



TDS2101 – PROJECT

TRIMESTER 1, 2022/2023

Predicting Flood Disasters in Kelantan During the next 3 months Using Historical Meteorological Records with Classification.

Assignment: Project (40%)

FULL NAME	STUDENT ID
Sarah Shahmina Binti Abdul Ra'uf	1201102479

Table of Contents

Table of Contents.....	2
Chapter 1: Introduction.....	3
1.1 Background.....	3
1.2 Project Objectives.....	3
1.3 Project Scope.....	4
Chapter 2: Data Cleaning Process.....	4
2.1: Cleaning Flood Cases Dataset.....	4
2.2: Cleaning Historical Weather Datasets.....	5
2.3: Merging Flood Cases Data and Historical Weather Data.....	6
2.3.1:Resampling technique.....	6
2.3.1: Check for missing data.....	6
2.4: Describing the Dataset.....	7
Chapter 3: Exploratory Data Analysis, Outlier Detection and Feature Selection.....	8
3.1 Descriptive statistic.....	9
3.2 Outlier Detection through Boxplot and Domain Analysis.....	10
3.3 Distribution.....	12
3.3.1 Comparison of distribution.....	15
3.4 Feature Selection using Heat Map.....	17
3.5 Min-Max Normalisation.....	18
Chapter 4: Model Construction and Comparison.....	18
4.1 Normalised vs Unnormalised data.....	19
4.2 Hyperparameter Tuning using Grid Search.....	21
4.3 Building Model with Optimised Hyperparameter.....	24
4.3.1 Testing Accuracy on Built Models.....	24
Chapter 5: 3 Month Prediction Using Built Model (Running Unseen New Data).....	26
Conclusion:.....	27
References:.....	28

Chapter 1: Introduction

1.1 Background

The 2014-2015 Malaysian floods, which impacted the northern and east coast states of Malaysia, including Kelantan, have been labelled as some of the worst in decades. With over 60,000 people forcefully displaced from their homes due to the high water levels, a majority of whom were from Kelantan, and significant loss of lives, the devastating consequences of the floods were deeply felt in the region.

The harsh impacts of hydrological disasters such as flooding often leave behind devastating damages to properties, essential services, human lives and the environment. Due to its geological location, Kelantan is a region in Malaysia that is especially vulnerable to flooding disasters. With many of Kelantan's major towns being situated near the Kelantan River Basin, which is prone to overspill, flooding highlights a major cause of concern for the local economy and local population.

Such risk calls attention to the importance of early flood detection and warning. Delaying rescue and aid missions to areas affected by a disaster can cause an increase in mortality and prolonged suffering to victims.

1.2 Project Objectives

Findings from this study as well as the prediction model can assist in decision making that can help in swift and well ordered rescue and aid in disaster response and recovery efforts for future disasters in Kelantan.

The main question that we aim to answer from this study is: Can we predict flood occurrences for 3 months based on different weather conditions?

1.3 Project Scope

1.3.1 The dataset

The Malaysian flood cases data was sourced from Em-Dat public database. The period of the dataset entries ranged from 1970-2023, though for this project, only data from 2000-2023 will be used. It is important to note that for an event to be entered into the database, it must have caused the deaths of 10 or more people, have caused 100 or more people to be affected/injured or homeless, or declared as an emergency by the country.

The data on the historical weather in Kelantan was collected through Open Meteo Historical Weather API. Historical Weather data was collected for 10 districts in Kelantan, which include Bachok, Gua Musang, Jeli, Kota Bharu, Kuala Krai, Machang, Pasir Mas, Pasir Puteh, Tanah Merah and Tumpat. The data covers a time interval ranging from January 1, 2000, to April 4, 2023.

Chapter 2: Data Cleaning Process

2.1: Cleaning Flood Cases Dataset

In this dataset, we are only concerned with extracting the date and location information of floods that happened in Kelantan and its various districts and the dates. Firstly, I merged the 'Start Year', 'Start Month' and 'Start Day' into a single column called 'Start Date' and did the same thing with the columns concerning the end date of the flood. I formatted the 'Geo Locations' column from a string list of locations to individual locations in its own row. Finally I filtered the 'Geo Locations' column to only show locations that correspond to 'Kelantan' or districts inside Kelantan.

Figure 1: Original dataset

Figure 2: Flood Case Dataset after data wrangling

The first step performed for this stage of data wrangling was to change the date format for the time column, check for any null cells in any of the rows, check for any duplications of dates or rows and check for any missing dates in the dataset. Because the datasets have the same attributes, a function was created that performs the check on all the datasets through loops.

Upon performing this check, it was found that the dataset that was obtained from the Open Meteo API was already very clean. There were no null rows, no duplication of rows and no missing dates. The only task that needed to be performed was to change the format of the data in the “time” column from the ISO 8601 format "YYYY-MM-DDTHH:MM" to "DD/MM/YYYY HH:MM".

time	temperature	relative_humidity	pressure	surface_precipitation
2000-01-01T00:00	21.8	99	1009.8	995.2
2000-01-01T01:00	22.2	99	1010.7	996.1
2000-01-01T02:00	22.8	99	1011.4	996.8
2000-01-01T03:00	24.1	96	1011.3	996.8
2000-01-01T04:00	24.8	89	1011	996.5
2000-01-01T05:00	25	88	1010.4	995.9
2000-01-01T06:00	24.5	86	1009.5	995
2000-01-01T07:00	24.5	88	1007.5	993.1
2000-01-01T08:00	24.1	96	1007.8	993.3
2000-01-01T09:00	23.5	98	1007.4	992.9
2000-01-01T10:00	23.4	99	1007.8	993.3
2000-01-01T11:00	22.9	99	1008.8	994.3
2000-01-01T12:00	22.4	99	1009.1	994.5
2000-01-01T13:00	22	99	1009.9	995.3
2000-01-01T14:00	22.5	95	1010.8	996.2
2000-01-01T15:00	22.1	99	1011.4	996.8
2000-01-01T16:00	21.9	100	1011	996.4
2000-01-01T17:00	22	99	1011	996.4
2000-01-01T18:00	22	99	1010.4	995.8

→

time	temperature	relative_humidity	pressure	surface_precipitation
1/1/2000 0:00	21.8	99	1009.8	995.2
1/1/2000 1:00	22.2	99	1010.7	996.1
1/1/2000 2:00	22.8	99	1011.4	996.8
1/1/2000 3:00	24.1	96	1011.3	996.8
1/1/2000 4:00	24.8	89	1011	996.5
1/1/2000 5:00	25	88	1010.4	995.9
1/1/2000 6:00	24.5	86	1009.5	995
1/1/2000 7:00	24.5	88	1007.5	993.1
1/1/2000 8:00	24.1	96	1007.8	993.3
1/1/2000 9:00	23.5	98	1007.4	992.9
1/1/2000 10:00	23.4	99	1007.8	993.3
1/1/2000 11:00	22.9	99	1008.8	994.3
1/1/2000 12:00	22.4	99	1009.1	994.5
1/1/2000 13:00	22	99	1009.9	995.3
1/1/2000 14:00	22.5	95	1010.8	996.2
1/1/2000 15:00	22.1	99	1011.4	996.8
1/1/2000 16:00	21.9	100	1011	996.4
1/1/2000 17:00	22	99	1011	996.4
1/1/2000 18:00	22	99	1010.4	995.8

Figure 3: (Left) Original Time Column Format, (Right) Modified Time Column Format

2.3: Merging Flood Cases Data and Historical Weather Data

At this stage, we need to merge the two datasets together so that the resulting dataset has a column called “Flood occurrence” with boolean value “True” and “False” corresponding to an occurrence of a flood and no occurrence of flood, respectively, at that particular time and geographic location. It was at this stage that we realised we needed more variables and so daily weather variables ‘temperature_2m_max (°C)’, ‘temperature_2m_min (°C)’, ‘rain_sum (mm)’ and ‘precipitation_hours (h)’ were collected from the Open Meteo API and merged with the rest of the data.

2.3.1: Resampling technique

Due to the time column in the Historical Weather Datasets being of the format "DD/MM/YYYY HH:MM" and Flood Cases Dataset being of the format “DD/MM/YYYY”, I needed to downsample the time column in the Historical Weather Dataset to from hourly to daily data. This was done by averaging the hourly weather data for each day.

2.3.1: Check for missing data

To check that the merging was successful, I first checked to see how many entries in the merged dataset. There are 8492 dates Between 01/01/2000 and 01/04/2023 inclusive, thus there needs to be $8492 \times 10 = 84920$ entries in the dataset, which my dataset exhibits.

