UNIVERSITY OF CENTRAL MISSOURI

A FINAL PROJECT INCREMENT ON MACHINE LEARNING – CS5710

RAINFALL PREDICTION USING MACHINE LEARNING TECHNIQUES

BY

SAI KRISHNA REDDY CHEVUTUKUR – 700743830

GITHUB Link:

https://github.com/csk17/ML-Final-Project-Increment

Abstract

Predicting the amount of rainfall is one of the most challenging endeavors possible since it is very unpredictable and has far-reaching repercussions for human civilization. It may be possible to avoid further loss of life and financial resources by making accurate predictions in a timely manner and using those predictions. This research presents a series of studies that employ popular machine learning methods to construct algorithms that predict whether or not it will rain the next day in major Australian cities, given meteorological data for the day in question. The research was carried out by the University of New South Wales and the Australian National University. The modeling input, the numerical modeling, and the pre-processing processes are going to be the key focuses of this examination. The findings include a comparative analysis of the performance of a number of different machine learning approaches for forecasting the weather using a variety of metrics of evaluation. It is also crucial that the prognosis be correct so that individuals may utilize it as a basis for taking preventative actions. It is imperative that nations like India, whose economies are so highly dependent on agriculture, have access to precise rainfall predictions. Many different types of machine learning models, including logistic Regression, Neural Networks, K-Means, and Naive Bayes, are utilized in the process of producing precipitation forecasts. Recognizing and forecasting rainfall with the use of these systems is one of the applications that may be used for these systems. Other applications include data extraction, training, and testing. A model that predicts future rainfall is presented here, and it makes use of logistic Regression (MLR) as well as neural networks. In this study, both machine learning and neural networks, as well as neural networks, were analyzed to determine which would be the most effective for predicting precipitation; the algorithm that performed the best was then utilized for the forecast. In order to calculate rainfall, the input data must include a wide variety of meteorological characteristics (including humidity, lowest temperature, highest temperature, elevation, cloud, cloud cover, wind speed, etc.). These meteorological characteristics are required in order to calculate rainfall. Measures of accuracy and correlation, in addition to Mean Absolute Error (MAE), are utilized in the process of verifying the suggested model. The performance of the suggested machine learning model is superior to that of earlier approaches described in the relevant body of research.

Keywords

Keywords: Rainfall forecasting; oversampling, under sampling, classifiers, assessing models

Introduction

The capacity to improve our ability to anticipate rainfall is something that continues to be of interest to a wide variety of stakeholders, including governments, corporations, risk management organizations, and scientists. It is common knowledge that precipitation has an effect on a wide variety of economic sectors, including

agriculture, construction, the generation of electricity, forestry, and even tourism. It is essential to have an accurate method of forecasting rainfall due to the strong correlation that exists between precipitation and potentially catastrophic natural events including landslides, flooding, mass movements, and avalanches. Even after so many years have passed, the effects of these happenings are still being felt. Because of this, it is feasible to take preventative measures as well as mitigation measures for these natural occurrences if an effective approach for rainfall prediction is utilized. In order to get rid of this uncertainty and come up with accurate projections, Users employed a variety of approaches and models from the field of machine learning. The goal of this work is to offer the entirety of the machine learning life cycle, beginning with the preparation of the data and finishing with the implementation and assessment of the model. A few examples of preprocessing stages for data include encoding categorizing features, selecting characteristics, scaling features, and imputing missing values. Other examples include scaling features, selecting characteristics, and converting features. Logistic Regression, Decision Tree, K-Nearest Neighbor, Rule-based, and Ensembles were some of the models that were implemented. During the analysis, metrics such as Accuracy, Precision, and Recall, as well as F-Score and Area under the Curve, were utilized. In the trials that users do, we train our classifiers using data on the weather that was gathered from all throughout Australia. The study will continue in the manner that is outlined here. In the section on processing the data, the dataset that is being examined is presented. The findings include graphics and a description of the experiments that were carried out, whilst the data preparation describes the methodologies and strategies that were employed.

