Can Data Analysis Techniques be applied to historical data to provide quantitative short term weather prediction?

Table of Contents

[Table of Figures 3](#_Toc517092292)

[Abstract 4](#_Toc517092293)

[1. Introduction 5](#_Toc517092294)

[2. Literature Review 6](#_Toc517092295)

[2.1 Overview of Existing Technologies 6](#_Toc517092296)

[2.1.1 J48 and Decision Trees 6](#_Toc517092297)

[2.1.2 K Nearest Neighbour 7](#_Toc517092298)

[2.1.3 Density Based Spatial Clustering of Applications with Noise 7](#_Toc517092299)

[2.1.4 K-Means Clustering 8](#_Toc517092300)

[2.1.5 Feed-Forward Neural Networks (Multilayer Perceptron’s) 8](#_Toc517092301)

[2.1.6 Recurrent Neural Network 9](#_Toc517092302)

[2.1.6.1 Elman recurrent neural network 9](#_Toc517092303)

[2.1.6.2 Hopfield Model 9](#_Toc517092304)

[2.1.7 Radial Based Function Network 10](#_Toc517092305)

[2.1.8 Support Vector Machines 10](#_Toc517092306)

[2.1.9 Principal Component Analysis 10](#_Toc517092307)

[2.2 Existing Data Mining Studies 11](#_Toc517092308)

[2.3 Current Numerical Weather Prediction Systems 16](#_Toc517092309)

[2.3.1 Numerical Weather Prediction Computing Power 19](#_Toc517092310)

[3. Methodology 21](#_Toc517092311)

[3.1 Dataset Description 21](#_Toc517092312)

[3.2 Cluster Analysis 21](#_Toc517092313)

[3.3 Rainfall Prediction 22](#_Toc517092314)

[3.3.1 Selected Algorithm 22](#_Toc517092315)

[3.3.2 Cross Validation Strategy 23](#_Toc517092316)

[3.3.3 Evaluation Metric 24](#_Toc517092317)

[References 25](#_Toc517092318)

# Table of Figures

[**Figure 1**. Graphical implementation of walk forward cross validation. Image sourced from (Nawara, 2017) 24](#_Toc517092253)

# Abstract

# Introduction

In this document an analysis of existing studies that utilise data science technologies for use in weather forecasting systems is outlined. Following this an exploratory data analysis of historical weather data obtained from Met Éireann will be completed. Using the results of the exploratory data analysis as well as knowledge obtained from the literary analysis stage, the aim is to apply multiple machine learning algorithms to the data to generate short term weather predictions.

Accurate short term weather prediction has the potential to greatly influence a number of industries with agriculture being one of them (Sivakumar, 2006). Up until recently, numerical weather prediction would have been considered a task for supercomputers, and to a certain extent, still is (Coiffier, 2011). With recent advances in computing power and algorithm development, this task can now be attempted without the excessive hardware requirements.

The remainder of this document will continue as follows; [Section 2](#Literature Review) will comprise of the literature analysis. This sections will be further partitioned into two subsections. [Section 2.1](#Overview of Existing Technologies) will provide a brief description of each of the algorithms that are used by existing studies. [Section 2.2](#Existing Data Mining Studies) will provide a broader outlook on each of these studies, such as any interesting findings that they make and how well the algorithms used perform. [Section 2.3](#2.3 Current Numerical Weather Prediction Systems) provides an insight into past and current numerical weather prediction practices as well as the weather models applied by Met Eireann to generate their forecasts.

# Literature Review

Agriculture is just one of the industries that can benefit from technological advances in weather prediction. (Sivakumar, 2006) outlines that accurate long term weather prediction can be used to mitigate risk in agriculture by helping predict the success or failure of an agricultural season. This would also lead to economic benefits, where inputs could be restricted in the event that the long term weather forecast is not agreeable. Similarly, accurate weather prediction could also be used to improve crop establishment and overall yields by utilising forecasts to determine optimal planting and harvesting times.

Precise weather prediction using predictive analytics would also assist local authorities in informing the population about incoming extreme weather events, potentially saving lives and livelihoods. (Huang and Ran, 2003) outline a traffic speed prediction model based on a neural network that determines the optimal speed under certain adverse circumstances such as severe weather events.

Given a dataset with the relevant information, weather prediction could be used to estimate the total hours of sunshine per day. This information could then be put to use in the solar energy industry for more efficient and cost-effective energy generation. These examples epitomise the range of applications that accurate weather prediction can influence.

Therefore the aim of this research is to determine if data analysis techniques can be used to accurately predict quantitative short term weather forecasts using past data. To implement this research, past data is programmatically retrieved from Met Éireann’s website to be analysed and used as a basis for the predictions.

## Overview of Existing Technologies

The intention of this section is to provide a brief synopsis on the technologies that are presented in [section 2.2](#Existing Data Mining Studies) for use in existing studies.

### J48 and Decision Trees

The J48 (Gholap, 2012) algorithm is a Java implementation of the C4.5 (Salzberg and Segre, 1994) decision tree algorithm. Decision trees can be used for both classification and regression, but as the existing studies in [section 2.2](#Existing Data Mining Studies) only used them from a classification perspective, this description will only describe them from a classification context. In its simplest form a decision tree can be thought of as a series of if-else statements that provide the best route to the outcome.

Decision tree’s look at each feature independently. As such, the decision tree algorithm searches over all possible values of the feature at the current node to find which value best splits the data into the desired classes. The algorithm will continuously split the data in this fashion and ultimately reach its leaf nodes where it will have correctly classified all observations.

### K Nearest Neighbour

Like decision trees K Nearest Neighbour (KNN) can be used for both classification and regression (James et al., 2013). Numerous studies found in [section 2.2](#Existing Data Mining Studies) use KNN mainly from a classification perspective. In terms of classification KNN determines the K closest observations to the current observation using a distance metric such as Euclidean distance, which is the straight-line distance between two points. Following this the algorithm then observes the class of each of these K closest observations and uses a voting mechanism whereby the most popular class is assigned to the unlabelled observation. To avoid having a tie in the voting mechanism, the value of K is generally set to an odd number. K = 3 or K = 5 often work quite well but are liable to overfit the data if the number of observations is large.

In terms of regression the algorithm works in virtually the same manner, except in this case, the test observation is given the mean or median value of the target variable from the K nearest observations. Every algorithm has its disadvantages and KNN is no different. The computation time with KNN increases significantly if the number of observations and dimension of the dataset is large (C. Müller and Guido, 2016).