time	temperature_2m (°C)	relativehumidity_2m (%)	pressure_msl (hPa)	surface_pressure (hPa)	rain (mm)	cloudcover_low (%)	cloudcover_mid (%)	cloudcover_high (%)	windspeed_100m (km/h)	winddirection	vapor_pressure_deficit (hPa)	soil_moisture_10cm	soil_moisture_20cm	soil_moisture_40cm	Flood occurrence	Geo Loc	temperature_2m_max (°C)	temperature_2m_sum (mm)	precipitation_hours (h)			
1/1/2000	22.6916667	94.708	1009.679167	980.47	0.4458	99.208	88	54.292	90.417	6.57916667	193.54	0.15416667	23.1	24.163	0.4732	0.505	FALSE	Gua Mus	25.4	21.2	8.7	13
1/1/2000	22.85	96.542	1009.6375	995.08	1.2125	97.792	72.917	53.333	95.583	13.02916667	74.042	0.102083333	23.6	24.9	0.5135	0.5111	FALSE	Jeli	25	21.5	24.8	23
1/1/2000	25.67083333	87.375	1009.429167	1008.3	1.1625	87.25	49.5	36.625	95.667	20.87916667	76.375	0.425	26.2	27	0.2867	0.3	FALSE	Tumpat	27.3	24.2	23.3	24
1/1/2000	23.57916667	95.667	1009.679167	1007	1.8333	94.75	125	55.083	94.375	12.1125	56.292	0.132916667	24.7	25.9	0.5135	0.5126	FALSE	Tanah M	25.5	22	37.9	23
1/1/2000	24.65683333	89.625	1009.6	1009.6	1.1917	92.5	62.833	45.252	77.25	16.8375	77.375	0.334166667	25.8	27	0.4976	0.502	FALSE	Backoh	26.4	23.4	28.6	24
1/1/2000	24.1625	93.042	1009.620833	1009.3	1.8917	94.75	69.333	47.667	93.417	18.1125	63.917	0.21875	25.3	26.5	0.5076	0.5166	FALSE	Pasir Put	26.1	22.5	42.4	24
1/1/2000	24.38333333	91	1009.445833	1009.4	1.35	92.833	64.125	44.375	94.583	18.52916667	70.375	0.284166667	25.4	26.6	0.5024	0.503	FALSE	Pasir Mai	26.1	22.6	28.6	24
1/1/2000	23.24583333	95.917	1009.679167	1000.3	1.6333	94.75	80.125	55.083	94.375	12.1125	56.292	0.124583333	24	25.2	0.4107	0.436	FALSE	Kuala Kri	25.9	21.9	37.9	23
1/1/2000	23.47916667	95.667	1009.679167	1007	1.8333	94.75	80.125	55.083	94.375	12.1125	56.292	0.131666667	24.7	25.9	0.5135	0.5126	FALSE	Madhang	25.4	22	37.9	23
1/1/2000	24.63333333	90.5	1009.445833	1006.6	1.35	92.833	64.125	44.375	94.583	18.52916667	70.375	0.305	25.5	26.7	0.506	0.5039	FALSE	Kota Bha	26.3	23	28.6	24
2/1/2000	24.1125	93.417	1009.291667	1008	0.9542	96.458	67.958	36.5	90.75	16.34583333	84.417	0.20875	25.3	26.458	0.4817	0.5178	FALSE	Pasir Put	25.9	23.1	23.2	24
2/1/2000	24.2125	92.292	1009.079167	1009.1	1.1333	93	62.917	35.875	89.833	17.23333333	90.458	0.240416667	25.4	26.6	0.513	0.5068	FALSE	Pasir Mai	26.2	23.1	30.3	24
2/1/2000	23.7375	94.375	1009.2875	1006.6	0.8708	99.25	81.917	41.75	92.542	10.425	80.542	0.172083333	24.7	25.9	0.4842	0.5174	FALSE	Tanah M	26.1	22.6	20.3	22
2/1/2000	25.35416667	89.542	1009.070833	1007.5	1.1292	89.292	56.25	33	96.708	20.20833333	94.875	0.34	26.2	27	0.2969	0.3066	FALSE	Tumpat	26.7	24.5	29.8	24
2/1/2000	24.44583333	92.208	1009.079167	1009.3	1.1333	93	62.917	35.875	89.833	17.23333333	90.458	0.246666667	25.5	26.7	0.5135	0.509	FALSE	Kota Bha	26.3	23.4	30.3	24
2/1/2000	24.59166667	91.708	1009.183333	1009.2	1.2625	93.708	69.667	37.208	87.625	17.65	79.625	0.264166667	25.8	27	0.514	0.5053	FALSE	Backoh	26.3	23.6	30.3	24
2/1/2000	23.57083333	96.875	1009.2875	999.93	0.8708	99.25	81.917	41.75	92.542	10.425	80.542	0.134583333	24.054	25.167	0.4246	0.436	FALSE	Kuala Kri	26.2	22.2	20.3	22
2/1/2000	22.65416667	95.667	1009.333333	980.13	0.6667	97.167	76.833	42.667	97.958	6.945833333	130.71	0.136833333	23.1	24.1	0.4811	0.5042	FALSE	Gua Mus	25.8	21.5	15.7	22
2/1/2000	23.02916667	96.25	1009.25	994.7	1.0375	87.542	77.375	41.667	89.542	12.40833333	94.042	0.117083333	23.6	24.821	0.5068	0.514	FALSE	Jeli	25.1	21.9	28.5	24
2/1/2000	24.12083333	93.458	1008.941667	1008.9	0.75	93.417	67.917	31.375	96	15.69583333	89.667	0.2025	25.4	26.538	0.5134	0.5125	FALSE	Pasir Mai	26	23	18.7	24
3/1/2000	24.00833333	94.667	1009.104167	1007.8	1.0208	94.375	60.75	35.708	98.292	16.425	81.625	0.166666667	25.3	26.4	0.486	0.5165	FALSE	Pasir Put	25.1	22.9	25	24
3/1/2000	23.57916667	95.667	1009.125	1006.4	0.9	97.042	71.417	41.333	97.917	9.3625	73.25	0.1325	24.7	25.821	0.4957	0.5164	FALSE	Tanah M	25.8	22.3	20	23
3/1/2000	24.36666667	93.333	1008.941667	1006.1	0.75	93.417	67.917	31.375	96	15.69583333	89.667	0.206666667	25.5	26.7	0.5098	0.5143	FALSE	Kota Bha	26	23.3	18.7	24
3/1/2000	22.8875	93.542	1009.0875	979.8	0.4708	97.083	70.333	37.5	98.667	6.391666667	105.5	0.201666667	23.1	24.1	0.4851	0.504	FALSE	Gua Mus	27	21	13	19
3/1/2000	23.45416667	94.875	1009.125	999.76	0.9	97.042	71.417	41.333	97.917	9.3625	73.25	0.160416667	24.1	25.1	0.4291	0.4365	FALSE	Kuala Kri	26.1	22.2	20	23
3/1/2000	24.4375	92.958	1009	1009	0.7792	94	66.25	31.833	98.542	15.62083333	92.458	0.217916667	25.8	26.917	0.5113	0.5109	FALSE	Backoh	25.7	23.6	18.7	24
3/1/2000	23.47916667	95.667	1009.125	1006.4	0.9	97.042	71.417	41.333	97.917	9.3625	73.25	0.131666667	24.7	25.821	0.4957	0.5164	FALSE	Madhang	25.8	22.2	20	23
3/1/2000	22.92083333	96.208	1009.104167	994.55	0.8708	97.875	77.125	34.792	92.375	11.325	93.667	0.112083333	23.638	24.8	0.5128	0.5146	FALSE	Jeli	25.2	22	20.3	22
3/1/2000	25.38333333	90.208	1009.933333	1007.7	0.7458	90.792	62.875	33.292	89.625	18.7375	96.375	0.320833333	26.2	27.008	0.3019	0.3016	FALSE	Tumpat	26.5	24.4	20.2	24
4/1/2000	23.13333333	94.292	1009.070833	980.5	0.4625	96.292	69.042	39.5	92.958	5.945833333	91.5	0.187083333	23.1	24.1	0.4908	0.504	FALSE	Gua Mus	27.1	20.8	10.3	17
4/1/2000	23.6625	94.792	1009.0625	1007	0.8458	97.292	74.125	37.042	92.667	8.316666667	83.958	0.164166667	24.7	25.8	0.4859	0.5149	FALSE	Tanah M	26	22.2	20.3	24
4/1/2000	24.57916667	92.167	1009.120833	1009.1	0.7083	92.125	60.083	33.917	95.833	16.52916667	94.125	0.247083333	25.8	26.9	0.5123	0.5153	FALSE	Backoh	26.2	23.7	17	24
4/1/2000	23.5625	92.833	1009.470833	1006.7	0.8292	93.917	63.5	33.5	92.083	15.16666667	102.29	0.227083333	25.5	26.683	0.5047	0.515	FALSE	Kota Bha	26.3	23.5	17	24
4/1/2000	24.09166667	93.667	1009.658333	1008.4	0.7208	94.75	60.458	32.875	94.708	14.5125	91.542	0.196666667	25.3	26.4	0.4873	0.5157	FALSE	Pasir Put	25.8	23.1	21.8	24
4/1/2000	23.13666667	94.458	1009.0625	1000.3	0.6458	97.292	74.125	37.042	92.667	8.316666667	83.958	0.174583333	24.1	25.1	0.4309	0.437	FALSE	Kuala Kri	27.1	21.8	20.3	24
4/1/2000	22.98333333	96.417	1009.025	995.06	0.5708	96.833	72.917	38.375	86.708	10.525	110.88	0.110833333	23.688	24.8	0.4848	0.5173	FALSE	Jeli	25.6	21.8	16.5	23
4/1/2000	23.5625	94.792	1009.0625	1007	0.8458	97.292	74.125	37.042	92.667	8.316666667	83.958	0.1625	24.7	25.8	0.4859	0.5149	FALSE	Madhang	25.9	22.1	20.3	24
4/1/2000	25.35	90.25	1009.45	1008.3	0.6083	91.125	66.167	30.792	85.958	17.475	109	0.323333333	26.258	27.004	0.306	0.3026	FALSE	Tumpat	26.8	24.5	15.9	24
4/1/2000	24.10833333	93.083	1009.470833	1009.5	0.8292	93.917	63.5	33.5	92.083	15.16666667	102.29	0.219583333	25.4	26.5	0.5056	0.515	FALSE	Pasir Mai	26.2	23.1	17	24
5/1/2000	23.775	94.333	1010.279167	1009	0.525	83.042	45.708	40.458	80.167	11.20416667	108.04	0.167916667	25.3	26.4	0.4923	0.515	FALSE	Pasir Put	26.1	22.8	13.7	21

Figure 4: The final dataset with a flood occurrence column

The resulting dataset shows daily average weather parameters for all 12 districts in Kelantan as well as a response variable called “Flood occurrence” for our binary classes.

