Weather Prediction Using Machine Learning Algorithms

by

Mohammed Abrar Ahasan Chowdhury
18201031
Soyelim Al Rozaik
19101602
Mahedi Hasan Shanto
18301185
Shah Md. Shakhawath Hossain
18101133

Department of Computer Science and Engineering
Brac University
May 2023

Abstract

Weather prediction refers to the use of scientific and technological methods to forecast weather conditions in a specific area. Despite being a highly complex and challenging task, this study aims to predict weather patterns by employing predictive analysis. To accomplish this objective, an extensive analysis of various data mining procedures is required before deployment. Specifically, this study proposes a classifier approach that utilizes artificial neural network and logistic regression techniques to forecast the weather. The proposed system utilizes web datasets to carry out two essential tasks - classification (training) and prediction (testing). The outcomes of this study demonstrate the capability of these data mining approaches to accurately forecast weather conditions.

Table of Contents

List of Figures

Chapter 1

Introduction

Over the last century, weather forecasting has remained a challenging and technologically complex subject. With the unpredictable shifts in environmental conditions, climate change has been a significant worldwide discussion topic. Despite the increasing advancements in technology, accurately predicting the weather remains a difficult task due to several limitations. This makes weather forecasting a crucial area of meteorology. Accurate weather alerts are necessary for safeguarding lives and property. Farmers and traders in product markets rely heavily on weather forecasts to make informed decisions about temperature, humidity, wind, and rainfall. Given the critical relationship between agriculture and rainfall, accurate predictions of rainfall are crucial for the growth and production of food crops, both at the germination and fruit development stages.

As the earth's climate continues to change, it is expected that some regions will experience rising temperatures and heavier rainfall, while others will experience less rainfall. Coastal flooding and other related phenomena may also reduce the amount of land available for agriculture. These climatic changes make it challenging for farmers to adapt to evolving weather conditions, so accurate weather forecasting is increasingly important.

Learning weather representation from vast amounts of weather data is one of the most significant challenges in weather forecasting. This requires using various data mining techniques to analyze data from different dimensions or viewpoints, classify it, and condense the correlations discovered using data mining algorithms. The key terms in data mining are classification, learning, and prediction.

This study focuses on developing a classifier for weather forecasting, which integrates the Logistic Regression method with an artificial neural network. The proposed system performs two essential tasks, namely prediction and classification (training and testing). The paper provides a comprehensive overview of the technical details, methodology, design architecture, and experimental findings. The related studies are discussed in section three, while sections five and six offer a detailed analysis of the experimental results. Finally, section seven summarizes the study's key findings and conclusions.

Chapter 2

Problem Statement

The primary motivation for embarking on this project is to develop a comprehensive and reliable economic model that can be utilized by agriculturalists to forecast rainfall patterns over time. The significance of rainfall to ranchers and farmers cannot be overstated, as it can either make or break their crops and undo the laborious work put into the fields. In contemporary times, with global warming posing an increasingly serious problem, understanding rainfall patterns has become more crucial than ever before.

Sudden rainfalls, which are both common and frequent, have played a pivotal role in shaping the character of the earth's landmasses over millions of years. However, with the negative impact of climate change and the increasing frequency of extreme weather events, it has become imperative to study rainfall patterns to mitigate their adverse effects. The ability to accurately reproduce recent rainfall patterns and analyze the impacts and aftereffects of erroneous rainfall that leads to flooding will be a significant step towards achieving this goal.

Through this initiative, we aim to investigate how rainfall has impacted agriculture over time. By developing a robust economic model, we can forecast rainfall patterns, providing farmers with valuable insights into the best time to plant their crops, manage irrigation, and harvest their yields. This study will enable us to understand how agricultural productivity has been affected by changes in rainfall patterns and how farmers can adapt their practices to cope with the evolving climate conditions.

To achieve our research objectives, we will employ sophisticated data analysis techniques, including machine learning algorithms and data mining methodologies. By leveraging these advanced analytical tools, we can examine historical rainfall data and identify trends and patterns that can inform our economic model. Moreover, we will analyze the impact of different variables on rainfall patterns, including geographical location, atmospheric pressure, and ocean currents, to develop a more comprehensive understanding of the complex interplay of factors that influence rainfall patterns.

In conclusion, this research project will provide valuable insights into the impacts of rainfall on agriculture and enable farmers to adapt their practices to cope with the evolving climate conditions.

Chapter 3

Literature Review

In the past ten years, there have been a number of significant efforts to address problems with weather forecasting using statistical models, including machine learning systems, with encouraging results. The Weather Prediction System has employed a variety of techniques. The beginning of the twenty-first century has proven to be important in the history of machine learning because of the arrival of vast data, efficient supercomputers with Graphics Processing Units (GPU), and research interest in emerging new methodologies. Despite the fact that several techniques have been extensively researched since the 1960s, recent years are regarded as the pinnacle of artificial intelligence and machine learning due to the extraordinary growth in data volume and computing power.

