**ISCG7426 Data mining**

**Assignment 2**

**Semester 1, 2019**

**Student** : Anna Henson 1434929

**Due Date :** 10/06/2019

**Lecturer :** Jane

**Deadline Time** : 23:50pm

Table of Contents

[Introduction 3](#_Toc11089745)

[Problem your trying to solve. 3](#_Toc11089746)

[Why is it an interesting problem. 3](#_Toc11089747)

[Research so far to solve this problem (Literature review). 3](#_Toc11089748)

[Proposed methods to solve the problem. 4](#_Toc11089749)

[Criteria used to evaluate performance of proposed methods. 4](#_Toc11089750)

[How workloads were divided. 4](#_Toc11089751)

[Methodology 4](#_Toc11089752)

[Describe your data set. 4](#_Toc11089753)

[How did you make your predictions. 4](#_Toc11089754)

[Evaluation 9](#_Toc11089755)

[Define quantitative metrics used 9](#_Toc11089756)

[Reason for choosing metrics used for the evaluation. 9](#_Toc11089757)

[How did you evaluate your method. 9](#_Toc11089758)

[Discussion 10](#_Toc11089759)

[Documented results for each technique. 10](#_Toc11089760)

[Best results for each technique. 16](#_Toc11089761)

[Analyse findings. 16](#_Toc11089762)

[What is learnt form the output 18](#_Toc11089763)

[What did work 18](#_Toc11089764)

[Conclusion 19](#_Toc11089765)

[Restate your problem statement. 19](#_Toc11089766)

[How do you propose to solve that problem. 19](#_Toc11089767)

[3 Result of proposed method. 19](#_Toc11089768)

[Summarize what was learnt and possible future works. 20](#_Toc11089769)

[References 22](#_Toc11089770)

# Introduction

A company called climate assist have been asked us to investigate through data analysis growing concerns from some sectors involving horticulture, farming, Auckland council, and ministry for the environment concerns regarding the changing weather patterns.

The report is fundamentally split into five sections which cover the design, development discussion,evaluation, and a conclusion. Results which will be gathered, and any conclusions made are guided the methodology used, whilst visualisation will allow any results to be graphically presented to a non-technical audience. In conclusion the report will show which method works best to solve the problem stated.

## Problem your trying to solve.

Prediction of weather imparticular out door temperature with datetime, wind speed and humidity as factors, which can affect the weather pattern.

## Why is it an interesting problem.

Some experts claim weather patterns are changing in New Zealand which could impact different sectors of the country’s economy. With this possible increase in temperature comes pressure from farmers and horticultural growers. Arable crops, animals and plants are struggling with some experts claiming that thw weather is getting hotter and drier. It has beed argued by some that a need is growing to analyse the weather and try to predict temperature fluctuations in the southern hemisphere.

## Research so far to solve this problem (Literature review).

The statistical analysis of trying to predict trends and patterns has divided data scientists around the world, which when put into context means data sets are fairly fresh samples.

The crux is that data is adaptive and technology is still trying to come up with the best model to compute historical data. Various forecasting models exist such as IPCCAR4 (Fu, Qian, & Wu, 2011) model, which computes a mismatch concerning ensembles and observational data.

Because the models use data repeatedly for machine learning to generalise on going work is still developing a model, which can address the issue of accuracy involving multi decadel variability(MDV). IPPCCAR4 was used in 2007 and classed as a multi model projection to simulate the mean temperature for the 21st century, but due to debates of multimodel accuracy in this field great interest has been to modify, or extend existing algorithms.

Staticians are still trying to resolve negative effects on models where prediction to move forwards has to go backwards. One issue is how does a model cope with temperature variability, but Extreme learning Machine(ELM) by researchers but results still differen. Not all models can eplicate multi decadel variability accurately and leads to overestimating temperature. Another factor is trying to avoid data overfitting and can be a problem as data can only be provided over a specific time frame when data was taken, therefore, models are still restricted by the size of a target in training data.

New concepts are emerging to distinguish between time intervals, and transition from one time step to another for temporal data. However, the perfect model which can be a hundred percent accurate is still being worked by researchers, and the ideal amount of data not yet quantified for accuracy.

## Proposed methods to solve the problem.

Utilise a base model called time series for prediction using a filtered classifier as a way to select different algorithms for machine learning. This will be the most suitable model to analyse the data as it can either output the results for single or multiple targets on a linear support vector machine for regression.

Time series model with filtered classifier involving regression:

* Used for applications which require statistical information outputted as a trend or prediction by a model using time dependency.
* Temporal ordering dependent on time for ordering of data points, and relevant to the output of data in graphical formats.
* Interpret output in graphical formats which is user friendly to non technical audiences.

## Criteria used to evaluate performance of proposed methods.

The criteria for evaluation of performance will be accuracy of the predicted trend for temperature according to the final evaluation to determine the best base model, and to compare performances with different metrics with a filtered classifier to train and test on the dataset.The results will then be collected, evaluated and analyzed to interpret performance.

Key factors are as follows:

* Allow the configuration for tuning parameters regarding the selected target fields.
* Each method applied to each model will use an evaluation.
* The method applied must process continuous data and targets chosen changes over time.

## How workloads were divided.

Due to the work not being done in a group it was undertaken by one person, and consequently the workload was not divided to solve the problem. The entire report and all its different components were done by myself who undertook additional research to assist with conclusions reached.

# Methodology

## Describe your data set.

### Original Data set

* Data over eight weeks
* 13 attributes: Date, Indoor temp, Humidity, Pressure, Description, Rain, Wind speed, Wind direction, Visibility, Cloud cover, Feels like, UV, Outdoor temp.
* Feels like Nominal as rest are numeric
* Formatted in CSV file

### Processed data set

* 4 attributes: Datetime, humidity, wind, and outdoor temperature all numeric.
* Training data comprised of all eight weeks but 30% kept aside as test data of the data set.
* Formatted in CSV file.

