Article

Data Preprocessing: Case Study On Australia Weather Dataset

Moubarak Mohamed Misbahou[1]
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM, Bangi Selangor, Malaysia; moubarak.mkouboi@gmail.com
* Correspondence: moubarak.mkouboi@gmail.com

Abstract: The weather remains one of the challenges in a country's development. It remains a mystery that meteorologists are still discovering new insights to understand it and make it use to our advantage to develop our country. Developing a predictive analysis to predict the next-day rain was suggested to overcome this problem. This paper will provide a detailed explanation of the dataset's pre-processing, with a focus on data reduction, cleaning, and transformation. We'll also talk about handling class imbalance problems and their associated data transformation.

Keywords: Predictive analysis, data pre-processing, imbalance data, SMOTENC.

1. Introduction

Weather is a natural phenomenon that impacts our daily activities and living conditions. The rain, wind, and temperature in the atmosphere above the earth, particularly when they occur over a specific region which is Australia in this study and at a specific time [1]. The temperature is one of the major factors of weather conditions whether it's by rain or snow and so on. Daily activities can be ruined because of the weather conditions and even accidents can be occurred because of it. Additionally, it is impacting other sectors as well such as the ecosystems and its biodiversity. Research by [2] reported that the water temperature change may alter the metabolism and physiology of aquatic animals thereby affecting the growth, fecundity, feeding behaviour, distribution, migration and abundance of fish as well as other aquatic animals. Some related work shows that using machine learning methods could reduce major incidents caused by the weather.

An imbalance in a typical dataset leads to a reduction in the generalisation of machine learning algorithms [3]. Data collected from diverse sources may exhibit missing data, noise, inconsistencies, excess volume, and class imbalance. To acquire the necessary understanding or information, this requires a stage of data preparation to clean and prepare the data for further analysis. To increase the overall performance of the model, this phase should be properly done, as it typically requires a large amount of time [3].

To reduce complexity, the data preprocessing activity involves several steps, such as data preparation, integration, cleaning, normalisation, scaling, and data reduction techniques. Feature selection and discretization are also used to handle noise in the data, outliers, and irrelevant components.

To provide an organised dataset for additional predictive analysis, the goal of this study is to carry out thorough data preprocessing work on the chosen dataset. To create a predictive model that performs better overall and has higher accuracy, a high-quality and clean dataset is needed.

2. Related Work

The use of machine learning for climate analysis and weather forecasting has advanced significantly in recent years. To increase the precision and dependability of weather forecasts, several approaches and strategies have been investigated. Here, we go over several noteworthy studies that have advanced this area.

A thorough summary of the uses and perspectives of machine learning in climate analysis and weather forecasting is given by Bochenek and Ustrnul [4]. Their research demonstrates how machine learning algorithms can improve our comprehension and forecasting of intricate meteorological occurrences.

Using a focus on smart city applications, Rahman et al. created a rainfall forecast system by combining various machine-learning approaches [5]. Their research shows how combining several machine learning techniques can improve rainfall event prediction ability, which is important for disaster management and urban planning.

Watson-Parris looked at the many difficulties and approaches involved in using machine learning in the weather and climate [6]. The paper makes the case that although machine learning is beneficial to both domains, there is a need for customized solutions in each because the requirements and methods vary significantly.

In a survey on deep learning-based weather prediction, Ren et al. offered a thorough examination of different deep learning architectures and how well they predict the weather [7]. Their survey is a useful tool for learning about the state of deep learning applications in weather prediction as well as their potential prospects.

Markovics and Mayer conducted a comparative analysis of different machine-learning techniques for forecasting solar power using numerical weather prediction [8]. Their research shows how precise weather forecasts may greatly increase the dependability of renewable energy sources, making it especially pertinent for fusing meteorological data with energy forecasting.

3. Material and Methods

For data scientists and analysts to transform and visualize the dataset in a way that the stakeholders will understand, preprocessing is a crucial step. The processed dataset is usually used for modelling or analysis to derive insights from the data and address a particular problem in the dataset in question. The preprocessing processes I used are illustrated in Figure 1.

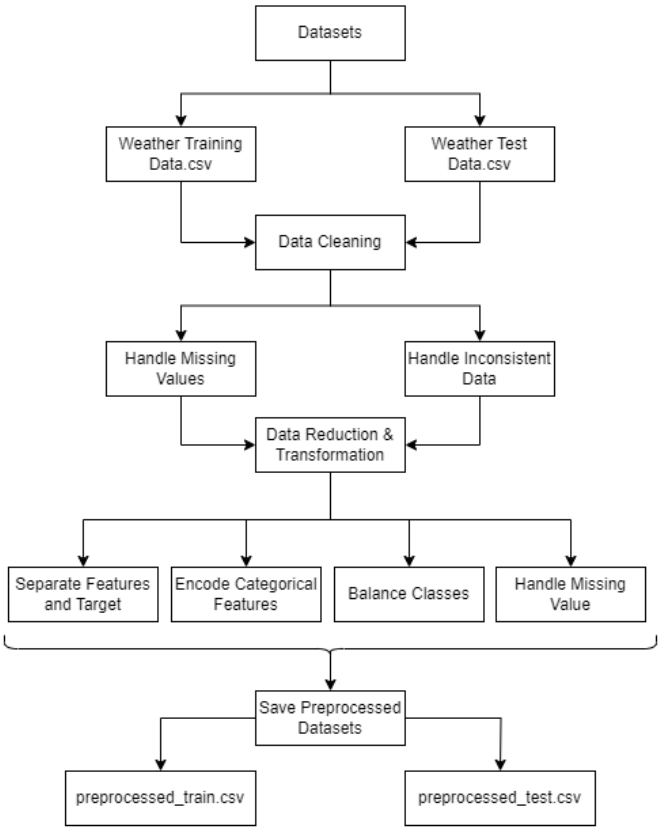


Figure 1. Preprocessing Steps

3.1. Tools

The Python programming language and the Google Colab Framework, which allowed me to code using Colab Notebook, were the two primary tools implemented in this study. These two tools were used for every process, and the procedure associated with the explanations can be explored in the subsections that follow.