Motivation

The last decade has seen unprecedented growth in both academic and commercial databases. Manually extracting sophisticated insight from these databases is arduous, expensive, and time consuming. When data grows beyond manageable proportions or becomes too complex to process, it's hopeless. Because of this, the automated analysis as well as visual analytics of massive multi-dimensional sets of data has been at the forefront of scientific inquiry in recent years. The primary objective is to detect patterns and associations in the data in order to gain access to hidden but potentially useful information. Among the many promising developments in this broad area are artificial neural networks. They are inspired by the progress in biomedical research to develop a category of algorithms with the intention of modeling brain neural structures. The reason for this is that the foundation of the ANN (Artificial Neural Network) framework is the ability to “predict” outcomes by “analyzing” the trend in a large body of historical data. All the other designs are either numerical or statistical, with ANN being the only exception. Very accurate calculations can be made with such models, but accurate predictions cannot be made because of their

inflexibility in dealing with data that varies in ways that cannot be captured by a function or derived from a formula. In order to better interpret these real-world scenarios, it has been found that 'artificial neurons' that can learn through experience, i.e. by training algorithm of errors in next guestimate and so on, are superior. While this could potentially decrease the quality of our results, it would give us a significant advantage in "understanding the problem," replicating it, and drawing conclusions from it. Rainfall is the most important climatic event because it sustains human life. Its regularity and amount on multiple scales are crucial to the survival of human civilization. To predict the occurrence of rainfall, to explore its seasonal variability, and to forecast annual amount of rain over some geographical area, numerous probability models have been attempted.

Main Contributions & Objectives

- To identify the proper dataset.
- To choose a machine learning prediction model that can help to predict the most accurate prediction result based on the chosen dataset.
- To implement the code using a python programming language.
- To check the accuracy of the predicted data with the training dataset.
- To check the data based on the other model and identify the best accuracy.
- To research to get the best prediction data.

Related Work

Predicting whether or not it will rain is one of the most crucial aspects of predicting weather conditions in any nation [1]. This research presents a logistic Regression-based rainfall forecast model for use with the Indian dataset (MLR). A more precise estimate of precipitation is possible because of the inclusion of many meteorological parameters in the input data. Mean Squared Error (MSE), Accuracy, as well as Correlation are used to verify the proposed model. Compared to previous methods in the literature, the suggested machine learning model performs exceptionally well.

The nonlinear and complex nature of weather and rainfall, sophisticated computer simulation and modeling are required for reliable forecasting [2]. Logistic Regression, Neural Networks, Random Forest, Naive Bayes, and other machine learning models are used to forecast precipitation. Applications of these systems include the extraction, training, and testing of data sets, as well as the identification and prediction of precipitation. This study demonstrates how to estimate rainfall using logistic regression as well as neural networks, as well as how to select crops using the Decision Trees technique. Therefore, users have determined that it is possible to recommend seeds based on the expected rainfall.

Heavy rainfall is strongly tied to the economy and human survival, making its prediction a top priority for meteorological departments [3]. It's responsible for yearly droughts and floods that affect millions of people throughout the globe. Countries like India,

where the agricultural sector is vital to the economy, need reliable rainfall forecasts. Since the atmosphere is always changing, statistical methods can't reliably predict when it will rain. Neural Networks are more effective than logistic methods because of the nonlinear nature of rainfall data. Tabular data displays the results of an examination and comparison of several approaches and algorithms used by scientists to forecast precipitation. This report attempts to do just that by explaining the tools and approaches that users re employed in this study to readers who are not familiar with them.

Precipitation forecasting is an essential part of meteorology. Precipitation forecasting and estimating are achieved by the use of a combination of factual processes and machine learning techniques [4]. This paper's goal is to help clients from a variety of fields, including farmers, scientists, and others, make sense of the significance of variations in weather patterns as well as atmospheric conditions like rainfall, temperature, as well as humidity.

Rainfall detection and forecasting are becoming increasingly important in developing nations as a result of rising sea levels and other effects of climate change since it may have far-reaching positive effects on agriculture, public health, water supply, and other sectors [5]. Future annual rainfall is estimated using SVR, SVM, as well as KNN machine learning algorithms, and this article compares the inferred results from each method.