## How did you make your predictions.

Used a data mining tool called Weka developed by the university of Waikato, which has a time series model with a forecast package to produce graphs, and visualise predicted data. It gives the ability to train and test future predictions on data sets using a two configurations with an ability to select a fields to analyse, therefore take past events and base predicted future events on known data. This starts with the underlying reasons for the predictions which were made to then understand what output the model has given to analyise.

### Techniques to pre-process data.

Data has to be clean as to reduce the impact of dirty data, which can influence the final results for graphical accuracy by removing outliers by visual identification using graphs.

Supervised techniques used a filter, which is configured to remove attributes according to the predictions which the model is trying to forecast. Another process was to look for any outliers, or anomelies for extreme values within the data such as wind speed, which could affect any upward trends of data. This was not correlated to other attributes and only a few instances missing would have little impact. The pre-process panel allowed a review of a selected attribute, and specific details such as missing data plus the mean and standard deviation by the current relation to the chosen attributes.

After looking at the visaulisation of data there was no particular strong groupings, or any striking patterns. Because of time constraints the data was loaded into Weka so each attribute could be evaluated through the graphs highlighting anomalies, and outlying data to change the accuracy of trends or predictions.

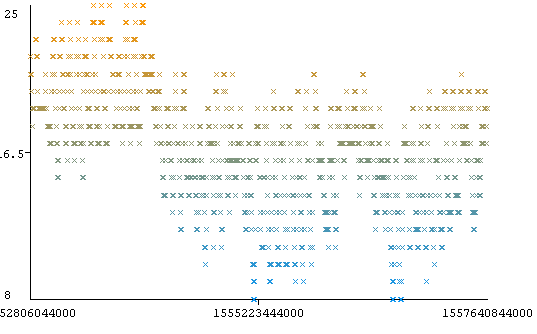
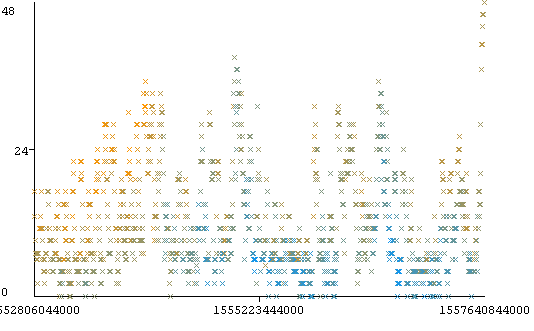


Figure 1 Wind speed vs date

Figure 2 Temperature vs date

### 

Figure 3 Humidity vs date

### Attributes included.

Non-numberic attributes like feels like and cloud cover were removed as they were not consistent with the remaining data. Initially due to using the remaining twelve it was apparent that the model was too complex despite having more instances then attributes.Therefore, datetime, windspeed, humidity and outdoor temperature were selected using the attribute removal filter for a simpler model. This was smaller in volume but can produce similar results and make trends easy to understand, so a wrapper based attribute selection didn’t make sense.

The number of attributes were reduced to four as Weka showed there were non missing. Secondly the selected attributes are appropriate as numeric values are calculated as part of the analysis for an approximation of targets. In essence the value of a dependent variable like temperature can change when independent varibles such as wind speed, humidity, and datetime are selected but not correlated.

Key factors are as follows:

* Temporal data utilises a natural ordering of data points from selected attribute(s)within the data set as they depend on time.
* Date has a time stamp field which is very specific to specifying the periodicity of data based on a time stamp.
* Time series is formed of discrete-time data contained within dataset.

### Unsupervised/Supervised techniques.

Support vectar machines are either linear or non linear supervised learning machine. However, the one which is most suitable depends on the data being transformed and which learning algorithms can be understood by a support vector machine. The selected techniques utlise SVM based learning and assume that we want some type of function to get a result from a numeric class labels. This allows class labels to use supervised techniques and suitable techniques for different types of linear regression, support vectors points help to build support vector machines and distinctly classify data points. This makes a SVM classifier as a powerful tool for supervised learning.

SMOreg is used in Weka by default in the analysis environment as a high performance classifier. The test was completed using a new model with 30% of the training set, and using 1 as the default value when tuning parameters, which is a default for SMOreg in Weka. However, the values of c can be changed for a hardness or softness of the hyperplane margin. If C is low then the margin is softer due to changes of the parameter, which is based on a hypothesis using a discriminator function of 1 or 0. C can be known as a free parameter as to influence features of x and y, however, choosing a value incorrectly can cause overfitting of data.

SMV linear is able to use time series model to fit the historical data provided to compare the past with present. These enable the existence of time to be transformed for machine learning from unsupervised to supervised. Therefore,the instance markers known as the data points identify the edge of a hyperplane which we call Support Vectors. These influence both the position and orientation of a hyperplane whilst maximising the margin.

Linear presumes that the function is to output a numeric value or number, and is simpler as it models a relationship between variables. The latter has to enable a plotting of every data point then find the best hyperplane that goes through all data points, but can be almost impossible to achieve this.

SVM Non-Linear assumes that the data is not sepererated using specific set intervals, but rather more dispersed causing a separation of data. It can utilise a kernel which means all data points can be separated linerally for straight lines when the algorithm is applied to the data set for trends and predictions.

SMOreg and regression can cope with the existence of sequenced observations such as datetime, and use lags as they are based on a measurement of inputs of the supervised machine learning.

Selected techniques allow the following for picking values:

* Advanced configuration model can be adjusted for variance as to override the default settings and specify specific lag lengths as units of time.
* Techniques utilise minimum and maximum lag to specify a minimum or maximum time step when creating lagged variables.
* Altering the number of time units allows the model to process how many steps into the future for predictions.