P.Goswami [1] and Srividya incorporated CN, RNN, and TDNN features, and their analysis concluded that, In terms of accuracy, composite models perform better than single models. Using multi-layer feed-forward neural networks, C. Venkate- san et al. predicted rainfall during the Indian summer monsoon (MLFNN). The Error Back Propagation (EBP) technique is taught and used to forecast rainfall. We looked at three distinct network models with two, three, and ten input parameters. They also contrasted the outcomes with statistical projections. A. Sahai et al. used monthly and seasonal time series to estimate the summer monsoon rainfall in India using an error back-propagation approach. They used information from the monthly and seasonal mean rainfall values of the five years prior to anticipating the amount of rainfall. Using a CN neural network, predicted annual rainfall in the Kerala area. According to their findings, CN performs better than Fourier analysis.

S. Nanda et al. [7] used A complex statistical model called ARIMA and three ANN models termed MLP, LPE (Legendre Polynomial Equation), and FLANN were used to forecast rainfall (Functional-Link Artificial Neural Network). FLANN has better forecast accuracy than the ARIMA model. A. Naik and S. Pathan employed an artificial neural network (ANN) model for the purpose of predicting rainfall. Instead of using the conventional back propagation method, they created one that is more durable. To predict the monthly rainfall in Assam, Pinky Saikia Dutta and Hitesh Tahbilder employed the traditional statistical method known as Multiple Linear Regression. Min-Max temperature, Mean Sea Level Pressure, Wind Speed, and Rainfall are some of the model's parameters. The multiple linear regression-based prediction model has a respectable level of accuracy.

M.Navon [2] claims to present a condensed review of different aspects of 4-D VAR and its evolution in the past 20 years in the paper. It also aims to present its evolution by highlighting its application and implementation. The 3-D VAR gives the optimal estimation of the true state of the atmosphere during the development of variational data assimilation at operational centers. Due to time and space limitations, no attempt has been made to cover the ensemble Kalman filter data assimilation. In a similar vein, this review is not exhaustive because it does not address all problems related to 4-D VAR applications. It is abundantly clear that in the past 15 years, significant advancements in NWP have been largely attributed to the development of observational sources, and that 4-D VAR may utilize these sources due to significant research efforts at both operational and research centers. Data assimilation ideas will undoubtedly be used more frequently as more geoscience-related scientific fields have access to larger data sets through sensor networks and satellite remote sensing and as Earth system models get more accurate and sophisticated. It is intended that this study would draw attention to a number of aspects of 4-D VAR data assimilation and pique the curiosity of scientists and practitioners in both the atmospheric and variational optimization fields.

F.Olaiya [4] (2012) proposed using the C5 decision tree algorithm to classify weather parameters such as rainfall, temperature, wind speed, and evaporation in terms of month and year. This approach revealed how these parameters affected weather patterns over the study period. By collecting enough data and time, it is possible to detect significant changes in climatic patterns. Artificial neural networks (ANNs) can then identify the relationship between meteorological parameters and use them to forecast future weather conditions. The study constructed predictive ANN models for forecasting Wind speed, Evaporation, Radiation, Minimum Temperature, Maximum Temperature, and Rainfall based on Month and Year, using both TLFN neural networks and Recurrent network architectures. The recurrent TLFD network equipped with the TDNN memory component provided better training and testing results than the TLFD network with the Gamma memory component. The study's findings were evaluated using training and testing data and found to be more acceptable given the small size of data available for training and testing. To achieve better results, it is necessary to collect data spanning many decades. Future research will employ neuro-fuzzy models to predict weather patterns. This work is crucial to understanding climatic changes as data mining approaches can be used to examine temperature, rainfall, and wind speed fluctuations.

In [3] this paper, Abhishek et al. (2012) proposed an artificial neural network model for weather forecasting. The study evaluated the performance of the model in predicting temperature and humidity in different seasons. The authors explain the limitations of traditional numerical weather prediction models and the potential of artificial neural network (ANN) models for weather forecasting. They then describe the architecture of their proposed model, which consists of three layers: input, hidden, and output. The input layer contains weather variables such as temperature, humidity, pressure, and wind direction. The hidden layer performs calculations on the input data, and the output layer produces a forecast of the weather variable of interest, such as temperature. The authors used data from the India Meteorological

Department (IMD) for training and testing their model. They tested their model using statistical metrics such as mean absolute error, mean bias error, and root mean square error. The results showed that the ANN model had better accuracy than traditional numerical weather prediction models for predicting temperature and rainfall.

Moreover, various authors tried to predict wind power by using a nonlinear prediction method, Gaussian Process rather than using traditional historic data from SCADA for better accuracy. By using this method, they improved the performance of NWP and made it possible to accurately predict wind power and speed so it becomes easy for maintenance workers to work on wind turbines. This paper contributes to maintaining the popular eco-friendly energy source of our time, wind power. So we can expect longer up time for those turbines. Data has been gathered from two wind farms in China. One is in the Gansu province of windy western China (denoted as Farm-G) and another one is from Jiangsu province, a coastal area (Farm-J). Both of these wind farms are 2400km apart. The methodology is where it gets interesting. Instead of using barebones numeric weather prediction, they added the Gaussian process as a correction tool to the model. In fact, they improved the Gaussian process by censoring it with factors only related to wind speed to make it more efficient. The following methods are used to measure accuracy; the root mean square error (RMSE), the normalized root mean square error (NRMSE), the mean absolute error (MAE), and the normalized mean absolute percentage error (NMAPE). The empirical evidence from the simulation demonstrates that the proposed GP-CSpeed model has substantially greater accuracy, ranging from 9% to 14%, compared to an MLP model when analyzing the voluminous datasets of Farm-G and Farm-J. These results conclusively validate the efficiency and efficacy of the GP-CSpeed model. Additionally, the GP-CSpeed model's superiority is particularly remarkable when confronted with limited training data, as it attains a remarkable 16.67% enhancement for the recently established Farm-R dataset. There was no significant limitation to this model but we can gather data from different parts of the world to build an international model which can help without the hassle of collecting data in each area.

