BIUST Research and Innovation Symposium 2019 (RDAIS 2019)
Botswana International University of Science and Technology
Palapye, Botswana, 4 - 7 June 2019

ISSN: 2521-2292

Machine learning for maximum and minimum temperature analytics and prediction at local level

Tsaone Swaabow Thapelo and Rodrigo S. Jamisola, Jr.
Simulation, Machine Learning, Robotics and Technopreneurship Laboratory (SMaRT Lab)
Department of Mechanical, Energy and Industrial Engineering
Botswana International University of Science and Technology
Palapye, Botswana

***Abstract*— Despite the fact that machine learning approaches have demonstrated to efficaciously model the perturbations tangled within the weather patterns, they are still under deployed in under represented countries. This proves the existence of gaps between the weather service providers and institutions that advocate for the data driven approaches of modelling stochastic systems like weather. For instance, the Botswana Department of Meteorological Services is currently looking for new avenues that can be deployed to compliment the conventional weather models; particularly for one-to-three months step-ahead of localised minimum and maximum temperature forecasts. Thereto, this work applies predictive analytics on local climatological data harvested, using Perl, from the Shakawe automated weather station starting from 01 July 2014 to 28 February 2019. First, statistical metrics such as scatter plots, box-plots, and correlation coefficients are used to infer patterns and relationships hidden within the collected numerical data. The same process, coupled with Random Forests, is deployed to reduce dimensions of the collected data, hence redundant variables are discarded. In the first phase, the models (Multi-Layer-Perceptron (MLP), k-Nearest neighbourhood, Random Forests) are built using the available data. In the second phase, the selected variables (average air temperature, diurnal temperature range, average wind speed, humidity, minimum temperature and barometer pressure) are used to build and compare the proposed models. The models were fit to 70 % of the training data, and validated on 30 % testing data. The results show that MLP outperforms other models based on the correlation coefficient, Root Mean Squared Error and Minimum Absolute Error.**

***Index Terms*— Machine learning models, data analytics, prediction, automated weather stations, climatological data**

I. Introduction

Weather systems are simulated using sophisticated physical based computations. Such models are imperfect for regional and local weather forecasts due to coherent biases in both spatial and temporal model resolution. As a result, these models have limited information about extremely local terrain conditions. Local phenomena like maximum and minimum temperatures in winter seasons are intricate to replicate by global models [1]. These variables are less handled by the size of the grids of the numerical models since they can occur as a result of small perturbations at a local level, or can be influenced by a large spatial event such as El-niño.

II. Background Information

This work is focused on the study of minimum and maximum air temperature trends and variations for the Shakawe Automated Weather Stations in Botswana. Botswana is a land-locked country in the southern part of Africa sharing borders with Namibia to the west, South Africa to the south, Zimbabwe to the north east and Zambia to the north.

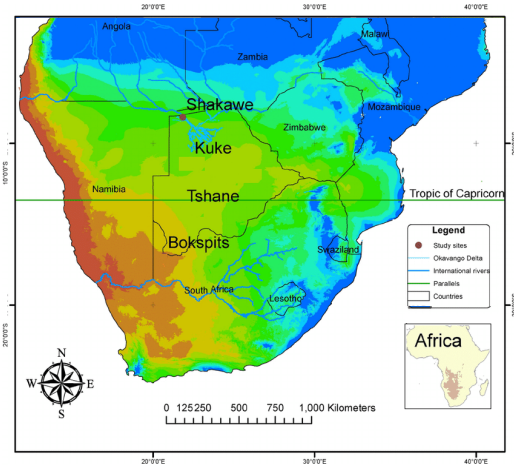

Fig. 1. Location of Shakawe in the northern west of Botswana

The country is characterised by a semi-arid climate, with great differences in day and night temperatures, low overall humidity and high pressure systems interacting with dry soils over the entire land [2], [3]. These have been reported to be associated with temperature trends, particularly high temperatures [4]. Botswana's mean monthly maximum temperatures range from 32 $^{\circ}C$ to 35 $^{\circ}C$. Daily maximums can reach as high as approximately 43 $^{\circ}C$; particularly from October to March. The country is predicted to experience harsh climate variability with local warming and drying above the global warming average of 1.5 $^{\circ}C$ [5].

ISSN: 2521-2292

III. Significance of Study and Envisioned Products

Small geographic changes in minimum and maximum air temperature can have significant impacts on human health, society and the economy [4]. Some researchers linked air temperature with mortality [6] resulting from severe weather and climate variations such as heat waves [4]. Botswana as a semi-arid country is also expected to have approximately "40 more days of heat waves at 1.5°C global warming, and about 75 more heatwave days at 2.0°C global warming" [7]. The re-occurrence of heat waves at recent has led to a number of studies regarding the analysis of temperatures. For instance, daily maximum temperature observational data was used in the study [4] to investigate the heat waves characteristics in the context of climate change.

Various human activities including electricity load forecasting [8], [9], agriculture and water management [10] depend much on these environmental systems. This underlines the value of weather modelling in operational services such as health sciences.

All these invite resilient avenues like the data driven approaches to compliment the physically based ones. Meanwhile, Botswana's environmental data science is still under-exploited despite the fact that climatological data collection has been going on in the country for long [11]. As a result, powerful models go un-explored [12] and potential data remain virgin; while the weather and climate change impacts continue to perpetuate the livelihoods of people.