Random Tree is classed as a decision tree method. This comes under a decision tree algorithm and classed as supervised for classification algorithms and regression problems, which is due to the higher number of trees in a forest gives you higher accuracy results. The algorithm uses the rules of each randomly created decision tree to predict outcome buy using the trained algorithm by passing the test features. It considers the trget your trying to predict with a subset of available features, so temperature is considered as a target prediction but other available features is a tree and cobined with all features selected to form a forest.

It is a algorithm which has a tree type structure which has an internal node, and each node is a feature or attribute to make a decision which splits training data into distinct classes. So a new case which is classified follows a matching path to the leaf node by utilising a process called Info-Gain selector for optimising the best feature to use at each internal node. The branches reproduce a class label based on the distribution at each node by repetition of selecting the best feature downwards, and maximise entrophy to express any useful information, which concludes with leaves on the tree being classifications. In time series by default there is no attempt to prune a decision tree and produces a unpruned model

Assumptions are as follows:

* Overfitting will not be a problem with the algorithm.
* The data will suit the algorithm for a regression task and identify the most important features out of the available features from the training set.
* Random forest can utilise both classification and regression whilst all the features represent useful information which may not be true.
* Will assume that you take the most important features out of the available features from a training set.

Because there is a random selection it can encounter problems like noise, or a higher risk of variability in the model. A method which may help is to tune parameters to affect the size of a random forest tree and properties randomely selected.

### Why those techniques.

SMOreg and linear regression techniques fit the criteria for the time series base model to compute an appropriate output. This system will use a time stamp as the data is periodic, and means that all the data selected is continously changing at specific but equally spaced points in time.

The selected techniques can compute time periods as the algorithms chosen assume that real numbers exist in a sequence. However, the techniques take into account that the next set of numbers are not present for machine learning. For each step ahead support vectar machines collect, and then summarise statistical data in order, which is done for algorithms using different methods to represent the patterns or trends according to what is trying to be interpreted by machine learning.

The time series model will use regression by a filtered classifier as each data point in the data set is not independent. These techniques facilitate the ordering of data points with time dependency via any added fields known as lagged variables.

Lagged variables are defined as “the main mechanism by which the relationship between past and current values of a series can be captured by a propositional learning algorithm which is created by a window or snapshot over a period of time and the number of lagged vartiables created determines the size of the window” (Hall, 2014).

In essence by creating lags you control the mechanism which affects a relationship between the past and current values, which the base model can manipulate using the advanced configuration to shift data back in time which enables supervised machine learning.

### What assumptions if any were made by each technique.

For unknown data to be predicted it can only process the known data in a consecutive order. This is based on the assumption that the system chosen can compute the confidence intervals when doing an evaluation performance for test predictions.

Assumptions made by each SVM:

* Problem to solve is a regression problem using complex or high level mathamatics.
* Change time stamp to hourly, monthly, daily, weekly or yearly as its numeric data is fairly represented in the data set.
* Confidence intervals have to be calculated based on the percentage of how many of the true target values fall within the interval.
* Able to predict Y if each value of X is within the range of values that are present in the training data.
* Linear use an instance to mark the edge of a hyper plane, and keep those instances which it needs for the training model.
* Data can be mapped by SVM to a higher dimension.
* Will need a classifier which utilises hyperplanes to separate data in a linear order.

The basic configuration allows an evaluation on the training data for performance, however once the forecaster is trained it sets up the ability for the system to make a supervised forecast over time to calculate future values. Using the advanced configuration allows further control of the underlying model parameters with higher complexity for further tuning and transforming learning to supervised.

Random Tree

Being a rule based system it will use the training set of data with targets or features, which in the case of time series is rules to perform a prediction on the test data.

Assumptions are as follows:

* Overfitting will not be a problem with the algorithm in a linear interaction regression problem.
* Random forest can utilise both classification and regression whilst all the features represent useful information which may not be true.
* Will assume that you take the most important features out of the available features from a training set.
* It will be able to handle missing values.

# Evaluation

## Define quantitative metrics used

When estimating the error we need to consider an under fit, over fit or another option where the fit is just right. So we need to train a model without any test data and calculaute a models performance when tested on test data. If the performance does not decline when tested on test data then he model is just right.

Mean absolute error(MAE):

* A quantity used to measure how close forecasts or predictions relate in their closeness to the actual outcome when processed by the selected alogrthim.

Root squared mean error(RMSE):

* The square root of the average otherwise known as the mean of the square of all the error once data has been processed by the algorithm.

By using these metrics we can compare and evaluate the output according to the the targets chosen. Secondly it gives a relative measure by expressing errors values as a percentage, which needs to be low as it’s a better prediction than just using the last known target. Quantative metrics calculaue then evaluate model values and the baseline error. This is relative to the forecast made by the steps ahead using the time series model, because it calculates the sum of squared error it means positive differences are offset by the negative ones. In effect find Y and c (complexity parameter) that produces the least squared errors.

## Reason for choosing metrics used for the evaluation.

The two metrics used by default give a standard calculation to give the lowest error rate on the data using a time series model, and with different metrics to assess algorithms for numeric data. The MEA and RSME increases greater accuracy for evaluating the required steps ahead regarding graphical output of the predicted forecast trend. So N is also key and integral to evaluating success of the base learner errors for both training and test data. Normally in time series the chosen metrics need to minimise the sum of squares errors.

## How did you evaluate your method.

Evaluation of performance on the chosen algorithms shows the actual error rate for each one on both the training and test data.

* Ensured evaluation by comparison using the metrics.
* A split used 30 percent of the data set for validation with a set number of instances, which tests the model with defined parameters after training data.
* Train on n-data points then test prediction on the next n-data points.
* Objectively compare the performance of each technique by utilising two configurations basic and advanced within the time series forecast package.