Here [9] authors try to predict ramps (sudden strong gusts of wind), that can make wind farms experience extreme increase or decrease in production within a space of a few hours using NWP ensemble wind power forecasts. This study holds significant relevance for contemporary wind farms, as the occurrence of ramps can lead to intricate fluctuations in power generation, resulting in an erratic supply of electricity. The word 'ramp' can have multiple interpretations, one of which pertains to fluctuations in wind power occurring at various temporal scales. It sometimes refers to intra-hourly variations, e.g. 10 minutes to an hour. A power system operator can use this to earn valuable intelligence to make management decisions. All the data for this research was gathered from an 8-megawatts wind farm in France. So they started off by mathematically defining a ramp using calculus power vs. time to find those quick changes in power signal. Clusters of ramps were used to detect ramps but it did not work out the way they expected. Then a probabilistic forecast was created to deal with that issue. Lastly, they used Nadarya–Watson estimator and logistic regression as well besides NWP ensembles. For testing purposes, decomposed

Brier score was used. This method provides much better accuracy in sharp turns in graphs which is necessary to predict ramps. This is a breakthrough finding rather than improving existing systems, so a lot can be improved both mathematically and technically. A wind power forecast for a limited time span may not provide adequate details to make crucial security management decisions regarding the possibility of ramp events. Ramps can have a correlation to many other factors. If we gather data on-ramps rather than manually defining it, we can achieve much higher accuracy. In further work, an evaluation of the methodology on more sites should help to validate their conclusions under different weather regimes.

Culclasure et al. [5] used neural networks to provide local weather forecasts. The thesis explores various machine learning techniques and evaluates their performance in predicting weather conditions. [6] Litta et al. proposed an artificial neural network model to predict meteorological parameters during pre-monsoon thunderstorms. The study focused on forecasting thunderstorms, which are critical weather phenomena in India. Rasp et al. [11] introduced WeatherBench, a benchmark data set for data-driven weather forecasting. The dataset includes high-resolution global atmospheric simulations and provides a standardized way to compare different forecasting models. Kumar et al. [10] proposed a method for verified uncertainty calibration in weather forecasting. The study aimed to provide a way to better estimate and communicate uncertainties in weather predictions, which can help decision-makers take appropriate actions. Föll et al. [8] proposed a deep recurrent Gaussian process model with a variational sparse spectrum approximation. The model is capable of learning long-term dependencies and can capture uncertainties in its predictions. Trebing et al. [12] introduced Smaat-UNet, a precipitation nowcasting model based on a small attention-UNet architecture. The model outperforms several existing methods in predicting short-term precipitation and can be used for real-time applications.

Weather forecasting is an essential aspect of human activity as it helps us plan our daily routines and make informed decisions. The early humans predicted the weather based on their observations of nature, such as cloud patterns, wind direction, and animal behavior. However, with the advancements in science and technology, the process of weather forecasting has become more sophisticated. Today, weather forecasting is done using computer-based models that use complex algorithms to predict the atmospheric conditions of a particular area and time. These predictive models collect comprehensive data on prevailing atmospheric conditions, encompassing crucial parameters such as temperature, barometric pressure, moisture content, wind velocity, and directional shifts. Based on this information, they forecast the future trajectory of weather patterns, enabling proactive measures to mitigate any potential impacts. While the use of computers in weather forecasting has improved the accuracy of predictions, it still requires human input to select the best predictive model for establishing the forecast. This involves pattern recognition skills, communication through telephone or other means, accessing model performance information, and considering model bias information. The benefits of using computers in weather forecasting are evident, as it allows for the analysis of large amounts of data and the use of sophisticated algorithms to make predictions. Moreover, computer-based models can identify patterns that humans may not detect, making it easier to predict weather patterns accurately.

Accurate daily weather prediction is crucial for making efficient decisions. However, traditional linear weather prediction models fail to account for non-linearity in input data, thereby necessitating the development of more advanced nonlinear models. Broadly, weather prediction models can be classified as data classification, data clustering, and data prediction. Data classification involves predicting the weather condition based on input parameters, while data clustering groups input data into clusters, each with a cluster centroid. Data prediction, also called regression analysis, predicts the output variable value using linear or nonlinear prediction models. This study proposes a fully connected neural network (FCNN) model for weather data classification. Unlike traditional weather classification models, the FCNN model considers the nonlinear relationship between input features and output conditions [10]. The FCNN model's learning and generalization abilities enable it to capture the nonlinear characteristics of input features in the dataset, outperforming similar models in terms of overall accuracy, user's accuracy, producer's accuracy, and kappa coefficient. The model achieved an OA of 87.83% as tested with the IMD dataset. Additionally, the model's flexibility makes it potentially useful for higher dimensional dataset classification.