This project aims to make use of the available open-source resources to add value to existing weather models in Botswana and the SADC region. The knowledge extraction process is particularly important for local to regional weather and climate studies. It presents cheaper and competitive modeling approaches to complement the convectional environmental models that deploy enormous computing systems with lower resolution than desired at regional and local level. Part of this work includes the implementation of recipes to analyse, visualize and assimilate climatological data. The data assimilation process will assist in filling the missing values in weather data.

Thus, the data driven models can be used to benchmark the level of detail that current regional models should be aiming for. In fact, ensembles models can improve model accuracy, robustness, confidence, efficiency, efficacy, reliability, and reduced costs. Lastly, but not least, the developed recipes and model products can be scaled to cater for the Southern African Community and hence facilitating knowledge sharing: all towards the transformation from resource based economies to knowledge based ones.

IV. Literature Review

Chauhan and Thakur presented a review of Data Mining Techniques to forecast several weather phenomena such as temperature, thunderstorms, and rainfall [13]. They found that major techniques like decision trees, lazy learning, artificial neural networks, clustering and regression algorithms are suitable to predict weather phenomena. They concluded that decision trees yield good results for this weather forecasting, followed by the ANN. They also suggested that DM can be considered as an alternative to traditional meteorological approaches.

Olaiya and Adesesan [14] applied decision trees (C4.5, CART) and artificial neural networks in their studies of weather prediction and climate change. They used a 10 years dataset with 4 variables (temperature, rainfall, evaporation, and wind speed) and 36000 instances to acquire an 82 % of accuracy via the percentage split. Meanwhile, their results show that the accuracy varies highly with the training dataset size. They concluded that given enough training data, data mining techniques can be efficiently used for weather prediction and climate change studies.

Petre used decision trees to predict the temperature values around Hong Kong. A 4 years data set with 48 instances was used, comprised of pressure, clouds quantity, humidity, precipitation, and temperature. The accuracy obtained was of 83 % [15]. They reported the need to increase weather variables for input, as well as the increase of data instances. The work also required the transformation of data. Maqsood presented the applicability of ensemble models [16] based on neural networks for one-day-ahead weather forecasting of temperature, humidity, and wind speed for winter, spring, summer and fall. The developed ensemble models generalized better than conventional regression with higher accuracy.

Kumar and Jha [17] used an ANN to predict minimum and maximum temperatures. They used a 100 year data set composed of monthly average min and max temperatures using the percentage split with 60 % for training and the remaining 40 % for testing. Their results showed that ANN may be an important tool for temperature forecasting. Meanwhile, they used only two input variables to compute two target variables.

V. Contribution of this work

Most of these studies discussed so far do not mention the methods used for: data collection, data cleaning and dataset generation as well as the data mining platforms used, or they use commercial platforms like Matlab [17]. The feature extraction and dimensionality reduction is also not explicit [15], [17]. All these make it difficult for reproducibility of results. For that main reason, the applicability of data driven approaches is lagged behind in operational meteorological services. If the data is not available, or the data mining platforms are not accessible, then no models will be developed [12]. Motivated by these, this work deploys the whole processes of predictive analytics or just data science to bridge this gap.

ISSN: 2521-2292

VI. Methodology

VI-A. Data Collection

The data used for modelling was harvested from http://www.sasscalweathernet.org/. Data cleaning and formatting was done using Perl scripting to transform it into consumable formats for `Weka`, `R` and `RWeka`. This was achieved using the `Gnumeric` spreadsheet application which comes with a command line utility called `ssconvert` to convert between a variety of file formats (`.xls` to `.csv`) as shown below:

$$ssconvert \quad InputFile.xlsx \quad OutputFile.csv$$

The basics of file handling were done by associating a named internal Perl structure, a *filehandle*, with an external entity (a .xls file). A variety of operators and functions within Perl were used to read and update the data stored within the data stream associated with the filehandle. The following syntax was used:

open(DATA, "< file.xlsx") or die "Couldn't open file file.txt!";

The source files included unwanted texts values, requiring some additional data processing to get rid of them. The array data structure was used to contain the data input and several Perl operating functions (`shift`, `pop`, `split` and `join`) were applied to manipulate the input data. A complete script was coded to automate the process, adding the shebang line to the code to extract a stream of data from an excel to `csv`, then prepare an `.arff` and `.csv` data sets. The code can then be used to extract data from the any file from any SASSCAL website then generate datasets.

VI-B. Data Pre-processing

Since some machine learning algorithms make assumptions about the data, the data transformation was done to map each point in the input data set X to selected functional output of that point.

$$output_i = Transformer(X_i) \tag{1}$$

The `Transformer()` functions are discussed below:

- The `Min-Max` (Equation 2) transforms the data to a new range [0,1] of values which is guaranteed by the existence of bounded minimum and maximum values.

$$X_{new1} = \frac{X - min(X_i)}{max(X) - min(X)} \tag{2}$$

- The `Z-Score Standardisation` (Equation 3) transforms the data to have zero mean and unit variance with the assumption that the data distribution is normal. It indicates how many standard deviations the data X is from the mean. It is sensitive to outliers [12].

$$X_{new2} = \frac{X - \mu}{\sigma} \tag{3}$$

- The `Magnitude Scaling` (Equation 4) transforms the variable by its maximum magnitude to create a maximum value of -1 or +1 depending on whether the maximum magnitude is negative or positive but it is not guaranteed to fill the entire range from -1 to +1.

$$X_{new3} = \frac{X}{max|X|}, \ in\ the\ range\ [-1, 1] \tag{4}$$

Data transformations can also to make machine learning algorithms such as ANN to train faster, avoid saturation.