2.4: Describing the Dataset

Variable	Continuous/Di crete	Comments
time	Continuous	
temperature_2m (°C)	Continuous	Average daily air temperature
relativehumidity_2m (%)	Continuous	Average relative humidity
pressure_msl (hPa) surface_pressure (hPa)	Continuous	Atmospheric air pressure at mean sea level or pressure at surface.
rain (mm)	Continuous	Sum of rainfall in the preceding hour
cloudcover (%) cloudcover_low (%) cloudcover_mid (%) cloudcover_high (%)	Continuous	Percentage of clouds covering the sky at different altitudes
windspeed_100m (km/h)	Continuous	Wind speed at 100 metres above ground.

winddirection_100m (°)	Continuous	Wind direction at 100 meters above ground.
vapor_pressure_deficit (kPa)	Continuous	The moisture difference between air and its saturation point.
soil_temperature_28_to_100cm (°C) soil_temperature_100_to_255cm (°C)	Continuous	Average temperature of different soil levels below ground.
soil_moisture_28_to_100cm (m ³ /m ³) soil_moisture_100_to_255cm (m ³ /m ³)	Continuous	Average soil water content
temperature_2m_max (°C) temperature_2m_min (°C)	Continuous	Maximum and minimum daily air temperature
rain_sum (mm)	Continuous	Sum of daily rainfall
precipitation_hours (h)	Continuous	The number of hours with rain
Geo Locations	Discrete	Location of the current entry
Flood occurrence	Discrete	Response column FALSE - No floods TRUE - Flood occurred

Chapter 3: Exploratory Data Analysis, Outlier Detection and Feature Selection

3.1 Descriptive statistic

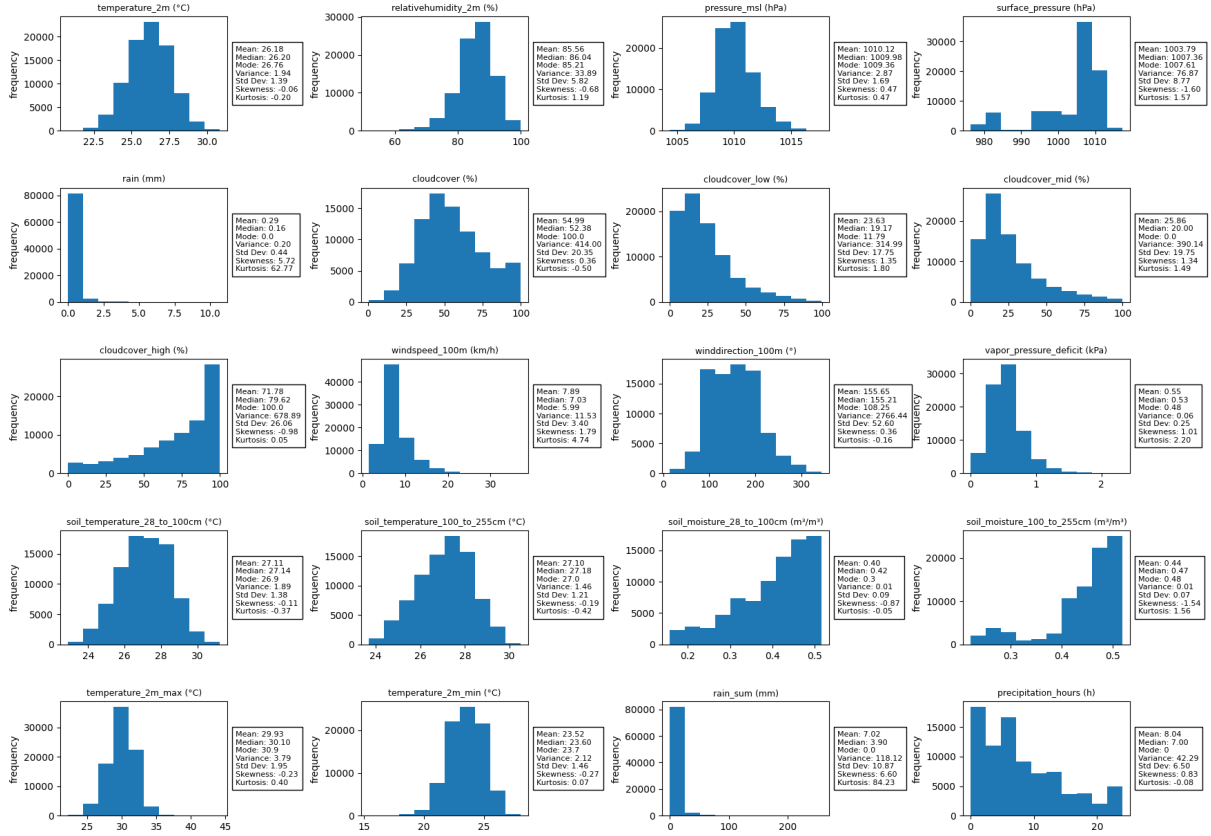


Figure 5: Histogram of Features

By plotting a histogram for all the continuous variables, we are able to gain some insights into our dataset. Some interesting observations to point out was the very high kurtosis value for rain (mm) and rain_sum(mm). A high kurtosis value suggests that these two variables exhibit a high amount of outliers that are far from the mean value. Other variables with relatively high kurtosis include windspeed_100m (km/h) and vapor_pressure_deficit (kPa).

Other observations found was that many of the variables in the dataset exhibit some degree of skewness. This is important to note as some machine learning models may be affected by skewed features (Jermain, 2019).

3.2 Outlier Detection through Boxplot and Domain Analysis

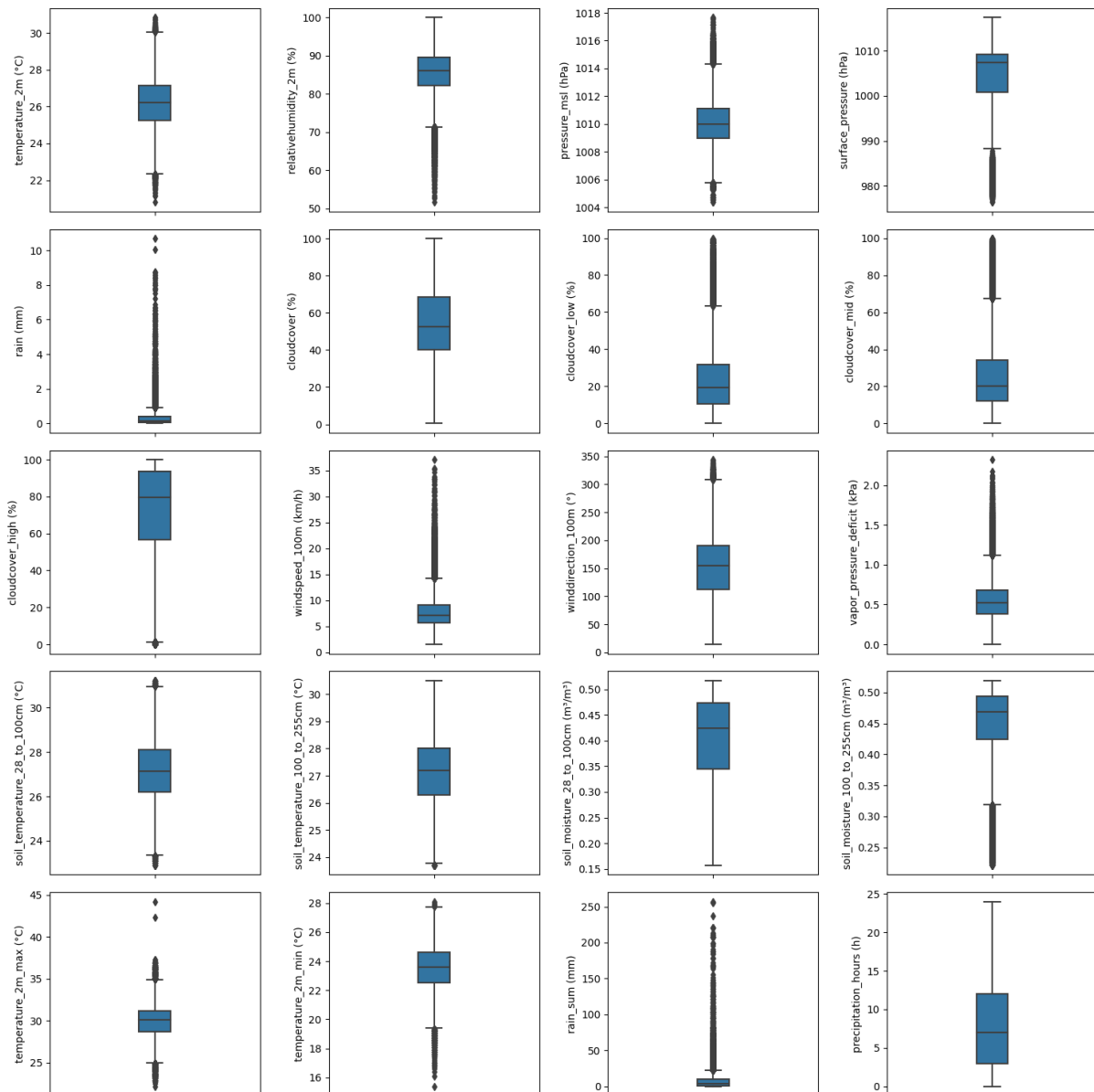


Figure 6: Box plots before the removal of outliers

index	temperature_2m_max (°C)	temperature_2m_min (°C)	rain (mm)	rain_sum (mm)
count	84920	84920	84920	84920
mean	29.92872939	23.5154357	0.2923237066	7.016131653
std	1.946052401	1.455932553	0.442270525	10.86846732
min	22.2	15.4	0	0
25%	28.7	22.5	0.0375	0.9
50%	30.1	23.6	0.1583333333	3.9
75%	31.2	24.6	0.3916666667	9.5
max	44.2	28.1	10.68333333	257.1

Figure 7: Descriptive statistics before removal of outliers

From the box plots, we are able to see that temperature_2m_max shows two significant outliers. Upon further analysis through finding the maximum value for this variable, these two outliers have values of over 40°C, which is higher than the highest recorded temperature in Malaysia (The Sun Daily, 2021). The two outliers were subsequently removed.