Neural network-based rain forecasting models were proposed in India. In this work, users evaluate the accuracy of two different approaches to predicting rain. The existing method for forecasting rainfall does not work well due to the complexity of the available data [6]. While statistical and numerical approaches have been adopted, they fail to provide reliable results in the presence of a non-linear trend. For increasingly complicated historical rainfall records, the current method breaks out. Researchers will examine either machine learning or neural networks to find the most effective strategy for predicting rainfall, and then implement the algorithm that produces great results. To construct a functional rainwater harvester, it is necessary to predict the amount of precipitation. Gaining more attention is as easy and quick as doing a weather prediction. The findings of this study can be used by national weather forecasting channels to improve the accuracy of their coverage across the nation.

Proposed Framework

1. Logistic regression

Logistic regression is a statistical method for showing the association between a dependent variable as well as one or more independent variables, and it is defined as a logistic approach. Logistic regression is considered “simple” when there is only one independent variable. Logistic regression is used when there is more than one independent variable. Different from logistic regression, in which a single variable is used to make predictions about a set of correlated dependent variables, this term has a completely different meaning. The model of logistic regression has many uses, including error reduction, prediction,

and forecasting. Logistic regression will be used to fit a predictive model to the collected data or new findings [7]. After the model has been fitted, predictions can be made from it. Here is the equation for logistic regression: Afterwards, the model can be used to make predictions.

2. Neural Networks

One of the most common types of deep learning and machine learning programs is the Neural Network. They are in awe of human neurons because, with the help of computations, they can produce decisions that are strikingly similar to those made by humans. When users train Neural Networks, for instance, users provide them with a wide variety of features, including humidity, temperature, pressure, etc., but they use the results of the training dataset to learn to recognize and analyze the rainfall on the basis of these features [8]. One analysis, one hidden neuron, one and output neuron could be all that's needed in the simplest neural network. It uses Activation functions as well as a unit step function on a set of dependent variables whose number is equal to the number of input parameters multiplied by coefficients proportional to the weights.

Step 1.

Import the CSV file containing the rainfall data set. The precipitation data for the UK spans the years 2008 through 2009. Variables included in the data set are minimum and maximum temperatures, air pressure, relative humidity, wind speed, direction, and so on. Rainfall, in millimeters, is the attribute [9].

Step 2.

Pre-processing A procedure performed on data before primary processing or further analysis begins. When there are numerous actions to arrange data for the user, the term could be applied to any initial or preparatory stage of the process. The average score of the data can be used to fill in the blanks.

Step 3.

One method for dealing with missing information is to apply a constant scaling factor to the data's features. Whether due to accidental deletion or intentional tampering, data corruption is always the culprit when missing values occur. Since many machine learning techniques do not support missing values, missing data handling is crucial during the pre-processing stage of the dataset.

Step 4.

Data Minimization via PCA-based Feature Reduction. Reducing the number of characteristics in a computation that is resource-intensive can be achieved through a technique known as feature reduction (or dimensionality reduction).

Step 5.

The training process is how users find out how well our model works. The data set is divided into two parts: the training set and the testing set, hence the name “Train/Test.” 80% for training, as well as 20% for testing. The training set is used to teach the model [10]. The information is split into two groups: a training set (which accounts for 80% of the total) as well as a testing set (which accounts for the remaining 20%).

Step 6.

Mean Absolute Error and r2 score are computed using the Prediction Model.

Step 7.

For each model, a scatter plot is generated between the predicted data and the tested

data; the errors are then compared, and the model with the lowest total error is chosen as the best candidate.

Step 8.

Display the outcomes and offers the required outcome with better precision [20].

Decision Tree:

In machine learning, the decision tree algorithm can be used for either classification or regression. Each branch node in a decision tree represents an option, while each leaf node represents a verdict. The parameter (rainfall) is a binary categorical, and the Decision's strong performance with both types of variables makes it a good fit for this investigation. Some common algorithms used when constructing decision trees are C4.5, ID3, Quest, Classification and Regression Trees (CART), C5.0, and Chi-squared Automatic Interaction Detection (CHAID). In this analysis, the C5.0 was chosen and used across all three of the study's training and testing combinations. C5.0 is a new methodology that improves upon its predecessors, C4.5 as well as ID3 [11].

