VAIBHAV HASU

vhasu@gmu.edu | https://www.linkedin.com/in/vaibhav-hasu-5414a1355/ | +1 (571) 663-7114 | Fairfax, VA – 22030

Summary

Data Analytics graduate student with strong hands-on experience in data analysis, machine learning, NLP, and big data technologies. Proficient in Python, SQL, and R with a proven ability to analyze large datasets, build predictive models, and generate actionable insights through data visualization and statistical analysis

Education

George Mason University, Fairfax, VA — Aug 2024 – May 2026
Master of Science, Data Analytics Engineering — GPA: 3.48/4.00
Coursework: Data Visualization, Big Data Technologies, Data Management, Natural Language Processing, Big Data Essentials, Health Data Integration.

CVR College of Engineering, Hyderabad, India — Jun 2020 – July 2024
Bachelor of Engineering, Computer Science Engineering — GPA: 7.15/10.00
Coursework: Database Management System, .Net framework for application development, Data Mining and Data warehousing, Data Structures, Discrete Mathematical Structures, Object Oriented Concepts, Cloud Computing and its application and Internet of things and applications.

Technical Skills

Core Skill	: Data and Statistical Analysis, Data Visualization, Data Cleaning, Machine Learning, Problem Solving
Programming Languages	: Python, SQL, R, Xml, Excel
Data Science	: Machine Learning, Data Visualization, EDA, Feature Engineering
Tools	: Visual Studio, Jupyter, Databricks, Hadoop, Spark, PySpark, Power BI, TensorFlow
Databases	: Microsoft SQL Server, MongoDB, MySQL, DB2
Big Data Platforms	: Databricks, Hadoop, Spark

Projects

Social Media Trend Analysis — Feb 2025 - May 2025

Python, NLP, BERTopic, BERT, NRC Lexicon

- An NLP pipeline was created to examine over 48,000 social media messages for mood, emotion, and popular subjects. 80% sentiment classification accuracy was attained with transformer and BERTopic models.
- Increased accuracy and scalability through the automation of preprocessing operations such as lemmatization and tokenization, combining transformer-based techniques with rule-based strategies like VADER.

Mental Health Care Utilization Trends During COVID-19 — Feb 2025 - May 2025

Python, R, SQL, Data Visualization, Statistical Analysis

- Analyzed U.S. mental health service usage using CDC survey data, focusing on prescription medication, therapy access, and demographic disparities from 2020 onward.
- Integrated multi-source data using SQL, and applied statistical analysis in R and Python to identify demographic and regional trends, revealing service access gaps and temporal changes in public health data.

Patient Data Integration and Matching Accuracy Analysis — Feb 2025 - May 2025

Python, Data Integration, Health Informatics, Record Linkage

- Conducted a large-scale patient record integration project involving over 4 million records to evaluate the effectiveness of deterministic matching methods. Identified limitations in demographic-based matching, resulting in over 3 million mismatches.
- Proposed implementation of a Master Patient Index (MPI) to improve match accuracy and reduce error rates by standardizing identifiers across healthcare systems.

COVID-19 Behavior Change Impact Study — Aug 2024 - Dec 2024

Python, Statistical Analysis, Regression, Data Visualization

- Identified differences in policy satisfaction among ethnic groups by conducting a sector-by-sector comparison examination of customer happiness during the epidemic.
- Strong effects of negative feedback were found when correlation analysis was used to link feedback with purchase decisions ($r > 0.8$). Provided useful information for consumer recovery plans and fair policy frameworks.

Movie Rating Prediction System — Aug 2024 - Dec 2024

Python, KNN, Collaborative Filtering, MovieLens Dataset

- K-nearest neighbors (KNN) and collaborative filtering were used to build a recommender system that predicts movie ratings.
- CF produced the lowest RMSE (0.9879), beating KNN in terms of accuracy. examined the trade-offs between runtime and accuracy for sparse data, emphasizing how well CF captures latent user-item interactions.

Chicago Public Library Event Analysis — Aug 2024 - Dec 2024

R, Statistical Modeling, Data Visualization

- In order to maximize community programming, an analysis of over 10,000 library events revealed that "Story Time" was the most popular event category (35% above average) and that evening hours were 25% more popular than morning ones.
- Significant geographic attendance differences ($p<0.05$) were found using ANOVA, allowing for data-driven resource allocation amongst branches.

Classification of Online Toxic Comments using ML Algorithms — Jan 2024 - Jun 2024

Python, NumPy, Pandas, TensorFlow, Logistic Regression, LSTM, GRU

- Developed a text classification system to detect toxic online comments using six machine learning models including logistic regression, SVM, LSTM, and GRU. Conducted performance evaluation across models; logistic regression achieved the highest accuracy.
- Demonstrated effective preprocessing and deep learning integration for real-world NLP applications.

Vitamin Deficiency Predictor — Aug 2023 - Dec 2023

HTML, CSS, JavaScript, JSP, MySQL

- Built a web-based health prediction platform enabling users to receive vitamin deficiency insights through symptom-based inputs. Implemented dynamic front-end design and connected to a MySQL database for secure data handling and result generation.
- Emphasized usability and accessibility in online healthcare applications.

Certifications

Oracle Database Programming with SQL - Coursera
IBM Data Analytics Professional Certificate - IBM
Python Basics for Beginners - E-Box Certification Course
Data Science Basics - 360DigiTMG
Google Data Analytics Professional - Coursera