Similarly with temperature_2m_min, an outlier with a value lower than the lowest recorded temperature in Kelantan (15.7°C) was removed from the dataset (The Strait Times, 2018).

The column "rain (mm)" exhibits two significant outliers with values greater than 10. Upon analysing the rows associated with these outliers, it was observed that they represented isolated events that did not correspond to any flooding incidents. Considering that high rainfall typically correlates with an increased likelihood of rain, we made the decision to remove these outliers from the dataset.

It is important to note that not all outliers located in the whiskers were removed from the dataset. This decision was made considering that some outliers may represent the natural variations of the weather variables, including cases that correspond to actual occurrences of floods, such as in the case for rain_sum(mm) where the an outlier of the value of (256.0) corresponded to an occurrence of a flood.

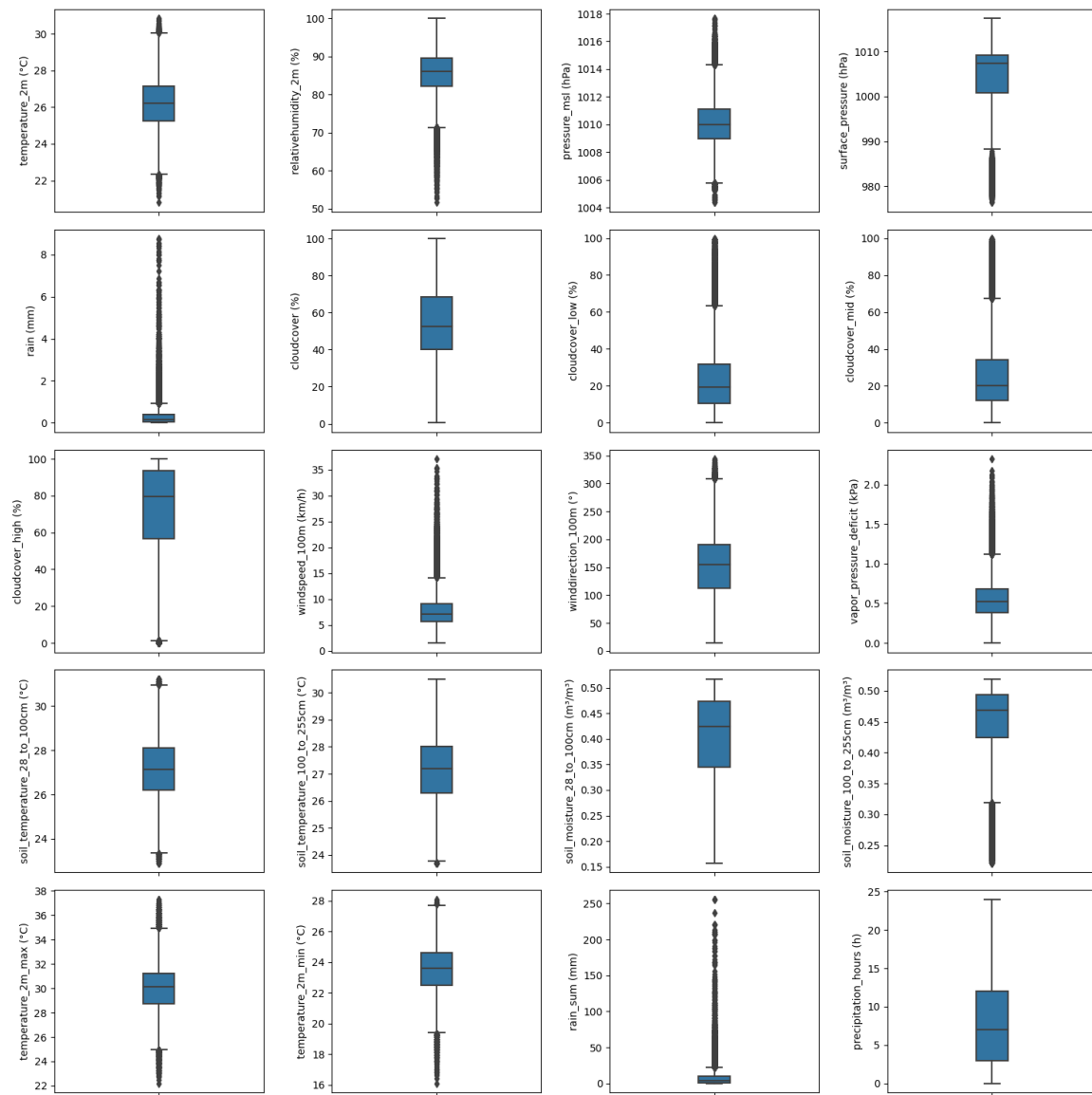


Figure 8: Box plots after the removal of outliers

index	temperature_2m_max (°C)	temperature_2m_min (°C)	rain (mm)	rain_sum (mm)
count	84915	84915	84915	84915
mean	29.92855581	23.51557793	0.2920750162	7.010478714
std	1.944832134	1.455686737	0.4395558751	10.80986414
min	22.2	18.1	0	0
25%	28.7	22.5	0.0375	0.9
50%	30.1	23.6	0.1583333333	3.9
75%	31.2	24.6	0.3916666667	9.5
max	37.3	28.1	8.7625	256

Figure 9: Descriptive statistics after the removal of outliers

3.3 Distribution

Next, we check and compare the histograms based on the flood occurrences. Upon creating the histogram, we noticed that there was an imbalance between the “true” class and the “false” class. There were significantly less occurrences of flooding cases (2606 occurrences) compared to occurrences of no flooding (82314 occurrences). To address this, data imbalance treatment was performed where we used undersampling to reduce the number of occurrences of no flooding to 2606. The figure below shows the histogram after performing data imbalance treatment.

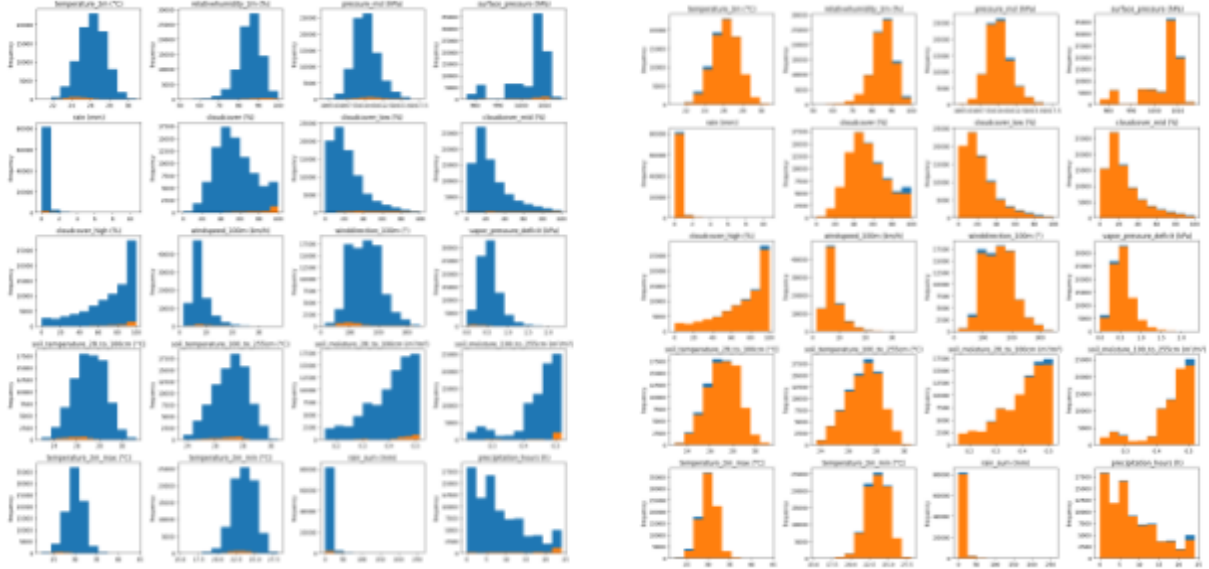


Figure 10: Histogram distribution of true observations (left) and false observations (right) before resampling

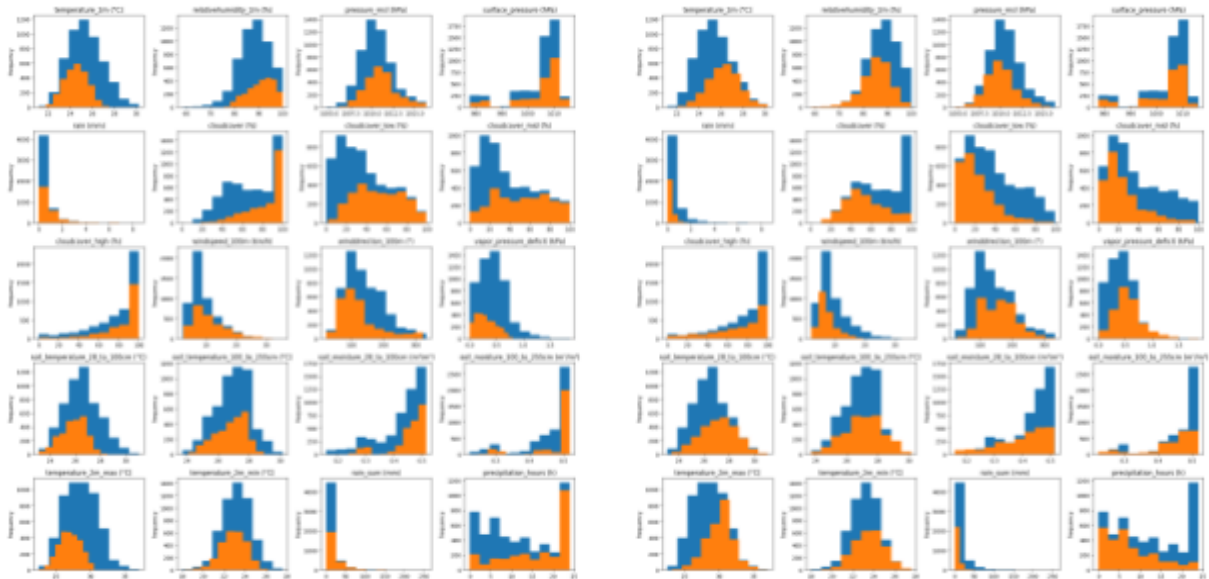


Figure 11: Histogram distribution of true observations (left) and false observations (right) after resampling