3.2. Data Source

The main dataset was collected from the Kaggle community, an online forum for machine learning and data science professionals. With Kaggle, users may collaborate with other data scientists and machine learning experts, discover and post data sets, explore and develop models in a web-based data science environment, and compete to solve data science problems[9]. Arunava Kr. Chakraborty provides this dataset, which contains daily weather measurements from several places in Australia covering almost ten years [9].

3.3. Data Description

The last attribute (RainTomorrow) in the Weather training data is the class attribute. The dataset contains 42677 instances of Weather test data with 22 attributes and 99516 instances of Weather training data with 23 attributes. Except for row ID, location, WindGustDir, WindDir9am, WindDir3pm, and RainToday, all attributes' data are numerical data type. The descriptions of each attribute are listed in Table 1. All statistical measurements of the numerical and categorical features are presented in Tables 2, 3, 4, and 5.

Table 1. Description of the dataset

Attributes	Description	Data Type
Row ID	A unique row identifier assigned to each row in the dataset	Nominal
Location	Name of the city from Australia	Nominal
MinTemp	The Minimum temperature during a particular day. (degree Celsius)	Numerical
MaxTemp	The maximum temperature during a particular day. (degree Celsius)	Numerical
Rainfall	Rainfall during a particular day. (millimeters)	Numerical
Evaporation	Evaporation during a particular day. (millimeters)	Numerical
Sunshine	Bright sunshine during a particular day. (hours)	Numerical
WindGusDir	The direction of the strongest gust during a particular day. (16 compass points)	Ordinal
WindGuSpeed	Speed of strongest gust during a particular day. (kilometers per hour)	Numerical
WindDir9am	The direction of the wind for 10 min prior to 9 am. (compass points)	Ordinal
WindDir3pm	The direction of the wind for 10 min prior to 3 pm. (compass points)	Ordinal
WindSpeed9am	Speed of the wind for 10 min prior to 9 am. (kilometers per	Numerical

	hour)	
WindSpeed3pm	Speed of the wind for 10 min prior to 3 pm. (kilometers per hour)	Numerical
Humidity9am	The humidity of the wind at 9 am. (percent)	Numerical
Humidity3pm	The humidity of the wind at 3 pm. (percent)	Numerical
Pressure9am	Atmospheric pressure at 9 am. (hectopascals)	Numerical
Pressure3pm	Atmospheric pressure at 3 pm. (hectopascals)	Numerical
Cloud9am	Cloud-obscured portions of the sky at 9 am. (eighths)	Numerical
Cloud3pm	Cloud-obscured portions of the sky at 3 pm. (eighths)	Numerical
Temp9am	The temperature at 9 am. (degree Celsius)	Numerical
Temp3pm	The temperature at 3 pm. (degree Celsius)	Numerical
RainToday	If today is rainy then 'Yes'. If today is not rainy then 'No'	Binary
RainTomorrow	If tomorrow is rainy then 1 (Yes). If tomorrow is not rainy then 0 (No)	Binary

Table 2. Dataset Quality Report for Training Data (Numerical Features)

Attributes	count	mean	std	min	25%	50%	75%	max
MinTemp	99073	12.17627	6.390882	-8.5	7.6	12	16.8	33.9
MaxTemp	99286	23.21851	7.115072	-4.1	17.9	22.6	28.2	48.1
Rainfall	98537	2.353024	8.487866	0	0	0	0.8	371
Evaporation	56985	5.46132	4.16249	0	2.6	4.8	7.4	86.2
Sunshine	52199	7.61509	3.783008	0	4.8	8.4	10.6	14.5
WindGustSpeed	93036	39.97697	13.58152	6	31	39	48	135
WindSpeed9am	98581	14.00485	8.902323	0	7	13	19	130
WindSpeed3pm	97681	18.65046	8.801827	0	13	19	24	87
Humidity9am	98283	68.86638	19.07495	0	57	70	83	100
Humidity3pm	97010	51.4333	20.77762	0	37	52	65	100
Pressure9am	89768	1017.685	7.110166	980.5	1013	1017.7	1022.4	1041
Pressure3pm	89780	1015.286	7.045189	978.2	1010.5	1015.3	1020	1039.6
Cloud9am	61944	4.447985	2.88658	0	1	5	7	9
Cloud3pm	59514	4.519122	2.716618	0	2	5	7	9
Temp9am	98902	16.97004	6.488961	-7	12.3	16.7	21.5	40.2
Temp3pm	97612	21.68134	6.931681	-5.1	16.6	21.1	26.4	46.7
RainTomorrow	99516	0.224677	0.417372	0	0	0	0	1

