



University of Regina

Latex Code

Submitted to

DR. Timothy Maciag

Lecturer

Faculty of Engineering

Software Systems Engineering

Submitted by

DHAVAL BHAILALBHAI PATEL

[200439819]

Faculty of Graduate Studies and Research

Faculty of Engineering and Applied Science

Software Systems Engineering


```

%% Link for the project https://www.overleaf.com/4821137153zvgbkcrfzcg

%% The first command in your LaTeX source must be the \documentclass
command.
\documentclass[acmsmall]{acmart}
%% NOTE that a single column version is required for

%% \BibTeX command to typeset BibTeX logo in the docs
\AtBeginDocument{%
  \providecommand\BibTeX{%
    \normalfont B\kern-0.5em{\scshape i\kern-0.25em b}\kern-0.8em\TeX}}

%% end of the preamble, start of the body of the document source.
\begin{document}

%%
%% The "title" command has an optional parameter,
%% allowing the author to define a "short title" to be used in page
headers.
\title{Poseidon  water infrastructure prediction system for communities }

%%
%% The "author" command and its associated commands are used to define
%% the authors and their affiliations.
%% Of note is the shared affiliation of the first two authors, and the
%% "authornote" and "authornotemark" commands
%% used to denote shared contribution to the research.
\author{Dhaval Bhailalbhair Patel}
\email{dpf761@uregina.ca}
\affiliation{%
  \institution{University of Regina}
  \streetaddress{3851 Retallack Street}
  \city{Regina}
  \state{Saskatchewan}
  \country{Canada}
}

%%
%% The abstract is a short summary of the work to be presented in the
%% article.
\begin{abstract}
  The project Poseidon leverage the data science and machine learning
  techniques to develop a novel solution for water crisis problem by
  predicting the status of water infrastructure of different communities
  across Tanzania. The developed model is deployed on a django based webapp
  to increase its global outreach. The project is develop to support UN
  sustainable goals of clean water and sanitation, good health and well-
  being, partnership for goals and sustainable communities and cities. One of
  the proposed method predicts the status of water infrastructure with
  training precision of 99 percent.
\end{abstract}

%%
%% The code below is generated by the tool at http://dl.acm.org/ccs.cfm.
\ccsdesc{Software and its engineering~Community oriented Design}
\ccsdesc{Human-centered computing~Collaborative and social computing }

```

```

\ccsdesc{Computing methodologies~Machine Learning}
\ccsdesc{Computing methodologies~Parallel computing methodologies}
\ccsdesc{Software and its engineering~Designing software}

%%
%% Keywords. The author(s) should pick words that accurately describe
%% the work being presented. Separate the keywords with commas.
\keywords{Stewardship, Separating the concerns, Mis-information handling, \wp}

%%
%% This command processes the author and affiliation and title
%% information and builds the first part of the formatted document.
\maketitle

```

Digital

```
\section{Introduction}
```

The key objective of this paper is to propose solution for solving water crisis issue in African country of Tanzania. According to UN report on water for life[1], more than half of the people living in Tanzania are without safe drinking water, there is mind boggling funding gap of 61 percent to develop current infrastructure, an average person has to travel more than 30 minutes to get access to clean drinking water and 4000 children death every year from water born disease. For a country like Tanzania facing major water crisis better managing their current water infrastructure is of paramount importance. The proposed solution revolves around the {\bfseries community centered design} using {\bfseries false-consensus effect} where solution goal is to design digital habitat for actual needs of the community. The remainder of this document is presented as follows: section 2 presents the proposed methodology; section 3 presents results and section 4 discuss the future work and conclusion.

```
\subsection{Community and stakeholders}
```

