



Contribution ID: 7

Type: **Talk**

Bridging the Data Divide: Towards a Clean Dataset Portal for Southern Africa's Machine Learning Future

Machine learning (ML) transformation in Africa faces a major barrier because of the insufficient availability of clean relevant datasets which are easily accessible. Developing algorithms for socioeconomic, environmental and health contexts of Southern Africa creates substantial challenges for machine learning developers and computer science students who are especially affected by this shortcoming. The study evaluates dataset availability through secondary data collection and direct student interviews in Namibia, Zimbabwe and South Africa. The research findings indicate that Africa lacks specific datasets while students struggle to locate suitable data for algorithm testing and training purposes. The absence of localised datasets restricts innovation while reducing model precision and forces dependence on Western-centric solutions which fail to address or dismiss African realities.

This study recommends establishing a centralised clean dataset portal which focuses on Southern African data needs. The proposed portal would gather structured and anonymised datasets from various sectors including public services and education and health and agriculture. The development of this platform would need collaboration between government organisations and academic institutions and open data projects. The research aims to initiate regional action which will enable future machine learning developers to address African issues using African data by documenting these experiences and revealing the systemic gap.

Key words: Data divide, Machine Learning, Datasets, Clean datasets, open data

Primary author: Prof. JERE, Nobert (University of Fort Hare)

Presenter: Prof. JERE, Nobert (University of Fort Hare)