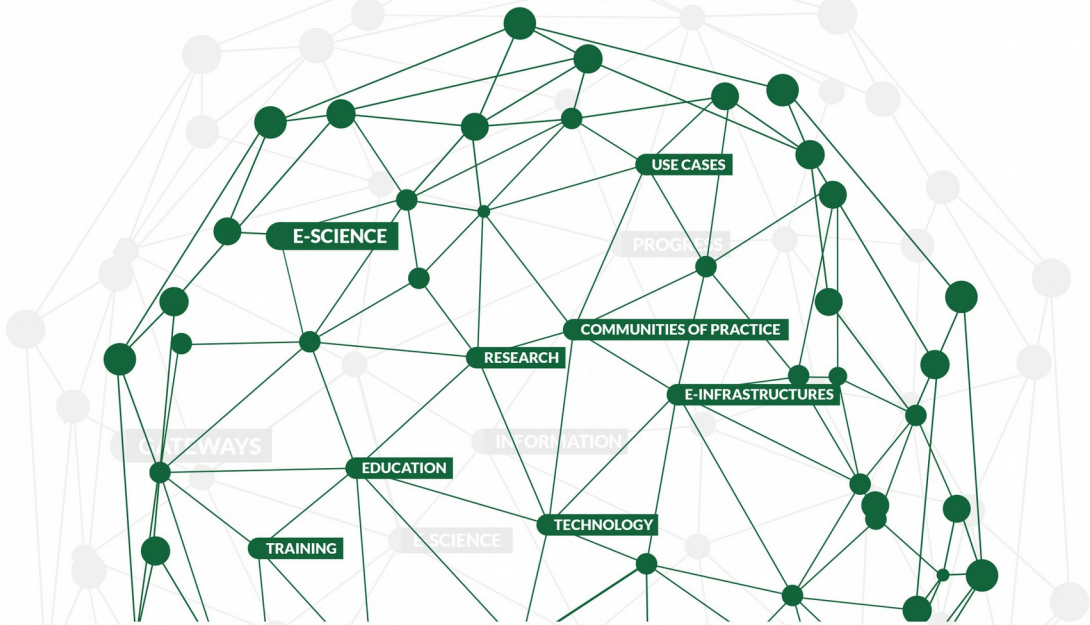
WEKA Machine Learning Use Case – Breast Cancer



Sci-GaIA
www.sci-gaia.eu

Stephan Mgaya – TERNET – Tanzania smgayanath@gmail.com
e-Research Summer Hackfest – Catania (Italy)

This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement n°

Outline

- Scientific Problem
- Computational and data model
- Implementation Strategy
- Conclusion

Scientific Problem

Breast cancer is a disease that affects many people world wide. In Tanzania, it is estimated that 19,008 females have cancer and breast cancer account for 14.4 % of all reported cases.

Using Wisconsin Breast Cancer datasets from the UCI Machine Learning Repository as use case is used to classify benign and malignant samples using WEKA. This aim to help doctors to distinguish breast cancer from benign samples.

Computing and data model

Waikato Environment for Knowledge Analysis (Weka) is data mining workbench which containing machine learning algorithms for data mining tasks,
written in Java, developed at the University of Waikato, New Zealand
-100+ classification algorithms
-75 data processing

Using WEKA as tool with various classification algorithm can be used to perfom this classification tasks by classify the malignant and benign using Naive Bayes classifier with 10-fold cross validation and any other classifier.

Implementation strategy

The main task is to develop web interface that can be used in Science gateway to interact with WEKA features to perform the above use case.

Technologies and tools.

- Technology used is Future gateway.
- Liferay Framework
- Docker container
- Language Java ,json
- Github -smgaya
- Onedata

Implementation strategy

Tasks

- Develop an interface with following features :
- Upload data or load data and convert to ARFF format
- Use Weka classifiers to classify tumors as benign or malignant using different classifications algorithms.
- Access Weka missing values tool to fill the missing values and observe the result and performance.
- check the probability of all diagnosed persons to have benign.

Summary and conclusions

- Output of the implementation must allow reuse in other use case with same scenario.
- Output can be easy extended future user case to be used in Weka.
- The Application have to be ported to in the science gateway and be tested it’s performance.
- Output should allow easy scalable to other platform.
-

Thank you!

sci-gaia.eu

info@sci-gaia.eu

Sci-GaIA
www.sci-gaia.eu