Can Data Analysis Techniques be used for accurate short term weather prediction using past data

Table of Contents

[Abstract 3](#_Toc514079421)

[Introduction 4](#_Toc514079422)

[Literature Review 5](#_Toc514079423)

# Abstract

# Introduction

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

Agriculture is just one industries that can benefit from these advances. (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.

In addition, the ability to accurately predict certain weather phenomenon would greatly impact particular industries. 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 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.

# Literature Review

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, the authors instead determine association rules for the weather i.e. under what circumstances particular weather events will occur.

(Sharma et al., 2014) use a combination of the Density Based Spatial Clustering of Applications with Noise (DBSCAN) and K-Nearest Neighbour (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.

(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.

(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 author evaluates the model 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, the authors decided to remove all outliers.