PHIẾU GHI NHẬN ĐIỀU CHỈNH ĐỒ ÁN MÔN HỌC

Môn học: Hệ Thống Thương Mại Thông Minh (504049)

Học kỳ: 1

Năm học: 2021 - 2022

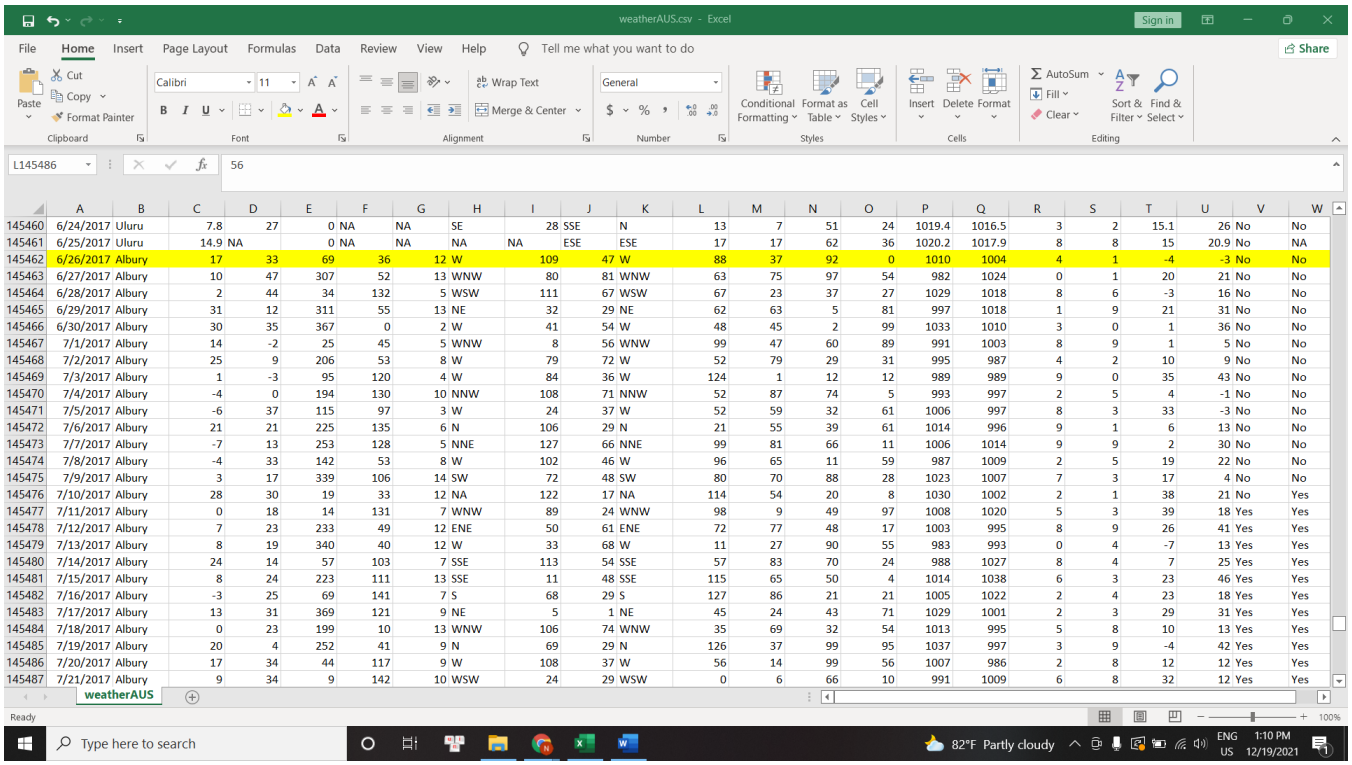
Mã nhóm: 12

Điều chỉnh 1: Thêm dữ liệu mới cho dataset weatherAUS.csv

```
[ ]  data.shape##Tổng giá trị dữ liệu và thuộc tính dataset cũ

     (145460, 23)

[ ]  data.shape##Tổng giá trị dữ liệu và thuộc tính thêm mới

     (161154, 23)
```

Dữ liệu weatherAUS.csv ban đầu được nghiên cứu bao gồm: Nhiệt độ, lượng mưa, độ bốc hơi, độ ẩm, áp xuất, hướng gió, vận tốc gió và mật độ mây được thu thập trong 10 năm (từ 1/12/2008 đến 25/6/2017) ở 49 trạm khí tượng của nước Úc bao gồm 145460 dòng với 23 thuộc tính.

