© Data Science Research Lab,
School Of Computing,
College of Art & Sciences,
University Utara Malaysia

Rain Prediction Intelligent System Based On Numerical Dataset

**Azib Fikri, N. H. Harun, Dayang Fatimah,
Muhammad Soleh, Muhammad Syafiq, Ezurin Farisha,**

Azib Fikri Bin Mohd Suhaimi,
Data Science Research Lab, School Of Computing, College of Art & Sciences, University Utara Malaysia

N. H. Harun,
Data Science Research Lab, School Of Computing, College of Art & Sciences, University Utara Malaysia

Dayang Fatimah Binti Zaki,
Data Science Research Lab, School Of Computing, College of Art & Sciences, University Utara Malaysia

Muhammad Soleh Bin Ab Majid,
Data Science Research Lab, School Of Computing, College of Art & Sciences, University Utara Malaysia

Muhammad Syafiq Bin Azhari,
Data Science Research Lab, School Of Computing, College of Art & Sciences, University Utara Malaysia

Nur Ezurin Farisha binti Jusshairi,
Data Science Research Lab, School Of Computing, College of Art & Sciences, University Utara Malaysia

Abstract: Rainfall prediction is one of the main areas of study in weather forecasting since it has a big influence on the environment and ecology in Australia. Rainfall has a big impact on natural phenomena like floods and droughts as well as meteorological indicators like relative humidity. An existing system develops a two-step prediction model using logistic regression. Several data sets variables are used to create the training phase. For feature selection, confusion matrix will be used to check the accuracy. The supervised learning approach is applied in the system and have significant impact on 0.84% is the accuracy score for the logistic regression model. Data on rainfall incidence is obtained using worldwide forecasts that can be transformed into a readable form using PYTHON functions. The model is effective at forecasting. Weather in Australia.

Keywords: Neural Network, Linear Regression, Confusion Matrix, Supervised Learning

1. Introduction

Forecasting rainfall is crucial because it may have a variety of effects, including the devastation of crops and farms and damage to property. Every year, people all across the world are impacted by natural catastrophes including floods and droughts. The reliability of the rainfall statement is essential for nations like Australia since applied mathematics approaches cannot provide consistent precision for a statement about precipitation due to the dynamic character of the environment. Machine learning is currently used more frequently in weather forecasting as a result of advancements in computer technology. The goal of this study is to estimate the likelihood of rain in Australia using an active learning system. Accurate rainfall forecasting is a challenging procedure that requires constant progress. To train and test our models, we will utilise classification, confusion matrices, and linear regression, and our method for predicting rainfall fits into the conventional framework for synoptic weather prediction, which entails collecting and evaluating a lot of data.

2. Methodology

Materials

This dataset includes around ten years' worth of daily weather observations made by the Australian Government's Bureau of Meteorology in several places all throughout Australia.

NAME	UNITS
Temperature	Degree Celsius (°C)
Rainfall	Millimeters (mm)
Evaporation	Millimeters (mm)
Sunshine	Hours (h)
Wind gust	The greatest wind gust's velocity (km/h)
Wind direction	Degrees (°)
Wind speed	Average wind speed (km/hr) over 10 minutes
Pressure	hectopascals (hPa)
Cloud	octas

TABLE 1. Data Description

The dataset was acquired from Kaggle and includes daily weather measurements from several places around Australia over the course of roughly ten years. 23 columns, 145460 rows, 22 independent columns, and 22 dependent columns make up the dataset. With the exception of the Date and Location columns, which have no values, the dataset has two different data types: float64 and object.

```
print(rain.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 145460 entries, 0 to 145459
Data columns (total 23 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   Date           145460 non-null  object
 1   Location       145460 non-null  object
 2   MinTemp        143975 non-null  float64
 3   MaxTemp        144199 non-null  float64
 4   Rainfall       142199 non-null  float64
 5   Evaporation    82670 non-null   float64
 6   Sunshine       75625 non-null   float64
 7   WindGustDir    135134 non-null  object
 8   WindGustSpeed  135197 non-null  float64
 9   WindDir9am     134894 non-null  object
 10  WindDir3pm     141232 non-null  object
 11  WindSpeed9am   143693 non-null  float64
 12  WindSpeed3pm   142398 non-null  float64
 13  Humidity9am    142806 non-null  float64
 14  Humidity3pm    140953 non-null  float64
 15  Pressure9am    130395 non-null  float64
 16  Pressure3pm    130432 non-null  float64
 17  Cloud9am       89572 non-null   float64
 18  Cloud3pm       86102 non-null   float64
 19  Temp9am        143693 non-null  float64
 20  Temp3pm        141851 non-null  float64
 21  RainToday      142199 non-null  object
 22  RainTomorrow   142193 non-null  object
dtypes: float64(16), object(7)
memory usage: 25.5+ MB
```