### Density Based Spatial Clustering of Applications with Noise

Density Based Spatial Clustering of Applications with Noise (DBSCAN) is a clustering algorithm often used as there is no requirement to specify the number of clusters to be found and it can also determine and mark outliers in the data (C. Müller and Guido, 2016). The algorithm operates by finding areas of the feature space that are dense with similar observations.

The two main parameters in DBSCAN (*min\_samples and eps*) determine the minimum number of samples and maximum distance of these samples from an arbitrary point for it to be considered a core sample. Core samples that are closer to each other than the value of eps form a cluster.

### K-Means Clustering

The K-Means algorithm was initially proposed by (MACQUEEN, n.d.). K-Means starts by picking a user specified amount of K arbitrary observations from the data and makes them the cluster centers. Following this, the algorithm finds the closest points to each of the cluster centers. Every time an observation is added to the cluster, the cluster center is updated by taking the mean of the observations within the cluster. This process repeats until all observations are assigned to a cluster (C. Müller and Guido, 2016).

The authors state the disadvantages of using K-Means are you need to know how many clusters you are looking for and even that may not be enough as each cluster is defined by its central observation. Therefore, the algorithm only finds clusters with an approximately spherical shape.

### Feed-Forward Neural Networks (Multilayer Perceptron’s)

Multilayer Perceptron’s (MLP) commonly referred to as Feed-forward neural networks or more generally as Artificial neural networks (ANN), were initially proposed by (McCulloch and Pitts, 1943). In their most simple state, Feed forward neural networks compute a weighted sum of the input features which is then passed to a hidden layer consisting of a user defined number of hidden nodes. The hidden layer then performs a weighted sum of its input features and applies a nonlinear function to the result (C. Müller and Guido, 2016). The result of this is then used as part of a weighted sum for the output node. Without the addition of the nonlinear function, the algorithm would only perform the same operations as linear or logistic regression. The more hidden layers included, the more coefficients the algorithm needs to learn.

(C. Müller and Guido, 2016) state an important property of neural networks is that the coefficients (weights) are randomly initialised. This affects the model that is learned. In such situations, changing the random seeds can increase or decrease the models accuracy. The authors also state that neural networks expect the data to have an approximately normal distribution so using the StandardScaler (sklearn.preprocessing.StandardScaler — scikit-learn 0.19.1 documentation, n.d.) which scales the data so it has a mean of 0 and standard deviation of 1 is advisable.

Often when reading about neural networks, feed-forward and back propagated architectures can be referred to interchangeably. Back propagated neural networks (Dreyfus, 1990) are different in that they provide a significant improvement on feed-forward neural networks as they allow for errors to be propagated back through the network so the weights can be updated to model the problem more accurately (Mazur, 2015).

### Recurrent Neural Network

Recurrent Neural Networks (RNN) are primarily used with sequential data (Goodfellow et al., n.d.). The authors state that the key difference between feed-forward neural networks and RNN’s is that recurrent networks share the same weights for observations over multiple time steps. One of the main benefits of RNN’s is they can be supplied with sequences of different lengths (Géron, 2017) which is why they are often used for speech and text recognition.

RNN’s tend to suffer from the exploding/vanishing gradients problem (Géron, 2017). This is the problem where the errors that are being propagated back through the network only make very small changes (vanishing gradients) or the opposite of very large changes (exploding gradients). Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) RNN’s are one method of combating this problem. In essence, an LSTM RNN can determine what information to store in long term memory, what information to store in short term memory and what information serves no purpose.

### Elman recurrent neural network

Elman RNN’s (Elman, 1990) are considered a simple form of RNN that use context units to provide the network with states from the previous time step (Jones, 2017). The outputs from each hidden layer neuron are stored in the context units for the next set of inputs but are also used in the output neurons.

### Hopfield Model

A Hopfield model (Hopfield, 1982) is another form of recurrent neural network. (Géron, 2017) refers to Hopfield networks as associative memory networks. The model works by learning patterns in the data and once it is given new input, it matches the new input to the closest learned pattern. The key feature of this architecture is that every neuron is connected to every other neuron in the graph.

### Radial Based Function Network

Radial based function networks (RBFN) (Broomhead and Lowe, 1988) are a form of neural network with three layers; input layer, hidden layer and output layer. RBFN’s are generally used to model non-linear problems. RBFN is very similar to other distance-based models such as KNN where it attempts classify observations based on their similarity to other observations in the dataset (McCormick, 2013). Each neuron in the hidden layer keeps a single observation that it uses for comparison between incoming observations. Generally, the gaussian radial basis function (RBF) is used as the activation function in the hidden layer. This function provides an output between 0 and 1 (Géron, 2017) which can be used as a probability of class membership. (McCormick, 2013) states that the stored observation is treated as the center which all distances are measured from.

### Support Vector Machines

Support Vector Machines (SVM) (Cortes and Vapnik, 1995) can be used for both classification (SVC) and regression (SVR) tasks. SVM’s work by finding the support vectors (observations) that are closest to a decision boundary that separates the observations appropriately (Géron, 2017). Aurélion Géron provides an interesting analogy where he compares SVM for classification to a street where the central road markings represent the decision function and the markings on each side of the road mark the distance the support vector (point that’s closest to the decision function) is from the decision function.

### Principal Component Analysis

Principal Component Analysis (PCA) is primarily used as a dimensionality reduction technique. PCA is often used in high dimensional datasets to find a certain number of principal components that best describe the maximum variance within the data (Géron, 2017). To find the principal components the data is projected onto a hyperplane that best splits the data which also preserves the maximum variance of the data, or as close as possible to it. PCA will find as many principal components as there are features in the dataset each of which is orthogonal to the previous principal components.

The main disadvantage of transforming a dataset to its principal components is the dataset is no longer interpretable. As a principal component is based on a complex combination of features, the user no longer knows what each column of the transformed data represents making it difficult to determine which features in the original dataset are most influential.

## Existing Data Mining Studies

There are multiple existing papers that engage in the task of weather prediction, each with their own unique outlook on the problem. (Talib et al., 2017) use the J48 and decision tree algorithm to perform an analysis on weather data from 2007 to 2016. Unlike many other studies that use machine learning algorithms to predict weather values or the occurrence of particular weather events, the authors instead determine association rules for the weather i.e. under what circumstances particular weather events will occur. For example, the authors system will determine that rainfall will occur if the temperature is 10 degrees Celsius and the wind speed is 30 km/h.

(Sharma et al., 2014) use a combination of DBSCAN and KNN algorithms to cluster similar data points then assign each cluster to a specific weather class. The authors state that their system will predict the occurrence of fog, rain and snow to within 90%, 67% and above 93% accuracy respectively. Although this is a relatively successful system, it is worth nothing that this system does not quantify the weather event, it simply classifies its occurrence.