Table 3. Dataset Quality Report for Training Data (Categorical Features)

Attributes	count	unique	top	freq
row ID	99516	99516	Row0	1
Location	99516	49	canberra	2393
WindGustDir	99516	16	W	13364
WindDir9am	99516	16	N	15058
WindDir3pm	99516	16	SE	10058
RainToday	99516	2	No	77460

Table 4. Dataset Quality Report for Test Data (Numerical Features)

Attributes	count	mean	std	min	25%	50%	75%	max
MinTemp	42483	12.21003225	6.43212164	-8.2	7.6	12	16.9	31.8
MaxTemp	42585	23.24606786	7.123596236	-4.8	18	22.6	28.3	47
Rainfall	42250	2.342861538	8.412105838	0	0	0	0.6	278.4
Evaporation	24365	5.489714755	4.248849509	0	2.6	4.8	7.4	145
Sunshine	22178	7.647831184	3.778018747	0	4.9	8.5	10.7	14.3
WindGustSpeed	39887	40.0013789	13.60591481	7	31	39	48	122
WindSpeed9am	42264	13.99531516	8.872444823	0	7	13	19	74
WindSpeed3pm	41882	18.60751636	8.806915849	0	13	19	24	83
Humidity9am	42136	68.79117619	18.9961152	1	57	70	83	100
Humidity3pm	41573	51.59767157	20.84452548	1	37	52	66	100
Pressure9am	38411	1017.581591	7.094069991	982.2	1012.9	1017.6	1022.4	1040.4
Pressure3pm	38432	1015.192792	7.016408003	977.1	1010.4	1015.2	1020	1038.9
Cloud9am	26592	4.412041215	2.887927461	0	1	5	7	8
Cloud3pm	25585	4.466054329	2.729640424	0	2	5	7	8
Temp9am	42387	17.02826574	6.501770926	-7.2	12.3	16.8	21.6	39.4
Temp3pm	41855	21.70098196	6.951426386	-5.4	16.6	21.1	26.5	45.4

Table 5. Dataset Quality Report for Test Data (Categorical Features)

Attributes	count	unique	top	freq
row ID	42677	42677	Row0	1
Location	42677	49	Canberra	1025
WindGustDir	39868	16	W	2937
WindDir9am	39670	16	N	3341
WindDir3pm	41547	16	SE	3253
RainToday	42250	2	No	32851

Data has been collected from multiple weather stations. The Australian Government's Climate Data Online, accessible on the Bureau of Meteorology website, provides access to the daily observations. The definitions are taken from the Australian Government's Bureau of Meteorology website.

3.4. Data Cleaning

The process of fixing or removing incorrect, corrupted, inappropriately formatted, duplicate, or incomplete data from a dataset is known as data cleaning. It is normal for data to be mislabeled or duplicated when integrating different data sources [10]. A check for missing values was performed and several missing values were found as shown in Figures 2 and 3. Also, no duplicate values were found.