Meteorological forecasting is the application of science and technology to predict the atmospheric conditions of a particular location and time. Since the 19th century, people have made ad hoc attempts to predict the weather for thousands of years. Weather forecasts are generated by gathering information about the current state of the atmosphere in a specific region and then using the weather to predict how the atmosphere will change in the future. Individual input is still necessary to select the most accurate predictive model and make a prediction. When it comes to human activities that are largely dependent on changes in barometric stress, current climate, and weather or cloud cover, weather forecasting now relies on computer-based models that examine a variety of celestial objects. Individual input, including pattern recognition skills, telephone communication, model performance information, and model bias information, is still necessary for determining the optimal predictive model. It is common knowledge that computers are utilized in the field of information management.

- Availability: The manual system did not provide us with this information.
- Time: Provides specifics (output) rapidly.
- Accuracy: With the aid of a computer, we will obtain more accurate information than if it were collected and written by hand.
- The Ideal: We were never given insufficient information by the computer. On a computer, we will always find comprehensive and exhaustive data.
- Effective and purposeful behavior: No matter what task we assign to a computer, it will only perform that one.

It indicates that the computer always performs a useful and user-friendly function. This system facilitates more accurate forecasting of the report. There are fewer instances of failure. The plan has reached a stable stage, but additional advancements are still necessary. The system operates at a high level of efficiency, and every user associated with it is aware of its benefits. It was created in response to a need. This system will be designed in the shape of a cross so that it can be utilized globally in the future. Every user can easily operate it due to its intuitive design. This application has been modified to consume less RAM and phone storage space.

Chapter 4

Research Methodology

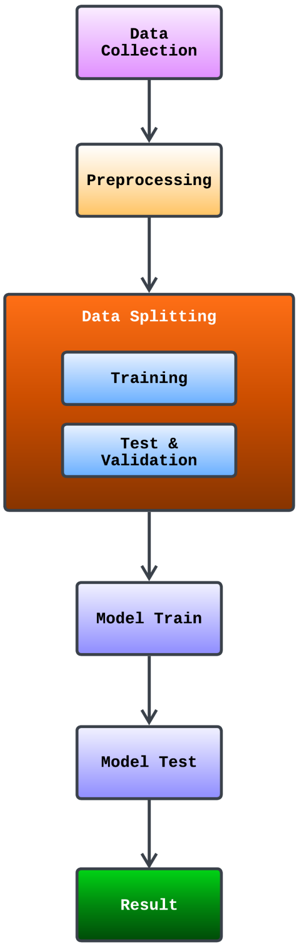


Figure 4.1: Workflow of Weather Prediction

This study utilizes data mining techniques, specifically Artificial Neural Network and Logistic Regression, to extract valuable information from a dataset and make weather predictions based on recent weather information. By using these techniques, the system is able to analyze and identify patterns in the data, allowing it to make accurate predictions about future weather conditions. The use of these advanced techniques ensures that the system is able to process and interpret complex data sets and produce reliable weather forecasts.

Chapter 5

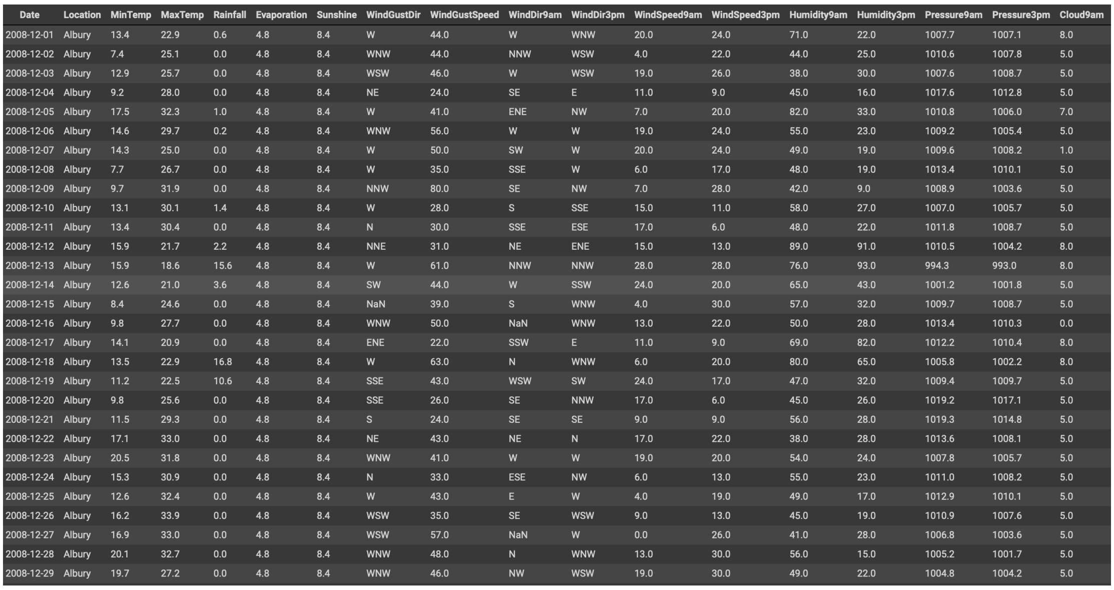
Model and description

Artificial Neural Network (ANN) is a type of machine learning model that is inspired by the structure and function of the human brain. It is a digitalized version of the human brain that can process information in a similar way to how the human brain works. Like humans, ANNs are trained with examples and data, rather than being programmed with specific instructions. By analyzing large datasets and using complex algorithms, ANNs can identify patterns and relationships in data that are difficult or impossible for humans to detect. This makes ANNs ideal for tasks such as image recognition, natural language processing, and weather prediction.