VII. Data Partitioning

The `percentage-split` [18, pg. 182] was used to partition the data into `training` (to build the model) and the `testing dataset` to test the model.

VIII. Machine Learning Models

Data modelling and visualisation was done using Weka, R [19] and RWeka by Kurt [20]. These are all open source platforms for statistical computing and graphics. The most important R and Weka packages used in this work include: "nnet"package [21], [21], [22], "kknn"package [23].

VIII-A. Artificial Neural Networks

Artificial Neural Networks (ANN) are non-linear universal function estimators inspired by the multi-tasking and parallel processing of the nervous system [12], [24]; which is supported by its massive amount of sensory data [25, pg. 197]. They are considered to be `distributional free` models that are also `robust` to handle `outliers` and `noisy` data [26]. The models consist of several neurons connected by weighted edges. The interconnections facilitate the exchange of data between neurons. Weka's `MLP` and `R`'s `nnet` and `neuralnet` packages were used in this work.

The learning process starts when the vector x_i is supplied to the `input layer`. The task involves the selection of connecting `weights` w_i between the neurons (depicted in circles). Each neuron has two components [12, pg. 242]: the `transfer function` (T_f) and the `activation function` (A_c). First, the `weights` are initialised, then a `transfer function` in Equation 5 is applied to extract linear combinations of w_i and x_i, coupled with some `bias` b.

$$T_f(x_i, w_i, b_i) = \sum_{i=1}^{n} w_i x_i + b_0 \approx \mathbf{x}^T \mathbf{w} \tag{5}$$

The weighted sum T_f is executed as an argument to $A_c()$ (see Equation 6) at each layer and then the function value of $A_c()$ is propagated to the next layer (another hidden layer or the output layer). There are various common choices for activation function $A_c()$.

$$A_c\big(T_f((x_i, w_i, b_i)\big) = A_c\big(\sum_{i=1}^{n} w_i x_i + b\big) \tag{6}$$

ISSN: 2521-2292

TABLE I
THE COMMON CHOICES FOR ACTIVATION FUNCTIONS

Squashing function	Formula	Range
`Sigmoid (logistic)`	$\phi_1(x) = \frac{1}{1+e^{-x}}$	0 to 1
`Hyperbolic tangent`	$\phi_2(x) = \frac{sinh(x)}{cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}}$	-1 to 1

VIII-B. k-Nearest Neighbourhood

The k-nearest neighbours (`k-NN`) algorithm is an instance-based type of learning model where new data are classified based on stored labelled instances (observations from the training dataset ℶ) [27, pg. 33]. The algorithm measures the distance between the input instance query and a set of instances residing in ℶ) to form $\hat{Y}$.

$$\hat{Y} = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i \tag{7}$$

where $N_k(x)$ is the neighbourhood of x defined by the k nearest points $x_i \in$ ℶ. The main work of the kNN happens during prediction time, where the prediction of a new test data instance is derived based on some similarity distance measure deployed to determine the distance between the stored data and the new instance. The commonly distance metric used are:

- `Minkowski distance function`

$$d(x, y) = (\sum_{i=1}^{n} |x_i - y_i|^r)^{\frac{1}{r}}, \tag{8}$$

  and when $r = 1$, the distance is referred to as Manhattan, while $r = 2$ gives the Euclidean distance.
- `Absolute distance measuring`

$$d_{ABS}(x, y) = \sum_{i=1}^{n} |x_i - y_i| \tag{9}$$

  The Euclidean distance is used to computed the distance among the numerical attributes in k-NN.

In all the cases, n is the number of dimensions equivalent to the number of attributes, and x_i and y_i are the i^{th} attributes of x and y. The k most similar training cases (i.e. neighbours) are used to obtain the prediction for the given test case. If k is too small, then the result can be sensitive to noise points, but it should not be too large, since the neighbourhood may include too many points from other classes. The value of k should also be odd to avoid discrepancy in determining the final class.

VIII-C. Ensembles

An ensemble can be seen as a set of multiple learning methods where a model decision is taken by averaging the results from various standalone models. Several models and their respective clones/configurations can be applied to a single data set (or several) to investigate their predictive capabilities. Bagging, Boosting and Random Forests are examples of ensembles [**?**], [12], [18]. This work deploys the Random forests for variable selection as well as modelling.

IX. RANDOM FOREST

Literature shows that this algorithm implements Breiman's FORTRAN random forest algorithm for classification and regression [**?**, pg. 17]; which adds some randomness to bagging by constructing each tree using a different bootstrap sample of the data. In standalone tree based models, each node is split using the best split among all variables; while in a random forest, each node is split using the best among a subset of predictors randomly chosen at that node.

This work deploys R's `randomForest` package to provide an interface for coding. The method function takes in as inputs: a formula interface, predictors - specified as data frame via the x argument, responses denoted as a vector via the y argument. The algorithm is very user-friendly, requiring only two parameters: the number of variables in the random subset at each node, and the number of trees in the forest). Here, the response for classification is a factor; and continuous for regressions.

X. EXPERIMENTAL SET UP

The methodology deployed is the percentage splitting, where the models are fit to some training data, then evaluated on some unseen testing data. various partition ratios are used in this work, and the best parameters are selected based on the errors in training and testing data.