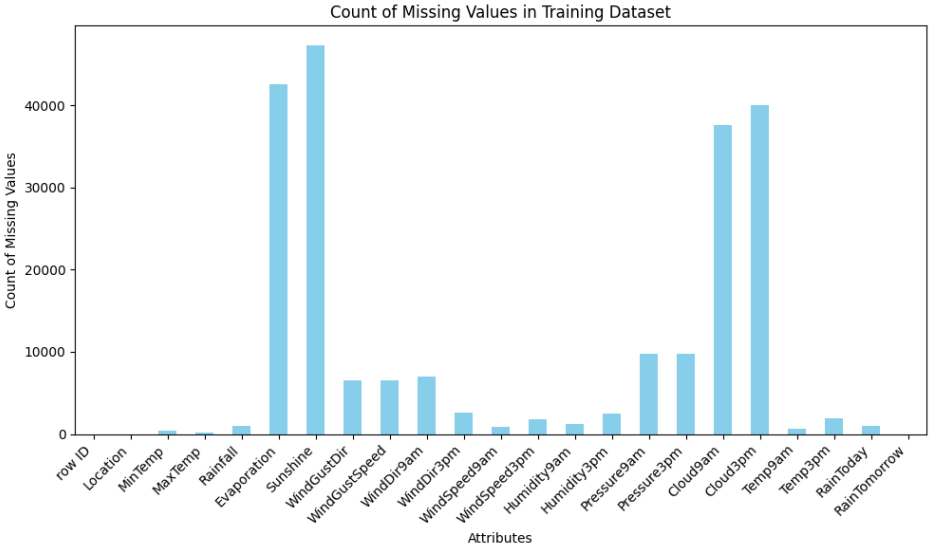


Figure 2. Missing Values in Training Dataset

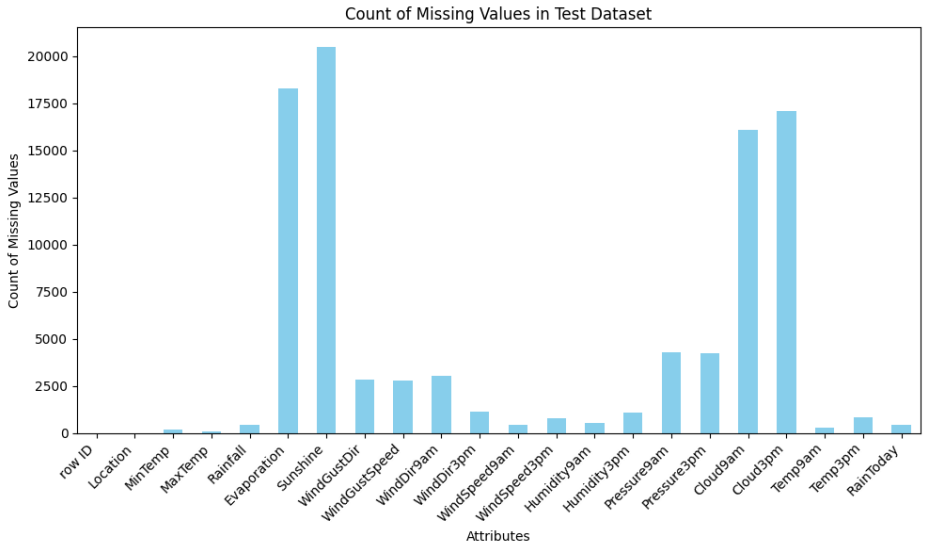


Figure 3. Missing Values in Test Dataset

To guarantee data quality, missing and inconsistent values must be addressed at the preprocessing stage of our data analysis. The functions used to deal with these problems are listed below:

- Handling Missing Values

  We used the handle_missing_values function to deal with missing values. This method prints the initial shape of the DataFrame first, then iterates through each column, adding the median in numerical columns and the mode in categorical columns to fill in missing values.

- Handling Inconsistent Data

  We used the handle_inconsistent_data method to make sure that there was consistency, especially in the Location column. This function prints the original shape of the DataFrame and then standardises the Location values by changing the text to lowercase and removing spaces.

- Implementation

On both the training and test datasets, the handle_missing_values function was executed to make sure there were no missing items. The Location column in both datasets was then normalised using the handle_inconsistent_data function.

After using the handle_missing_values function, Figures 4 and 5 demonstrate the success of our data-cleaning procedure by demonstrating that the dataset has no missing values.

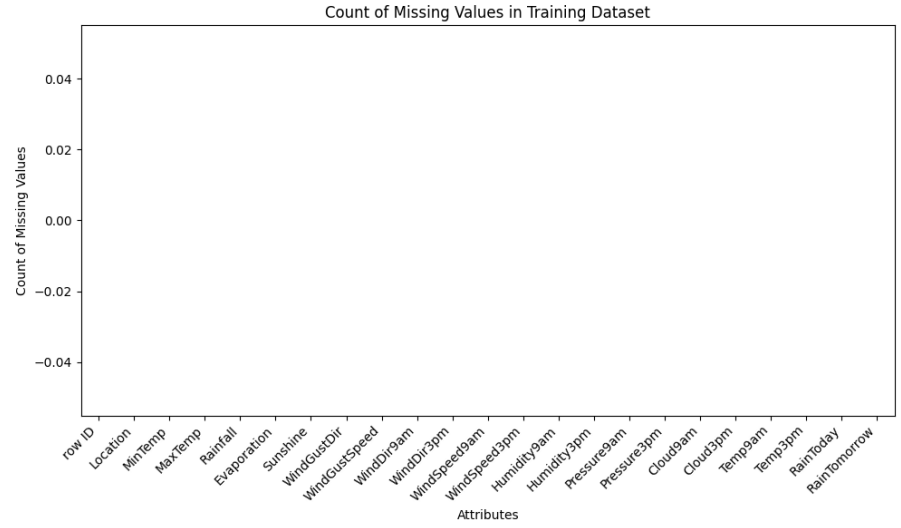


Figure 4. Missing Values in the Training Dataset after cleaning

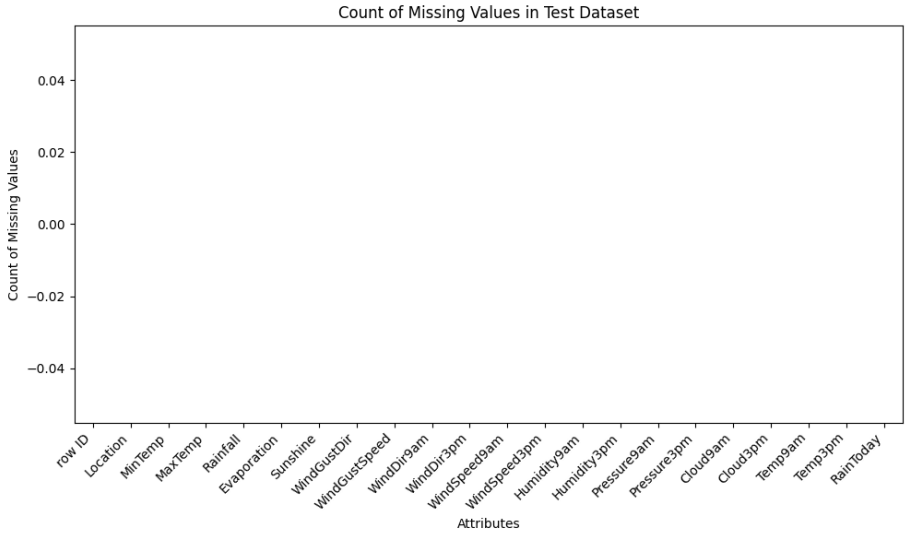


Figure 5. Missing Values in the Training Dataset after cleaning

To guarantee that the data used for ensuing analysis and model training procedures is clean and consistent, these preprocessing procedures are essential for preserving data integrity.

