**Manoj Kumar Surabhi**

9062992809 mksurabh@mtu.edu [GitHub](https://github.com/Manoj-kumar-surabhi)-[Profile](https://github.com/Manoj-kumar-surabhi) [LinkedIn-Profile](https://www.linkedin.com/in/manoj-kumar-surabhi/)  Houghton, MI

**SUMMARY**

Data Science graduate student with expertise in Python, R, and SQL, skilled in machine learning, statistics, predictive modeling, and data engineering. Experienced in ETL, data pipelines (Azure Databricks, Data Factory), and visualization (Power BI) to drive insights and decision-making.

**EDUCATION** 

**Michigan Technological University (MTU)**  Houghton, MI

*MS in Data Science - GPA 3.5/4.0 Jan 2024 – Dec 2025*

**Guru Nanak Institutions Technical Campus (GNITC)**  Hyderabad, India

*B.Tech in Mechanical Engineering — GPA 3.3/4.0 June 2019 - July 2022*

**TECHNICAL SKILLS**

**Data Analysis & Engineering**: Python, R, SQL, Power BI**,** Azure Databricks, Data Lake Storage, Data Factory, Big Query, & Spark.

**Web Development:** HTML, CSS, JavaScript, React JS, Node JS, Bootstrap, REST API’s, VS Code, Git, & GitHub.

**Coursework:** Bayesian Statistics, Predictive Modelling, Statistical Methods, Data Mining, & Machine Learning.

**Soft Skills:** Persistence, Written Communication, Problem Solving, Curiosity, & Adaptability

# WORK EXPERIENCE

# Project Intern at Bharat Dynamics Limited (BDL), India Dec 2021 - Jan 2022

- Leveraged Python and R for exploratory data analysis (EDA) on 1,000+ process data points, identifying trends and optimizing workflows. Achieved a 5% downtime reduction through data-driven parameter adjustments and enhanced operational efficiency.

- Conducted ANOVA and hypothesis testing on 200+ production samples to evaluate process performance variations. Identified key factors affecting quality, ensuring a 99% compliance rate and driving targeted improvements in production workflows.

- Designed visualizations and reports using Matplotlib, Seaborn, and SQL, analyzing weekly inspection data from 200+ samples. Presented insights to stakeholders, improving quality control, reducing defects, and guiding strategic operational enhancements.

- Processed and maintained datasets of 10,000+ records using ETL techniques and data wrangling. Applied machine learning models with scikit-learn to predict quality outcomes, streamline reporting, and enhance decision-making processes.

# PROJECTS

**A Machine Learning Pipeline for Brain Tumor Segmentation (Ongoing)**

- Designing and implementing a cloud-based machine learning pipeline, where MRI/CT scans are uploaded through a web application, stored in ADLS, and processed via ADF-triggered Databricks workflows for data preprocessing.

- Developing a web application to enable seamless dataset uploads, automatically storing raw data in ADLS, performing preprocessing tasks in Databricks (resizing, normalization, augmentation), and storing processed data for deep learning model training.

- Building and training deep learning models (CNNs/U-Net) in Azure ML to classify brain tumors, leveraging GPU-powered instances for model training and optimizing data pipelines to streamline data processing, storage, and real-time inference.

**A Data Pipeline for F1 Motorsport Historical Data Analysis and Reporting Aug 2024 - Jan 2025**

- Developed a scalable data platform on Azure Databricks, leveraging PySpark and the Ergast API to ingest, transform, and store historical and real-time F1 data for analytics, reporting, and machine learning workflows.

- Built an end-to-end ETL pipeline using Azure Data Factory, automating data extraction from the Ergast API, ingestion into a raw ADLS container, preprocessing in Databricks, and structured storage in a presentation layer for optimized reporting.

- Designed and optimized real-time and batch processing pipelines, utilizing Delta Lake for efficient storage, version control, and time travel while integrating Power BI to visualize key racing insights, enabling data-driven decision-making for historical F1 analysis.

**Prediction Of Breast Cancer Diagnostic**   **Aug 2024 – Dec 2024**

**-** Developed and implemented a variety of machine learning models, including neural networks, Linear Discriminant Analysis (LDA), KNN, and Support Vector Machines (SVM), achieving 98**%** accuracy in classifying cancerous vs. non-cancerous tumours.

**-** Applied advanced data preprocessing techniques such as Box-Cox transformation, spatial sign transformation, and Principal Component Analysis (PCA) to standardize features, reduce dimensionality, and improve model performance.

**-** Enhanced model performance using R by applying statistical modeling and predictive analytics, with Support Vector Machines (SVM) emerging as the best-performing model.

**Bayesian Mixture Normal Model of Plasma Glucose Data Aug 2024 – Dec 2024**

**-** Implemented a Bayesian mixture model with two normal distributions to analyze plasma glucose levels from 532 females near Phoenix, Arizona, effectively capturing data skewness and outliers that a single normal distribution could not represent.

**-**  Applied Gibbs samplingusingMarkov Chain Monte Carlo and utilized JAGS for efficient posterior approximation, ensuring model convergence and validating results through diagnostic metrics such as trace plots, effective sample size, and posterior summaries.

**-** Used Bayesian methods to uncover two distinct subpopulations with different means and variances, emphasizing the effectiveness of mixture models in analyzing complex data patterns and showcasing the Bayesian techniques for handling intricate datasets.

**Health Harbour**   **Jan 2024 – May 2024**

**-** Led the development of Health Harbour, a health and fitness website, managing a Scrum team of five to ensure timely feature development aligned with user needs and business goals. Spearheaded the Physical Fitness feature, creating four tailored fitness programs (gym workouts, cardio, calisthenics, and yoga) to engage a diverse user base with personalized recommendations.

**-** Implemented RESTful APIs and integrated a dynamic blog feature to enable seamless communication between fitness programs, user profiles, and workout tracking. Designed the blog to provide daily health tips, fitness routines, and expert advice for enhanced user engagement.