

# Rahul Veerapur

(215) 651-0869 | mv.rahul9@gmail.com | rxv5218@psu.edu | www.linkedin.com/in/rahul-v-rvce999/

## Education

### Pennsylvania State University

Master of Data Analytics,

Big Data Analytics , Database Design , Large Scale Database , System Design, Statistics, Data Warehousing , Foundation of AI, Data Mining, Machine Learning, NLP, Predictive Analysis.

Aug. 2024 – Expected Dec. 2025

Philadelphia, PA, United States

### R.V. College of Engineering

Bachelor of Engineering in Computer Science,

Coursework: Data Structures, Algorithms, Operating Systems, System Design, Software Engineering, Machine learning

Aug. 2018 – Aug. 2023

Bengaluru, India

## Experience

### Pennsylvania State University, Research Assistant (Trade Market Analyst)

Python, Deep Learning, sickit-learn , Power Bi.

Oct. 2024 – Dec. 2025

Pennsylvania, United States

- **Drove data analysis and predictive modeling initiatives** by developing Spatio-Temporal Graph Neural Networks (GNNs) for financial time-series forecasting, delivering a **15% improvement in prediction accuracy** to inform business strategy and planning. Partnered with cross-functional teams to **translate analytical findings into actionable business insights**, supporting key scorecard metrics and long-term strategy.
- **Engineered and optimized large-scale models (100M+ rows)** using GATv2Conv and LSTM architectures, achieving a **20% reduction in validation loss** through advanced hyperparameter tuning with Ray Tune and Optuna. **Enhanced analytical workflows and process efficiency**, improving model deployment agility and enabling more data-driven decision-making.

### CME Group, Intern

Java, UC4, SQL, Data Warehouse, Automated Reporting System, Power Bi, Tableau, Random Forest, .

Jan. 2022 – Jul. 2022

Bengaluru, India

- **Optimized processing in RDBMS environment:** Refined SQL query structure and **implementing automated schema for data warehouse achieving a 14% boost** in server efficiency also reducing database lookup.
- **Dynamic Dashboard:** Designed a **reporting system that provides statistical tracking** such as data visits, connection time, and the type of data retrieved.
- **Predictive Intelligence:** Integrated a **predictive analysis using Random forest with gradient boosting** and also using DBSCAN to determine the information required by employees before initiating a search.

## Projects

### Real Time Distributed System for trend Analysis

Python, Kafka, PostgreSQL, Cassandra, MongoDB, Spark, Apache Flink , Powerb Bi

- **Real-time Distributed Data Pipeline with ETL :** Built a real time ETL pipeline for transforming 2.5M+ unstructured news articles into structured, analysis ready datasets ensuring accuracy consistency and scalable for decision making
- **Data Integration & Storage :** Implemented polyglot databases by combining SQL based with NOSQL system ( mongodb , cassandra) by using data streams such as Flink / Kafka supporting exploratory analysis and large scale reporting.
- **Business Intelligence & Visualization :** Designed interactive Power Bi dashboards to visualize the merging of new trends using complex DAX to evaluate emerging news trends, sentiment patterns and time series metrics by enabling actionable insights for stakeholders. Integrated Streaming outputs with BI tools to provide near real-time monitoring and trend detection.

### Data-Driven Insights & Predictive Modeling for Opioid-Related Crash Risk

Python, SVM, Hyperparameter Tuning, LightGBM, Random Forest, SHAP, PowerBi, Scikit-learn, Tableau, Plotly

- **Data Engineering & Database Management:** Designed, implemented, and managed robust Python/MySQL data pipelines, integrating and cleaning over 2.5 million records from 5 state and national opioid overdose datasets. This process involved transforming 130+ raw fields into 70+ analytical features, ensuring high data quality and integrity for advanced analytics.
- **Advanced Analytics & Predictive Modeling:** Developed and validated machine learning models (Logistic Regression, Random Forest, XGBoost) to predict opioid-related crash risk and severity, achieving AUC scores up to 0.99. Conducted in-depth statistical and geospatial analyses of over 16,000 opioid-related crashes, identifying key demographic risk segments and high-risk geographic hotspots.
- **Strategic Insights & Quantifiable Business Impact:** Translated complex analytical findings into actionable business strategies, demonstrating potential cost savings exceeding \$114 billion with an ROI up to 8000%. Presented data-driven recommendations for optimizing insurance premiums, underwriting policies, and targeted interventions to mitigate opioid-related crash risks.

## Technical Skills

**Languages and Frameworks:** Java, Python, C, C++, SQL, MongoDB, MATLAB, Pandas, Numpy, Tensorflow, Pytorch, Keras, Seaborn.

**Technologies:** Big Data(Hadoop, Spark, Hive, Kafka, flink), Cloud (AWS S3, Docker, Kubernetes), Predictive Analytics(Regression, Classification, Time Series Analytics), Reporting System(Power BI, Knime, Tableau), Machine Learning(Pytorch, sickit-learn, ), NLP.