(Kalyankar and Alaspurkar, 2013) use the K-means clustering algorithm to create clusters of weather data which they then perform an analysis on for knowledge discovery. For the purpose of knowledge discovery any clustering algorithm should be appropriate but using the DBSCAN clustering algorithm has two distinct benefits; the number of clusters does not need to be supplied to the algorithm and it also determines and marks outliers in the data meaning it can be used for outlier analysis.

(Olaiya and Adeyemo, 2012) apply two forms of neural networks and a decision tree on data spanning ten years to predict a combination of weather phenomenon such as maximum temperature, rainfall, evaporation and wind speed. In summary the authors used the decision tree to determine association rules resembling the work of (Talib et al., 2017). Following this they implement a time lagged feed forward neural network and recurrent neural network to make predictions. Overall the time lagged feed forward neural network performed better with an error of ~24%.

(Jan et al., 2008) perform seasonal climate prediction using the KNN algorithm based on ten years of past data. In this scenario, the data consisted of seventeen features based on ten locations. The authors found that when predicting a Boolean attribute, such as the presence of fog or snow, they could achieve accuracies of 96.6% and greater.

(Petre, n.d.) uses a decision tree for temperature prediction modelled as a classification problem where the output temperatures are transformed into certain ranges determined by the author. The author uses data collected from 2002 to 2005 for Hong Kong. The model is evaluated under numerous classification metrics for each of the temperature range classes. The system resulted in training accuracies of 83.33%. Unfortunately the author did not evaluate the trained model on an independent test set meaning the true significance of the model can not be determined.

(Al­Roby and Alaa M, 2011) performed numerous data mining techniques to determine wind speed, which was again treated as a classification problem. The authors used ten years of historical daily data for use in their case study. The authors perform some interesting transformations of the data so each observation contains the windspeed values for the previous two days. Following this the target column of windspeed is discretized. The authors approach the problem using multiple techniques such as association rule mining, classification and clustering. In terms of classifying future wind speeds, the authors use two algorithms KNN and a feed forward neural network. The authors note that KNN and the feed forward neural network reach 62.70% and 67.37% accuracy respectively.

Similarly (Kohail and El-Halees, 2011) perform numerous data mining techniques on weather data from 1977 to 1985. The techniques performed include outlier analysis, clustering, numerical prediction, classification and association rule mining. The authors perform an interesting operation called linear interpolation which is used to fill in missing values between a known amount of data points by fitting a polynomial curve to the data. This operation is often used to fill in missing data in time series problems. Like (Al­Roby and Alaa M, 2011), the authors create three new variables in the dataset that represent the previous three days temperatures, for each observation. After performing an outlier analysis, the results indicate that the outliers contain both real and input error observations. Instead of removing the incorrect observations only, the authors decided to remove all outliers.

In terms of prediction (Kohail and El-Halees, 2011) use an ANN and least median squares linear regression to predict daily average temperature. This results in the ANN having a lower correlation coefficient between the actual and predicated temperatures. The authors also perform daily temperature prediction from a classification perspective using four models. Like previous studies, the data is discretized so temperatures are classed as cold, warm or hot. The algorithms used are Naïve Bayes, KNN, decision trees and an ANN. The individual accuracies range from 81.40% to 85.77% where the best accuracy was produced by the ANN.

(PAL et al., n.d.) outline how they used a back-propagated ANN to predict minimum and maximum ground level temperatures. The authors perform some basic feature engineering by shifting the data so it includes measurements from the previous two days like many of the previous authors have already done. By doing this, the ANN should be able to look at the effects of previous weather events on each observation. The authors found that the optimal ANN had an error rate of 2 °C 80% of the time.

(Nagalakshmi et al., 2013) provide a description of existing papers and provide some recommendations for future works. The authors note that a radial based function network was the best form of ANN but they also say to get the best results overall an ensemble approach should be used. Interestingly meteorologists use an ensemble approach when determining the optimal forecast using their existing statistical numerical weather prediction methods (Flynn, n.d.).

(De and Debnath, 2009) use three back-propagated ANN’s to predict minimum and maximum temperature in the months of June, July and August. The ANN’s were trained on data from the months of December to May. The authors note that the ANN for predicting Augusts temperatures was very accurate with a prediction error of approximately 5%.

(Maqsood et al., 2004) use an ensemble approach to predict weather forecasts 24 hours ahead. The predicted weather forecast consists of temperature, wind speed and humidity. The authors ensemble consists of a multi-layered perceptron (MLP), Elman recurrent neural network (ERNN), radial basis function network (RBFN) and Hopfield Model (HFM). The authors train the models on hourly data based on all four seasons but use an interesting validation strategy where they remove one extreme weather observation from each season and place them in their test set. The authors compare the models on an individual basis and as part of two ensemble methods. The fist ensemble method uses a weighted average to determine the result whereas the second ensemble model uses a winner takes all approach. RBFN was deemed to be the best individual model in terms of accuracy and training time. Regarding the ensemble approach, the authors found the winner takes all ensemble to have the smaller prediction error of the two.

(Radhika and Shashi, 2009) use an SVM to predict maximum atmospheric temperature 24 hours ahead. The authors compare the results of an SVM to the results achieved by a back-propagated MLP. Data spanning from 2003 to 2007 was used to train the models. Data from January to July in 2008 is treated as the test data. Unlike previous studies that have used linear interpolation to fill in missing values, the authors populate empty cells with the average temperature for that specific month. The authors found that the SVM consistently performed better than the MLP at predicting the maximum temperature for the following day.

(Wang and Sheng, 2010) compare a generalised regression neural network which is similar to an RBFN, with a back-propagated neural network for long term rainfall prediction between the years 1955 and 2009. The authors state that one of the main challenges with rainfall prediction is its characteristic of being non-linear over time. The generalised regression model is determined as the superior of the two, as it consistently has a lower mean standard error (MSE) than the back-propagated neural network.

(Gumaste and Kadam, 2016) propose a weather prediction system using a genetic algorithm and fast Fourier transform (FFT) aimed at assisting the agricultural community. In essence the proposed system implements the genetic algorithm alongside FFT to observe previous weather events for the same day in previous years, by taking averages of these past values and comparing them to the actual outcomes.

An interesting plant monitoring system is created by (Kurniawan et al., 2017) for use in the agricultural sector. Through the use of past weather data and data obtained from sensors for soil moisture levels for example, the authors created a system that determined whether a plant would need irrigation or not. The authors implement a system which uses fuzzy logic to determine the weather. The system is tested 33 times in comparison to weather recordings from external sources. On all occasions the system is claimed as being 100% accurate.