# Discussion

## Documented results for each technique.

Evaluation for models selected using a filtered classifier. They were linear regression, SMOreg with c = 0.5 and c= 1, and random tree using seed 1 and seed 2.

### Linear regression

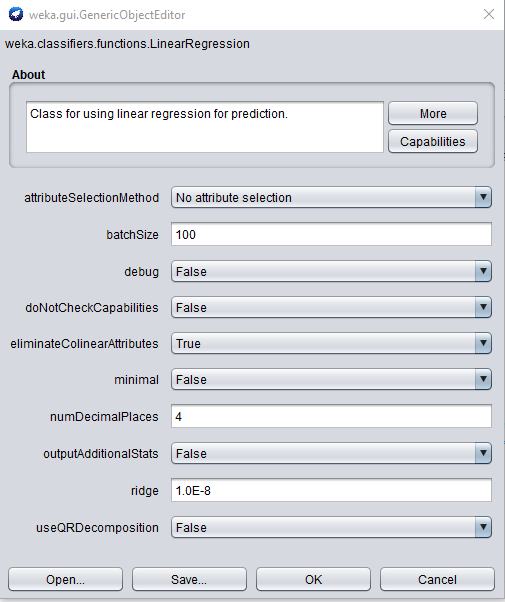
A close up of a map

Description automatically generated

|  |  |  |
| --- | --- | --- |
| Model Linear regression | Training(%) 1 step ahead | Testing(%) 1 step ahead |
| Humidity | | |
| N | 873 | 400 |
| Mean absolute error | 1.6154 | 2.7039 |
| Root mean squared error | 2.1072 | 3.4378 |

| Model Linear regression | Training(%) 1 step ahead | Testing(%) 1 step ahead |
| --- | --- | --- |
| Wind Spped | | |
| N | 873 | 400 |
| Mean absolute error | 5.4713 | 8.664 |
| Root mean squared error | 7.0132 | 12.1767 |

|  |  |  |
| --- | --- | --- |
| Model Linear regression | Training(%) 1 step ahead | Testing(%) 1 step ahead |
| Outside temperature | | |
| N | 873 | 400 |
| Mean absolute error | 1.539 | 3.9325 |
| Root mean squared error | 1.9636 | 4.5356 |

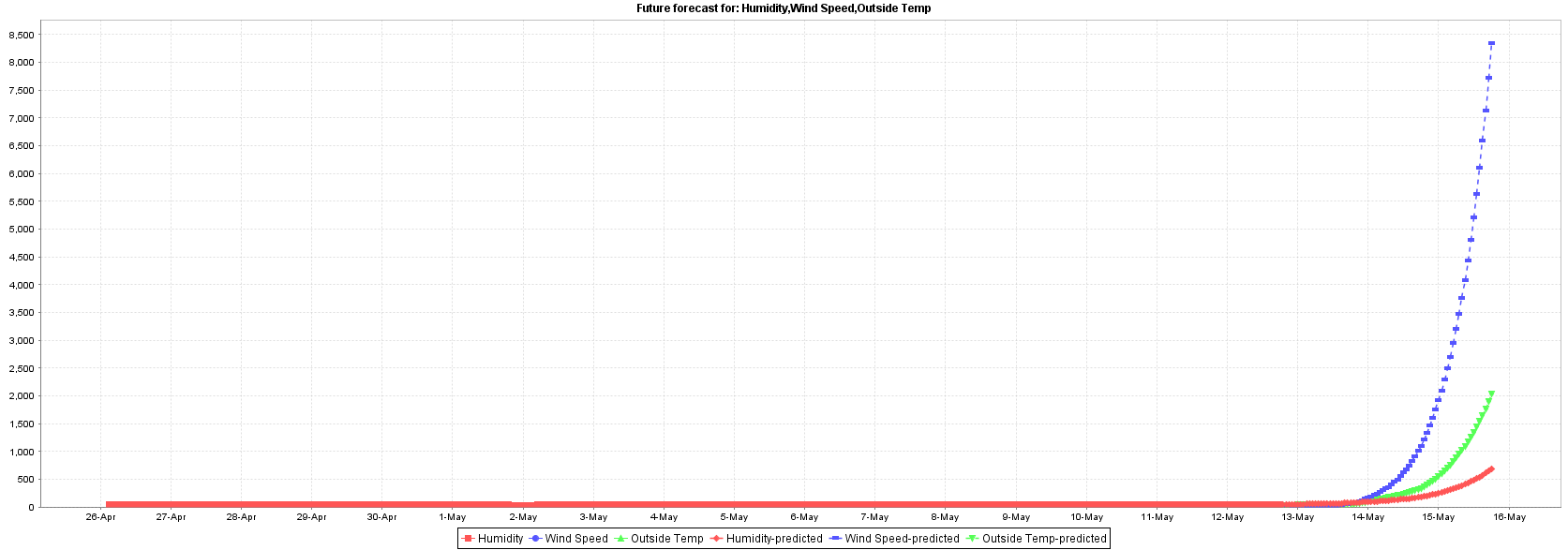
Turning by setting attribute selection method to No attribte selection made a slight difference. Changing other attributes made no difference to the results.