Summary of a dataset

Equations

The relationship between the independent variable X and the dependent variable y is explained using regression models. Predictive or explanatory variables are additional names for independent variables, whereas the dependent variable can sometimes be referred to as the response variable. Continuous predictor variables may be referred to as covariates, whilst categorical predictor variables may alternatively be referred to as factors. The design matrix, abbreviated as X, is a matrix of observations on predictor variables.

A multiple linear regression model is

$$yi=\beta 0+\beta 1xi1+\beta 2xi2+\ldots+\beta pxip+\epsilon i,\ i=1,\ldots n$$

yi is the ith answer, where.

The model's constant term is denoted by 0, while the kth coefficient is denoted by k. The constant term may occasionally be mentioned in design matrices. You cannot insert a column of 1s into your design matrix X because LinearModel.fit or LinearModel.stepwise by default contain a constant term in the model.The ith noise term, or random error, is I and Xij is the ith observation on the jth predictor variable, where j = 1,..., p. A model of the following generalisation can be a linear regression model:

Data Preparation

In our method, we predict rainfall using a modified version of linear regression. The steps that follow provide an explanation of how this strategy works.A review of the supplied datasets. The training set's input data were collected between 2008 and 2017. Implement the suggested system and verify the procedure.

1. The input datasets are used to create the training and test data. For the years 2023 to 2027, the training set includes average temperature values from the input datasets, together with information on precipitation, evaporation, sunshine, wind gusts, wind directions, wind speeds, pressure, cloud, and rain temperatures. This training set is used using the suggested strategy. The test data includes information from 2023 to 2027 that is used to test the model.
2. Drop the date column that has information, as a training parameter and find the numerical variable. Drop the Rain Tomorrow column because it's the output and not part of the variables we need.

	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	WindSpeed9am	WindSpeed3pm
0	Albury	13.4	22.9	0.6	NaN	NaN	W	44.0	W	WNW	20.0	24.0
1	Albury	7.4	25.1	0.0	NaN	NaN	WNW	44.0	NNW	WSW	4.0	22.0
2	Albury	12.9	25.7	0.0	NaN	NaN	WSW	46.0	W	WSW	19.0	26.0
3	Albury	9.2	28.0	0.0	NaN	NaN	NE	24.0	SE	E	11.0	9.0
4	Albury	17.5	32.3	1.0	NaN	NaN	W	41.0	ENE	NW	7.0	20.0

3. Finding the categorical and numerical features in a dataset.
4. Insert the mode of each column into the categorical variables. In a numerical variable, replace the null values with the corresponding median.

```
MinTemp          1485
MaxTemp          1261
Rainfall         3261
Evaporation     62790
Sunshine        69835
WindGustSpeed   10263
WindSpeed9am     1767
WindSpeed3pm     3062
Humidity9am      2654
Humidity3pm      4507
Pressure9am     15065
Pressure3pm     15028
Cloud9am        55888
Cloud3pm        59358
Temp9am          1767
Temp3pm          3609
year                0
month               0
day                 0
dtype: int64
```

5. Encode the string type variables into numerical data by using preprocessing LabelEncoder.

	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	WindSpeed9am	WindSpeed3pm
0	0	13.4	22.9	0.6	5.318667	7.611178	0	44.0	0	0	20.0	24.0
1	0	7.4	25.1	0.0	5.318667	7.611178	1	44.0	1	1	4.0	22.0
2	0	12.9	25.7	0.0	5.318667	7.611178	2	46.0	0	1	19.0	26.0
3	0	9.2	28.0	0.0	5.318667	7.611178	3	24.0	2	2	11.0	9.0
4	0	17.5	32.3	1.0	5.318667	7.611178	0	41.0	3	3	7.0	20.0

These category data must be encoded into numerical data using the replace() method in order to be used in modelling data.

6. Split the dataset by 80% Training and 20% Testing. Train the model using logistic regression by taking X train and Y train and fit into a logistic regression.

```
X = rain.drop(['RainTomorrow'],axis=1)

y = rain['RainTomorrow']
```

i) Split data into independent and dependent features.

X – Independent Features or Input features   Y – Dependent Features or target label

ii) Data division into a test and training set

iii) Model evaluation using the class of logistic regression to train and evaluate models or classifiers.

iv) Confusion Matrix used to assess how well the categorization task performed. It presents an overall assessment of the model's effectiveness.