Random Forest:

Breiman created Random Forest (RF) in 2001 to aid in classification. In an ensemble machine learning algorithm, random forests are built from many individual classification trees. To facilitate regression computation, it combines multiple decision trees that have previously been used for classification and regression. In spite of its apparent favoritism for high-level variables among categorical variables of varying levels, the RF algorithm has come to be seen as an extremely stringent learning algorithm as of late [12]. Gathering random selection from of the provided data is the first step in the RF algorithm's process. In the second step, a clustering algorithm will be built for each of the samples. Next, the predicted outcomes are voted on, and the classification with the most votes is selected. All three ratios of training to testing data were run through the RF algorithm. In this analysis, we used a model with a configuration of 100 weak learners and a maximum tree depth of 16 [19].

Extreme Gradient Boosting:

Based on the gradient boosting algorithm introduced, Extreme Gradient Boosting (XGBoost) is a cutting-edge machine learning method. By formally describing the model, XGBoost is better equipped to deal with overfitting. The fast running time of this method made it a prime candidate for our investigation. All three ratios of training to testing were subjected to XGB [13].

Data Description

Data for the United Kingdom are taken from Kaggle datahub Online and span the years 2008 and 2009. Those numbers are a 12-years' worth of daily weather reports from Kolhapur District. To be more specific, Rain Today is the target variable that will be evaluated to forecast using machine learning algorithms, and a number of characteristics indicate whether or not it rains the same day.

- Minimum and maximum temperatures, barometric pressure, humidity, wind speed and direction, cloud cover, and precipitation totals are all examples of the types of factors that may be found in the data.
- In practice, only 20% of data will be used for the actual testing, while the

remaining 80% is used for actual training [14].

- Rain is the dependent variable, whereas minimum/maximum temperatures, pressure/altitude, humidity/wind speed/direction, and clouds are indeed the independent variables.
- The characteristics are the precipitation totals in millimeters.

Data Preparation

1. Missing data

It is estimated that roughly half of the tests would need to be discarded if those without data in either of their variables were removed. In order to avoid throwing away a large amount of data, analyses were conducted on the variables that had no records, with the records for each city treated separately [18].

2. Data normalization

The mix-max data were normalized so that all of the variables could have values ranging from 0 and 1, as this range was determined by the researchers. This way, we were able to avoid the use of algorithms for machine learning that would have been negatively impacted by variables taking on extremely large values [15].

3. Detection of outliers

To improve accuracy, users first identify and eliminate any outliers in the graph, as their presence leads to a drop in predictive power for the entire model.

Results/ Experimentation & Comparison/Analysis

The implementation process based on the chosen dataset and the implemented result with the code and outcomes snips is included in this section.

This image aims to share the process of the dataset importation method using Python pandas library.

```
import pandas as pd
full_data = pd.read_csv('weatherAUS.csv')
full_data.head()
```

	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	...	Humidity3pm	Pressure9am	Pressure3
0	2008-12-01	Albury	13.4	22.9	0.6	NaN	NaN	W	44.0	W	...	22.0	1007.7	100
1	2008-12-02	Albury	7.4	25.1	0.0	NaN	NaN	WNW	44.0	NNW	...	25.0	1010.6	100
2	2008-12-03	Albury	12.9	25.7	0.0	NaN	NaN	WSW	46.0	W	...	30.0	1007.6	100
3	2008-12-04	Albury	9.2	28.0	0.0	NaN	NaN	NE	24.0	SE	...	16.0	1017.6	101
4	2008-12-05	Albury	17.5	32.3	1.0	NaN	NaN	W	41.0	ENE	...	33.0	1010.8	100

Figure1 : Import dataset

(Source: Created by the learner)

After importing the full dataset need to apply the EDA process to understand the data details with the data correlation.