3.5. Data Transformation

The method of transforming, cleaning, and organizing data into a format that can be used for analysis and assisting decision-making processes to further the growth of a business is known as data transformation [11]. A check for noise and outliers was performed and not many noises and outliers were found as shown in Figures 6, 7, 8 and 9. Also, no duplicate values were found.

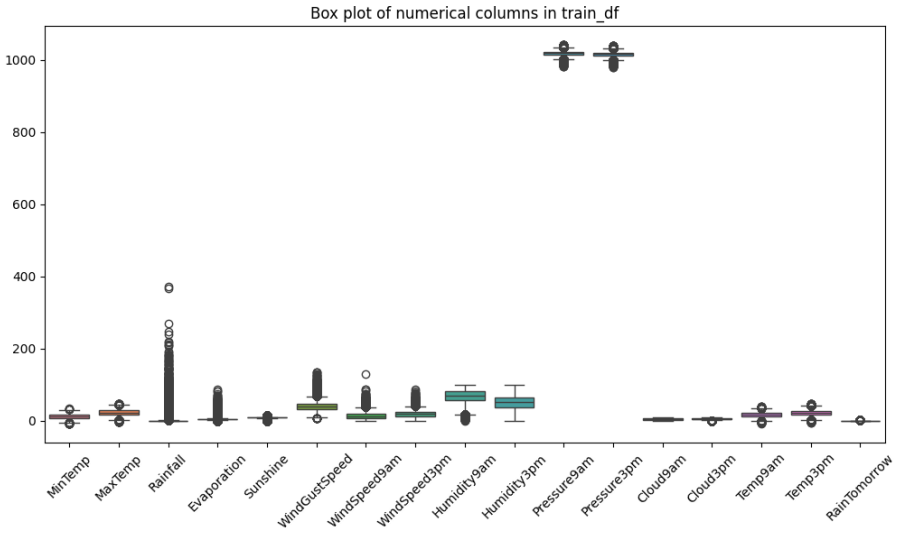


Figure 6. Box plot of numerical columns in training data

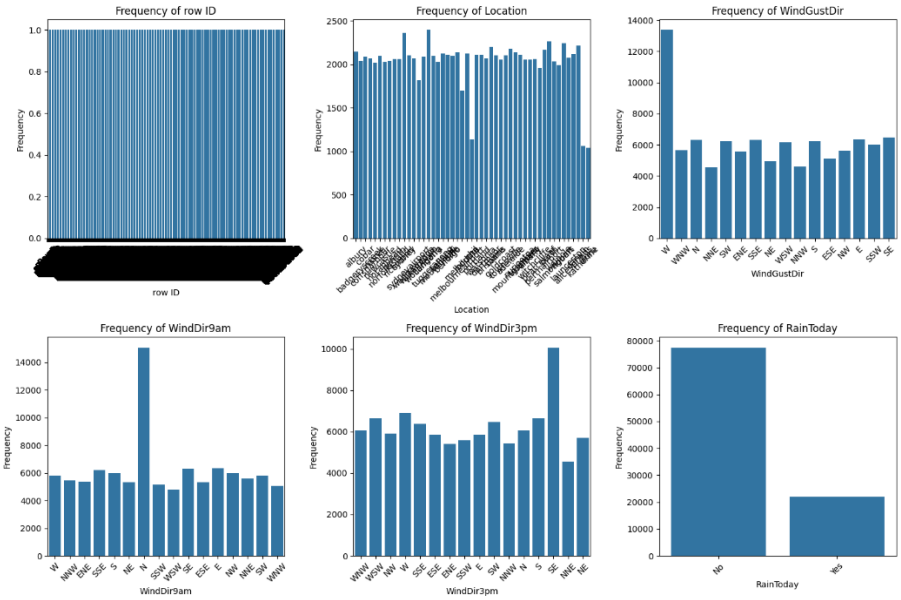


Figure 7. Bar plots for categorical variables in the training dataset

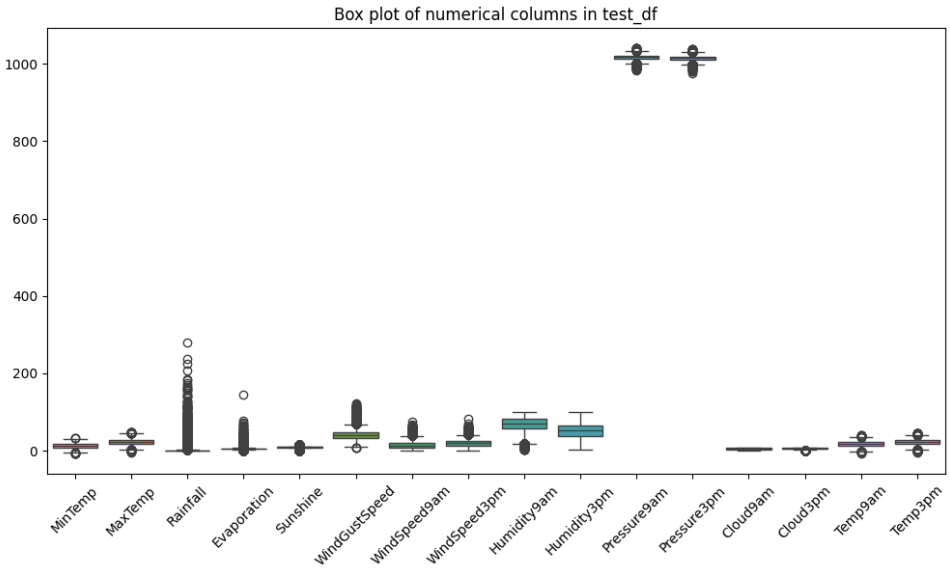


Figure 8. Box plot of numerical columns in test data

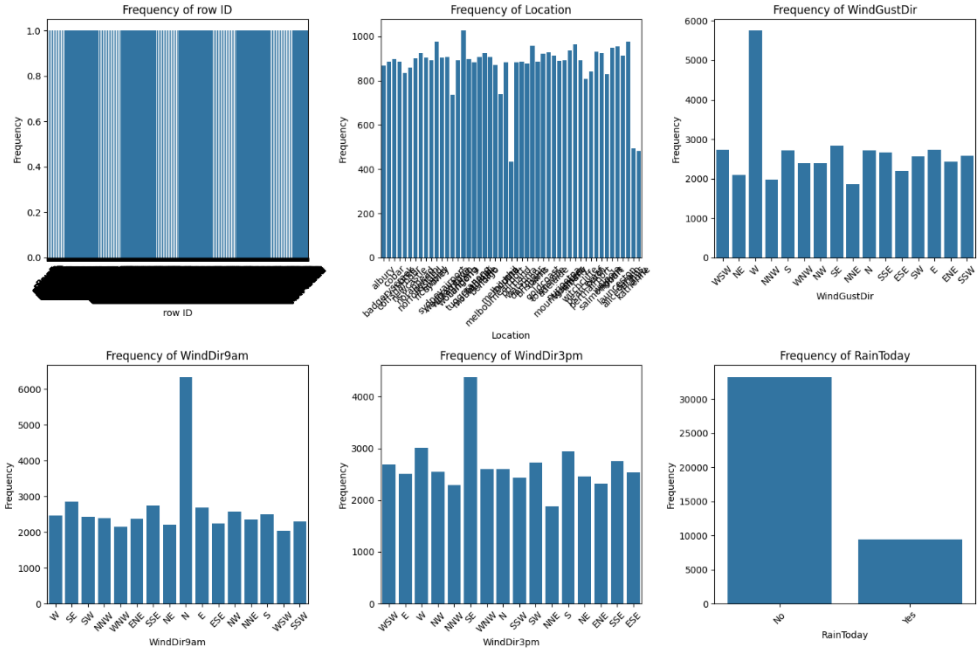


Figure 9. Bar plots for categorical variables in the test dataset

We preprocess the input datasets for machine learning tasks after examining noise and outliers. The target variable ('RainTomorrow') and features are first separated. LabelEncoder is used to encode the 'RainToday' attribute and convert its values to a numerical representation. Given that the wind direction features are ordinal, OrdinalEncoder is used to encode them ('WindGustDir,' 'WindDir9am,' and 'WindDir3pm'). To maintain uniformity, numerical