X-A. Statistical metrics of model assessment

Marques de Sá defines a `statistic` is a function, t_n, of n samples values of x_i [28]. The central idea in statistics is to find the ünderlying law.ºf the data by averaging out measurements errors from each single experiment.
Thus, in each of statistical metrics deployed, one first computes the error of an estimate- the actual value minus the predicted estimate, and then computes the appropriate statistic based on those errors.

1. The `Mean Squared Error` provides a gross idea of the magnitude of error.

$$MSE = \frac{\sum_{i=1}^{n} (y_{predicted} - y_{actual})^2}{n} \tag{10}$$

2. The root-mean-square deviation (RMSD) was also used to measure the differences between values (sample or population values) predicted by a model and the values observed.
3. The correlation was also used. The correlation coefficient between X_1 and X_2 is the co-variance of the standardisation of X_1 and X_2; and it is defined by

$$Cor(X_1, X_2) = \rho = \frac{Cov(X_1, X_2)}{\sigma_{X_1}, \sigma_{X_2}} \tag{11}$$

   where $-1 \leq \rho \leq 1$. The parameter ρ measures the linear relationship between between variables. If the correlation is positive then when X_1 is large, X_2 will tend to be large as well. If the correlation is negative then when X_1 is large, X_2 will tend to be small.

ISSN: 2521-2292

XI. Summarising the Data via descriptive statistics

This sections articulates some descriptive statistics indices used to give a global picture regarding where and how the data is concentrated and as well as its shape of distribution. The following indices will be analysed for the purpose of summarising the SAWS dataset.

Measures of Location

The following measures of location are used in order to determine where the data distribution is concentrated.

1. The `arithmetic mean` of the data x is the sample estimate of the mean of the associated random variable [28] and is denoted by
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{12}$$
2. The `median` of a dataset is that value of the data below which lie 50 % of the cases. Other measures of location include the mode, and the quantiles (see [28, pg. 59]).

Measures of Spread

The measures of spread (or dispersion) give an indication of how concentrated a data distribution is. The most usual measures of spread are presented below.

1. The `range` of a dataset is the difference between its maximum and its minimum,
$$range = x_{max} - x_{min} \tag{13}$$
It has a draw back that it is dependent on the extreme cases of the dataset; and it also tends to increase with the sample size [28, pg. 62].
2. The `inter-quartile range` is defined as:
$$IQR = x_{0,75} - x_{0,52}. \tag{14}$$
It has the advantage that it is less influenced by outliers, extreme cases nor the sample size as compared to the `range`.
3. The variance of a dataset can be interpreted as the mean square deviation (or `mean square error`, MSE) of the sample values from their mean. $Var(x)$ (sample variance) is defined as:
$$Var(x) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} \tag{15}$$
The variance has $df = n-1$ degrees of freedom; while the mean, on the other hand, has n degrees of freedom [28, pg. 62].
4. The `standard deviation` of a dataset is the root square of its variance. It is, therefore, a `root mean square error` (RMSE) defined as:
$$\sigma = \sqrt{Var(x)} = [\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}]^{\frac{1}{2}} \tag{16}$$
The standard deviation is preferable than the variance as a measure of spread, since it is expressed in the same units as the original data.

Measures of Shape

- The coefficient of skewness was determinded, which is the asymmetry measure around the mean, defined as
$$\gamma = \frac{E[(X-\mu)^3]}{\sigma^3} \tag{17}$$
This measure uses the fact that any central moment of odd order is zero for symmetrical distributions around the mean. For asymmetrical distributions γ reflects the unbalance of the density or probability values around the mean.
- The work also determined the coefficient of excess, kurtosis, which is just the degree of flatness of a probability or density function near the center of the data distribution.
$$\kappa = \frac{E[(X-\mu)^4]}{\sigma^4} - 3 \tag{18}$$
The factor 3 is introduced in order that $\kappa = 0$ for the normal distribution. Distributions flatter than the normal distribution have $\kappa < 0$; distributions more peaked than the normal distribution have $\kappa > 0$. More on this can be found in [28, pg. 65]

Measures of Association for Continuous Variables

The `correlation coefficient` (Pearson correlation) is the most popular measure of association for continuous type data. For a dataset with two variables, $V1$ and $V2$ the sample estimate of the `correlation coefficient` ρ_{V1V2} is computed as
$$r = r_{V1V2} = \frac{S_{V1V2}}{S_{V1}S_{V2}}, \tag{19}$$
where
$$S_{V1V2} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1} \tag{20}$$
This dimensionless measure of the degree of linear association of the two random variables has [-1,1] as the interval; with:

- -1 $\Longrightarrow$ Total linear association, with V1 and V2 varying in the opposite direction,
- 0 $\Longrightarrow$ No linear association; V1 and V2 are linearly uncorrelated,
- 1 $\Longrightarrow$ Total linear association, with V1 and V2 varying in the same direction.