Logistic regression, on the other hand, is a statistical analysis approach that uses previous observations from a data set to predict a binary result, such as yes or no. It is a powerful tool for analyzing data and making predictions based on patterns and correlations between variables. By examining the correlation between one or more independent variables, a logistic regression model forecasts a dependent data variable. Logistic regression is widely used in fields such as finance, healthcare, and marketing, where accurate predictions are essential for decision-making. In the context of weather prediction, logistic regression can be used to analyze patterns in previous weather data and make predictions about the likelihood of specific weather events occurring in the future.

Chapter 6

Dataset



Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am
2008-12-01	Albury	13.4	22.9	0.6	4.8	8.4	W	44.0	W	WNW	20.0	24.0	71.0	22.0	1007.7	1007.1	8.0
2008-12-02	Albury	7.4	25.1	0.0	4.8	8.4	WNW	44.0	NNW	WSW	4.0	22.0	44.0	25.0	1010.6	1007.8	5.0
2008-12-03	Albury	12.9	25.7	0.0	4.8	8.4	WSW	46.0	W	WSW	19.0	26.0	38.0	30.0	1007.6	1008.7	5.0
2008-12-04	Albury	9.2	28.0	0.0	4.8	8.4	NE	24.0	SE	E	11.0	9.0	45.0	16.0	1017.6	1012.8	5.0
2008-12-05	Albury	17.5	32.3	1.0	4.8	8.4	W	41.0	ENE	NW	7.0	20.0	82.0	33.0	1010.8	1006.0	7.0
2008-12-06	Albury	14.6	29.7	0.2	4.8	8.4	WNW	56.0	W	W	19.0	24.0	55.0	23.0	1009.2	1005.4	5.0
2008-12-07	Albury	14.3	25.0	0.0	4.8	8.4	W	50.0	SW	W	20.0	24.0	49.0	19.0	1009.6	1008.2	1.0
2008-12-08	Albury	7.7	26.7	0.0	4.8	8.4	W	35.0	SSE	W	6.0	17.0	48.0	19.0	1013.4	1010.1	5.0
2008-12-09	Albury	9.7	31.9	0.0	4.8	8.4	NNW	80.0	SE	NW	7.0	28.0	42.0	9.0	1008.9	1003.6	5.0
2008-12-10	Albury	13.1	30.1	1.4	4.8	8.4	W	28.0	S	SSE	15.0	11.0	58.0	27.0	1007.0	1005.7	5.0
2008-12-11	Albury	13.4	30.4	0.0	4.8	8.4	N	30.0	SSE	ESE	17.0	6.0	48.0	22.0	1011.8	1008.7	5.0
2008-12-12	Albury	15.9	21.7	2.2	4.8	8.4	NNE	31.0	NE	ENE	15.0	13.0	89.0	91.0	1010.5	1004.2	8.0
2008-12-13	Albury	15.9	18.6	15.6	4.8	8.4	W	61.0	NNW	NNW	28.0	28.0	76.0	93.0	994.3	993.0	8.0
2008-12-14	Albury	12.6	21.0	3.6	4.8	8.4	SW	44.0	W	SSW	24.0	20.0	65.0	43.0	1001.2	1001.8	5.0
2008-12-15	Albury	8.4	24.6	0.0	4.8	8.4	NaN	39.0	S	WNW	4.0	30.0	57.0	32.0	1009.7	1008.7	5.0
2008-12-16	Albury	9.8	27.7	0.0	4.8	8.4	WNW	50.0	NaN	WNW	13.0	22.0	50.0	28.0	1013.4	1010.3	0.0
2008-12-17	Albury	14.1	20.9	0.0	4.8	8.4	ENE	22.0	SSW	E	11.0	9.0	69.0	82.0	1012.2	1010.4	8.0
2008-12-18	Albury	13.5	22.9	16.8	4.8	8.4	W	63.0	N	WNW	6.0	20.0	80.0	65.0	1005.8	1002.2	8.0
2008-12-19	Albury	11.2	22.5	10.6	4.8	8.4	SSE	43.0	WSW	SW	24.0	17.0	47.0	32.0	1009.4	1009.7	5.0
2008-12-20	Albury	9.8	25.6	0.0	4.8	8.4	SSE	26.0	SE	NNW	17.0	6.0	45.0	26.0	1019.2	1017.1	5.0
2008-12-21	Albury	11.5	29.3	0.0	4.8	8.4	S	24.0	SE	SE	9.0	9.0	56.0	28.0	1019.3	1014.8	5.0
2008-12-22	Albury	17.1	33.0	0.0	4.8	8.4	NE	43.0	NE	N	17.0	22.0	38.0	28.0	1013.6	1008.1	5.0
2008-12-23	Albury	20.5	31.8	0.0	4.8	8.4	WNW	41.0	W	W	19.0	20.0	54.0	24.0	1007.8	1005.7	5.0
2008-12-24	Albury	15.3	30.9	0.0	4.8	8.4	N	33.0	ESE	NW	6.0	13.0	55.0	23.0	1011.0	1008.2	5.0
2008-12-25	Albury	12.6	32.4	0.0	4.8	8.4	W	43.0	E	W	4.0	19.0	49.0	17.0	1012.9	1010.1	5.0
2008-12-26	Albury	16.2	33.9	0.0	4.8	8.4	WSW	35.0	SE	WSW	9.0	13.0	45.0	19.0	1010.9	1007.6	5.0
2008-12-27	Albury	16.9	33.0	0.0	4.8	8.4	WSW	57.0	NaN	W	0.0	26.0	41.0	28.0	1006.8	1003.6	5.0
2008-12-28	Albury	20.1	32.7	0.0	4.8	8.4	WNW	48.0	N	WNW	13.0	30.0	56.0	15.0	1005.2	1001.7	5.0
2008-12-29	Albury	19.7	27.2	0.0	4.8	8.4	WNW	46.0	NW	WSW	19.0	30.0	49.0	22.0	1004.8	1004.2	5.0