|  |  |  |
| --- | --- | --- |
| Model Linear regression | Training(%) 1 step ahead | Testing(%) 1 step ahead |
| Humidity | | |
| N | 873 | 400 |
| Mean absolute error | 1.6155 | 2.6852 |
| Root mean squared error | 2.1071 | 3.4181 |

|  |  |  |
| --- | --- | --- |
| Model Linear regression | Training(%) 1 step ahead | Testing(%) 1 step ahead |
| Wind Speed | | |
| N | 873 | 400 |
| Mean absolute error | 5.4713 | 8.664 |
| Root mean squared error | 7.0132 | 12.1767 |

|  |  |  |
| --- | --- | --- |
| Model Linear regression | Training(%) 1 step ahead | Testing(%) 1 step ahead |
| Outside temperature | | |
| N | 873 | 400 |
| Mean absolute error | 1.5394 | 3.9067 |
| Root mean squared error | 1.9633 | 4.4973 |

### Linear regression with no lag



|  |  |  |
| --- | --- | --- |
| Model Linear regression | Training(%) 1 step ahead | Testing(%) 1 step ahead |
| Humidity | | |
| N | 921 | 400 |
| Mean absolute error | 0.8065 | 1.4051 |
| Root mean squared error | 1.0995 | 1.8599 |

|  |  |  |
| --- | --- | --- |
| Model Linear regression | Training(%) 1 step ahead | Testing(%) 1 step ahead |
| Wind Speed | | |
| N | 921 | 400 |
| Mean absolute error | 2.5898 | 5.8162 |
| Root mean squared error | 3.512 | 7.198 |

|  |  |  |
| --- | --- | --- |
| Model Linear regression | Training(%) 1 step ahead | Testing(%) 1 step ahead |
| Outside temperature | | |
| N | 921 | 400 |
| Mean absolute error | 0.5933 | 1.3337 |
| Root mean squared error | 0.7769 | 1.7341 |

While the values here look better, the graph looks completely wrong. I do not have enough knowledge about this to be able to say if this is correct or not.

### SMOreg C = 1

A close up of a map

Description automatically generated

| Model SMOreg | Training(%) 1 step ahead | Testing(%) 1 step ahead |
| --- | --- | --- |
| Humidity | | |
| N | 873 | 400 |
| Mean absolute error | 1.6004 | 4.1831 |
| Root mean squared error | 2.0918 | 4.9625 |

|  |  |  |
| --- | --- | --- |
| Model SMOreg | Training(%) 1 step ahead | Testing(%) 1 step ahead |
| Wind speed | | |
| N | 873 | 400 |
| Mean absolute error | 5.249 | 7.3206 |
| Root mean squared error | 7.1938 | 10.5779 |

| Model SMOreg | Training(%) 1 step ahead | Testing(%) 1 step ahead |
| --- | --- | --- |
| Outside temperature | | |
| N | 873 | 400 |
| Mean absolute error | 1.5157 | 5.4416 |
| Root mean squared error | 1.9467 | 6.199 |

### SMOreg C = 0.5

A close up of a map

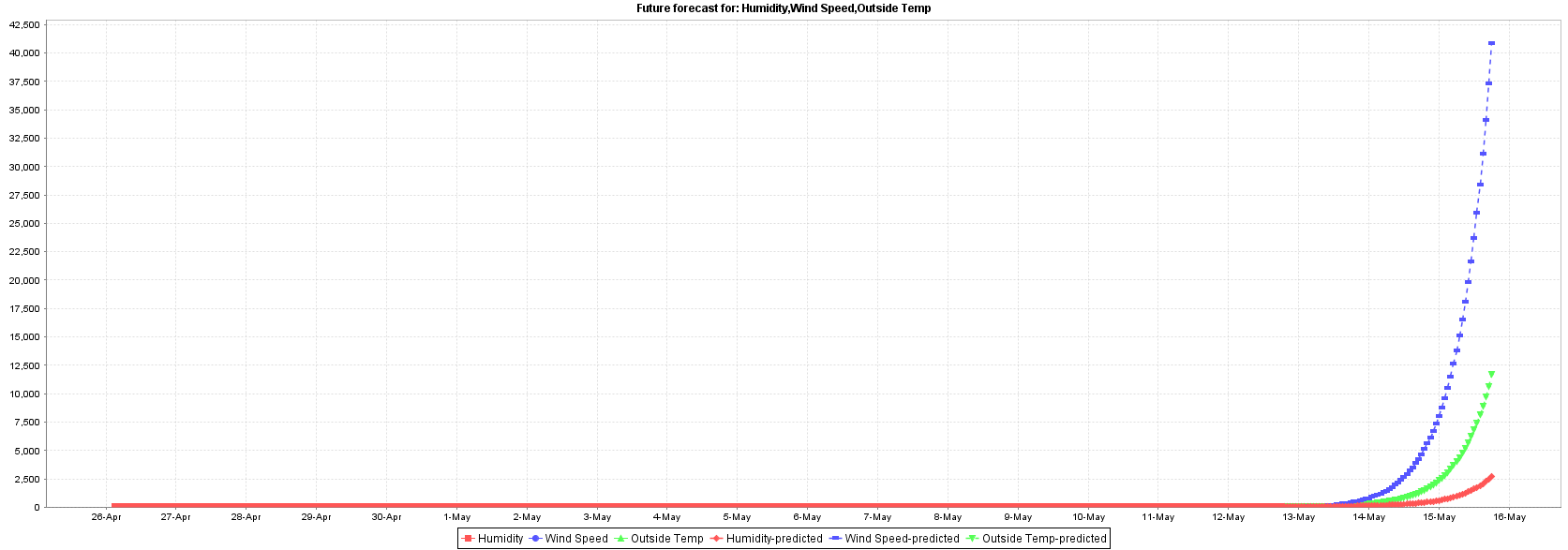
Description automatically generated

|  |  |  |
| --- | --- | --- |
| Model SMOreg | Training(%) 1 step ahead | Testing(%) 1 step ahead |
| Humidity | | |
| N | 873 | 400 |
| Mean absolute error | 1.6055 | 3.3442 |
| Root mean squared error | 2.0987 | 4.0695 |

|  |  |  |
| --- | --- | --- |
| Model SMOreg | Training(%) 1 step ahead | Testing(%) 1 step ahead |
| Wind speed | | |
| N | 873 | 400 |
| Mean absolute error | 5.2843 | 9.1686 |
| Root mean squared error | 7.205 | 12.4279 |

| Model SMOreg | Training(%) 1 step ahead | Testing(%) 1 step ahead |
| --- | --- | --- |
| Outside temperature | | |
| N | 873 | 400 |
| Mean absolute error | 1.524 | 4.338 |
| Root mean squared error | 1.9594 | 4.9981 |