(Pandey et al., 2017) implement an adaptive neuro fuzzy inference system (ANFIS) to perform weather classification. As the system is using fuzzy logic, the problem is treated as a multi-class classification task. The authors take an unusual approach to cleaning the dataset by applying the wordcount program from Hadoop on the dataset. The ANFIS model achieves a relatively low MSE of approximately 1.42 indicating it is a reasonable model.

(Saha and Chauhan, 2017) attempt to predict low temperature, high temperature, humidity and wind speed using a non-linear autoregressive neural network. The data used in this study is based on daily observations over a time span of 45 years. As the model was evaluated using multiple metrics, the authors noted that the optimal number of hidden neurons ranged between 3 and 5 for each of the targets.

(Qu Xiaoyun et al., 2016) compare an SVM to a Long Short Term Memory (LSTM) neural network for next day wind power prediction. Principal component analysis (PCA) is used as a dimensionality reduction technique in this study which finds the directions of maximum variance within the data. Generally when PCA is used the dataset contains a lot of features. In this case the original dataset contained five features, and after the PCA transformation only two components are used as input. The authors determine that these components are related to wind speed and direction. The authors compare the results of the LSTM with raw NWP input and PCA input. The results showed that the LSTM neural network with PCA had a lower normalised RMSE. When compared with the SVM’s output, the PCA LSTM neural network performed significantly better.

(Nurunnahar et al., 2017) implement wind speed prediction using Support Vector Regression (SVR) and back-propagated ANN’s. The authors performed quite an extensive study by building multiple models for each algorithm to provide forecasts for different time intervals. A sliding window validation strategy was used which is specific to time series prediction. In principle for each training iteration the training set increases in size. As the dataset used in this study was sparse, the Leaky ReLu function was used to replace the zero’s with some values that did not disturb the existing relationship in the data. In terms of feature engineering a relatively straightforward approach was taken where the average, minimum and maximum wind speeds were included, although the timescale these features related to was not supplied. The results of this study were superb. Bearing in mind that the authors created models for 1, 3, 5, 7, 10 and 15 day ahead forecasts, SVR had accuracies ranging from 90.10% to 99.60%. In this case the 99.60% represented 3 day ahead forecasts. The back-propagated ANN’s on the other hand had accuracies ranging from 64.90% to 99.80%. In this scenario the accuracy score of 99.80% represented a 5 day ahead forecast.

(Sanusi and Corne, 2015) address the topics of feature selection and feature engineering for short term wind speed forecasting. The authors look at the effect of derived features on the overall outcome, i.e. the difference in wind speed between the current and previous timestep. The authors found that using derived features on data with a 10 minute timestep did not improve the performance of the models in this scenario. This could be due to the increase in dimensionality causing an increase in model complexity. Numerous experiments using different time intervals were used. As the time intervals increased, the influence of the derived features also increased although this influence was not considered overly significant.

## Current Numerical Weather Prediction Systems

As one could imagine numerical weather prediction (NWP) is a vast and complicated field. The principle of NWP was realized by an English meteorologist, Lewis Fry Richardson (Flynn, n.d.). The system implemented by Richardson was fundamentally based on using past data, although a very small amount, to predict the air pressure in the next six hours. The fact that even the earliest attempts at NWP used past data to perform predictions highlights the need for incorporating data analysis techniques, both descriptive and predictive, into modern NWP systems. The calculations required to make a prediction took six weeks further highlighting the complexity of the task at hand.

NWP is used in order to get approximations of the atmosphere in numerical format, as the name suggests (Coiffier, 2011). The equations used to get these approximations are beyond the scope of this document but are strongly linked to fluid mechanics. A model of the atmosphere is formed using these calculations which is then discretised, so the model no longer reflects the continuous variables. This process reflects the series of actions some of the aforementioned data mining applications have taken to make weather prediction a classification problem rather than a regression problem. Although the NWP systems are still predicting a numeric value, the range of values is restricted in a sense.

In essence NWP is based on modelling the atmosphere which is treated as a fluid, hence why it is strongly linked to fluid mechanics. As NWP is considered an initial value problem, the success of a forecast as largely related to the accuracy of the initial values used to represent the atmospheres current state.

Improving weather prediction is highly correlated with improving NWP processes (Shuman, 1978). The atmosphere and its constituents are at the heart of NWP. Rainfall is possibly the most important weather factor from an individual’s perspective. Rainfall is largely related with the radiative properties of oceans and seas. As heat takes longer to dissipate from a body of water than a mass of land, evaporation can continue for longer over a body of water. This increase in evaporation over oceans leads to large quantities of clouds that become dense with water vapour. As these cloud systems are pushed in certain directions by the prevailing winds, the cloud systems make landfall. When these clouds, dense with water vapour, meet the configuration of the earth’s surface, rainfall occurs.

By utilising NWP processes to model atmospheric conditions these weather phenomena can be forecasted in advance. (Shuman, 1978) outlines that heating and friction are important for determining an atmospheric forecast which extends a week in advance. Interestingly the author gives an overview of what they perceive to be the most influential features of NWP systems; three of the five are related to advances in technology. The effectiveness of technological advances on NWP systems is further highlighted when the author states that these advances are heavily reliant on progress in computational domains.

The main analysis of this document, both descriptive and predictive, will be based on hourly data obtained from Met Éireann. Hence this section will continue by looking at the NWP processes and models employed by Met Éireann in their weather forecasting systems.

Fortunately, Met Éireann provide a brief overview of their entire weather forecasting process (*How Met Éireann produces a forecast - Met Éireann - The Irish Meteorological Service*, n.d.). As is the case with all NWP systems, the first step is to collect observations of current and past weather conditions. Following this, the process enters the data assimilation stage. Numerous papers referred to in this document, such as (Shuman, 1978) , state that NWP is an initial value problem. Therefore, Met Éireann use the data assimilation stage to compare previous forecasts with the actual observations. This process of data assimilation is very similar to model training techniques in the data science world. In essence when creating a predictive model, an initial model is created with basic parameters and a subset of the data. When the model is run and evaluated under some performance metric it is often re-run with updated parameters to increase the predictive accuracy of the model which is approximately the purpose of data assimilation.

Given the complexity of NWP and its heavy reliance on computing power, its intriguing to know that forecasters in Met Éireann still adopt a manual approach when creating forecasts on certain occasions. Forecasters will sometimes draw charts by hand to determine what the weather observations are indicating (Forecasting Centre - Met Éireann - The Irish Meteorological Service, n.d.).

Met Éireann uses a combination of two forecasting models for their weather predictions, HIRLAM and HARMONIE. HIRLAM, which stands for High Resolution Local Area Model, is used to provide short term weather forecasts. This model has been developed by a number of countries as part of a consortium. A medium range forecasting model implemented by the ECMWF is used to provide weather forecasts approximately ten days in advance.

