**Venkata Sai Phaneesha Chilaveni**

[vchilave@buffalo.edu](mailto:vchilave@buffalo.edu) | [linkedin.com/in/phaneesha-chilaveni](http://linkedin.com/in/venkata-sai-phaneesha-chilaveni) | <https://github.com/phanee16> | +17164269247

**DATA SCIENTIST**

Experienced Data Scientist skilled in analyzing large data sets, extracting actionable insights, and driving data-driven decision-making for early-stage technology companies. Possess five years of expertise collaborating within 8-12 member teams to deliver impactful results. Currently working as Data Research Analyst for a Non-profit Organization.

**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 hyper parameter tuning for data preprocessing, feature engineering, and model selection, and utilized Apache Spark and Tensor Flow for distributed computing and deep learning to handle massive amounts of data.

**Data Scientist, ValueMomentum, Hyd, IND. May 2019 – July 2021**

**Article Recommendation for Whitecoats Application**

* Developed a content-based recommendation engine for the Whitecoats application, a widely used tool among healthcare professionals, aimed at suggesting relevant articles.
* Conducted in-depth data analysis using nltk and spacy packages, exploring millions of data records.
* Generated weighted embeddings using TFIDF scores and PubMed embeddings.
* Implemented cosine similarity to measure the similarity between the user's current article and the recommended article.
* Demonstrated strong analytical skills and proficiency in data exploration, embedding generation, and similarity measurement techniques.
* Contributed to enhancing the user experience and knowledge acquisition of healthcare professionals through personalized article recommendations.

**Framework to classify Business into different NAICS Categories**

* Developed a framework for classifying businesses into different NAICS categories.
* Utilized web scraping to gather predefined NAICS class code descriptions, which were then used to train the model.
* Conducted extensive data analysis on millions of records using nltk and spacy packages to gain insights.
* Employed word embeddings to train the text classification model.
* Implemented the model to classify business descriptions into NAICS codes, enabling the determination of risk, coverage, and premiums.
* Contributed to streamlining the categorization process and facilitating accurate risk assessment and insurance decisions.
* Demonstrated proficiency in web scraping, data analysis, text classification, and business categorization.

**Research Intern, Indian Institute of Technology, Hyderabad, India Feb 2017 - Nov 2019**

* Collaborated with a PhD Scholar to design and fabricate a Passive Dynamic Walker using Solid Works 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.

**SKILLS**

***Programming Languages:*** Python, R, Java, C, SQL, HTML/CSS, MATLAB

***Databases:*** MySQL, MongoDB

***Tools:*** Tableau, Excel, GCP stack, Apache Spark, Jupyter Notebooks, GCP, Git, Jenkins, Docker,Kubernets, Apache Airflow

***Frameworks:*** Flask, Keras, TensorFlow, Scikit-learn, Streamlit,Numpy,Pandas,

**RESEARCH PROJECTS**

**Travelling Salesman Problem (TSP)** [GitHub Link](https://github.com/phanee16/Traveling_Salesman_Problem_real-time)

***Tech Stack*:** *Python (Scikit-Learn, Folium, Geo-Py, Image Io), OpenStreetMap API, Google Colab*

* Engineered an ingenious solution for the Travelling Salesman Problem by leveraging cutting-edge technologies such as the Nearest Neighbour algorithm and OpenStreetMap API.
* Crafted a visually stunning and intuitive geographical map to showcase the optimized route, elevating the understanding of complex optimization problems to a new level.

**Forecasting Risk Gene discovery in Autism with Genome Scale Data** [GitHub Link](https://github.com/phanee16/-Comparative-Analysis-of-Boosting-Algorithms-for-Autism-Detection-using-Genome-Data-)

***Tech Stack:*** *R (Caret, GGplot2, Random Forest, 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](https://github.com/phanee16/-Web-Application-for-Analyzing-NYC-Collision-Data-)

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

* Extracted NYC collision data 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.

**Revolutionizing Clothing Categorization with CNNs** [GitHub Link](https://github.com/phanee16/-Revolutionizing-Fashion-Classification-with-CNN-A-Deep-Learning-Approach-to-Enhance-Clothing-Categ)

***Tech Stack:*** *Python (Pandas, NumPy, Matplotlib, TensorFlow, Keras)*

* Created and trained Convolutional Neural Networks with TensorFlow on a Fashion-MNIST dataset. Initiated a base model with 3- layer neural network architecture and tuned the hyper-parameters for better results.

**Reinforcement Learning in Grid World: A SARSA Approach** [GitHub Link](https://github.com/phanee16/-Comparative-Analysis-of-Boosting-Algorithms-for-Autism-Detection-using-Genome-Data-)

***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.

**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.

**CERTIFICATIONS**

* Google Data Analytics Professional Certificate
* Data Visualization with Tableau Specialization.
* Applied Data Science with Python Specialization.
* Apache Spark for Data Engineering and Machine Learning.

**EDUCATION**

**University at Buffalo, The State University of New York, MS in Data Science** **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.