columns are rounded to one decimal place. In categorical columns, missing values are represented by -1 to denote unmapped values, which we will remove if we discover any before modelling. SMOTENC is used to consider both numerical and categorical variables to balance the classes in the training dataset.

4. Results of data pre-processing

The final dataset after the preprocessing process is shown in Figure 10. A total of 22 attributes and 154314 instances will be used for developing a predictive model. A total of 21 attributes and 42677 instances will be used for testing the model. A summary of data preprocessing tasks for this dataset is described in Table 6.

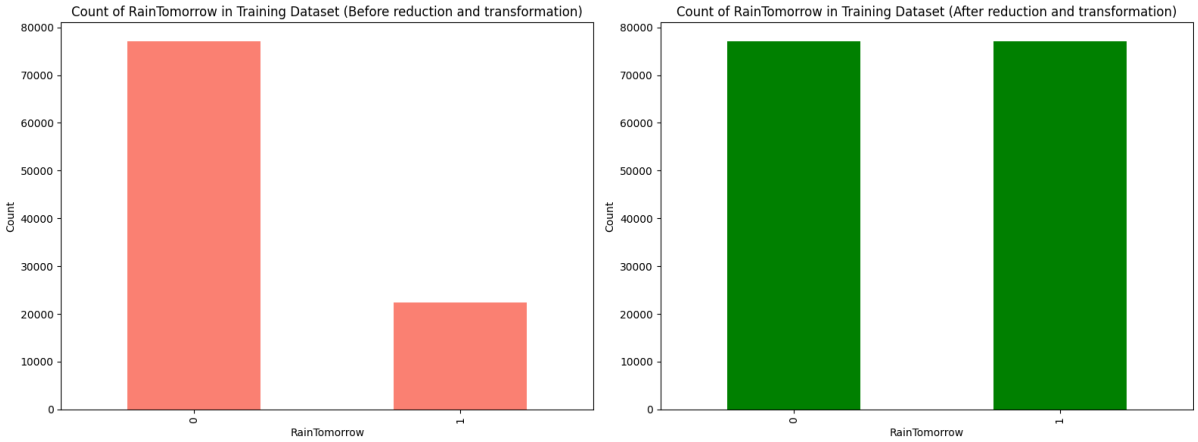


Figure 10. Plot of Before and After reduction and transformation

Table 5: Summary of data preprocessing tasks

Data Preprocessing Tasks	Description
Data Cleaning	a) Fill in missing values b) Make 'Location' features have the same format
Data Transformation	c) Removed 'row ID' attributes from both datasets. d) Changing 'RainToday' features from Binary to Nominal features e) Changing 'WindGustDir,' 'WindDir9am,' and 'WindDir3pm' features from Binary to Nominal features f) Performed SMOTENC on all the attributes. g) Saved preprocessed data