3.3.1 Comparison of distribution

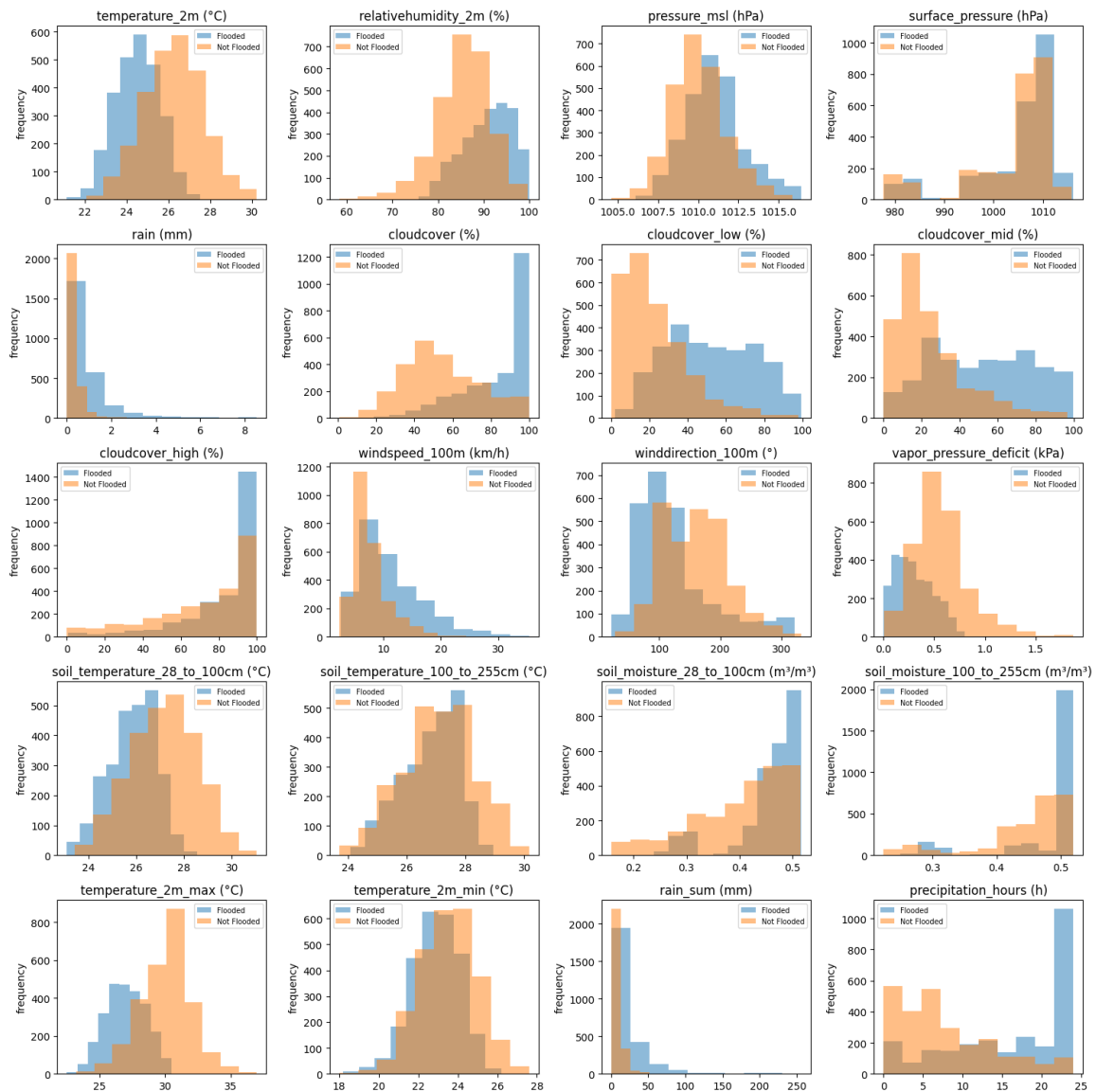


Figure 12: Histogram distribution comparison of true observations (blue) and false observations (orange)

The findings from the comparison of the histograms can be summarised as follows:

- Lower temperatures are more likely to be associated with flood cases.
- Higher levels of relative humidity tend to correlate with flood cases.
- There seems to be minimal correlation between sea level pressure or surface level pressure and the occurrence of floods.
- Flood cases are more common during periods of higher rainfall, indicating a positive correlation between rainfall levels and flooding.

- A higher percentage of cloud cover is observed during flood events, suggesting that increased rain clouds contribute to the occurrence of floods.
- Flood cases tend to coincide with higher wind speeds, indicating that stronger winds play a role in the occurrence of floods.
- Flood cases are more likely to occur when the wind direction degree is lower or is oriented towards the northeast, corresponding to the North East Monsoon (Satari et al., 2015).
- Lower values of vapour pressure deficit are associated with a higher incidence of floods.
- Flood cases tend to be associated with lower soil temperatures at depths ranging from 28-100 cm below the ground, while there appears to be little correlation with soil temperatures at depths of 100-255 cm.
- Floods are more likely to occur when soil moisture levels are higher.
- Lower maximum temperatures are typically recorded during flood events, whereas there seems to be little correlation between minimum recorded temperatures and floods.
- Longer durations of rainfall are associated with a higher number of flood cases, indicating a positive relationship between rainfall duration and flooding incidents.

3.4 Feature Selection using Heat Map

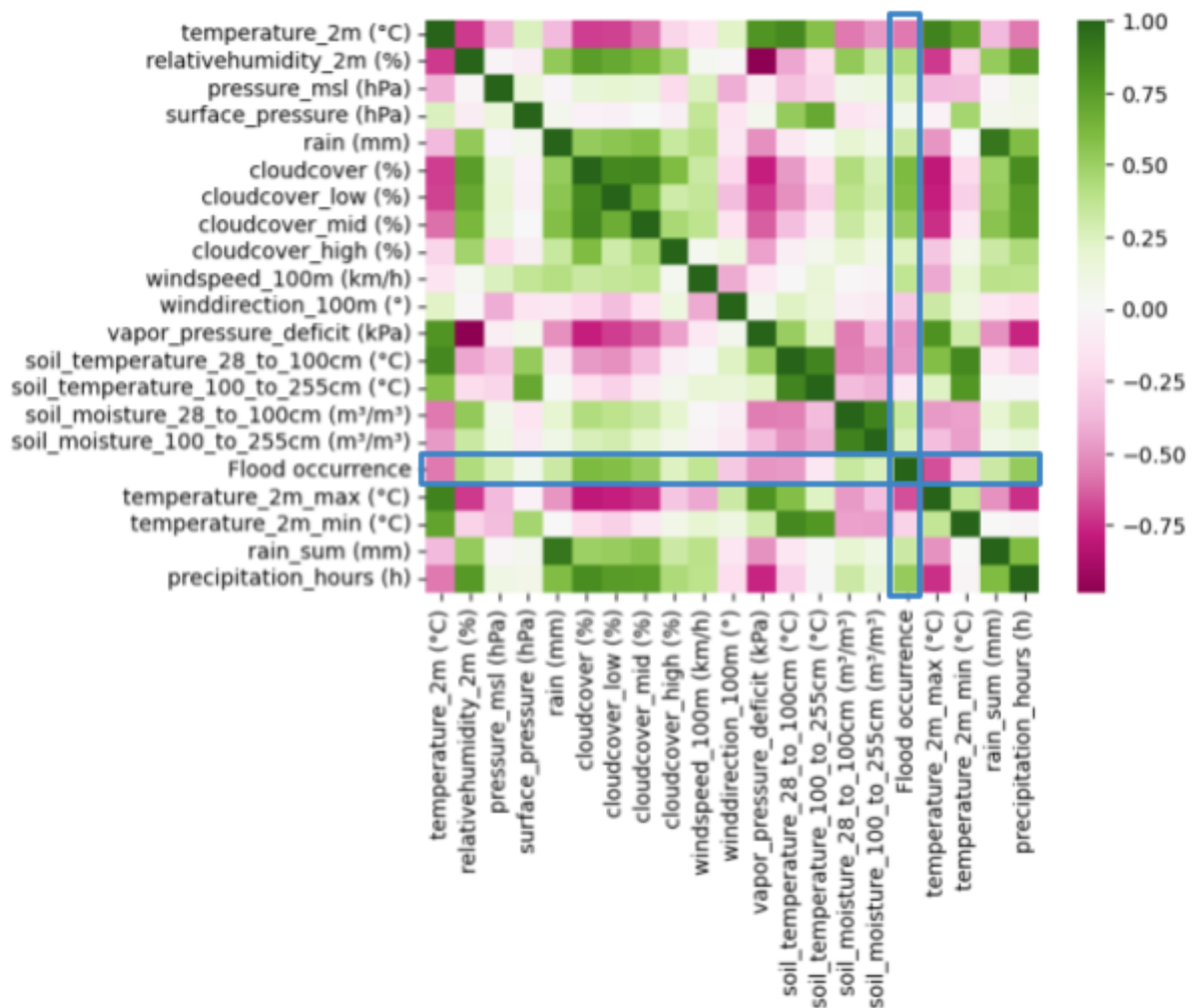


Figure 13: Heatmap with boxed variable being our target variable

To increase the accuracy of our machine learning models and prevent overfitting, we need to perform feature selection. To do so, a heatmap was created to identify variables that exhibit stronger correlations, as indicated by darker colours. By selecting these variables, we aim to focus on the most influential features for our models.