ISSN: 2521-2292

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
TimeStamp	1	1704	NaN	NA	NA	NaN	NA	Inf	-Inf	-Inf	NA	NA	NA
AVG.AT	2	1693	22.58	4.43	23.35	22.75	4.74	11.59	32.80	21.21	-0.35	-0.83	0.11
Min.AT	3	1693	14.29	5.98	15.80	14.67	6.57	-0.50	26.88	27.38	-0.45	-1.07	0.15
DTR	4	1693	17.34	5.42	17.84	17.50	6.15	0.09	30.76	30.67	-0.23	-0.70	0.13
AVG.ST	5	1693	24.93	4.01	26.27	24.76	3.63	17.72	33.94	16.22	0.04	-0.94	0.10
P	6	1693	1.41	6.26	0.00	0.09	0.00	0.00	80.60	80.60	7.69	72.78	0.15
Avg.WS	7	1693	1.38	0.62	1.31	1.35	0.61	0.05	3.70	3.65	0.48	0.11	0.01
AVG.WD	8	1693	126.70	69.72	130.59	121.18	43.76	0.10	359.85	359.75	1.06	2.25	1.69
Max.WS	9	1693	4.47	1.08	4.33	4.38	0.98	1.76	10.10	8.34	0.93	1.59	0.03
Max.WD	10	1693	125.37	72.44	116.70	116.75	58.27	0.00	358.50	358.50	1.09	1.15	1.76
H	11	1693	54.83	18.12	54.45	54.32	19.72	16.67	100.00	83.33	0.20	-0.75	0.44
BP	12	1693	903.06	3.42	902.40	902.90	3.71	895.40	912.60	17.20	0.38	-0.68	0.08
AVG.SR	13	1693	281.26	58.68	276.91	282.19	56.22	36.87	528.03	491.16	-0.12	0.30	1.43
Sum.SR	14	1693	24.21	5.05	23.89	24.34	4.82	0.07	34.69	34.62	-0.31	0.37	0.12
LW	15	1693	164.27	10.55	160.01	161.91	2.90	154.96	259.60	104.65	3.18	14.46	0.26
DP	16	1693	10.94	6.88	10.94	11.11	9.31	-7.39	22.46	29.85	-0.12	-1.21	0.17
WB	17	1693	15.65	4.55	16.01	15.85	6.14	5.02	22.86	17.84	-0.25	-1.25	0.11
SD	18	1691	537.34	163.08	604.48	559.44	112.53	0.00	741.83	741.83	-1.17	0.71	3.97
Max.AT	19	1693	31.63	4.17	31.37	31.64	4.77	19.79	41.50	21.71	0.01	-0.75	0.10

Fig. 2. Summary statistics of the SAWS dataset. The mean value for Precipitation (P) seem to be strange though; its maximum is 80,80 mm while the minimum is 0,0 mm, yet the mean is 1,41 mm larger than the median, 0 mm.

Data visualisation via correlation plots

A correlation analysis provides insights into the independence of the numeric input variables. In this work, the Pearson's product-moment correlation was used to measure the numerical relationship of one variable to another within a 95 % confidence interval. Variables with high correlations have values close to 1 for positive correlations, and close to -1 one for negative correlations. The Figure 3 presents the numerical plot articulated so far.

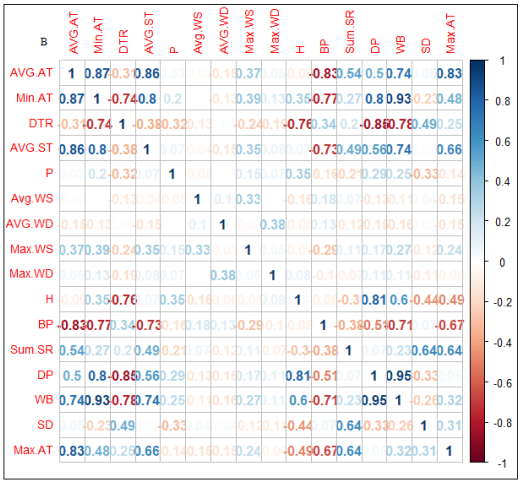

Fig. 3. The correlation coefficients for the SAWS dataset.

Visualisation via box-and-whiskers plots

A box-plot facilitated exploration of data distribution (centre and spread of a numeric variables). It uses the five-number statistics (Minimum, Q1 (first quartile), Q2 (median), Q3 (third quartile), and Maximum) of a vector. It visualises the range, outliers and skew of numerical variables, as well as the comparisons of such variables. This was used to diagnose the problems that are encountered within the data. Since the numerical variables in the dataset were from different instruments, they were normalised to allow a fair comparisons of the variables on the same scale. Figure 6 presents box-plots for our dataset. Only those variables selected via random forests are displayed. Among the variables with some outliers, Leaf Wetness and Precipitation had extreme outliers to the right; while Average Solar Radiation (AVG Solar R) had some outliers in both extremes. These values can be corrected to make the IQR box easier to visualise [29, pp. 62] .

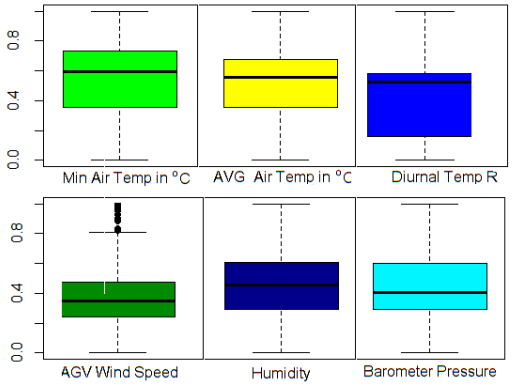

Fig. 4. Thw box-plots for the selected variables for maximum air temperature. We can note that the variable air temperature has some outliers above the extreme maximum whisker denoted by some dots

XI-A. Variable selection via Random Forest in R

Determining the influence measure of a variable is a time consuming and challenging task due to intricate interaction within the modelling variables. Fortunately, the `randomForest` package optionally provide access to two additional pieces (measures) of information: a measure of the importance of the predictor variables, and a measure of the internal structure of the data (the proximity of different data points to one another).