### SMOreg no lag



|  |  |  |
| --- | --- | --- |
| Model SMOreg | Training(%) 1 step ahead | Testing(%) 1 step ahead |
| Humidity | | |
| N | 921 | 400 |
| Mean absolute error | 0.7709 | 1.6795 |
| Root mean squared error | 1.1306 | 2.1114 |

|  |  |  |
| --- | --- | --- |
| Model SMOreg | Training(%) 1 step ahead | Testing(%) 1 step ahead |
| Wind speed | | |
| N | 921 | 400 |
| Mean absolute error | 2.4684 | 4.3504 |
| Root mean squared error | 3.6005 | 5.4862 |

|  |  |  |
| --- | --- | --- |
| Model SMOreg | Training(%) 1 step ahead | Testing(%) 1 step ahead |
| Outside temperature | | |
| N | 921 | 400 |
| Mean absolute error | 0.5672 | 1.2123 |
| Root mean squared error | 0.7976 | 1.5787 |

Again this looks wrong, but I have not been able to find out enough to know why.

### Random tree

A close up of a map

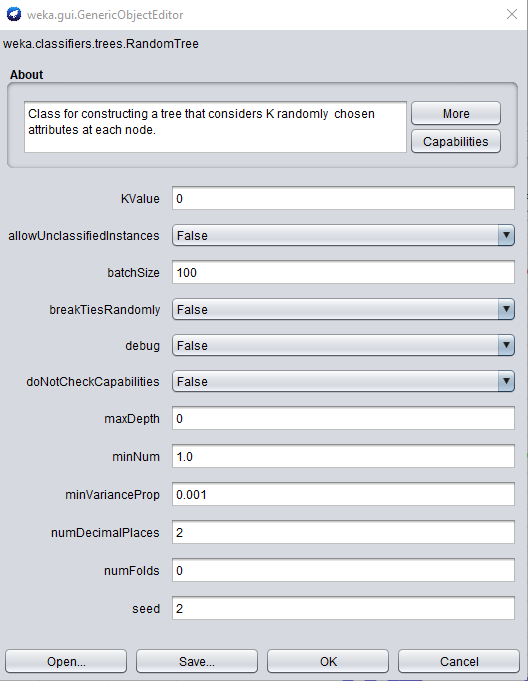
Description automatically generated

| Model Random tree | Training(%) 1 step ahead | Testing(%) 1 step ahead |
| --- | --- | --- |
| Humidity | | |
| N | 873 | 400 |
| Mean absolute error | 0.0108 | 2.3672 |
| Root mean squared error | 0.0317 | 3.1164 |

|  |  |  |
| --- | --- | --- |
| Model Random tree | Training(%) 1 step ahead | Testing(%) 1 step ahead |
| Wind speed | | |
| N | 873 | 400 |
| Mean absolute error | 0.0048 | 13.1425 |
| Root mean squared error | 0.0744 | 14.7744 |

|  |  |  |
| --- | --- | --- |
| Model Random tree | Training(%) 1 step ahead | Testing(%) 1 step ahead |
| Outside temperature | | |
| N | 873 | 400 |
| Mean absolute error | 0.0069 | 3.23 |
| Root mean squared error | 0.1057 | 3.9357 |

Random true using a seed of 2



|  |  |  |
| --- | --- | --- |
| Model Random tree | Training(%) 1 step ahead | Testing(%) 1 step ahead |
| Humidity | | |
| N | 873 | 400 |
| Mean absolute error | 0.0097 | 2.4255 |
| Root mean squared error | 0.0229 | 3.1153 |

|  |  |  |
| --- | --- | --- |
| Model Random tree | Training(%) 1 step ahead | Testing(%) 1 step ahead |
| Wind speed | | |
| N | 873 | 400 |
| Mean absolute error | 0.015 | 13.085 |
| Root mean squared error | 0.1233 | 14.9355 |

|  |  |  |
| --- | --- | --- |
| Model Random tree | Training(%) 1 step ahead | Testing(%) 1 step ahead |
| Outside temperature | | |
| N | 873 | 400 |
| Mean absolute error | 0.0053 | 2.69 |
| Root mean squared error | 0.0487 | 3.3679 |

## Best results for each technique.

Best implementation of each filtered classifier table for collected accuracies:

Random Tree with a seed of 1 was best for humidity.

| Model Random tree | Training(%) 1 step ahead | Testing(%) 1 step ahead |
| --- | --- | --- |
| Humidity | | |
| N | 873 | 400 |
| Mean absolute error | 0.0108 | 2.3672 |
| Root mean squared error | 0.0317 | 3.1164 |

SMOreg with C = 1 was best for wind speed.

|  |  |  |
| --- | --- | --- |
| Model SMOreg | Training(%) 1 step ahead | Testing(%) 1 step ahead |
| Wind speed | | |
| N | 873 | 400 |
| Mean absolute error | 5.249 | 7.3206 |
| Root mean squared error | 7.1938 | 10.5779 |

Random Tree with a seed of 2 was best for temperature

|  |  |  |
| --- | --- | --- |
| Model Random tree | Training(%) 1 step ahead | Testing(%) 1 step ahead |
| Outside temperature | | |
| N | 873 | 400 |
| Mean absolute error | 0.0053 | 2.69 |
| Root mean squared error | 0.0487 | 3.3679 |

## Analyse findings.

### Temperature

|  |  |  |
| --- | --- | --- |
| **Classifier** | **Mean absolute error** | **Root mean squared error** |
| Linear regression | 3.9325 | 4.5356 |
| Linear regression - tuned | 3.9067 | 4.4973 |
| SMOreg C = 1 | 5.4416 | 6.199 |
| SMOreg C = 0.5 | 4.338 | 4.9981 |
| Random tree seed = 1 | 3.23 | 3.9357 |
| Random tree seed = 2 | 2.69 | 3.3679 |

From the above we can see that the best results for temperature where the Random tree with a seed of 2.

