PHIKOLOMZI SAMKA

+27 82 048 7105 | samkaphiko@gmail.com | linkedin.com/in/phikolomzi-samka-9504041a0 |
https://github.com/PhikoSamka/Medication-Adherence

PROFESSIONAL SUMMARY

Data Scientist with hands-on experience in optimizing data pipelines, automating ETL processes, building and deploying machine learning models. Proficient in Python, SQL, cloud platforms GCS, AWS, and data visualization with Tableau and Power BI. Skilled in working across teams to deliver actionable insights and drive process improvements.

PROFESSIONAL EXPERIENCE

Sand Technologies **Remote**
Data Scientist Intern *March 2024 - June 2024*

- Contributed in developing a predictive digital twin model for groundwater level management, improving water quality monitoring and decision-making for Thames Water utility services
- Extracted, cleaned, and analyzed over 20 years of hydrological data, including rainfall, river flow, and groundwater infiltration using Python, QGIS, and Google Earth Engine.
- Achieved a Nash-Sutcliffe Efficiency (NSE) score of 0.73, improving predictive accuracy from a baseline of 0.07.
- Implemented a weighted voting ensemble model that reduced prediction error with an RMSE of 0.0425, surpassing CatBoost (0.0442) and Random Forest (0.0472).
- Developed an interactive Streamlit web application hosted on AWS EC2, featuring dashboards and categorization of groundwater levels into actionable ranges for pumping recommendations.
- Increased operational efficiency by reducing application load times by 75% (from 8 seconds to 2 seconds) through global data caching optimization.
- Experimented with water quality parameters such as ammonia, conductivity, dissolved oxygen, pH, and temperature, highlighting fluctuations in groundwater levels that improved decision-making accuracy by 15% for water extraction strategies.
- Minimized purification costs and reduced regulatory fines by improving real-time detection of groundwater contamination risks.
- Collaborated with data scientists and water utility representatives to refine model objectives and address implementation challenges.
- Delivered a scalable, user-friendly tool that minimized purification costs and regulatory fines by enhancing water quality and reducing contamination incidents.

Younglings Africa **Cape Town, WC, South Africa**
Data Scientist Trainee *November 2021 - October 2022*

- Ingested 3 million NYC taxi records into PostgreSQL using Python, Docker, and Docker-Compose. Reduced processing time by 58.18% (from 55 to 23 minutes) through ETL pipeline automation using Mage AI.
- Migrated ETL processes to Google Cloud Storage (GCS) and built pipelines partitioning data by date. Optimized data ingestion to BigQuery, achieving a 76.36% reduction in processing time (13 minutes) compared to local testing.
- Developed a daily data ingestion schedule after stakeholder consultations, reducing cloud resource usage and operational costs.
- Created interactive dashboards and reports using Looker, providing stakeholders with real-time insights to support decision-making.
- Developed and implemented object detection and tracking models using YoloV4 & DeepSORT, and deployed as a web application using Streamlit for stock counting of sheep, goats, cows, chickens and pigs in the agricultural sector, applying machine learning techniques to analyze complex datasets.

PROJECTS & OUTSIDE EXPERIENCE

Medication Adherence **Remote**

- This project aims to predict whether patients will adhere to their prescribed treatment plans based on factors like age, prescription period, and health conditions. By identifying at-risk patients, healthcare providers can improve outcomes, reduce costs, and optimize resource use.
- The dataset includes 180,000 records with features such as age, gender, health conditions (e.g., diabetes, hypertension), and adherence status. Using various machine learning models like Random Forest, Logistic Regression, and CatBoost, the Random Forest model was chosen for its balance between recall and computational efficiency.
- Project was developed with Flask (back-end) and Boostrap(fron-end)

- Deployment instructions include both local (Docker) and cloud (AWS EC2) setups, ensuring reproducibility and scalability. This project demonstrates the potential of predictive modeling to enhance patient compliance, thereby improving healthcare delivery and reducing costs.
- *Link to project*

EDUCATION

Explore AI Academy **July 2023 - June 2024**
Certification, Data Science
- Gathering and Managing data
- Analytical programming
- Data visualisation
- Machine learning
- Model Deployment

Deviare Africa **August 2020 - March 2021**
Certification, data engineering
- Introduction to data analysis
- Introduction to data science
- Mathematics for data science
- Workflow orcherstration
- Data warehousing
- Analytics engineering

University of South Africa
Bachelor's, Mathematics and Statistics

SKILLS

Soft-Skills: Problem-solving, Stakeholder-management, Communication(Verbal and Writtern), Time Management, Adaptability, Collaboration & Teamwork, Analytical Thinking, Process Optimization, Project Management, Decision-Making, Resource Management
Technical Skills: Data Ingestion and ETL Automation, Cloud Integration (GCS, BigQuery), Database Management (PostgreSQL), Machine Learning (Object Detection, Tracking Models, Predictive Analytics), Data Analysis and Reporting, Data Pipeline Automation, Web Application Development, Model Evaluation and Optimization, Process Monitoring and Scheduling
Tools/Technologies: Python, Docker & Docker-Compose, Mage AI, Git/GitHub, PowerBI, Tableau, Flask, Streamlit, HTML/CSS
Languages: Xhosa, Zulu