Venkata Sai Phaneesha Chilaveni

venkatasaiphaneesha@gmail.com | linkedin.com/in/phaneesha-chilaveni | https://github.com/phanee16 | +17162998640

EDUCATION

University at Buffalo, The State University of New York, Master's in Data Science and Applications **Aug 2021 – Feb 2023**

- Maintained a **4.0 GPA** while completing relevant coursework in Statistics, Data Analysis, Data Mining (R programming Language), Databases (SQL), Machine Learning, Numerical Analysis, and Data structures and Algorithms.

RESEARCH PROJECTS

Highway Traffic Data Integration and Real-time Streaming Pipeline for Toll Plaza Analysis

Tech Stack : Apache Airflow, Bash, Apache Kafka, Zookeeper, Simulators, Python, DAG's

- Developed and implemented a data pipeline using **Apache Airflow** to download, extract, transform, and consolidate data from various file formats (CSV, TSV, fixed width) with DAG definition, data extraction, transformation, and pipeline submission.
- Configured and managed a streaming data pipeline using Apache Kafka, including setting up **Zookeeper**, starting **Kafka** server, creating a topic, downloading, and configuring the Toll Traffic Simulator, and running the streaming data reader script and performed health checks to ensure the smooth functioning of the pipeline.

ETL and Machine Learning (Edx)

Tech Stack :PySpark, Apache Spark, Elyra, IBM Watson , GitHub

- Utilized the HMP dataset to develop a machine learning model, leveraging the open source CLAIMED library for data extraction, transformation, and loading; model creation was facilitated by **Apache Spark** and stored to Cloud Object Store.
- Employed the **Elyra JupyterLab** extension for editing notebooks and pipeline design, showcasing the utility of **IBM's Watson Studio Orchestration Flow** tool for cloud-based, end-to-end data science workflows.

Forecasting Risk Gene discovery in Autism with Genome Scale Data GitHub Link

Tech Stack : R (Caret, GGplot2, RandomForest, GBM, XGBoost, AdaBoost), RStudio

- Replicated the methodology described in the research paper "Forecasting risk gene discovery in autism with machine learning and genome-scale data" by Brueggeman et al. (2018) and identified a gap in the methodology and took the **initiative** to replicate and improve the analysis.
- Conducted a comparative analysis of ensemble learning algorithms including **bagging** and **boosting** methods, such as XGBoost, AdaBoost, and Gradient Boosting, to identify the most effective approach for predicting autism risk genes using High confidence **SFARI** genes, resulting in a significant improvement in accuracy and performance metrics, such as a training error of 0.02795 and AUC-ROC score of 84%.

Web application on NYC Collision Analysis GitHub Link

Tech Stack :Python, Streamlit, PyDeck Google Cloud Platform, BigQuery, SQL, Looker Studio

- Extracted NYC collision data(**~2 Million observations**) from NYC Open Data and analyzed it using **GCP BigQuery**. Visualized the results with **Looker Studio** and built a web application with **Streamlit** to share the insights, later deployed in **Heroku**.

Reinforcement Learning in Grid World: A SARSA Approach GitHub Link

Tech Stack : Python(Numpy, gym, google_colab)

- Implemented a **flexible** and adaptable reinforcement learning project that involved creating a Grid Environment class, **SARSA**_Agent function, and render function to enhance the navigation and reward outcomes for an agent navigating through a 5x5 grid environment using SARSA algorithm.

SKILLS

- **Programming Languages**: Python, R, Java, C, SQL, HTML/CSS, MATLAB
- **Databases**: MySQL, MongoDB
- **Tools**: Tableau, Excel, GCP stack, Apache Spark,AirFlow,Kafka,ZooKeeper, Jupyter Notebooks, GCP BigQuery, Git
- **Frameworks**: Flask, Keras, TensorFlow, Scikit-learn, Streamlit

EXPERIENCE

Data Scientist Intern (Remote), Marvel Technology Solutions, USA **Jun 2022 - Aug 2022**

- Developed a **robust recommendation** system for an e-commerce company using Python and its associated data science libraries, including Pandas, NumPy, and Scikit-learn.
- Utilized advanced recommendation algorithms such as collaborative filtering, content-based filtering, and hybrid methods to deliver personalized product recommendations to millions of customers and used Tableau for data visualization and reporting to make insights easily understandable for stakeholders.
- Leveraged cutting-edge techniques such as dimensionality reduction and hyperparameter tuning for data preprocessing, feature engineering, and model selection, and utilized **Apache Spark** and **TensorFlow** for distributed computing and deep learning to handle massive amounts of data.

Research Intern, Indian Institute of Technology, Hyderabad, India **Nov 2018 - Jul 2020**

- Collaborated with a PhD Scholar to design and fabricate a Passive Dynamic Walker using SolidWorks and Laser beam machine, respectively, culminating in a successful analysis and optimization of the design, resulting in a 20% increase in efficiency.
- Conducted simulation of the walker's behavior on a slope with an inclination of 3 degrees using **MATLAB**, leading to successful analysis and optimization of the design.
- Presented the project at the Connaisance Conference, where the improved efficiency and innovative design were showcased to a diverse audience of over 50 professionals.

CERTIFICATIONS

- Google Data Analytics Professional Certificate