To create a digital habitat that is habitable and thriving for its users a comprehensive study was under taken to understand communities in Tanzania and their orientations. The inference from the study [2] are, the current water infrastructure is managed by ministry of water, the {\bfseries north star customers} for project is communities across Tanzania, the computer literacy rate is low in the country thus we expect bulk of the customer base comprised of {\bfseries lurkers}, the future work suggests creation of chat forum for more {\bfseries prosumer-based } collaborative approach moving forward and welcoming contributions from {\bfseries innovators}. Stakeholders includes communities of Tanzania as NorthStar customers, water infrastructure management Government agency, various local and global NGO's.

```
\subsection{Tools and Technologies}
```

The software development life cycle for project Poseidon is based on {\bfseries Agile} approach. The agile SDLC for this project consist of {\bfseries 3 sprints of one week each resulting in 3 MVP}. Using the {\bfseries Bazaar ideology}, technology and license used to develop the {\bfseries end-end solution} were {\bfseries open-source} for supporting collaborative work. This technology includes python programming language and its various open-source modules, Anaconda development environment, Jupiter notebook Kernel running on Visual Studio code editor, creately and photo-shop for creating documentation. GitHub for version control, Django library deployed in anaconda virtual environment for hosting web

application. The approach uses machine learning algorithm of random forest to predict the status of water infrastructure of the community in Tanzania.

\section{Proposed methods}

Every machine learning model needs data to make prediction. To leverage the machine learning technique to transform {\bfseries raw data into knowledge} we required a non-synthetic dataset to solve real-world problem and not to limit proposed method by using synthetic data. After browsing for the dataset on various data platforms like Kaggle, data driven, etc. The selected dataset that was used to develop the predictive model was based on data available from Government of Tanzania - Ministry of water and hosted by data driven organization as a {\bfseries public dataset}. This selection was in accordance of bazaar approach used in this project. For the purpose of {\bfseries separating the concerns} at various development levels prediction model and web interface are developed differently and if there are any future improvement in the model it can be easily deployed on webapp by simply uploading the model file into webapp. This feature also consider {\bfseries UFFFAA} [4] approach in which flagging the shortcoming of the current model can be overcome by framing knowledge, fixing knowledge and later assuring knowledge in later versions of upcoming models also preserving knowledge from previous versions as well.

\subsection{Dataset}

The dataset comprised of 40 features with total of 74,000 listed rows of data in total. The 40 features include, amount-tsh, date-recorded, funder, gps-height, installer, longitude, latitude, wpt-name, num-private, basin, subvillage, region, region-code, district-code, lga, ward, population, public-meeting, recorded-by, scheme-management, scheme-name, permit, construction-year, extraction-type, extraction-type-group, extraction-type-class, management, management-group, payment, payment-type, water-quality, quality-group, quantity, quantity-group, source, source-type, source-class, waterpoint-type, waterpoint-type-group. The data provided is in {\bfseries raw CSV format} with many missing values for the features like funder, scheme-name and permit which contributed 3635, 28166 and 3056 values respectively.

The label to be predicted has three possible values

\begin{itemize}

\item {\verb|functional|}: the waterpoint is operational and there are no repairs needed

\item {\verb|functional needs repair|}: the waterpoint is operational, but needs repairs

\item {\verb|non functional|}: the waterpoint is not operational

\end{itemize}

\subsection{Feature Engineering}

The feature engineering plays an important role in designing the model. Basically, all machine learning algorithms uses some input data to create outputs. This input data comprises of features, which are usually in the form of structured columns as discussed in dataset. Algorithms require features with some specific characteristic to accurately predict the result.

