**SSIED Projekt**

Pump it Up

Martin Czakański

Adam Krawiec

Kuba Łada

Michał Slabik

## Tomasz Ścigała

Wykonane Kroki:

1. Rejestracja oraz utworzenie zespołu w konkursie: [Pump it Up](https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/team/)
2. Utworzenie repozytorium [GitHub](https://github.com/martinCzakanski/SSIED-DrivenData-PumpItUp)
3. Pobranie danych do analizy
4. Wstępna analiza danych
5. Wybranie technologi w której bedzię realizowany projekt - RapidMiner
6. Utworzenie procesu w RapidMiner
7. Opis zawartości plików źródłowych

Dalsze kroki:

1. Porównanie zakresu cech między Danymi treningowymi a testowymi
2. Oczyszczenie danych (wartości domyślne, duplikacje itp)
3. Analiza cech pod kątem konwersji typu danych
4. Wyznaczenie kluczowych cech
5. TODO

# **Wprowadzenie**

Dane zostały dostarczone przez Taarifa oraz Wodne Ministerstwo w Tanzani. Celem projektu jest przewidzenie stanu pomp wodnych na terenie kraju w Tanzani. Pompy wodne mogą być w jednym z trzech następujących stanów:

* Functional - sprawna
* Functional needs repair - sprawna ale wymaga naprawy
* Non Functional - nie sprawna

Projekt jest realizowany na potrzeby przedmiotu Sieci Społeczne i Eksploracja Danych.  
Konkurs jest darmowy i każdy może wziąć w nim udział. Dodatkowo za najlepsze rozwiązanie problemu nie ma żadnych przewidzianych nagród.

Więcej informacji na temat konukrsy można znaleźć pod poniższym linkiem: <https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table>

# **Opis zawartości plików źródłowych**

W konkursie są udostepniane czterych pliku źródłowe w formacie .csv. Pliki **SubmissionFormat.csv** oraz **Training set values.csv** zawierają etykiety dla zbioru testowego oraz dla pliku do odesłania.

# SubmissionFormat.csv oraz Training set values.csv

* + - Id – Id pompy wodnej
    - status\_group - stan pompy wodnej możliwe wartości:
      * functional
      * non functional
      * functional needs repair

Pliki **Test set values.csv** oraz **Training set values.csv** zawierają dane testowe oraz dane do przeprowadzenia analizy.

Test set values.csv oraz Training set values.csv

* Id – identyfikator pompy wodnej
* amount\_tsh - ilośc dostępnej wody
* date\_recorded - data wprowadzenia danych
* funder - fundator pompy wodnej
* gps\_height - wysokość geograficzna
* installer - organizacja odpowiedzialna za instalacje
* longitude - długość geograficzna
* latitude - szerokość geograficzna
* wpt\_name - nazwa
* num\_private - ?
* basin - zbiornik wodny
* subvillage - lokalizacja - wioska
* region - lokalizacja - region
* region\_code - lokalizacja - kod regionu
* district\_code - lokalizacja - kod dzielnicy
* lga - lokalizacja - ?
* ward - lokalizacja - oddział
* population – liczba populacji
* public\_meeting - czy jest to w miejscu publicznym
* recorded\_by – organizacja odpowiedzialna za wpisany rekord
* scheme\_management - kto jest operatorem
* scheme\_name – nazwa operatora
* permit – czy punkt wodny jest dozwolony
* construction\_year – rok produkcji
* extraction\_type – rodzaj wydobycia wody
* extraction\_type\_group – grupa wydobycia wody
* extraction\_type\_class – klasa wydobycia wody
* management – nazwa zarządcy
* management\_group – nazwa grupy zarządzającej
* payment – koszt wody
* payment\_type – typ płatności
* water\_quality – jakość wody
* quality\_group – grupa jakości wody
* quantity – ilośc wody
* quantity\_group – grupa ilości wody
* source - źródło
* source\_type – typ źródła
* source\_class – klasa źródła
* waterpoint\_type – typ pompy wodnej
* waterpoint\_type\_group – grupa typu pompy wodnej