Figure 6.1: Sample Dataset

The initial stages of the data mining process include gathering and preparing the data. The dataset used in this project was obtained from the Kaggle website. Since the accuracy of the outcomes is dependent on the reliability of the data, data preparation is a crucial step. The project utilizes user data. Even though the dataset had a large number of attributes, only the most important ones were focused on during the data preparation stage, while the rest were disregarded. After modifying the data, a format suitable for data mining was created. The weather forecasting was identified based on four specific characteristics.

Chapter 7

Model Creation

The study employs two data mining techniques, namely Artificial Neural Network (ANN) and Logistic Regression, to categorize and analyze the weather data. The data used for the analysis is collected from a training dataset, which is integrated with test data to predict the weather. Both ANN and Logistic Regression algorithms search for correlations between the predictor values and the goal values and then use this information to make predictions on test data.

ANN is a type of computer system that imitates the structure and functioning of animal brains' neural networks. It is comprised of nodes or connected units that simulate the behavior of biological neurons. In an ANN, the input is processed by a hidden layer, which then communicates the output. The computation in an ANN involves the weighted sum of the inputs, which is then modified by a bias. This computation is represented mathematically as a transfer function.

Logistic Regression, on the other hand, is a popular algorithm used in Machine Learning and falls under the category of Supervised Learning. It is used to predict a categorical dependent variable using a predetermined set of independent factors. In binary logistic regression, there is a single binary dependent variable, indicated by values 0 or 1. The independent variables can either be binary or continuous. Logistic Regression identifies the relationship between the dependent and independent variables and uses this information to make predictions.

Chapter 8

Implementation

We had to create appropriate libraries needed for neural networks, such as TensorFlow, Keras, etc. for the construction of ANN and Logistic Regression. The dataset was loaded in the second step using the Pandas library. About ten years' worth of daily weather measurements from various points across Australia are included in the dataset. Several weather stations were used to gather observations. Correlation between numerical properties, date-time parsing, and encoding of days and months as continuous cyclic features were all used in the implementation.