The main objective for using feature engineering can be summarized as follow

\begin{itemize}

```
\item {}To prepare the proper input data set compatible with the machine
learning algorithm requirements
\item {}For improving the performance of machine learning model
\end{itemize}
```

To develop additional input parameters in the data features were engineered. Some of these features are

```
\begin{itemize}
\item {}Age of the pump is a key feature but it was missing from the
original dataset which was computed using the values from date-recorded and
construction year
```

```
features['age'] = features['date-recorded'] - features['construction-year']
\item {}Population served per age of the pump
```

```
features['pop/year'] = features['population'].replace({0:1}) /
features['age']
\end{itemize}
```

```
\subsection{Imputation}
```

Imputation is the process of replacing the missing values. They pose a huge challenge in creating data pipelines and adversely affect the accuracy of the model. The missing value in the selected dataset was replaced by mean values of respective columns which also helps to create normal distribution of values for these features which is ideal for obtaining better overall accuracy.

```
\subsection{Encoding }
```

The encoding is performed for categorical data as it is difficult for any machine learning algorithm to understand the categorical data. This process simply changes the data with cardinality more than 150 values into numerical format thus enabling grouping of data without any data loss and rest of the low cardinality variables were ordinarily encoded in with string values were simply replaced by a numeric value representing that class of value. This was performed to build pipelines architecture using SKlearn module.

```
\subsection{Scaling}
```

The scaling helps to bring different values to the same numeric range so that a machine learning algorithm can compare different features and identify important features from the feature group. For scaling, transform functionality of SKlearn module was used.

```
\subsection{Parallely Implemented random forest classifier model }
```

The prediction model is based on random forest classifier algorithm implemented using parallel programming. The model is hyper-tuned using gridsearchCV method of SKlearn. The hyper-tuned parameters for the model only consist of max depth thus resulting cluster of decision trees with max depth 25. There are many methods available for hyper-tuning but to make the model simpler and decrease the training time only max depth was used as tuning it drastically impacted the accuracy. The n-tier which controls the number of combinations is set to 5 thus covering a wider search space. The value of CV is set to 5 for stopping model from over-fitting the data.

The project has two versions of model with same hyper-tuned parameters.

```
\begin{itemize}
```

\item {}{\bfseries Version 1}[Deep Model] was trained using 40 input parameters developed in Sprint 1. It has 322 computed important features to make predictions

\item {}{\bfseries Version 2}[Shallow model] was trained using 8 input parameters developed in Sprint 2

It has 9 computed important features selected using base model evaluation by using SAS miner tool to make predictions.

\end{itemize}

\subsection{Flow Diagram of the proposed method }

\begin{figure}[h]

\centering

\includegraphics[width=\linewidth]{MVP3 flow diagram .png}

\caption{Flow Diagram of project Poseidon}

\Description{A diagram depicting the pipeline architecture for training the prediction model and deploying the model on a django web app}

\end{figure}

\subsection{Web Portal}

To increase the global outreach of the model and to create a software based on {\bfseries people centric approach} instead of product-based approach, I have deployed the prediction model on a django based web application so as to test the model with real world data. This approach would also help to improve the product based on feed backs from people about its accuracy.

The django webapp have 2 primary components controller and views. Users interact with the views implementing controller as a backend. The concept of {\bfseries separations of concern} is used to separate the model from the controller code. This result in a smooth mechanism to deploy and test models with different configuration and hyper-tuning parameters with ease. {\bfseries Version 2} model was considered for testing purpose for the application due to time constraints between each sprints.

\section{Results}

The developed models were compared using confusion matrix analysis. The criteria used for evaluation of the developed models are precision, recall, f1 score and accuracy.

\begin{table}[h]

\caption{Version 1 [Deep Model] Training evaluation matrix}

\label{tab:freq}

\begin{tabular}{cccc}

\toprule

Predicted outcome&Precision&Recall&F1-score&Accuracy\\

\midrule

Functional &0.98&1.00&0.99&0.99\\

Needs Repair &1.00&0.95&0.98&0.99\\

Not-Functional &1.00&0.99&0.99&0.99\\

\bottomrule

\end{tabular}

\end{table}

\begin{table}[h]

\caption{Version 2 [shallow Model] Training evaluation matrix}

\label{tab:freq}

\begin{tabular}{cccc}

\toprule

```

    Predicted outcome&Precision&Recall&F1-score&Accuracy\\
    \midrule
    Functional  &0.60&0.97&0.74  &0.63\\
    Needs Repair  &0.00  &0.00  &0.00  &0.63\\
    Not-Functional  &0.82  &0.27  &0.41&0.63\\
\bottomrule
\end{tabular}
\end{table}

```

We can observe **from** table 1 and table 2 that version 2 have lower accuracy than version 1. It was expected **as** version 2 only uses 8 features to predict the value **as** compared to original 40 features used by version 1. We can observe a **{\bfseries** trade off between accuracy **and** model size **} while** comparing accuracy of the two versions of model **for** this dataset.