### Humidity

|  |  |  |
| --- | --- | --- |
| **Classifier** | **Mean absolute error** | **Root mean squared error** |
| Linear regression | 2.7039 | 3.4378 |
| Linear regression - turned | 2.6852 | 3.4181 |
| SMOreg C = 1 | 4.1831 | 4.9625 |
| SMOreg C = 0.5 | 3.3442 | 4.0695 |
| Random tree seed = 1 | 2.3672 | 3.1164 |
| Random tree seed = 2 | 2.4255 | 3.1153 |

There is very little between the two random seed classifiers. With a seed of 1 the MAE marginally lower, but with a seed of 2 the RMSE is marginally better, with Linear regression not too far behind. With this I would have to say I would choose Random tree with a seed of 1.

Wind speed

|  |  |  |
| --- | --- | --- |
| **Classifier** | **Mean absolute error** | **Root mean squared error** |
| Linear regression | 8.664 | 12.1767 |
| Linear regression - tuned | 8.664 | 12.1767 |
| SMOreg C = 1 | 7.3206 | 10.5779 |
| SMOreg C = 0.5 | 9.1686 | 12.4279 |
| Random tree seed = 1 | 13.1425 | 14.7744 |
| Random tree seed = 2 | 13.085 | 14.9355 |

Finally with wind the results where a lot worse, probably due to the high variation in readings. The best classifier for this was SMOreg with C = 1.

From the above we can see that there is no best classifier that suits the three attributes used. Each attribute has a different classifier that works best for it.

What is learnt form the output.

Patterns, trends and findings to justify output gained.

There was a pattern which showed a majority of the data points between the actual and predicted forecast had a majority of proximity with each other. The first four weeks was the end of summer and the next four weeks the start of autumn, although the summer was a closer prediction than the trend showing the latter four weeks heading into autumn.

The trend was skewed significantly when using the basic configuration as no lags were applied so predictions way out to what was expected. However, when the unsupervised data had lags specified in the advanced configuration to make it supervised it did make the predictions for temperature more reasonable. The gap between the actual and predicted trend got bigger although one step ahead was still propagating errors in a graph format.

What did work

### Advantages

Time series model was the most appropriate for utilising the chosen supervised algorithms under the classified filter functions as data supplied was numeric. Two algorithms worked most naturally with the type of data to predict and fitted the criteria for regression.

Facilitated two levels of configuration as to fine tune specific parameters, which mostly showed a prediction between the actual and predicted temperature over eight weeks. It facilitated the ability to create lags for a snap shot over a specified time period, and being consecutive were then averaged by the system. This helped to show trends as it did calculate using hyperplanes the relationship in a simple model relationships between variables.

### Limitations

Limitations were the amount of data which could be collected over that specific time frame, and the change of seasons in Auckland due to a sub tropical climate, which is different from the rest of New Zealand so a direct comparison in the country is difficult. New Zealand can be influenced by Australia’s east coastline primarily Sydney down to Melbourne which may of influenced the upper north island weather temperatures.

### What didn’t work.

After reviewing the graphical output of the predicted temperature it was not a fully accurate prediction, and errors were propagated as to affect the trend further out in the prediction.

It would be better to take eight consecutive weeks either just summer or just autumn then track exactly the same data over a three year period. This would allow a comparison year on year which this data set can’t achieve, and result in less errors as to skew the result for the time series temperature prediction.

### Future direction

One factor is the need to collect lots more data using the same numeric attributes for more data points, which could allow for a much bigger data set. Secondly to carry out the same experients and evaluate those results as a comparison still using Weka, which would require more time to conduct for evaluation using the same techniques and analyse their effectiveness. Time series has a higher level of mathematical complexity which needs more understanding.

Another option would be to investigate further tuning of the parameters, and with more data try using different lags and statistical transformation by applying various other machine learning algorithms for better forecasting. This could make further reviews to improve the baseline error verses the tuned model error, and reduce the error percentage.

# Conclusion

## Restate your problem statement.

Using time series simple model to predict temperature fluctuations either lower or higher than the actual recorded temperature.

## How do you propose to solve that problem.

Try refining the results by reserving 40% of the data set by taking and running the same experimental analysis on Weka using time series forecasting, and using the advanced configurations available to tune the algorthim parameters.

## 3 Result of proposed method.

A close up of a map

Description automatically generated

|  |  |  |
| --- | --- | --- |
| Model Linear regression | Training(%) 1 step ahead | Testing(%) 1 step ahead |
| Humidity | | |
| N | 737 | 536 |
| Mean absolute error | 0.0106 | 4.4666 |
| Root mean squared error | 0.0324 | 5.1011 |

|  |  |  |
| --- | --- | --- |
| Model Linear regression | Training(%) 1 step ahead | Testing(%) 1 step ahead |
| Wind speed | | |
| N | 737 | 536 |
| Mean absolute error | 0.006 | 10.5159 |
| Root mean squared error | 0.0811 | 13.8262 |

| Model Linear regression | Training(%) 1 step ahead | Testing(%) 1 step ahead |
| --- | --- | --- |
| Temperature | | |
| N | 737 | 536 |
| Mean absolute error | 0.002 | 2.0681 |
| Root mean squared error | 0.0157 | 2.5921 |

This made the results for temperature and wind speed better, but made humidity worse.