The selected features are:

- temperature_2m (°C)
- temperature_2m_max (°C)
- precipitation_hours (h)
- rain_sum (mm)
- cloudcover (%)
- relativehumidity_2m (%)

3.5 Min-Max Normalisation

	time	Flood occurrence	Geo Location	Normalized_temperature_2m (°C)	Normalized_relativehumidity_2m (%)	Normalized_cloudcover (%)	Normalized_temperature_2m_max (°C)	Normalized_rain_mm (mm)	Normalized_precipitation_hours (h)
0	2017-09-01	True	Tanah Merah	0.461292	0.584	0.541876	0.489655	0.000000	0.000000
1	2004-12-15	True	Pasei Mas	0.371049	0.575	0.478224	0.379319	0.000000	0.000000
2	2022-12-20	True	Jati	0.190519	0.957	0.867755	0.124138	0.168531	1.000000
3	2006-12-20	True	Mechang	0.354100	0.643	0.754606	0.344828	0.001172	0.003333

Figure 13: Image of data after normalisation

Finally, min-max normalisation is performed on our chosen features. The accuracy score of our normalised data will be compared to the accuracy score of our unnormalized data when comparing the algorithms we will use.

Chapter 4: Model Construction and Comparison

The three algorithms that were chosen for analysis and comparison are Logistic Regression, Decision Tree Classifier and K Nearest Neighbour.

Logistic Regression is amongst the chosen algorithms as it is widely used for binary classification problems. In our case, where we aim to predict whether a flood will occur (the "true" class) or not (the "false" class), logistic regression is a suitable choice. A potential disadvantage to using this model is that it is sensitive to skewed data (Jermain, 2019).

Decision Tree Classifier is a very popular machine learning model that we have also chosen for analysis and comparison. The reason that we have chosen the Decision Tree Classifier model is that it can be used for classification problems. It can also capture non-linear relationships and is resistant to outliers in our data that we have failed to capture.

Lastly, K Nearest Neighbor (KNN) was selected because it is a non-parametric algorithm that performs effectively with skewed distributions, which is a characteristic that was found with a lot of our variables.

4.1 Normalised vs Unnormalised data

To test out how normalised or unnormalised data affects the predictions, we run both the normalised data set and unnormalised data set on all three algorithms. Findings from this experiment conclude that the Logistic Regression model works best with normalised data as the accuracy score for the normalised dataset is higher compared to the unnormalised dataset.

It was found that with the Decision Tree Classifier, the unnormalised dataset produced a slightly higher accuracy score compared to the normalised dataset. Similarly with kNN, the unnormalised dataset produced a substantially higher accuracy score compared to the normalised dataset. This phenomenon could be attributed to normalised data removing or distorting some hidden relationship or pattern. Another explanation could be that the model has been overfitted on the dataset, as indicated by the accuracy scores of close to 90% or more, and thus may not produce accurate results on a different dataset.

Due to this uncertainty, we have decided to proceed with the normalised dataset for further experimentation to ensure that our model has better generalisation and to avoid potential issues resulting from overfitting.

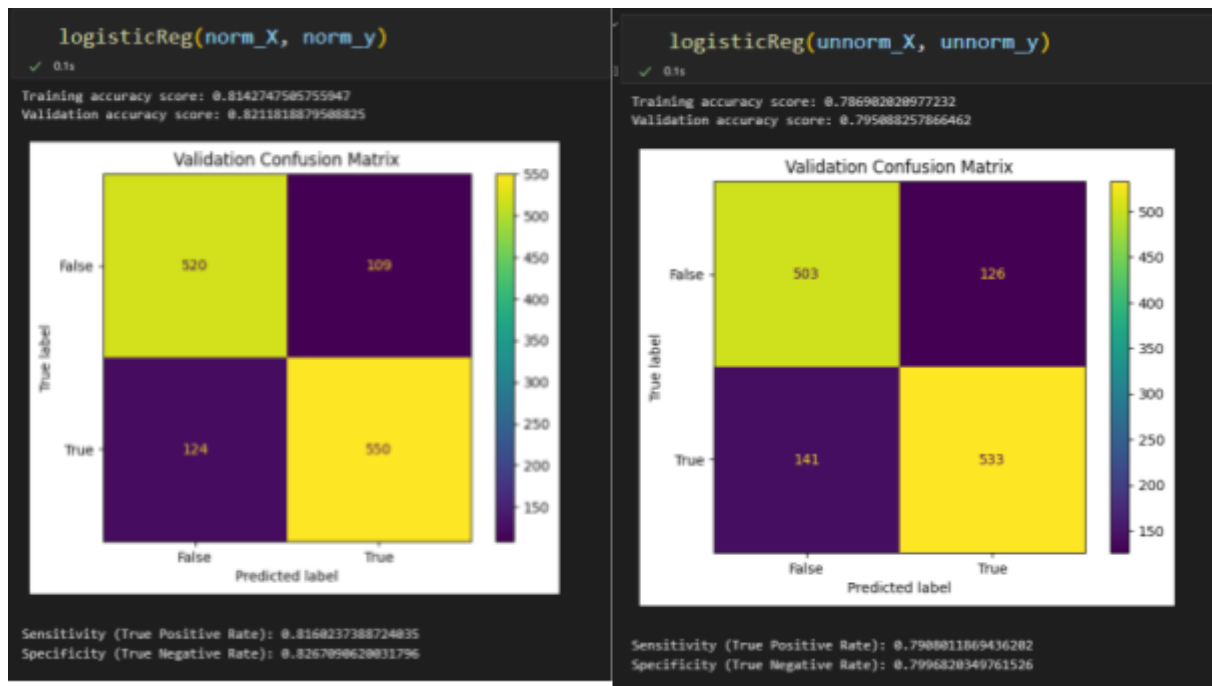


Figure 14: Logistic Regression: Normalised Data (left) vs Unnormalised Data (right)

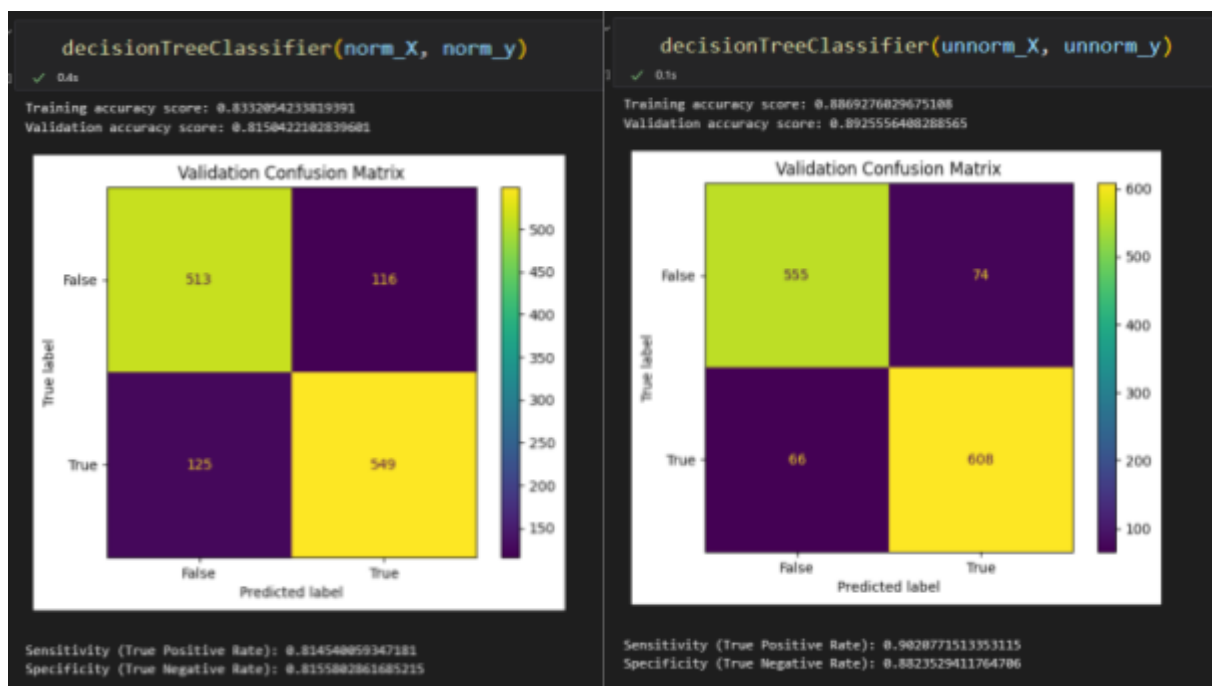


Figure 15: Decision Tree Classifier: Normalised Data (left) vs Unnormalised Data (right)