Using the shape function, the shape of the dataset is displayed [16].

```
full_data.shape
```

```
(142193, 26)
```

Figure2 : Find the dataset shape

(Source: Created by the learner)

The info function in the pandas library helps to find out the data attributes data types.

```
full_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 142193 entries, 0 to 142192
Data columns (total 26 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   Location       142193 non-null  object
 1   MinTemp        141556 non-null  float64
 2   MaxTemp        141871 non-null  float64
 3   Rainfall       140787 non-null  float64
 4   Evaporation    81350 non-null   float64
 5   Sunshine       74377 non-null   float64
 6   WindGustDir    132863 non-null  object
 7   WindGustSpeed  132923 non-null  float64
 8   WindDir9am     132180 non-null  object
 9   WindDir3pm     138415 non-null  object
 10  WindSpeed9am   140845 non-null  float64
 11  WindSpeed3pm   139563 non-null  float64
 12  Humidity9am    140419 non-null  float64
 13  Humidity3pm    138583 non-null  float64
 14  Pressure9am    128179 non-null  float64
 15  Pressure3pm    128212 non-null  float64
 16  Cloud9am       88536 non-null   float64
 17  Cloud3pm       85099 non-null   float64
 18  Temp9am        141289 non-null  float64
 19  Temp3pm        139467 non-null  float64
 20  RainToday      140787 non-null  object
 21  RISK_MM        142193 non-null  float64
 22  RainTomorrow   142193 non-null  object
 23  year           142193 non-null  int64
 24  month          142193 non-null  int64
 25  day            142193 non-null  int64
dtypes: float64(17), int64(3), object(6)
memory usage: 28.2+ MB
```

Figure3: Find the dataset info

(Source: Created by the learner)

This image aims to share the prediction data for the tomorrow's rain prediction or chances.

```
import matplotlib.pyplot as plt
fig = plt.figure(figsize = (20,5))
ax=full_data.RainTomorrow.value_counts(normalize = True).plot(kind='bar', color= ['skyblue','navy'], alpha = 0.9, rot=0)
plt.title('RainTomorrow Indicator No(0) and Yes(1) in the Imbalanced Dataset')
for p in ax.patches:
    ax.annotate(str(round(p.get_height(),2)), (p.get_x() * 1.01 , p.get_height() * 1.01))
plt.show()
```

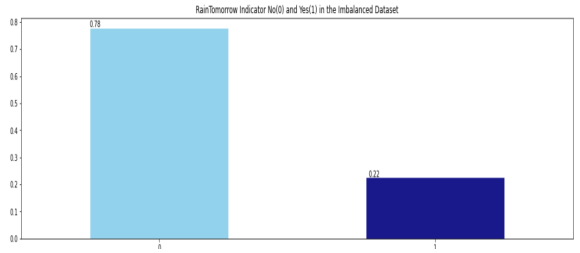

Figure 4: Rain data prediction for the next day

(Source: Created by the learner)

The missing data pattern for the training data is visualized in this section [17].

```
import seaborn as sns
plt.figure(figsize = (20,5))
sns.heatmap(oversampled.isnull(), cbar=False, cmap='PuBu')
plt.show()
```

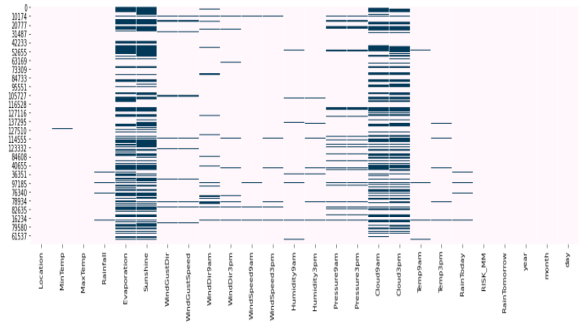

Figure 5: missing data pattern

(Source: Created by the learner)

The correlation of all data attributes using the heatmap using the seaborn data visualization python library is included in this section.

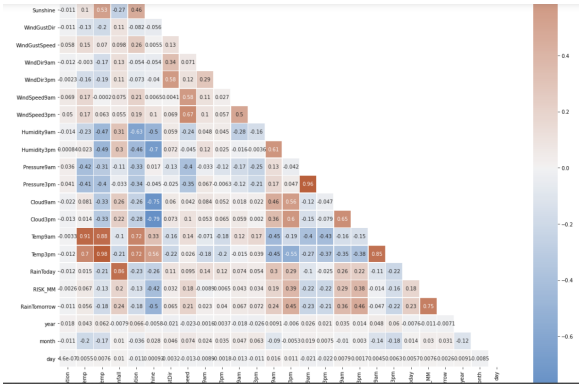

Figure 6: data correlation using heatmap

(Source: Created by the learner)

To understand the data details or correlation between all data attributes, this section aims to find out the as well as plot the details of the all data attributes relation using the pairplot visualization concept.