When reading about NWP, resolution is a term that surfaces more often than not. The resolution of a weather forecasting model is not necessarily an accuracy measure but an indicator of how localised a weather model can be. Met Eireann’s NWP models for example have a resolution of approximately 2.5km (Numerical Weather Prediction - Met Éireann - The Irish Meteorological Service, n.d.). This means that the model splits the country into grids of 2.5km2 and then approximates the atmospheric conditions for each of those grids. Although this resolution seems quite good from a newcomer’s point of view, it is stated that observations would need to be recorded over every couple of meters to have a completely accurate measure of the atmospheric conditions. It is also stated that, as these recordings are not attainable, assumptions need to be made. This characteristic again resembles that of any machine learning problem where assumptions need to be made about the data and problem being undertaken in order to perform predictions.

Mitigating the errors introduced by making assumptions is crucial in machine learning. This is no different in terms of NWP models. To avoid the errors introduced by these assumptions an ensemble approach is used whereby numerous forecasting models are built using a different initial state of the atmosphere for each model. The idea of using an ensemble approach is again common in the machine learning and predictive analytics world whereby numerous models are built using the same (with different hyperparameter values) or different algorithms. The results of these algorithms are then averaged, using a weighted average for example, to get an overall result.

(Andersson, 2014) states that the ECMWF medium range prediction is based on an ensemble of 52 models. One of these models is at a much higher resolution than the other 51 models meaning the initialisation values used for that model are the most accurate description for the current state of the atmosphere. A second model takes the opposite approach where it uses a much lower resolution than all the rest. This results in an individual model that has a less accurate but broader outlook. The remaining models are then initialised with values within the range of the high resolution and low resolution models.

As opposed to the ensemble machine learning approach, where the results are effectively averaged, the result of ensemble weather forecasting is to provide multiple forecasts that account for the error introduced in making assumptions about the atmosphere (Jeppesen, 2017). This approach of producing multiple forecasts enables the meteorologists to refrain from making a deterministic forecast which could potentially be incorrect. The importance of ensemble prediction models is accentuated by (Barratt, 2017) whereby ensembles are considered the way forward for accurate weather prediction.

### Numerical Weather Prediction Computing Power

As outlined by (Shuman, 1978) increases in computing power is largely associated with advances in NWP. Even by today’s standards, basic CPU’s alone are not sufficient for weather forecasting. Continuous developments of GPU’s has allowed problems which are largely concerned with complex calculations to become more mainstream. The reasons that GPU’s have enabled this is their innate ability to perform linear algebraic calculations anywhere from 10 to 100 times faster than CPU’s. This, combined with their ability to parallelise operations makes them an optimal solution in any computing intensive operation.

(Michalakes and Vachharajani, 2008) demonstrate the power of GPU’s by running a computationally intensive NWP program on multiple GPU’s and provide and analysis of the results in comparison to a traditional high-powered CPU based system. The authors experiment was run on a cluster containing 64 CPU-GPU pairs. The authors found that incorporating GPU’s into a system decreased computation time by orders of magnitude. Interestingly the NWP models used by Met Éireann are run on supercomputers that contain multiple high powered GPU’s further highlighting the scale of NWP problems.

(Coiffier, 2011) outlines the scale of the calculations and computational requirements to make a 24 hour forecast using the ARPEGE model that is used by the French. The computer used in this system is powerful enough to create a 24 hour forecast in under 15 minutes.

# Methodology

The purpose of this section of the document is to outline the technologies and techniques that will be applied to the data in an attempt to generate weather predictions as outlined below in [section 3.2](#Cluster Analysis) onwards.

## Dataset Description

As previously mentioned in section 2, the data used in this study will be programmatically retrieved from the Met Éireann website. The dataset will consist of the following variables:

* **wetb:** Wet bulb Air Temperature (**°C**) is used to measure the extent of cooling as moisture dries on a surface (Temperatures - Dry Bulb/Web Bulb/Dew Point, n.d.).
* **dewpt:** Dew point Air Temperature(**°C**) is used to determine the temperature at which the air can no longer retain moisture. This figure should always be less than or equal to the value of the actual temperature. If air temperature cools to dew point, or the dew point rises to the current temperature then fog or clouds appear (Weather Questions & Answers, 2012).
* **rhum:** Relative Humidity describes how close the air is to saturation with moisture at a given temperature (Graham, 2014). Therefore, a high relative humidity (100%) indicates there’s more moisture in the air.
* **msl:** Mean Sea Level Pressure (**hpa**) is the atmospheric pressure recorded at sea.
* **vappr:** Vapour Pressure (**hpa**) exerted by water vapour in the air (Water Vapour - Met Éireann - The Irish Meteorological Service, n.d.).
* **date:** Date and Time (**utc**).
* **rain:** Rainfall (**mm**).
* **wdsp:** Windspeed (**kt**).
* **wddir:** Wind direction (**degrees**).
* **Station Name:** Name of weather station where the observation was recorded.
* **Station Height:** Height of the weather station above sea level.
* **Longitude & Latitude:** Coordinates of the weather stations location.

## Cluster Analysis

Cluster Analysis, using the K-Means clustering algorithm, will be performed to determine any seasonal patterns that occur in the data. This process will be performed to gather an understanding of how the weather elements change from season to season. Some of these changes will be expected, for example temperature values in the winter months will be lower than temperature values in the summers months, on average. It is expected that the clustering algorithm will correctly cluster the observations from the four seasons of spring, summer, autumn and winter. This processing step may also serve as a form of outlier analysis. In the event that the algorithm does correctly separate the majority of observations belonging to each season, there are likely to be errors in each of the clusters which may be considered as outliers. An analysis of these outliers will follow, assuming they are present.

As the aim of this processing step is to cluster the data into the four weather seasons, an algorithm that will generate a user defined number of clusters is required. Therefore, as previously alluded to, the K-Means clustering algorithm will be used as it fits this requirement fully.

## Rainfall Prediction

Following the implementation and analysis of section [3.2](#Cluster Analysis) the next step will be to attempt rainfall prediction. Initially the prediction will provide rainfall forecasts an hour in advance. As the data is already in one-hour intervals, this is a natural starting point. It is possible that the dataset size may influence the training time of the selected algorithm meaning that hourly prediction will not be possible. In that case, the data will be aggregated into intervals that are divisible by 24 (i.e. predictions less than a day in advance) until the algorithm is capable of training and providing predictions within that chosen interval.