Additionally, you can see parts of the original datasets' raw data as well as data that has been cleaned and transformed in Figures 11, 12, 13, and 14.

row ID	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday	RainTomorrow
Row0	Albury	13.4	22.9	0.6			W	44	W	WNW	20	24	71	22	1007.7	1007.1	8		16.9	21.8	No	0
Row1	Albury	7.4	25.1	0			WNW	44	NNW	WSW	4	22	44	25	1010.6	1007.8			17.2	24.3	No	0
Row2	Albury	17.5	32.3	1			W	41	ENE	NW	7	20	82	33	1010.8	1006	7	8	17.8	29.7	No	0
Row3	Albury	14.6	29.7	0.2			WNW	56	W	W	19	24	55	23	1009.2	1005.4			20.6	28.9	No	0
Row4	Albury	7.7	26.7	0			W	35	SSE	W	6	17	48	19	1013.4	1010.1			16.3	25.5	No	0
Row5	Albury	13.1	30.1	1.4			W	28	S	SSE	15	11	58	27	1007	1005.7			20.1	28.2	Yes	0
Row6	Albury	13.4	30.4	0			N	30	SSE	ESE	17	6	48	22	1011.8	1008.7			20.4	28.8	No	1
Row7	Albury	15.9	21.7	2.2			NNE	31	NE	ENE	15	13	89	91	1010.5	1004.2	8	8	15.9	17	Yes	1
Row8	Albury	12.6	21	3.6			SW	44	W	SSW	24	20	65	43	1001.2	1001.8		7	15.8	19.8	Yes	0
Row9	Albury	9.8	27.7				WNW	50	NA	WNW		22	50	28	1013.4	1010.3	0		17.3	26.2	NA	0
Row10	Albury	14.1	20.9	0			ENE	22	SSW	E	11	9	69	82	1012.2	1010.4	8	1	17.2	18.1	No	1

Figure 11. Portion of raw data of the original Weather train data

row ID	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday
Row0	Albury	12.9	25.7	0			WSW	46	W	WSW	19	26	38	30	1007.6	1008.7		2	21	23.2	No
Row1	Albury	9.2	28	0			NE	24	SE	E	11	9	45	16	1017.6	1012.8			18.1	26.5	No
Row2	Albury	14.3	25	0			W	50	SW	W	20	24	49	19	1009.6	1008.2	1		18.1	24.6	No
Row3	Albury	9.7	31.9	0			NNW	80	SE	NW	7	28	42	9	1008.9	1003.6			18.3	30.2	No
Row4	Albury	15.9	18.6	15.6			W	61	NNW	NNW	28	28	76	93	994.3	993	8	8	17.4	15.8	Yes
Row6	Albury	11.5	29.3	0			S	24	SE	SE	9	9	56	28	1019.3	1014.8			19.1	27.3	No
Row7	Albury	16.2	33.9	0			WSW	35	SE	WSW	9	13	45	19	1010.9	1007.6		1	23.2	33	No
Row8	Albury	12	24.4	0.8			W	39	WNW	WNW	17	17	48	28	1006.1	1005.1	1		16.9	22.7	No
Row9	Albury	11.3	26.5	0			WNW	56	W	WNW	19	31	46	26	1004.5	1003.2			19.7	25.7	No
Row10	Albury	12.9	35.8	0			WNW	41	ENE	NW	6	26	41	9	1012.6	1009.2			22.4	34.4	No

Figure 12. Portion of raw data of the original Weather test data

Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday	RainTomorrow
albury	13.4	22.9	0.6	4.8	8.4	12	44	12	13	20	24	71	22	1007.7	1007.1	8	5	16.9	21.8	0	0
albury	7.4	25.1	0	4.8	8.4	13	44	15	11	4	22	44	25	1010.6	1007.8	5	5	17.2	24.3	0	0
albury	17.5	32.3	1	4.8	8.4	12	41	3	14	7	20	82	33	1010.8	1006	7	8	17.8	29.7	0	0
albury	14.6	29.7	0.2	4.8	8.4	13	56	12	12	19	24	55	23	1009.2	1005.4	5	5	20.6	28.9	0	0
albury	7.7	26.7	0	4.8	8.4	12	35	7	12	6	17	48	19	1013.4	1010.1	5	5	16.3	25.5	0	0
albury	13.1	30.1	1.4	4.8	8.4	12	28	8	7	15	11	58	27	1007	1005.7	5	5	20.1	28.2	1	0
albury	13.4	30.4	0	4.8	8.4	0	30	7	5	17	6	48	22	1011.8	1008.7	5	5	20.4	28.8	0	1
albury	15.9	21.7	2.2	4.8	8.4	1	31	2	3	15	13	89	91	1010.5	1004.2	8	8	15.9	17	1	1
albury	12.6	21	3.6	4.8	8.4	10	44	12	9	24	20	65	43	1001.2	1001.8	5	7	15.8	19.8	1	0
albury	9.8	27.7	0	4.8	8.4	13	50	0	13	13	22	50	28	1013.4	1010.3	0	5	17.3	26.2	0	0
albury	14.1	20.9	0	4.8	8.4	3	22	9	4	11	9	69	82	1012.2	1010.4	8	1	17.2	18.1	0	1

