**DATA ANALYSIS PROJECT**

**Unit Chair: Dr Musa Mammadov**

**Submission Date: 5:00PM Friday of Week 6**

**Table of Contents**

[**Section 1: Introduction and Data Description**](#_k1td6m53ikm0) **2**

[**Section 2: Exploratory Data Analysis and Results**](#_apfts8lpkmyd) **3**

[**Section 3: Conclusion**](#_iecfvi4z0ppz) **4**

[**Section 4: References**](#_3cm7bptggzp1) **5**

**Dataset Name**: Rain in Australia

**Group Name**: Mon-13 (FANH) **On Campus/Cloud**: On Campus

|  |  |  |
| --- | --- | --- |
| **STUDENT ID** | **STUDENT FULL NAME** | **Individual contribution\*** |
| 218401269 | ALEXANDER PAK YU LAI |  |
| 218241616 | HARRY WILLIAM LODGE |  |
| 218459058 | VIET NAM NGUYEN |  |
| 218271795 | JARROD KENG YEN YONG |  |

**\*** 5 - Contributed significantly, attended all meetings

4 - Partial contribution, attended all meetings

3 - Partial contribution, attended few meetings

2 - No contribution, attended few meetings

1 - No contribution, did not attended any meetings

|  |
| --- |
| **Section 1: Introduction and Data Description** |

The Australian weather broadcast or the appropriate agencies comprehends the tomorrow weather, specifically whether it will be raining or not, to deliver updated weather news to citizens or propose proper plans for social activities respectively. Therefore, the dataset of day-to-day weather was collected across various stations over a period of 10 years. The observations extracted from the Australian weather dataset aim to return prediction of tomorrow rain by producing the possibilities of it or two binary labels (Yes and No, 1 and 0 in that order).

The “Rain in Australia” dataset consists of 24 variables in total. Among these variables, there are two variable types, categorical variable and numerical variable.

For the categorical variable, five variables are normal, including “Date”, “Location”, “WindGustDir”, “WindDir9am” and “WindDir3pm” while the remaining two binary variables are “RainToday” and “RainTomorrow”, especially “RainTomorrow” is a target variable. Also, these categorical variables are defined with nominal type. In addition, the “WindDir9am” and “WindGustDir” variables experience the biggest percentages of missing value with around 7 percent (Young 2017). Furthermore, the variable having the highest cardinality, which means that a variable has the largest number of labels, is “Date” with 3436 labels and the second is “Location” with 49.

For the numerical variable, there are seventeen variables, which refer to continuous type, namely “MinTemp”, “MaxTemp”, “Rainfall”, “Evaporation”, “Sunshine”, “WindGustSpeed”, “WindSpeed9am”, “WindSpeed3pm”, “Humidity9am”, “Humidity3pm”, “Pressure9am”, “Pressure3pm”, “Cloud9am”, “Cloud3pm”, “Temp9am”, “Temp3pm” and “RISK\_MM”. According the figures in the table indicated the dataset provided by Young (2017), the variables witnessing the most missing values are “Evaporation” and “Sunshine” with roughly 43 and 48 percent respectively, followed by “Cloud9am” and “Cloud3pm” with nearly 38 and 40 percent in that order. Also, the “RISK\_MM” variable is in consideration for dropping out of the dataset. According to Young (2017), this variable should be eliminated if “RainTomorrow” is considered as a target and it aims to train a regression model instead of classification one because including this variable which indicates the further information of rain impacts negatively to the predicted values of the trained model or a lower accuracy score.

The observation about the “Rain in Australia” is that the method of training model to return a high accuracy score can be a regression algorithm instead of classification, but the null accuracy, where the accuracy is gained by preferring the most frequent value, should be compared with it to guarantee that the applied model results in the higher accuracy score. Another observation is that many pairs of variables would have strong or high correlation in the positive trend since they are intuitively correlated in terms of related fields of rain prediction.

Before the dataset is analyzed thoroughly and its comprehensive patterns are extracted, the dataset ensures to be evaluated carefully and cleaned to get rid of missing values or incorrect data by finding the frequency of null values for each variable, calculating z scores or exploring the inner problem of categorical and numerical variables. After finishing the process of data cleansing, the cleaned dataset should be tested, followed by discovering patterns which define relationships between variables due to their correlation coefficient. Eventually, the comprehensive analysis is finalized to sum up as well as conclude findings. Throughout the exploratory data analysis, the appropriate visualizations of data evaluation or results are illustrated.

|  |
| --- |
| **Section 2: Exploratory Data Analysis and Results** |

|  |
| --- |
| **Section 3: Conclusion** |

|  |
| --- |
| **Section 4: References** |

Bureau of Meteorology n.d., *Note to accompany Daily Weather Observations*, Australian Government, retrieved 18 April 2020, <<http://www.bom.gov.au/climate/dwo/IDCJDW0000.shtml>>.

Young, J 2017, *Rain in Australia*, Kaggle, retrieved 18 April 2020, <<https://www.kaggle.com/jsphyg/weather-dataset-rattle-package>>.