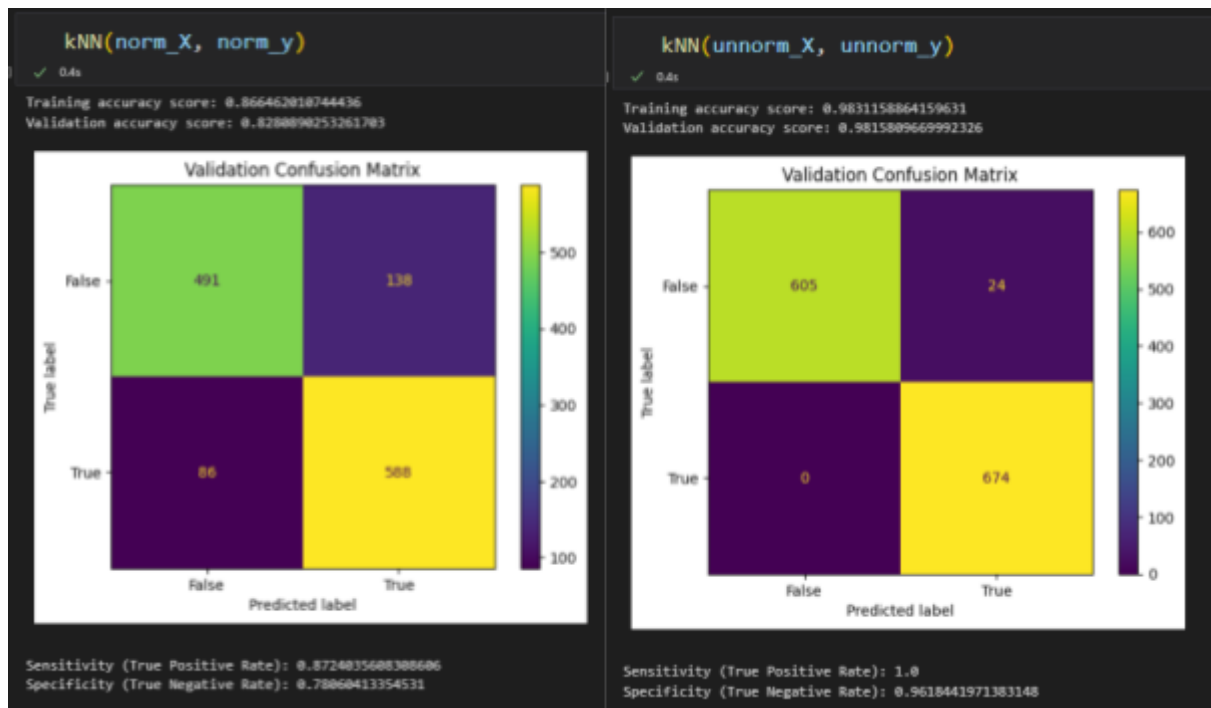


Figure 16: kNN: Normalised Data (left) vs Unnormalised Data (right)

4.2 Hyperparameter Tuning using Grid Search

The next step in building our model is to discover the hyperparameters that can produce the best accuracy score. To do so, we use the Grid Search technique, which evaluates a model's performance based on a combination of predefined parameters. The predefined parameters for each model can be seen in the figures below.

```

def logisticRegressionGridSearch():
    model = LogisticRegression()
    parameters = {
        'solver': ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'],
        'penalty': ['l1', 'l2'],
        'C': [0.1, 1, 10, 100]
    }

    gridSearch = GridSearchCV(
        estimator=model,
        param_grid=parameters,
        cv=10,
        n_jobs=-1
    )

    gridSearch.fit(X_train, y_train)
    results = pd.DataFrame(gridSearch.cv_results_)
    return results

```

Figure 17: Logistic Regression

```

def decisionTreeGridSearch():
    model = DecisionTreeClassifier()
    parameters = {
        'criterion': ['gini', 'entropy'],
        'splitter': ['best', 'random'],
        'max_depth': [1, 10, 100, 1000]
    }

    gridSearch = GridSearchCV(
        estimator=model,
        param_grid=parameters,
        cv=10,
        n_jobs=-1
    )

    gridSearch.fit(X_train, y_train)
    results = pd.DataFrame(gridSearch.cv_results_)
    return results

```

Figure 18: Decision Tree Classifier

```

def knnGridSearch():
    model = KNeighborsClassifier()
    parameters = {
        'n_neighbors': [3, 5, 7],
        'weights': ['uniform', 'distance'],
        'p': [1, 2]
    }

    gridSearch = GridSearchCV(
        estimator=model,
        param_grid=parameters,
        cv=10,
        n_jobs=-1
    )

    gridSearch.fit(XX_train, yy_train)
    results = pd.DataFrame(gridSearch.cv_results_)
    return results

```

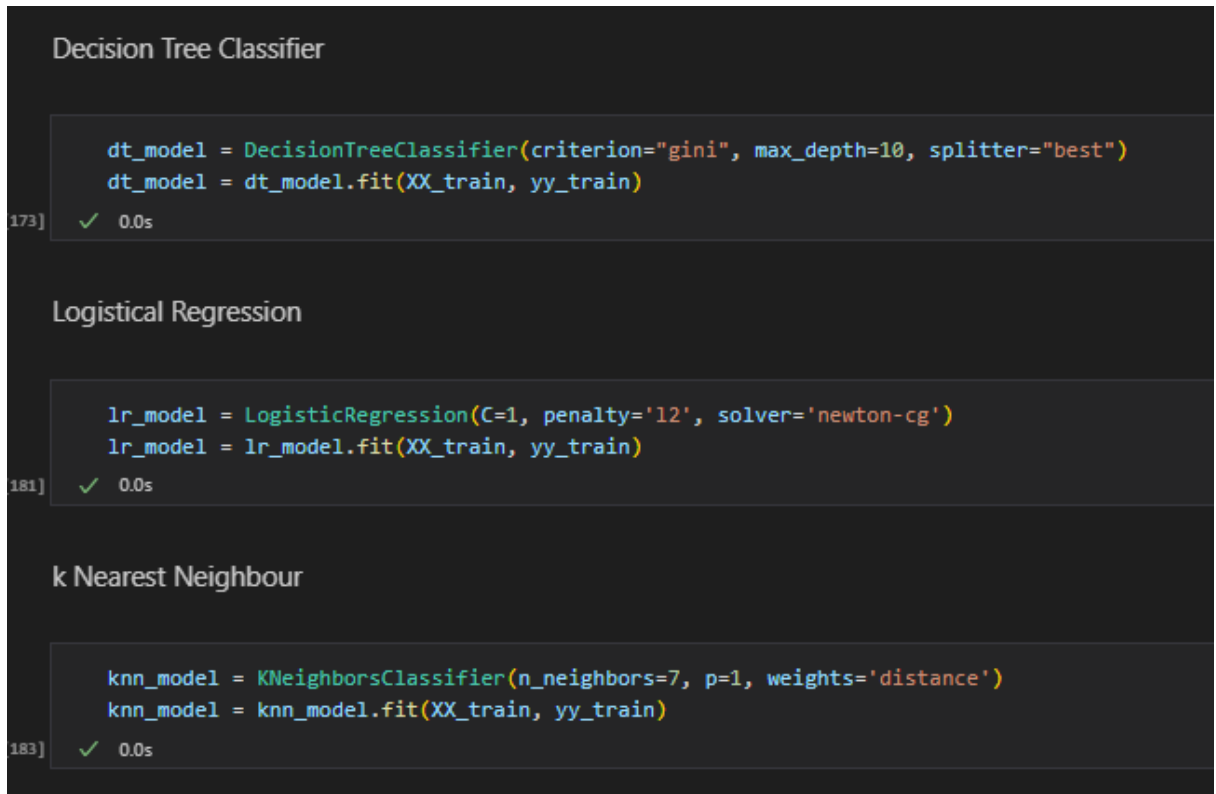
Figure 19: kNN

To find the most optimal parameters, we need to search for the set of parameters with the lowest “rank_test_score” and the highest “mean_test_score”. From these two criterias, the optimal parameters can be summarised in the following table:

	Logistic Regression	Decision Tree Classification	K Nearest Neighbour
Optimal Parameters	“C” : 1 “penalty”: 12 “solver”: newton-cg	“Criterion” : gini “max_depth”: 10 “splitter”: best	“n_neighbors”: 7 “p”: 1 “weights”: distance
Mean Test Score	0.811328	0.808929	0.846416
Rank Test Score	1	1	1

4.3 Building Model with Optimised Hyperparameter

We will then use the optimal parameters to build the model. Because the training data was used to fit the Grid Search, the validation and test data is needed to check the accuracy of the model after using the optimised parameters to avoid overfitting.



```
Decision Tree Classifier

dt_model = DecisionTreeClassifier(criterion="gini", max_depth=10, splitter="best")
dt_model = dt_model.fit(XX_train, yy_train)
173] ✓ 0.0s

Logistical Regression

lr_model = LogisticRegression(C=1, penalty='l2', solver='newton-cg')
lr_model = lr_model.fit(XX_train, yy_train)
181] ✓ 0.0s

k Nearest Neighbour

knn_model = KNeighborsClassifier(n_neighbors=7, p=1, weights='distance')
knn_model = knn_model.fit(XX_train, yy_train)
183] ✓ 0.0s
```

Figure 20: Building the models with using the optimal parameters

4.3.1 Testing Accuracy on Built Models

Based on the comparison of the accuracies of each model, the model that will be chosen is the logistic regression model as it demonstrates consistent performance when predicting both true and false cases (due to sensitivity and specificity being around the same value). Another reason logistic regression was chosen is the difference between training accuracy and validation and test accuracy being very minimal, indicating that the model is not overfitted.