### Selected Algorithm

The algorithm that will be used to formulate rainfall predictions is XGBoost (XGB) (*Introduction to Boosted Trees — xgboost 0.72 documentation*, n.d.). XGB is a gradient boosted tree machine learning model. Gradient boosted trees function by creating multiple decision trees sequentially but each tree fits to the residuals of the previous tree which combines multiple weak learners into one strong learner (Géron, 2017) .

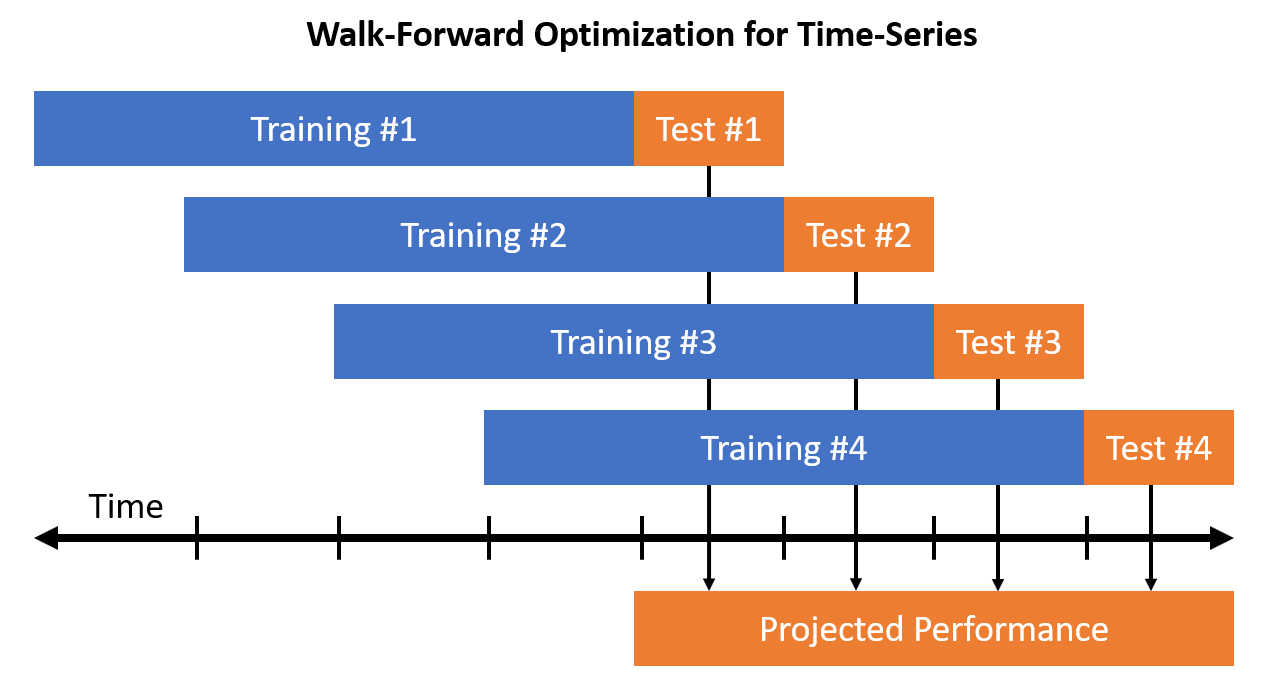
There are several reasons for selecting XGB for rainfall prediction. Firstly, despite the relatively large amount of rainfall experienced in Ireland, rainfall does not occur every day. Hence there will be a significant amount of zero values in the data. Most machine learning models have difficulty dealing with sparse data but XGB and tree based models in general do not suffer from this problem.

Furthermore, tree based models are also capable of finding linear and non-linear patterns in the data. This makes them a popular selection in many machine learning tasks. The final reason for choosing XGB is it has not been implemented in any of the existing studies in section [2.2](#Existing Data Mining Studies) despite the previously outlined advantages.

### Cross Validation Strategy

Cross validation is an important element in training a model. Cross validation is used to increase a models generalisation performance by training on a subset of the data and testing against another unseen subset of the data (*3.1. Cross-validation: evaluating estimator performance — scikit-learn 0.19.1 documentation*, n.d.).

As rainfall prediction will be treated as a time series problem based on the dataset described in section [3.1](#Dataset Description), most conventional cross validation techniques such as K-Fold cross validation for example, will not suit as they do not maintain temporal order during the training process or alternatively, may suffer from data leakage. Based on these requirements, a method of cross validation commonly referred to as walk forward cross validation or walk forward optimisation, will be used. Walk forward cross validation operates by partitioning the data into a train a test set like numerous other forms of validation techniques, but in this case only a small sample of the entire dataset is used for both the training set and test set. Traditionally the training and test set are split 75:25. After each validation round completes the training and test sets are offset by the number of elements in the first test set, thus maintaining temporal order of the data. A graphical example of walk forward cross validation can be found in figure 1.



**Figure 1**. Graphical implementation of walk forward cross validation. Image sourced from (Nawara, 2017)

### Evaluation Metric

# References

*3.1. Cross-validation: evaluating estimator performance — scikit-learn 0.19.1 documentation* (no date). Available at: http://scikit-learn.org/stable/modules/cross\_validation.html (Accessed: 18 June 2018).

Al­Roby, M. F. and Alaa M, . El­Halees (2011) ‘Data Mining Techniques for Wind Speed Analysis’, 2(1), p. 6. Available at: https://www.researchgate.net/publication/276144000\_Data\_Mining\_Techniques\_for\_Wind\_Speed\_Analysis?enrichId=rgreq-3f2673005d987f1865d014062c0f1695-XXX&enrichSource=Y292ZXJQYWdlOzI3NjE0NDAwMDtBUzoyMjgwMjUzMTI3MzkzMjlAMTQzMTM3Njg3OTg0Mg%3D%3D&el=1\_x\_2&\_esc=publicationCoverPdf (Accessed: 1 June 2018).

Andersson, E. (2014) *Medium-range forecasts*, *ECMWF*. Available at: https://www.ecmwf.int/en/forecasts/documentation-and-support/medium-range-forecasts (Accessed: 28 May 2018).

Barratt, M. (2017) *25 years of ensemble forecasting at ECMWF*, *ECMWF*. Available at: https://www.ecmwf.int/en/newsletter/153/meteorology/25-years-ensemble-forecasting-ecmwf (Accessed: 28 May 2018).

Broomhead, D. S. and Lowe, D. (1988) ‘Multivariable Functional Interpolation and Adaptive Networks’, p. 35. Available at: https://web.archive.org/web/20140714173428/https://www.complex-systems.com/pdf/02-3-5.pdf (Accessed: 11 June 2018).

C. Müller, A. and Guido, S. (2016) *Introduction to Machine Learning with Python. A Guide For Data Scientists.* First Edition. O’Reilly.