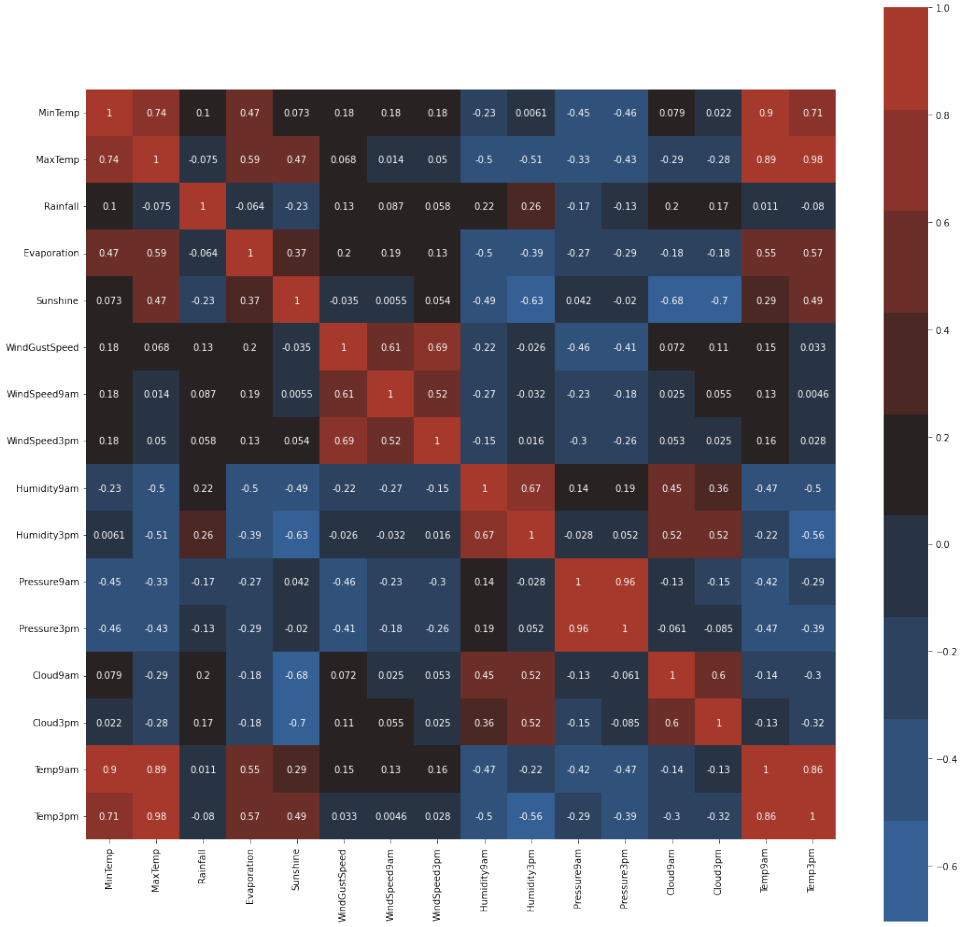


Figure 8.1: Correlation matrix

In order to test the accuracy of the data mining algorithms, several steps were taken to prepare the dataset. First, preprocessing techniques were applied to the data, which involved cleaning and organizing the data to remove any errors or inconsistencies. Next, scaling was performed to standardize the range of values in the dataset, which makes it easier for the algorithms to process the data. Outliers, which are values that are significantly different from the rest of the dataset, were also eliminated to improve the accuracy of the algorithms.

After preprocessing, the feature and goal variables were included in the dataset, and the modeling process began. This involved separating the data into train and test sets, which are used to train and validate the accuracy of the algorithms. The train set is used to train the ANN and Logistic Regression models, while the test set is used to evaluate the accuracy of the models.

Finally, the train and test sets were input into the ANN and Logistic Regression algorithms for output. The algorithms use the data to identify patterns and correlations between the independent and dependent variables, which they use to make predictions about the target variable, in this case, weather forecasting. The output of the algorithms provides insight into the accuracy and reliability of the predictions made by the models.

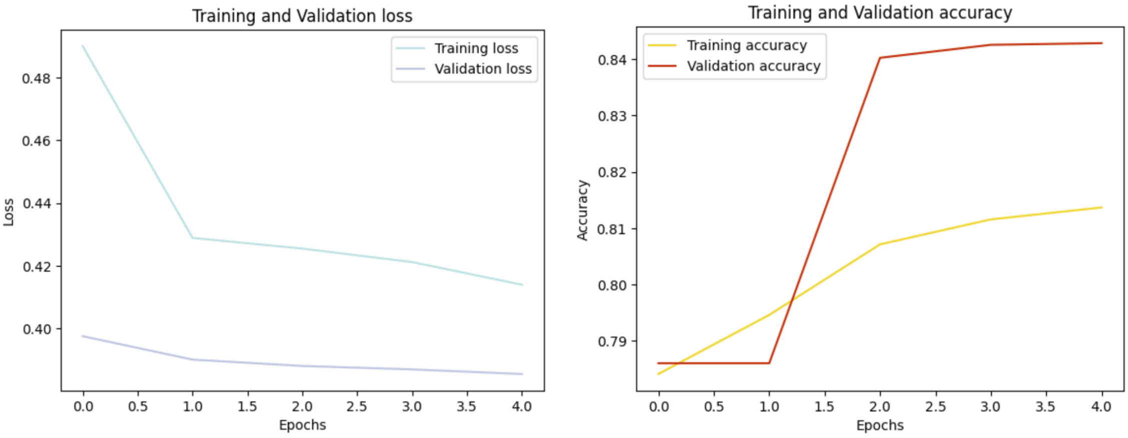
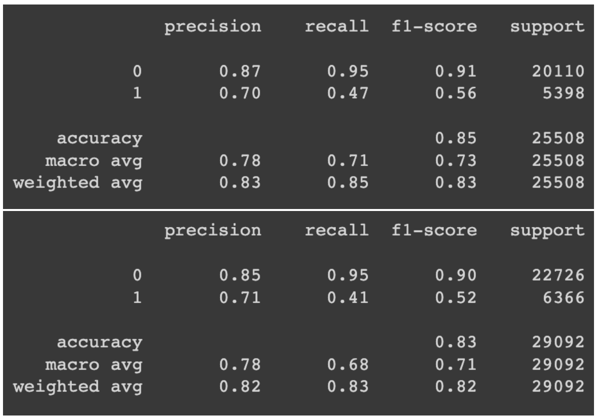


Figure 8.2: Training and Validation Loss and Accuracy



```
              precision    recall  f1-score   support

           0       0.87      0.95      0.91     20110
           1       0.70      0.47      0.56      5398

    accuracy                           0.85     25508
   macro avg       0.78      0.71      0.73     25508
weighted avg       0.83      0.85      0.83     25508
```

```
              precision    recall  f1-score   support

           0       0.85      0.95      0.90     22726
           1       0.71      0.41      0.52      6366

    accuracy                           0.83     29092
   macro avg       0.78      0.68      0.71     29092
weighted avg       0.82      0.83      0.82     29092
```

Figure 8.3: Test Results of ANN and Logistic Regression

Chapter 9

Challenges

Implementing a project for the first time can be challenging, and our project had several obstacles and hurdles to overcome. One of the primary challenges was identifying the appropriate dataset that would be suitable for our project. This was a difficult task, as we had to search through multiple sources before finding a good dataset that fit our needs.