Thus, this work deploys a `randomForest` algorithm to estimate the importance of a variable regarding the target variables. The algorithm works by looking at how much prediction error increases when the modelling data for that variable is permuted. The necessary calculations are carried out tree by tree as the random forest is constructed.

To determine the importance of each variable, the measure is computed from permuting Out-Of-Bag (OOB) data. For each tree, the prediction error on the OOB portion of the data is recorded. For Regression, two metrics are used: the mean decrease in accuracy and the mean decrease in Mean Squared Error. In this work, the mean of squared residuals is computed as

$$MSE_{OOB} = n^{-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (21)$$

ISSN: 2521-2292

where $\hat{y}_i$ is the average of the model predictions. The smaller the result value, the better the model; depending on the problem. It is good to compare to a reference model though.

```
Call:
 randomForest(formula = TMAX ~ ., data = SAWS, ntrees = 500, mtry = 13,
   importance = TRUE, keep.forest = TRUE, subset = train)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 13

          Mean of squared residuals: 0.4703323
                    % Var explained: 97.27
```

Fig. 5. The random tree model with 500 trees, and 13 predictors tried at each tree split. It can noted that 97.27 % of the variation is .[ex]plained"by our model, and the mean of squared errors is ≈ 0.47; quite good.

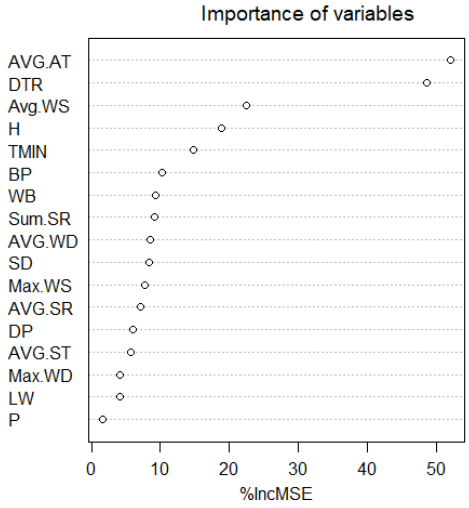

Fig. 6. Graph showing the importance of predictors on Maximum Air Temperature

We can see that the most contributing variables are average temperature (AVG.AT), Diurnal Temperature Range (DT), average wind speed (Avg.WS), and humidity (H).

XII. Results and Discussions

Different configurations were tested by varying the learning rate, momentum, number of iterations and the network structure as seen in Figure 8. Though MLP 2 gave better results during training compared to MLP 3, the later generalised well during the testing phase.

The k-NN was evaluated on the training (70 %) and testing (30 %) data. Different values of 'k' were experimented, and the model statistical results were recorded as shown in Table II It can be noted that the correlation coefficient reduces from 0.98 to 0.96 as more points are used as the number of neighbourhoods (k) increases; while other statistical metrics increases the training phase. Meanwhile, the opposite was observed during the testing phase where all the model replicas of k-NN showed some good performance on the

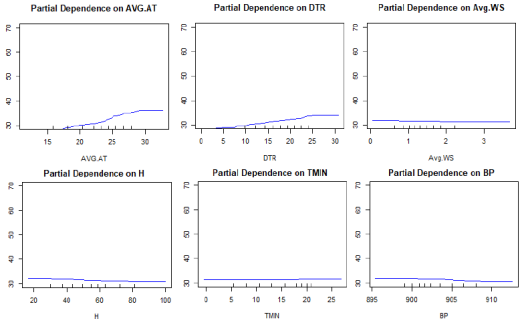

Fig. 7. Partial dependence of the first six elements selected based on random Forest. The first three elements shows some strong dependency with the Maximum air temperature

Training phase	Learning rate	Momentum	Iterations	Architecture	CC	MAE	RMSE
MLP 1	0.2	0.2	500	7,15,1	0.9998	0.0688	0.0904
MLP 2	0.4	0.2	500	7,15,1	0.9999	0.0438	0.06
MLP 3	0.8	0.6	700	7,15,1	0.9999	0.0678	0.0797
MLP 4	0.8	0.7	700	7,25,1	0.9986	0.5357	0.5765
MLP 5	0.8	0.4	700	7,15,1	0.9999	0.0951	0.112
Testing phase	**Learning rate**	**Momentum**	**Iterations**	**Architecture**	**CC**	**MAE**	**RMSE**
MLP 1	0.2	0.2	500	7,15,1	0.9998	0.0707	0.0917
MLP 2	0.4	0.2	700	7,15,1	0.9999	0.0868	0.1002
MLP 3	0.8	0.6	700	7,15,1	0.9999	0.0502	0.066
MLP 4	0.8	0.7	700	7,25,1	0.9979	0.5867	0.6517
MLP 5	0.8	0.4	700	7,25,1	0.9999	0.084	0.1039

Fig. 8. The MLP with a sigmoid activation function. We note that the MLP 3 is more stable for both training and testing phase as compared to the other configurations based on MAE and RMSE

testing data. A slight observation to the results from both the training and testing phase revealed that the model with 15 nearest neighbourhoods was more stable than the rest of the replicas; and with a MAE of 0:9 in the testing phase.