7. A linear regression is used to forecast the amount of precipitation, with the average temperature and cloud cover acting as independent variables and the rainfall from the training datasets as the dependent variable.
8. Check accuracy and have the confusion matrix
9. The test data are forecast using the most recently updated coefficients, which results in the most precise prediction values.

Cardinality check for categorical features

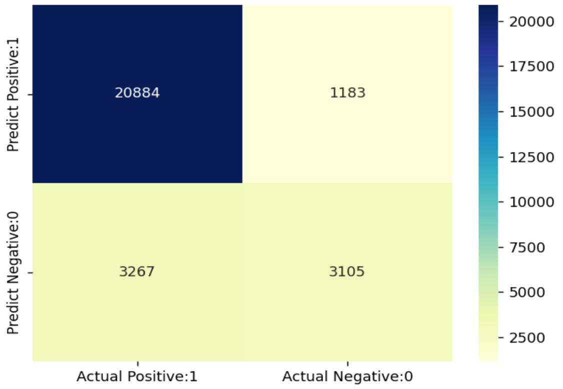
In addition to the model we employ, how data is pre-processed and the kind of data you give the classifier when it is learning to affect the accuracy and performance of the classifier. Date columns have a large cardinality, which causes the model to perform poorly and causes data dimensions to grow when they are represented as numerical data.

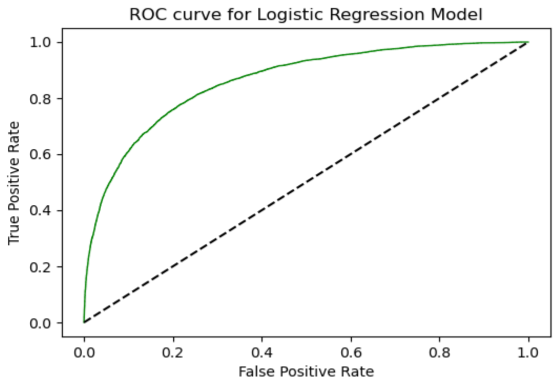
3. Result

The outcomes of using the scikit-learn library's Python Logistic Regression function are presented in this section. The goal is to accurately estimate the variable "RainToday" using the rainfall database for Australia. The information obtained is shown below:

	Precision	**Recall**	**F1-Score**	**Support**
0	0.86	0.95	0.90	22067
1	0.72	0.49	0.58	6372
Accuracy			0.84	28439
Macro average	0.79	0.72	0.74	28439
Weighted average	0.83	0.84	0.83	28439

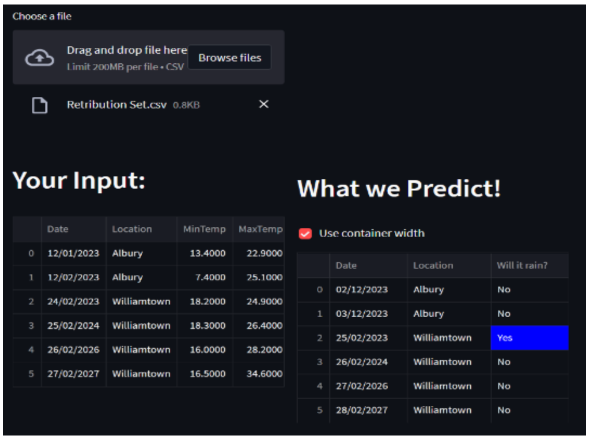
Classification Report of the Logistic Regression algorithm.





Confusion Matrix of the Model ROC ***Curve of Logistic Regression Model***

Based on our findings, it can be said that we have achieved a staggering 84% as our accuracy score for the logistic regression model. The model is quite effective at forecasting the weather in Australia based on this dataset.Both underfitting and overfitting are not evident in the model. This illustrates how effectively the model generalises to new data. Nearly the same as the mean accuracy score of the original model, the cross-validation accuracy score. Therefore, cross-validation may not be able to increase the model's accuracy.