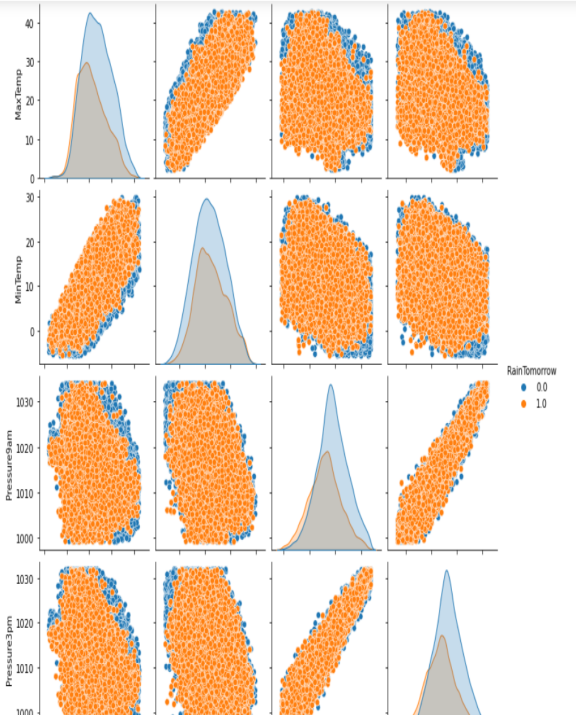

Figure 6: data correlation using pair plot

(Source: Created by the learner)

The training and testing of the dataset spliting concept is included in this section.

```
features = MiceImputed[['Location', 'MinTemp', 'MaxTemp', 'Rainfall', 'Evaporation', 'Sunshine', 'WindGustDir',
                        'WindGustSpeed', 'WindDir9am', 'WindDir3pm', 'WindSpeed9am', 'WindSpeed3pm', 'Humidity9am',
                        'Humidity3pm', 'Pressure9am', 'Pressure3pm', 'Cloud9am', 'Cloud3pm', 'Temp9am', 'Temp3pm',
                        'RainToday']]
target = MiceImputed['RainTomorrow']

# Split into test and train
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(features, target, test_size=0.2, random_state=42,
                                                    shuffle=True, stratify=target)

# Normalize Features
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.fit_transform(X_test)
```

Figure 7: Train and testing data split

(Source: Created by the learner)

The various model is included in the prediction section to get and check the best model to predict the rainfall based on the data as well as get the accuracy. This section aims to share the details of the various types of model performance in the bar chart visual concept.

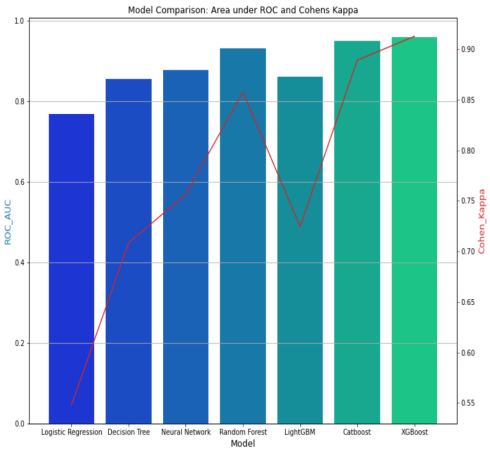

Figure 7: Model performance

(Source: Created by the learner)

References

[1] Schultz, M.G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L.H., Mozaffari, A. and Stadtler, S., 2021. Can deep learning beat numerical weather prediction?. Philosophical Transactions of the Royal Society A, 379(2194), p.20200097.

[2] Van Klompenburg, T., Kassahun, A. and Catal, C., 2020. Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture, 177, p.105709.

[3] Kashinath, K., Mustafa, M., Albert, A., Wu, J.L., Jiang, C., Esmaeilzadeh, S., Azizzadenesheli, K., Wang, R., Chattopadhyay, A., Singh, A. and Manepalli, A., 2021. Physics-informed machine learning: case studies for weather and climate modelling. Philosophical Transactions of the Royal Society A, 379(2194), p.20200093.