TABLE II
The k-Nearest neighbourhood with 75 % used for training

Training phase	**K = 2**	**K = 3**	**K = 4**	**K = 15**
Correlation coefficient	0.9841	0.9796	0.9769	0.9639
Mean absolute error	0.571	0.6407	0.6835	0.8783
Root mean squared error	0.7432	0.8432	0.899	1.1445
Time (seconds)	0.92	0.97	1.16	1.11
Testing phase	**K = 2**	**K = 3**	**K = 4**	**K = 15**
Correlation coefficient	0.9397	0.9502	0.9519	0.958
Mean absolute error	1.0956	0.986	0.9775	0.952
Root mean squared error	1.4153	1.2938	1.2777	1.2552

Lastly, the ANN model configurations and the Random Forest ensemble were selected to perform some final trends since they demonstrated better predictive capabilities compared to other models based on the proposed statistical metrics. The results are shown below.

XII-A. Base line model construction via Random Forest in R

In Figure 14, we can observe that the Red line is the Out of Bag Error Estimates and the Yellow Line is the Error calculated on Test Set. Both curves have a quite similar trend

ISSN: 2521-2292

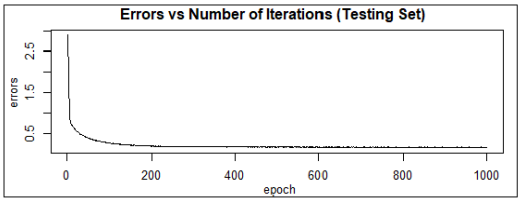

Fig. 9. The graph of the epoch versus error for minimum air temperature based on ANN. We note that the models converge at roughly 25 iterations

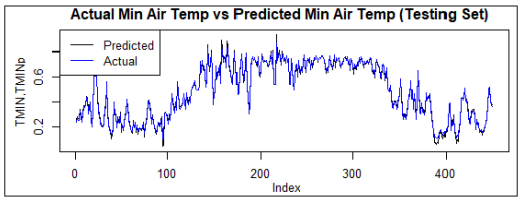

Fig. 10. The graph of the actual minimum air temperature vs predicted minimum air temperature based on ANN showing some perfect fit

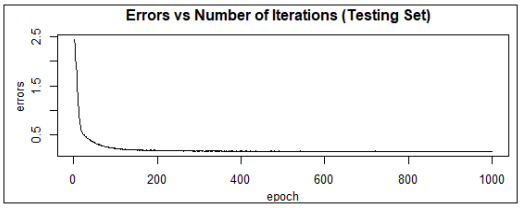

Fig. 11. The graph of the epoch versus error for maximum air temperature based on ANN

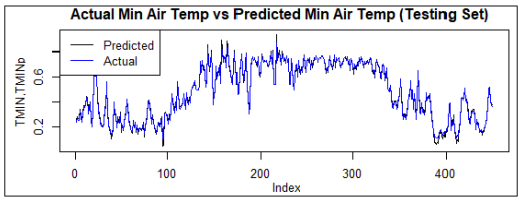

Fig. 12. The graph of ANN showing a perfect fit for the actual maximum air temperature vs predicted maximum air temperature

and the error estimates are somewhat correlated too. The Error Tends to be minimised at around $mtry = 6$. On the Extreme Right Hand Side of the above plot, we considered all possible 18 predictors at each Split which is only Bagging.

XIII. Conclusions

In this paper, we built predictive models for temperature analytics and modelling. Variable selection was done using Random forests and the modelling was done using ANN, kNN and Random forests for the Shakawe weather station. The results show that machine algorithms, can be add value to the convectional weather models. The patterns and trends were perfectly resembled by our models, with ANN outperforming the other two method though the difference was insignificant; all less than one in terms of Mean Squared Errors. Thus the overall errors are within the threshold used by the Botswana Department of Meteorological Services, which is plus or minus two for summers and plus or minus three for winters. Meanwhile, more weather stations should be tried to validate the experiments.

XIV. Future Work

To expand the research to cater for on-line real time machine learning and predictive analytics; and ultimately design a user friendly regional weather forecasting application based on the designed models.

References

[1] T. Tsaone, Swaabow, *Técnicas de aprendizaje automatizado para el pronóstico de temperaturas minímas en el Centro Meteorológico de Villa Clara, Santa Clara*. PhD thesis, Universidad Central "Marta Abreu"de Las Villas, 2014.

[2] T. Nkemelang, *Evaluating temperature and precipitation extremes under 1.5° C and 2.0° C warming above pre-industrial levels: Botswana case study*. PhD thesis, University of Cape Town, 2018.

[3] O. D. Kolawole, P. Wolski, B. Ngwenya, and G. Mmopelwa, "Ethno-meteorology and scientific weather forecasting: Small farmers and scientists' perspectives on climate variability in the okavango delta, botswana," *Climate Risk Management*, vol. 4-5, pp. 43 – 58, 2014.

[4] O. Moses, "Heat wave characteristics in the context of climate change over past 50 years in botswana," 2017.

[5] M. New, "What the latest assessment on global warming means for southern africa," *Quest*, vol. 14, no. 4, pp. 12–13, 2018.

[6] T. Li, R. M. Horton, D. A. Bader, F. Liu, Q. Sun, and P. L. Kinney, "Long-term projections of temperature-related mortality risks for ischemic stroke, hemorrhagic stroke, and acute ischemic heart disease under changing climate in beijing, china," *Environment International*, vol. 112, pp. 1 – 9, 2018.

[7] M. New, "What the latest assessment on global warming means for southern africa," *Quest*, vol. 14, no. 4, pp. 12–13, 2018.