Result of Rain Prediction Dataset

4. Discussion

Based on our findings, an accuracy of 84% is relatively a good value and can be said that logistic regression is a suitable algorithm to fit into this model. However, due to use of only one algorithm in this study which is the Python Logistic Regression function defined in the scikit learn library. By doing so, we limit our quantitative comparisons as we are not able to compare our findings to other algorithms. We make up for this limitation by comparing our findings to the results of other research papers using the same dataset as the data set that we have chosen is that of the Australian Government. The data from this research were compared to Cabezuelo, (2022) [11] to see whether the training outcomes were particularly good. From his paper, it can be seen that the percentage of accuracy for 3 different algorithms; Random Forest, Decision Tree and KNN are the same at 83% accuracy. This shows that Logistic Regression has the highest percentage of accuracy even when compared to other algorithms. This might be caused by logistic regression having the slight upper hand in terms of its nature of being a parametric test whereas the other 3 algorithms are non-parametric tests. Further research can be done into finding out which algorithm would be the most suitable to be used in predicting the rain.

5. Conclusion

In summary, this paper proposes a technique for predicting the chance of rain in Australia using supervised machine learning and logistic regression. The dataset covers elements including cloudiness, wind, sunshine, humidity, pressure, temperature, and whether or not it rained on the sample day. It also includes information on these and other factors. The study applies various machine learning models to the pre-processed data to make rainfall predictions. The method used in this study is a modified version of linear regression. The data preparation process involved reviewing the supplied datasets and using them to create the training and test data. The input datasets were collected between 2008 and 2017 and were used to train the model. The test data includes information from 2023 to 2027 that is used to test the model. The dataset was pre-processed by dropping the date column, filling in missing values and encoding string type variables. Next, the dataset was divided into 20% for testing and 80% for training. The model was trained using logistic regression, utilising the average temperature and cloud cover as independent variables and the rainfall from the training datasets as the dependent variable. The accuracy of the model was then checked and a confusion matrix was generated. The test data is forecast using the most recent updated coefficients to provide the most precise prediction values.

This work can be continued in several possible ways. By using an updated dataset from the same source, the model can be tested again to see if it still holds up. Not only that, but the model then can be improved with this new updated dataset as it is new information being fed into the artificial intelligence. Finally, a separate study can run multiple algorithms with this dataset to see which will be the most efficient and accurate by using better measurements and metrics.

6. Acknowledgment

For our research in artificial intelligence, we are extremely grateful to the Data Science Research Lab at the School of Computing, College of Art & Sciences, University Utara Malaysia. Without the perseverance of our coworkers, with whom we worked closely to accomplish this project after extensive research, discussion, and thought, we would not have been able to travel this route. Last but not least, we would like to express our gratitude to those who, in spite of their hectic schedules, assisted in completing this project within the allotted time period and occasionally provided advice to help us create it..

7. References

[1] A., & Fotovatikhah, F. (2018). Survey of computational intelligence as a basis to Big Flood Management: Challenges, Research Directions, and future work. Taylor & Francis.

[2] Balamurugan, M. S., & Manojkumar, R. (2019, October 26). Study of short-term rain forecasting using machine learning-based approach - wireless networks. SpringerLink.

[3] Benedict. (2022, March 28) Predicting rain with machine learning - towards data science.

[4] Cabezuelo, Antonio. (2022). Prediction of Rainfall in Australia Using Machine Learning. Information. 13. 163. 10.3390/info13040163.

[5] Liu, J., Cho, H.-S., Osman, S., Jeong, H.-G., & Lee, K. (2022). Review of the status of urban flood monitoring and forecasting in TC region. Tropical Cyclone Research and Review, 11(2), 103–119.

[6] Mohammed, M., Kolapalli, R., Golla, N., & Maturi, S. (n.d.). Prediction Of Rainfall Using Machine Learning Techniques.

[7] Pham, Q. B., Abba, S. I., Usman, A. G., Linh, N. T. T., Gupta, V., Malik, A., Costache, R., Vo, N. D., & Tri, D. Q. (2019). Potential of Hybrid Data-Intelligence Algorithms for Multi-Station Modelling of Rainfall. Water Resources Management, 33(15), 5067–5087.

[8] Parmar, Aakash, Kinjal Mistree, and Mithila Sompura. "Machine learning techniques for rainfall prediction: A review." 2017 International Conference on Innovations in information Embedded and Communication Systems. 2017.

[9] Raval, M., Sivashanmugam, P., Pham, V., Gohel, H., Kaushik, A., & Wan, Y. (2021). Automated predictive analytics tool for rainfall forecasting. Scientific Reports, 11(1).

[10] Sarasa-Cabezuelo, A. (2022). Prediction of Rainfall in Australia Using Machine Learning. Information, 13(4), 163.

[11] Z. He, "Rain Prediction In Australia With Active Learning Algorithm," 2021 International Conference on Computers and Automation (CompAuto), Paris, France, 2021, pp. 14-18.