	Decision Tree Classification	Logistic Regression Model	K Nearest Neighbour
--	---------------------------------	------------------------------	------------------------

	Model		
Training accuracy	0.91605	0.80962	1.0
Validation accuracy	0.80276	0.82425	0.835
Validation Sensitivity/Specificity	0.81751/ 0.78696	0.81899/0.82988	0.88576/0.78060
Test accuracy	0.79447	0.82413	0.83538
Test Sensitivity/Specificity	0.83910/ 0.74948	0.83503/0.81314	0.93075/0.73922

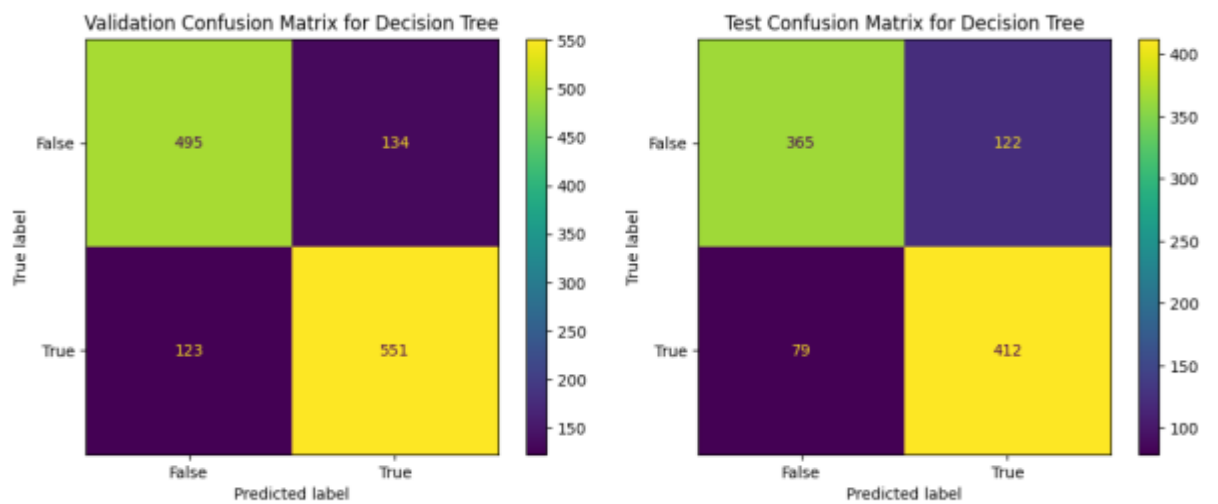


Figure 21: Confusion Matrix for Decision Tree Classifier

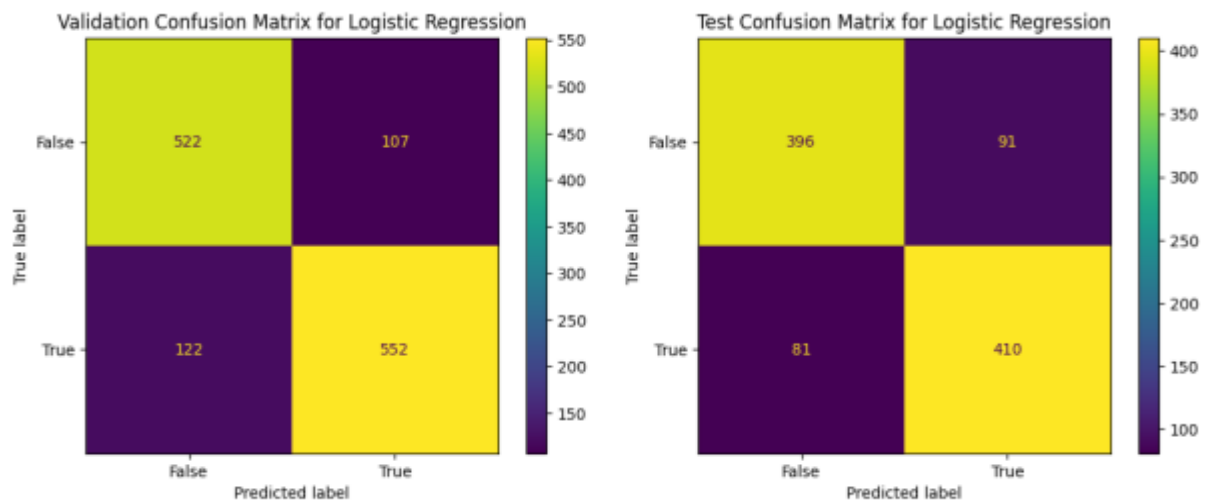


Figure 22: Confusion Matrix for Logistic Regression

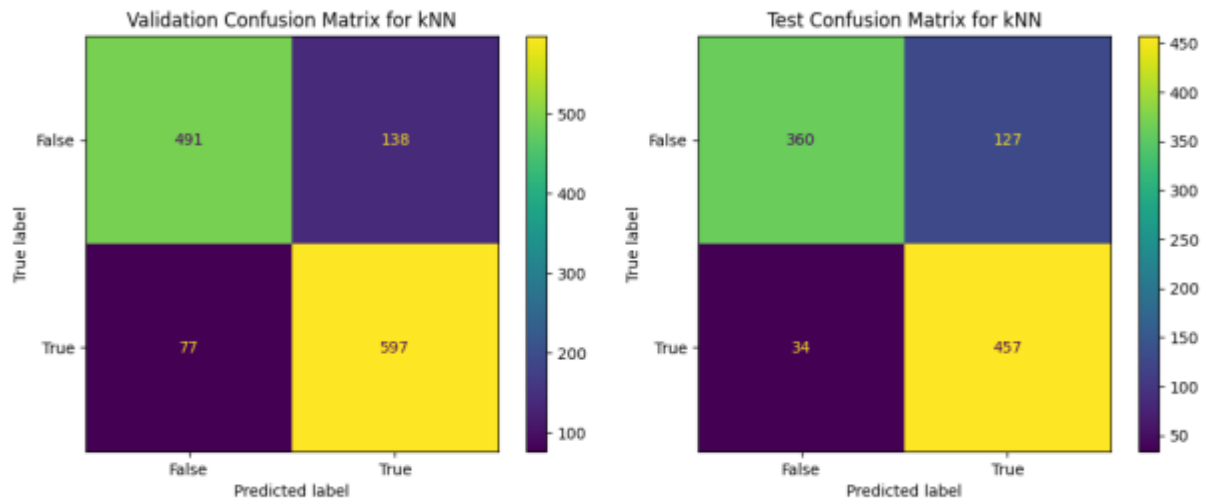


Figure 23: Confusion Matrix for kNN

Chapter 5: 3 Month Prediction Using Built Model (Running Unseen New Data)

Weather data for Kota Bharu from the dates that were not collected earlier (02/04/2023 to 25/06/2023) has been obtained from the Open Meteos Weather API. The gathered data has undergone the same preprocessing steps as the previous dataset used for analysis and model training.

During the span of 85 days, the model made predictions for 12 flood cases and 73 no flood cases. However, upon searching news articles, it was found that there were no major flood incidents during those 85 days. This indicates that the model achieved an accuracy rate of 85.9% for this unseen data.

	time	Normalized_temperature_2m (°C)	Normalized_relativehumidity_2m (%)	Normalized_cloudcover (%)	Normalized_temperature_2m_max (°C)	Normalized_rain_sum (mm)	Normalized_precipitation_hours (h)	Flood occurrence
0	02/04/2023	0.300000	0.569191	0.570659	0.588235	0.120879	0.421053	False
1	03/04/2023	0.200000	0.744125	0.717362	0.509804	0.347253	0.421053	True
2	04/04/2023	0.400000	0.509138	0.283311	0.627451	0.024176	0.368421	False
3	05/04/2023	0.200000	0.728460	0.330417	0.352941	0.274725	0.631579	True
4	06/04/2023	0.200000	0.610966	0.378869	0.509804	0.072527	0.210526	False
...
80	21/06/2023	0.633333	0.321149	0.246972	0.588235	0.092308	0.210526	False
81	22/06/2023	0.666667	0.268930	0.222073	0.568627	0.054945	0.263158	False
82	23/06/2023	0.633333	0.490862	0.389637	0.686275	0.043956	0.263158	False
83	24/06/2023	0.500000	0.686684	0.329071	0.607843	0.191209	0.789474	False
84	25/06/2023	0.433333	0.634465	0.323015	0.607843	0.380220	0.526316	False

Figure 24: Sample results from unseen new data

Conclusion:

In this project, we aimed to predict flood occurrences in Kelantan, Malaysia, based on different weather conditions. We started by cleaning and merging the flood cases dataset and the historical weather datasets to get a resulting dataset which contained daily weather parameters for 10 districts in Kelantan, along with a binary response variable indicating flood occurrences.

Exploratory data analysis revealed interesting insights, such as the correlation between temperature, relative humidity, rainfall, cloud cover, wind speed, wind direction, vapour pressure deficit, soil temperature, and soil moisture with flood occurrences. Outlier detection was done through boxplots and feature selection was done using a heatmap. Normalisation of data was also done so that we can compare the accuracy of using min-max normalised data vs unnormalised data.

Three machine learning algorithms, Logistic Regression, Decision Tree Classifier, and K Nearest Neighbour (KNN) - were chosen for model comparison, but ultimately, the Logistic Regression model was chosen due to a good sensitivity and specificity score and was not overfitted compared to the results from other models.

Finally, we were able to use the model on unseen data spanning 3 months (85 days), and through analysis, concluded that the results from the model had an 85.9% accuracy rate.

Overall, this study provides insights into the relationship between weather conditions and flood occurrences in Kelantan. The developed prediction model can aid decision-making processes in early flood detection and warning, enabling swift and well-ordered rescue and aid efforts during future disasters in the region.

References:

The Straits Times. (2018, February 13). Cold snap hits central Kelantan as temperature dips to 16 deg C on Sunday. *The Straits Times*.

<https://www.straitstimes.com/asia/se-asia/cold-snap-hits-central-kelantan-as-temperature-dips-to-16-deg-c-on-sunday#:~:text=%22But%20this%20is%20not%20the>

The Sun Daily. (2020). *Malaysia to experience hotter and drier climate until mid-September*.
Www.thesundaily.my.

<https://www.thesundaily.my/home/malaysia-to-experience-hotter-and-drier-climate-until-mid-september-DY8098005>

Satari, S. Z., Zubairi, Y. Z., Hussin, A. G., & Hassan, S. F. (2015). *Some Statistical Characteristic of Malaysian Wind Direction Recorded at Maximum Wind Speed: 1999-2008*. UKM.

http://www.ukm.my/jsm/pdf_files/SM-PDF-44-10-2015/18%20S.Z.%20Satari.pdf

Jermain, N. (2019, June 24). *Transforming Skewed Data for Machine Learning*. Open Data Science - Your News Source for AI, Machine Learning & More.

<https://opendatascience.com/transforming-skewed-data-for-machine-learning/>