[8] M. Dahl, A. Brun, and G. B. Andresen, "Using ensemble weather predictions in district heating operation and load forecasting," *Applied Energy*, vol. 193, pp. 455–465, 2017.

[9] T. Ahmad and H. Chen, "Short and medium-term forecasting of cooling and heating load demand in building environment with data-mining based approaches," *Energy and Buildings*, vol. 166, pp. 460 – 476, 2018.

[10] "Predictability of daily precipitation using data from newly established automated weather stations over notwane catchment in botswana," *Biodiversity & Ecology*, vol. 50-52, pp. 46–51, 2018. 12th WaterNet/WARFSA/GWP-SA Symposium: Harnessing the rivers of knowledge for socio-economic development, climate adaptation & environmental sustainability.

[11] G. Muche, S. Kruger, T. Hillmann, K. Josenhans, C. Ribeiro, M. Bazibi, M. Seely, E. Nkonde, W. de Clercq, B. Strohbach, K. Piet Kenabatho, R. Vogt, F. Kaspar, J. Helmschrot, and N. Jürgens, "Sasscal weathernet: present state, challenges, and achievements of the regional climatic observation network and database," *Biodiversity & Ecology*, vol. 6, 03 2018.

[12] D. Abbott, *Applied predictive analytics: Principles and techniques for the professional data analyst*. John Wiley & Sons, 2014.

[13] W. Boonsiritomachai, G. Michael McGrath, S. Burgess, and S. Liu, "Exploring business intelligence and its depth of maturity in thai smes," *Cogent Business & Management*, vol. 3, 08 2016.

[14] O. Folorunsho and A. Adeyemo, "Application of data mining techniques in weather prediction and climate change studies," vol. 4, 02 2012.

[15] E. G. Petre, "A decision tree for weather prediction," *PP*, vol. 77, p. 82, 2009.

ISSN: 2521-2292

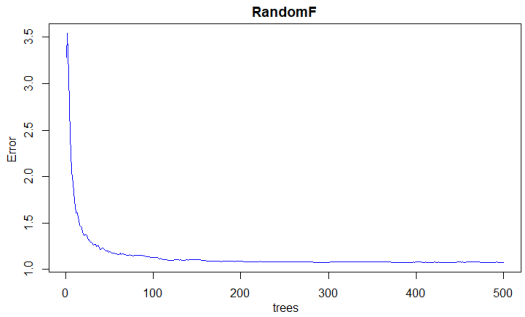

Fig. 13. The plot showing Error versus the Number of Trees. It can be noted that the Error drops gradually as the number of trees increases.

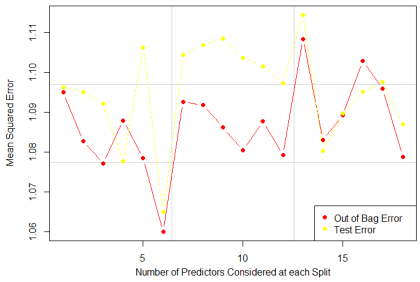

Fig. 14. Plot showing both Test Error and Out of Bag Error using random Forest. We can note that the overall pattern is well resembled, and the error is good, less than 1.2. Which is also less than the threshold for the convectional models of plus or minus two.

[16] I. Maqsood, M. R. Khan, and A. Abraham, "An ensemble of neural networks for weather forecasting," *Neural Computing & Applications*, vol. 13, no. 2, pp. 112–122, 2004.
[17] N. Kumar and G. K. Jha, "A time series ann approach for weather forecasting,"
[18] E. F. Ian H. Witten, *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann, 2 ed., 2005.
[19] R. Ihaka and R. Gentleman, "R: a language for data analysis and graphics," *Journal of computational and graphical statistics*, vol. 5, no. 3, pp. 299–314, 1996.
[20] K. Hornik, C. Buchta, T. Hothorn, A. Karatzoglou, D. Meyer, A. Zeileis, and M. K. Hornik, "Package 'rweka'," 2019.
[21] B. Ripley, W. Venables, and M. B. Ripley, "Package 'nnet'," *R package version*, vol. 7, pp. 3–12, 2016.
[22] S. Fritsch, F. Guenther, and M. F. Guenther, "Package 'neuralnet'," *Training of Neural Networks*, 2019.
[23] K. Schliep, K. Hechenbichler, and M. K. Schliep, "Package 'kknn'," 2016.
[24] C. Giuseppe and V. Balaji, *Neural Networks with R*. Springer Texts in Statistics, 2017.
[25] P. C. Edwin, K and Z. Stanislaw, H, *An Introduction to Optimization*. John Wiley & Sons, INC., fourth edition ed., 2013.
[26] K. Gibert, J. Izquierdo, M. Sànchez-Marrè, S. H. Hamilton, I. Rodríguez-Roda, and G. Holmes, "Which method to use? an assessment of data mining methods in environmental data science," *Environmental Modelling and Software*, 2018.
[27] H. Trevor, T. Robert, and F. Jerome, *The Elements of Statistical Learning*. Springer Series in Statistics, second edition ed., 2017.
[28] M. d. S. Joaquim, P, *Applied statistics using SPSS, STATISTICA,MATLAB and R*. Springer, 2007.
[29] J. Abbot and J. Marohasy, "Application of artificial neural networks to rainfall forecasting in queensland, australia," *Advances in Atmospheric Sciences*, vol. 29, pp. 717–730, Jul 2012.