## Summarize what was learnt and possible future works.

It is a difficult subject but classifiers which utilise linear regression seem to be a better option with temporal data, which needs algorithms able to cope with specific intervals for prediction or trends. Due to the aspect of using Datetime a support vectar machine based classifier seems to cope quite well with the need to use a higher complexity of maths and the relationship of independent variables being continuous and changing over a time frame.

Key factors as to what was learnt are as follows:

* Time series has to have a lot more instances than attributes in order to give a accurate output for future prediction or predictions.
* It can handle singular or mutlpile targets if you want to vizualise data in a graphical format for a non technical group of people.
* By default time series analysis uses linear regression(Weka’s SMOreg) for numeric data.
* The model is able to use multiple types of regression which utilises error performance as percentages for direct comparisons between training and testing data.
* System can model and compensate for variances within the data as it increases or decreases over time, but more data over a longer period is better for this type of prediction.
* Be aware of what results by evaluation the training data gives out as in time series errors can propagate.
* Time series is classed as advance and as such takes more experiments, and two configuration panels in order to achieve satisfactory results.
* Need to understand that basic configuration is unsupervised but lag creation transforms machine learning into supervised machine learning when tuning any underlying pararmeters.

After researching the topic of time series it can be difficult to predict weather due to values, and data needing to have a high level of accuracy. What has been attempted is a basic approach with with Weka using time series forecasting, and due to time constraints for fine tuning reaching a more accurate a conclusion could be achievable. To try and get accurate values is a big challenge on one attempt, and needs to refined with different parameter settings in a regression model.

Useing a full model may cause overfitting of data, but a simple model can reduce the risk and seemed a good idea as opposed to a wrapper based approach of using all the attributes. In general the results for Linear regression ad SMOREG seemed very similar as regression models.

### Possible future works

Try to get information from another weather data set either from a climate similar to New Zealand, which would offer a comprehensive and direct evaluation. Getting access to additional data from Sydney would be a decent comparison as it lies a bit furthernorth in latitiude to Auckland, and utilise the same time series models in Weka with the same techniques. Another option would be tailor the same models for an area such a east Auckland as to try a different geographical area, but try to improve the result and reduce bias from a trends.

From research it appears that seasonality has been part of evaluation, and sufficient data to cover this problem is needed as realistic factor for accurate predictions or trends.These models couldn’t factor this problem, and alot more analysis through experimemtal modelling with filtered classification is needed but sufficient knowledge is limited for this area.

I could also try and improve the models more by trying other values for the tuned parameters. In conclusion further analysis could determine if any additional attributes for seasonality that may be missing may or could useful or necessary.

# References

Brown, M. (2012, December 11th). *Data mining techniques*. Retrieved from IBM Developer: https://developer.ibm.com/articles/ba-data-mining-techniques/

*Creating and Implementing a Data Collection Plan*. (n.d.). Retrieved from http://www.strengtheningnonprofits.org/resources/e-learning/online/datacollection/Print.aspx

*Data Structure and Algorithms - Tree*. (n.d.). Retrieved from tutorialspoint: https://www.tutorialspoint.com/data\_structures\_algorithms/tree\_data\_structure.htm

Deloitte Development LLC. (2013). *Developing an effective governance operating model.* Retrieved from Deloitte: https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Financial-Services/dttl-fsi-US-FSI-Developinganeffectivegovernance-031

Elgabry, O. (2019, March 1st). *The ultimate guide to data cleaning*. Retrieved from Towards Data Science: https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4

Fu, C., Qian, C., & Wu, Z. (2011, 07 23). *Projection of global mean surface air temperature changes in next 40 years: Uncertainties of climate models and an alternative approach*. Retrieved from Springer Link: https://link.springer.com/article/10.1007/s11430-011-4235-9

Hall, M. (2014, 3 24). *Time Series Analysis and Forecasting with Weka*. Retrieved from Pentaho Community Wiki: https://wiki.pentaho.com/display/DATAMINING/Time+Series+Analysis+and+Forecasting+with+Weka

han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques.* Massachusetts: Morgan Kaufmann.

IBM. (n.d.). *Data integration*. Retrieved 03 29, 2019, from IBM: https://www.ibm.com/analytics/data-integration

Liu, Z., Loo, C., Masuyama, N., & Pasupa , K. (2018). Recurrent Kernel Extreme Reservoir Machine for Time Series Prediction. *IEEE Access*, 19583 - 19596. Retrieved from IEEE Xplore: https://ieeexplore.ieee.org/abstract/document/8331837

M.Ramageri, B. (2010, December). *Data mining techniques and applications*. Retrieved from ResearchGate: https://www.researchgate.net/publication/49616224\_DATA\_MINING\_TECHNIQUES\_AND\_APPLICATIONS

Niwa. (n.d.). *Fact sheet for HPCF*. Retrieved from Niwa: https://www.niwa.co.nz/our-services/high-performance-computing-facility/hpcf-fact-sheet

Niwa. (n.d.). *High Performance Computing Facility*. Retrieved from Niwa: https://www.niwa.co.nz/our-services/high-performance-computing-facility

predictiveanalyticstoday.com. (2018). *What is data mining*. Retrieved from Pat Research: https://www.predictiveanalyticstoday.com/what-is-data-mining/

Witten, I. H., Frank, E., & Hall, M. (2011). *Data Mining: Practical Machine Learning Tools and Techniques (3rd Edition) .* Burlington, MA: Morgan Kaufmann.