Coiffier, J. (2011) *Fundamentals of Numerical Weather Prediction*. Cambridge University Press. Available at: https://books.google.ie/books?id=vY29h\_Bvqk0C&printsec=frontcover&source=gbs\_atb#v=onepage&q&f=false (Accessed: 1 June 2018).

Cortes, C. and Vapnik, V. (1995) ‘Support-vector networks’, *Machine Learning*, 20(3), pp. 273–297. doi: 10.1007/BF00994018.

De, S. S. and Debnath, A. (2009) ‘Artificial Neural Network Based Prediction of Maximum and Minimum Temperature in the Summer Monsoon Months over India’, *Applied Physics Research*, 1(2). doi: 10.5539/apr.v1n2p37.

Dreyfus, S. E. (1990) ‘Artificial neural networks, back propagation, and the Kelley-Bryson gradient procedure’, *Journal of Guidance, Control, and Dynamics*, 13(5), pp. 926–928. doi: 10.2514/3.25422.

Elman, J. L. (1990) ‘Finding structure in time’, *Cognitive Science*, 14, pp. 179–211. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.298.8354&rep=rep1&type=pdf (Accessed: 11 June 2018).

Flynn, C. (no date) *Forecasts in retrospect: A history of Numerical Weather Prediction | MetService Blog*. Available at: https://blog.metservice.com/HistoryNWP (Accessed: 17 May 2018).

*Forecasting Centre - Met Éireann - The Irish Meteorological Service* (no date). Available at: https://www.met.ie/science/forecasting-centre (Accessed: 28 May 2018).

Géron, A. (2017) *Hands-On Machine Learning with Scikit-Learn & TensorFlow.* First Edition. O’Reilly.

Gholap, J. (2012) ‘Performance Tuning Of J48 Algorithm For Prediction Of Soil Fertility’, *arXiv:1208.3943 [cs, stat]*. Available at: http://arxiv.org/abs/1208.3943 (Accessed: 4 June 2018).

Goodfellow, I., Bengio, Y. and Courville, A. (no date) *Deep Learning*, *Deep Learning*. Available at: http://www.deeplearningbook.org/ (Accessed: 7 June 2018).

Graham, S. (2014) *Relative Humidity vs Dewpoint | WeatherWorks*. Available at: https://www.weatherworksinc.com/humidity-vs-dewpoint, https://www.weatherworksinc.com/node/718 (Accessed: 16 June 2018).

Gumaste, S. S. and Kadam, A. J. (2016) ‘Future weather prediction using genetic algorithm and FFT for smart farming’, in. IEEE, pp. 1–6. doi: 10.1109/ICCUBEA.2016.7860028.

Hochreiter, S. and Schmidhuber, J. (1997) ‘Long Short-Term Memory’, *Neural Computation*, 9(8), pp. 1735–1780. doi: 10.1162/neco.1997.9.8.1735.

Hopfield, J. J. (1982) ‘Neural networks and physical systems with emergent collective computational abilities’, *Proceedings of the National Academy of Science USA*, 79. Available at: http://www.pnas.org/content/pnas/79/8/2554.full.pdf (Accessed: 11 June 2018).

*How Met Éireann produces a forecast - Met Éireann - The Irish Meteorological Service* (no date). Available at: https://www.met.ie/education/how-met-eireann-produces-a-forecast (Accessed: 21 May 2018).

Huang, S.-H. and Ran, B. (2003) ‘Application of Neural Network on Traffic Speed Prediction Under Adverse Weather Conditions’, p. 22. Available at: https://www.researchgate.net/publication/265318230\_An\_Application\_of\_Neural\_Network\_on\_Traffic\_Speed\_Prediction\_Under\_Adverse\_Weather\_Condition?enrichId=rgreq-01d3ff2af55c5896c3643d9c3e2180da-XXX&enrichSource=Y292ZXJQYWdlOzI2NTMxODIzMDtBUzoxNzc1OTU3NDg0NjI1OTNAMTQxOTM1MzUzNDU2MQ%3D%3D&el=1\_x\_2&\_esc=publicationCoverPdf (Accessed: 1 June 2018).

*Introduction to Boosted Trees — xgboost 0.72 documentation* (no date). Available at: https://xgboost.readthedocs.io/en/latest/model.html (Accessed: 16 June 2018).

James, G. *et al.* (eds) (2013) *An introduction to statistical learning: with applications in R*. New York: Springer (Springer texts in statistics, 103).

Jan, Z. *et al.* (2008) ‘Seasonal to Inter-annual Climate Prediction Using Data Mining KNN Technique’, in Hussain, D. M. A. et al. (eds) *Wireless Networks, Information Processing and Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 40–51. doi: 10.1007/978-3-540-89853-5\_7.

Jeppesen, J. (2017) *Fact sheet: Ensemble weather forecasting*, *ECMWF*. Available at: https://www.ecmwf.int/en/about/media-centre/fact-sheet-ensemble-weather-forecasting (Accessed: 28 May 2018).

Jones, T. (2017) *Recurrent neural networks deep dive*. Available at: http://www.ibm.com/developerworks/library/cc-cognitive-recurrent-neural-networks/index.html (Accessed: 11 June 2018).

Kalyankar, M. A. and Alaspurkar, S. J. (2013) ‘Data Mining Technique to Analyse the Metrological Data’, *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(2), p. 5. Available at: https://s3.amazonaws.com/academia.edu.documents/32823336/V3I2-0177.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1527857625&Signature=YosWsDxgv3%2FClOqkSUm1%2FOyyMIo%3D&response-content-disposition=inline%3B%20filename%3D2013\_IJARCSSE\_All\_Rights\_Reserved\_Data.pdf (Accessed: 1 June 2018).

Kohail, S. N. and El-Halees, A. M. (2011) ‘Implementation of Data Mining Techniques for Meteorological Data Analysis’, 1(3), p. 5. Available at: https://www.researchgate.net/profile/Sarah\_Kohail/publication/235665139\_Implementation\_of\_Data\_Mining\_Techniques\_for\_Meteorological\_Data\_Analysis\_A\_case\_study\_for\_Gaza\_Strip/links/09e4151269e3be4d48000000.pdf (Accessed: 1 June 2018).

Kurniawan, A. P., Jati, A. N. and Azmi, F. (2017) ‘Weather prediction based on fuzzy logic algorithm for supporting general farming automation system’, in. IEEE, pp. 152–157. doi: 10.1109/ICA.2017.8068431.

MACQUEEN, J. (no date) ‘SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS’, *MULTIVARIATE OBSERVATIONS*, p. 17.

