

# HARSHAL KAKAIYA

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

Detail-oriented Data Scientist with hands-on experience in data analysis, machine learning, and deep learning gained through academic experience, internship and diverse projects. Proficient in Python, SQL, and R, with a proven ability to develop predictive models and extract meaningful insights from complex datasets.

## TECHNICAL SKILLS

<b>Programming Tools:</b>	Python, R, SQL
<b>Data Science:</b>	Data Visualization, Feature Engineering, Machine Learning, Deep Learning
<b>Mathematical Concepts:</b>	Linear Algebra, Calculus, Statistics, Probability
<b>Packages/Frameworks:</b>	Numpy, Pandas, Matplotlib, Seaborn, Scikit-learn, Tensorflow, Keras, Tidyverse, RShiny
<b>Databases :</b>	MySQL, MongoDB
<b>Cloud Deployment:</b>	Netlify, Git/Github

## INTERNSHIP EXPERIENCE

### Machine Learning Intern

January 2023 - June 2023

*Tops Technologies*

*Surat, GJ, India*

- Contributed to the development of a cancer classification model using ensemble ML techniques, achieving over 90% accuracy to enhance a robust solution for healthcare providers.
- 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%.

## PROJECTS

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

December 2024

- Engineered **NLP** Features from ingredients lists using tokenization, TF-IDF, and custom preprocessing steps.
- Trained and fine-tuned **SVC** and XGBoost models, achieving 85% F1-score placing our team's score as **Top 15%**.
- Implemented stratified cross-validation and data cleaning to improve robustness by 20% in final submission.

### Co-operators Housing Data Price Prediction | [LINK](#)

October 2024

- Executed a detailed ML workflow on AMES dataset, including data analysis & robust preprocessing for prediction.
- Identified Ridge Regression as the top model, achieving a lowest RMSE of 18,407 in the class leaderboard.
- Analyzed the bias-variance trade-off in **Ridge Regression** by plotting a U-shaped curve to deepen insights.

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

June 2022

- Segmented business customers based on Recency, Frequency, and Monetary (RFM) to understand behavior.
- Utilized Python and libraries like Seaborn, Matplotlib, and Squarify to visualize key insights and trends.
- Developed tailored strategies for different customer segments, achieving 30% more customer retention.

## EDUCATION

### Master of Data Science

Expected September 2025

*University of Guelph | GPA: 4.0/4.0*

*Guelph, ON*

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

### Bachelor of Engineering in Information Technology

July 2023

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

*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