Another challenge we faced was working with a large amount of data, which can slow down the processing time. To address this issue, we divided the larger dataset into smaller pieces and processed it in stages to make the project go more quickly and smoothly.

We also encountered hardware dependency issues, as the ANN technique we used required high-configuration hardware to compute. This added an extra layer of complexity to the project, as we had to ensure that our hardware was up to the task.

Finally, we had to overcome overfitting issues with the logistic regression model. This required us to carefully fit the model to ensure that it was not overfitting the data.

Despite these challenges, we were able to overcome them and successfully complete the project. Through our hard work and perseverance, we were able to implement an effective weather prediction system using data mining techniques.

Chapter 10

Future Works

Feature scaling is indeed a critical step in ML models, as it helps to ensure that all input variables are on a similar scale. This is important because some models may give higher weightage to variables that have larger values, even if they are not necessarily more important. By scaling the features, we can avoid this issue and make sure that the model is not biased toward certain variables. Outliers can also be problematic for ML models as they can skew the results and lead to inaccurate predictions.

It is true that weather forecasting is a complex and nonlinear system, and traditional linear regression models may not be the most accurate for predicting weather patterns. However, as you mentioned, with the help of AI and machine learning, we can improve the accuracy of weather forecasting models by analyzing large amounts of data and identifying patterns that humans may not be able to discern on their own.

The use of IoT sensors to gather weather data is also an interesting approach, as it can help to increase the amount of data available for analysis and improve the accuracy of predictions. As technology continues to advance, we may see even more sophisticated AI systems being developed to analyze weather data and make predictions with even greater accuracy.

Chapter 11

Conclusion

To summarize, we employed Artificial Neural Network (ANN) and Logistic Regression techniques to classify rainfall and compared their accuracy scores to determine the suitable method for developing a weather forecasting model. Our results showed that the ANN model outperformed the logistic regression model, achieving an 84% accuracy score. Moreover, we proposed a methodology for creating weather forecasts using machine learning algorithms that are more efficient than traditional physical models. These intelligent models require minimal resources and can operate on various platforms, including mobile devices.

Bibliography

[1] P. Goswami and Srividya, "A novel neural network design for long range prediction of rainfall pattern," *Current Science*, vol. 70, no. 6, pp. 447–457, 1996.

[2] I. Navon, "Data assimilation for numerical weather prediction: A review," in *Data assimilation for atmospheric, oceanic and hydrologic applications*, Springer, 2009, pp. 21–65.

[3] K. Abhishek, M. Singh, S. Ghosh, and A. Anand, "Weather forecasting model using artificial neural network," *Procedia Technology*, vol. 4, pp. 311–318, 2012.

[4] F. Olaiya and A. B. Adeyemo, "Application of data mining techniques in weather prediction and climate change studies," *International Journal of Information Engineering and Electronic Business*, vol. 4, no. 1, p. 51, 2012.

[5] A. Culclasure, "Using neural networks to provide local weather forecasts," Electronic Theses and Dissertations, 2013.

[6] A. J. Litta, S. Mary Idicula, and U. Mohanty, "Artificial neural network model in prediction of meteorological parameters during premonsoon thunderstorms," *International Journal of atmospheric sciences*, vol. 2013, 2013.

[7] S. K. Nanda *et al.*, "Prediction of rainfall in india using artificial neural network (ann) models," *International Journal of Intelligent Systems and Applications*, vol. 5, no. 12, p. 1, 2013.

[8] R. Föll, B. Haasdonk, M. Hanselmann, and H. Ulmer, "Deep recurrent gaussian process with variational sparse spectrum approximation," *arXiv preprint arXiv:1711.00799*, 2017.

[9] D. N. Fente and D. K. Singh, "Weather forecasting using artificial neural network," in *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, IEEE, 2018, pp. 1757–1761.

[10] A. Kumar, P. Liang, and T. Ma, "Verified uncertainty calibration," *arXiv preprint arXiv:1909.10155*, 2019.

[11] S. Rasp, P. D. Dueben, S. Scher, J. A. Weyn, S. Mouatadid, and N. Thuerey, "Weatherbench: A benchmark data set for data-driven weather forecasting," *Journal of Advances in Modeling Earth Systems*, vol. 12, no. 11, e2020MS002024, 2020.

[12] K. Trebing, T. Stanczyk, and S. Mehrkanoon, "Smaat-unet: Precipitation nowcasting using a small attention-unet architecture," *Pattern Recognition Letters*, vol. 145, pp. 178–186, 2021.