition, Data Compression, Software Engineering.

PROJECTS

Kaggle Competition – What's Cooking? | [LINK](#)

December 2024

- Processed **35k+** multilingual recipe datasets using advanced NLP techniques (tokenization, TF-IDF, lemmatization, stop-word removal).
- Built and optimized **SVC** and **XGBoost** pipelines with grid search and stratified cross-validation, achieving **85%** F1-score (**Top 15%**) on the global leaderboard.
- Implemented feature importance analysis and automated preprocessing pipelines with scikit-learn, improving the reproducibility and scalability.

RFM Analysis for Customer Segmentation | [LINK](#)

June 2022

- Analyzed **525k+** e-commerce customer records by calculating Recency, Frequency, and Monetary (RFM) scores.
- Applied clustering techniques and **visualized behavioral segments** using Seaborn, Matplotlib, and Squarify.
- Delivered actionable insights that increased simulated customer retention by **30%**, supporting marketing