Figure 13. Portion of raw data of the Weather train data after data cleaning and transformation

Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday
albury	12.9	25.7	0	4.8	8.5	11	46	12	11	19	26	38	30	1007.6	1008.7	5	2	21	23.2	0
albury	9.2	28	0	4.8	8.5	2	24	6	4	11	9	45	16	1017.6	1012.8	5	5	18.1	26.5	0
albury	14.3	25	0	4.8	8.5	12	50	10	12	20	24	49	19	1009.6	1008.2	1	5	18.1	24.6	0
albury	9.7	31.9	0	4.8	8.5	15	80	6	14	7	28	42	9	1008.9	1003.6	5	5	18.3	30.2	0
albury	15.9	18.6	15.6	4.8	8.5	12	61	15	15	28	28	76	93	994.3	993	8	8	17.4	15.8	1
albury	11.5	29.3	0	4.8	8.5	8	24	6	6	9	9	56	28	1019.3	1014.8	5	5	19.1	27.3	0
albury	16.2	33.9	0	4.8	8.5	11	35	6	11	9	13	45	19	1010.9	1007.6	5	1	23.2	33	0
albury	12	24.4	0.8	4.8	8.5	12	39	13	13	17	17	48	28	1006.1	1005.1	1	5	16.9	22.7	0
albury	11.3	26.5	0	4.8	8.5	13	56	12	13	19	31	46	26	1004.5	1003.2	5	5	19.7	25.7	0
albury	12.9	35.8	0	4.8	8.5	13	41	3	14	6	26	41	9	1012.6	1009.2	5	5	22.4	34.4	0
albury	13.7	37.9	0	4.8	8.5	12	52	6	13	4	26	33	8	1010.9	1006.7	5	5	23.1	36.8	0

Figure 14. Portion of raw data of the Weather test data after data cleaning and transformation

5. Conclusion

For the community and development sectors, such as businesses and governments, it can be expensive and disruptive if there is no rain forecast or if it falls during regular activities. First, there are 42677 instances of Weather test data with 22 attributes and 99516 instances of Weather training data with 23 attributes in these two datasets. 22 attributes and 154314 instances will be used to build a predictive model after performing data preprocessing tasks. Additionally, 42677 instances and 21 attributes will be used to evaluate the model. Hopefully, this model can be used for a prediction of the likelihood of the next day's rain in everyday activities in Australia and the use of appropriate preventive measures.

References

[1] "WEATHER | English meaning - Cambridge Dictionary." Accessed: May 24, 2024. [Online]. Available: https://dictionary.cambridge.org/dictionary/english/weather

[2] S. Prakash, "IMPACT OF CLIMATE CHANGE ON AQUATIC ECOSYSTEM AND ITS BIODIVERSITY: AN OVERVIEW," *International Journal Biological Innovations*, vol. 03, no. 02, 2021, doi: 10.46505/ijbi.2021.3210.

[3] V. Werner De Vargas *et al.*, "Imbalanced data preprocessing techniques for machine learning: a systematic mapping study," *Knowl Inf Syst*, vol. 65, pp. 31–57, 2023, doi: 10.1007/s10115-022-01772-8.

[4] B. Bochenek and Z. Ustrnul, "Machine Learning in Weather Prediction and Climate Analyses—Applications and Perspectives," *Atmosphere (Basel)*, vol. 13, no. 2, Feb. 2022, doi: 10.3390/atmos13020180.

[5] A. U. Rahman *et al.*, "Rainfall Prediction System Using Machine Learning Fusion for Smart Cities," *Sensors*, vol. 22, no. 9, May 2022, doi: 10.3390/s22093504.

[6] D. Watson-Parris, "Machine learning for weather and climate are worlds apart," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 379, no. 2194. Royal Society Publishing, Apr. 05, 2021. doi: 10.1098/rsta.2020.0098.

[7] X. Ren *et al.*, "Deep Learning-Based Weather Prediction: A Survey," *Big Data Research*, vol. 23, p. 100178, Feb. 2021, doi: 10.1016/J.BDR.2020.100178.

[8] D. Markovics and M. J. Mayer, "Comparison of machine learning methods for photovoltaic power forecasting based on numerical weather prediction," *Renewable and Sustainable Energy Reviews*, vol. 161, p. 112364, Jun. 2022, doi: 10.1016/J.RSER.2022.112364.

[9] "Australia Weather Data." Accessed: Jun. 01, 2024. [Online]. Available: https://www.kaggle.com/datasets/arunavakrchakraborty/australia-weather-data?select=Weather+Training+Data.csv

[10] "Data Cleaning: Definition, Benefits, And How-To | Tableau." Accessed: Jun. 01, 2024. [Online]. Available: https://www.tableau.com/learn/articles/what-is-data-cleaning

[11] "What is Data Transformation? | TIBCO." Accessed: Jun. 01, 2024. [Online]. Available: https://www.tibco.com/glossary/what-is-data-transformation