Maqsood, I., Khan, M. and Abraham, A. (2004) ‘An ensemble of neural networks for weather forecasting’, *Neural Computing and Applications*, 13(2). doi: 10.1007/s00521-004-0413-4.

Mazur (2015) ‘A Step by Step Backpropagation Example’, *Matt Mazur*, 17 March. Available at: https://mattmazur.com/2015/03/17/a-step-by-step-backpropagation-example/ (Accessed: 11 June 2018).

McCormick, C. (2013) *Radial Basis Function Network (RBFN) Tutorial*. Available at: http://mccormickml.com/2013/08/15/radial-basis-function-network-rbfn-tutorial/ (Accessed: 11 June 2018).

McCulloch, W. S. and Pitts, W. (1943) ‘A logical calculus of the ideas immanent in nervous activity’, *The bulletin of mathematical biophysics*, 5(4), pp. 115–133. doi: 10.1007/BF02478259.

MICHALAKES, J. and VACHHARAJANI, M. (2008) ‘GPU ACCELERATION OF NUMERICAL WEATHER PREDICTION’, p. 18. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.142.5554&rep=rep1&type=pdf (Accessed: 1 June 2018).

Nagalakshmi, R., Usha, M. and Ramanathan, R. A. N. (2013) ‘International Journal of Advance Research in Computer Science and Management Studies’, p. 9. Available at: http://ijarcsms.com/docs/paper/volume1/SplDecember2013/V1I7-0002.pdf (Accessed: 1 June 2018).

Nawara, S. (2017) ‘Avoiding Data Leakage in Machine Learning’, *Conlan Holdings LLC*, 6 October. Available at: https://conlan.io/blog/data-science/avoiding-data-leakage-machine-learning/ (Accessed: 18 June 2018).

*Numerical Weather Prediction - Met Éireann - The Irish Meteorological Service* (no date). Available at: https://www.met.ie/science/numerical-weather-prediction (Accessed: 28 May 2018).

Nurunnahar, S. *et al.* (2017) ‘A short term wind speed forcasting using SVR and BP-ANN: A comparative analysis’, in *2017 20th International Conference of Computer and Information Technology (ICCIT)*. *2017 20th International Conference of Computer and Information Technology (ICCIT)*, pp. 1–6. doi: 10.1109/ICCITECHN.2017.8281802.

Olaiya, F. and Adeyemo, A. B. (2012) ‘Application of Data Mining Techniques in Weather Prediction and Climate Change Studies’, *International Journal of Information Engineering and Electronic Business*, 4(1), pp. 51–59. doi: 10.5815/ijieeb.2012.01.07.

PAL, S. *et al.* (no date) ‘Short Term Prediction of Atmospheric Temperature using Neural Networks’, p. 1. Available at: http://metnet.imd.gov.in/mausamdocs/15347.pdf (Accessed: 1 June 2018).

Pandey, A. K., Agrawal, C. P. and Agrawal, M. (2017) ‘A hadoop based weather prediction model for classification of weather data’, in. IEEE, pp. 1–5. doi: 10.1109/ICECCT.2017.8117862.

Petre, E. G. (no date) ‘A Decision Tree for Weather Prediction’, p. 6. Available at: http://demo.pohonkeputusan.com/files/A%20Decision%20Tree%20for%20Weather%20Prediction.pdf?i=1 (Accessed: 1 June 2018).

Qu Xiaoyun *et al.* (2016) ‘Short-term prediction of wind power based on deep Long Short-Term Memory’, in. IEEE, pp. 1148–1152. doi: 10.1109/APPEEC.2016.7779672.

Radhika, Y. and Shashi, M. (2009) ‘Atmospheric Temperature Prediction using Support Vector Machines’, *International Journal of Computer Theory and Engineering*, pp. 55–58. doi: 10.7763/IJCTE.2009.V1.9.

Saha, G. and Chauhan, N. (2017) ‘Numerical weather prediction using nonlinear auto regressive network for the Manaus region, Brazil’, in. IEEE, pp. 1–4. doi: 10.1109/IPACT.2017.8245061.

Salzberg, S. and Segre, A. (1994) ‘Review of C4.5: Programs for Machine Learning by J. Ross Quinlan.’, *Machine Learning - ML*, 16, pp. 235–240. doi: 10.1007/BF00993309.

Sanusi, U. and Corne, D. (2015) ‘Feature selection for accurate short-term forecasting of local wind-speed’, in *2015 IEEE 8th International Workshop on Computational Intelligence and Applications (IWCIA)*. *2015 IEEE 8th International Workshop on Computational Intelligence and Applications (IWCIA)*, pp. 121–126. doi: 10.1109/IWCIA.2015.7449474.

Sharma, A., Chaturvedi, S. and Gour, B. (2014) ‘A Semi- Supervised Technique for Weather Condition Prediction using DBSCAN and KNN’, *International Journal of Computer Applications*, 95(10), pp. 21–26. doi: 10.5120/16631-6500.

Shuman, F. G. (1978) ‘Numerical weather Prediction’, 59(1), p. 13. Available at: https://journals.ametsoc.org/doi/pdf/10.1175/1520-0477%281978%29059%3C0005%3ANWP%3E2.0.CO%3B2 (Accessed: 1 June 2018).

Sivakumar, M. (2006) ‘Climate prediction and agriculture: current status and future challenges’, *Climate Research*, 33, pp. 3–17. doi: 10.3354/cr033003.

*sklearn.preprocessing.StandardScaler — scikit-learn 0.19.1 documentation* (no date). Available at: http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html (Accessed: 7 June 2018).

Talib, M. R. *et al.* (2017) ‘Application of Data Mining Techniques in Weather Data Analysis’, 17(6), p. 7. Available at: http://paper.ijcsns.org/07\_book/201706/20170604.pdf (Accessed: 1 June 2018).

*Temperatures - Dry Bulb/Web Bulb/Dew Point* (no date). Available at: https://www.weather.gov/source/zhu/ZHU\_Training\_Page/definitions/dry\_wet\_bulb\_definition/dry\_wet\_bulb.html (Accessed: 15 June 2018).

Wang, Z. and Sheng, H. (2010) ‘Rainfall Prediction Using Generalized Regression Neural Network: Case Study Zhengzhou’, in. IEEE, pp. 1265–1268. doi: 10.1109/ICCIS.2010.312.

*Water Vapour - Met Éireann - The Irish Meteorological Service* (no date). Available at: https://www.met.ie/climate/what-we-measure/water-vapour (Accessed: 16 June 2018).

*Weather Questions & Answers* (2012) *Weather Questions*. Available at: http://www.weatherquestions.com (Accessed: 15 June 2018).