The front end design of the web application **as** displayed in **{\bfseries** figure 2 **} is** responsive **and** designed to minimise **{\bfseries** gulfs of interaction **} for** users. The website has a simple form design **with** precursors **for** aiding **in** filling the form. The button clearly state what action **is** to be expected once it **is** clicked. The predicted result clearly provide prediction **with** accuracy of the information thus handling **{\bfseries** misinformation **}in** the application.

```

\begin{figure}[h!]
    \centering
    \includegraphics[width=0.8\textwidth]{UI.png}
    \caption{User interface of web portal }
    \Description{image displaying the user interface for web app}
\end{figure}

```

\section{Future Work **and** Conclusion}

The present document provide insight into predictive machine learning approach to provide a novel solution **for** solving water crisis **in** Tanzania. The proposed method discuss various design aspects **and** technology concepts used to develop two versions of predictive model **and** webapp used to deploy these models. The result **from** the study prove that version 1 - Deep model has better accuracy than version 2 - Shallow model.

The future work include functionality of chat forum, announcement page **for** NGO, Donate page **for** the community, Award section **for** recognizing efforts **from** community members, live sensor data integration **for** providing real time status prediction **and** deploying a deep model on web interface. Using concepts of gamification i.e. guessing which pump will be non operational today, can help to rise awareness **in** the community to forge a stronger resolution **in** working towards solving water crisis through collaborative efforts of communities.

```

%% The acknowledgments section is defined using the "acks" environment
%% (and NOT an unnumbered section). This ensures the proper
%% identification of the section in the article metadata, and the
%% consistent spelling of the heading.

```

```

\begin{acks}
\begin{itemize}
\item {}To Government of Tanzania - ministry of water for providing the dataset.
\item {}To Datadriven organization for hosting the dataset
\item {}To Dr.Timothy Maciag for reviewing my idea.
\end{itemize}
\end{acks}

```

```

%%
%% The next two lines define the bibliography style to be used, and
%% the bibliography file.
\bibliographystyle{ACM-Reference-Format}
\bibliography{sample-base}

\begin{enumerate}
  \item United Nations. (n.d.). Water for Life. Retrieved June 20,
2021, from https://www.un.org/
  \item Patel, D. (n.d.). Activity 1 - ENSE 885. Retrieved June 20, 2021,
from https://github.com/Dhaval-B-Patel/ENSE-885---spring-
2021/blob/522763601394b79914e59ee50c9aba06bdad1a8d/Activity%20%20-
%20Requirement%20Analysis/Activity%20%20-
%20Requirement%20analysis%20complete%20merged%20document.pdf
  \item Macaig, T, 2021, Class Notes, People Centered Design, University of
Regina, delivered May 2020
  \item Macaig, T, 2021, Class Notes, Topics in Computer-Supported
Collaborative Work, University of Regina, delivered May 2021
  \item Towards Data Science . (n.d.). Fundamental Techniques of Feature
Engineering for Machine Learning. Retrieved June 20, 2021, from
https://towardsdatascience.com/feature-engineering-for-machine-learning-
3a5e293a5114
  \item Towards Data Science . (n.d.). Hyperparameter Tuning the Random
Forest. Retrieved June 20, 2021, from
https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-
python-using-scikit-learn-28d2aa77dd74

\end{enumerate}
\end{document}
\endinput
%%
%% End of file 'sample-acmsmall-conf.tex'.

```