[4] Akhter, M.N., Mekhilef, S., Mokhlis, H. and Mohamed Shah, N., 2019. Review on forecasting of photovoltaic power generation based on machine learning and metaheuristic techniques. IET Renewable Power Generation, 13(7), pp.1009-1023.

[5] Lim, B. and Zohren, S., 2021. Time-series forecasting with deep learning: a survey. Philosophical Transactions of the Royal Society A, 379(2194), p.20200209.

[6] Lim, B. and Zohren, S., 2021. Time-series forecasting with deep learning: a survey. Philosophical Transactions of the Royal Society A, 379(2194), p.20200209.

[7] Sharma, A., Jain, A., Gupta, P. and Chowdary, V., 2020. Machine learning applications for precision agriculture: A comprehensive review. IEEE Access, 9, pp.4843-4873.

[8] Yagli, G.M., Yang, D. and Srinivasan, D., 2019. Automatic hourly solar forecasting using machine learning models. *Renewable and Sustainable Energy Reviews*, *105*, pp.487-498.

[9] Mekonnen, Y., Namuduri, S., Burton, L., Sarwat, A. and Bhansali, S., 2019. Machine learning techniques in wireless sensor network based precision agriculture. *Journal of the Electrochemical Society*, *167*(3), p.037522.

[10] Zhang, L., Wen, J., Li, Y., Chen, J., Ye, Y., Fu, Y. and Livingood, W., 2021. A review of machine learning in building load prediction. *Applied Energy*, *285*, p.116452.

[11] Elavarasan, D., Vincent, D.R., Sharma, V., Zomaya, A.Y. and Srinivasan, K., 2018. Forecasting yield by integrating agrarian factors and machine learning models: A survey. *Computers and Electronics in Agriculture*, *155*, pp.257-282.

[12] Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J. and Carvalhais, N., 2019. Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), pp.195-204.

[13] Sharma, R., Kamble, S.S., Gunasekaran, A., Kumar, V. and Kumar, A., 2020. A systematic literature review on machine learning applications for sustainable agriculture supply chain performance. *Computers & Operations Research*, *119*, p.104926.

[14] Ahmed, R., Sreeram, V., Mishra, Y. and Arif, M.D., 2020. A review and evaluation of the state-of-the-art in PV solar power

forecasting: Techniques and optimization. *Renewable and Sustainable Energy Reviews*, *124*, p.109792.
[15] Wang, Z., Hong, T. and Piette, M.A., 2020. Building thermal load prediction through shallow machine learning and deep learning. *Applied Energy*, *263*, p.114683.
[16] de Freitas Viscondi, G. and Alves-Souza, S.N., 2019. A Systematic Literature Review on big data for solar photovoltaic electricity generation forecasting. *Sustainable Energy Technologies and Assessments*, *31*, pp.54-63.
[17] Liu, Z., Wu, D., Liu, Y., Han, Z., Lun, L., Gao, J., Jin, G. and Cao, G., 2019. Accuracy analyses and model comparison of machine learning adopted in building energy consumption prediction. *Energy Exploration & Exploitation*, *37*(4), pp.1426-1451.
[18] Salcedo-Sanz, S., Ghamisi, P., Piles, M., Werner, M., Cuadra, L., Moreno-Martínez, A., Izquierdo-Verdiguier, E., Muñoz-Marí, J., Mosavi, A. and Camps-Valls, G., 2020. Machine learning information fusion in Earth observation: A comprehensive review of methods, applications and data sources. *Information Fusion*, *63*, pp.256-272.
[19] Salcedo-Sanz, S., Cornejo-Bueno, L., Prieto, L., Paredes, D. and García-Herrera, R., 2018. Feature selection in machine learning prediction systems for renewable energy applications. *Renewable and Sustainable Energy Reviews*, *90*, pp.728-741.
[20] Yaseen, Z.M., Sulaiman, S.O., Deo, R.C. and Chau, K.W., 2019. An enhanced extreme learning machine model for river flow forecasting: State-of-the-art, practical applications in water resource engineering area and future research direction. *Journal of Hydrology*, *569*, pp.387-408.