Tổng giá trị được thêm mới gồm 15694 dòng với 23 thuộc tính vẫn được chỉ định ở 49 trạm ở Úc. Được thu thập thêm từ 26/6/2017 đến 25/6/2018.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
145460	6/24/2017	Uluru	7.8	27	0	NA	NA	SE	28	SSE	N	13	7	51	24	1019.4	1016.5	3	2	15.1	26	No	No
145461	6/25/2017	Uluru	14.9	NA	0	NA	NA	NA	NA	ESE	ESE	17	17	62	36	1020.2	1017.9	8	8	15	20.9	No	NA
145462	6/26/2017	Albury	17	33	69	36	12	W	109	47	W	88	37	92	0	1010	1004	4	1	-4	-3	No	No
145463	6/27/2017	Albury	10	47	307	52	13	WNW	80	81	WNW	63	75	97	54	982	1024	0	1	20	21	No	No
145464	6/28/2017	Albury	2	44	34	132	5	WSW	111	67	WSW	67	23	37	27	1029	1018	8	6	-3	16	No	No
145465	6/29/2017	Albury	31	12	311	55	13	NE	32	29	NE	62	63	5	81	997	1018	1	9	21	31	No	No
145466	6/30/2017	Albury	30	35	367	0	2	W	41	54	W	48	45	2	99	1033	1010	3	0	1	36	No	No
145467	7/1/2017	Albury	14	-2	25	45	5	WNW	8	56	WNW	99	47	60	89	991	1003	8	9	1	5	No	No
145468	7/2/2017	Albury	25	9	206	53	8	W	79	72	W	52	79	29	31	995	987	4	2	10	9	No	No
145469	7/3/2017	Albury	1	-3	95	120	4	W	84	36	W	124	1	12	12	989	989	9	0	35	43	No	No
145470	7/4/2017	Albury	-4	0	194	130	10	NNW	108	71	NNW	52	87	74	5	993	997	2	5	4	-1	No	No
145471	7/5/2017	Albury	-6	37	115	97	3	W	24	37	W	52	59	32	61	1006	997	8	3	33	-3	No	No
145472	7/6/2017	Albury	21	21	225	135	6	N	106	29	N	21	55	39	61	1014	996	9	1	6	13	No	No
145473	7/7/2017	Albury	-7	13	253	128	5	NNE	127	66	NNE	99	81	66	11	1006	1014	9	9	2	30	No	No
145474	7/8/2017	Albury	-4	33	142	53	8	W	102	46	W	96	65	11	59	987	1009	2	5	19	22	No	No
145475	7/9/2017	Albury	3	17	339	106	14	SW	72	48	SW	80	70	88	28	1023	1007	7	3	17	4	No	No
145476	7/10/2017	Albury	28	30	19	33	12	NA	122	17	NA	114	54	20	8	1030	1002	2	1	38	21	No	Yes
145477	7/11/2017	Albury	0	18	14	131	7	WNW	89	24	WNW	98	9	49	97	1008	1020	5	3	39	18	Yes	Yes
145478	7/12/2017	Albury	7	23	233	49	12	ENE	50	61	ENE	72	77	48	17	1003	995	8	9	26	41	Yes	Yes
145479	7/13/2017	Albury	8	19	340	40	12	W	33	68	W	11	27	90	55	983	993	0	4	-7	13	Yes	Yes
145480	7/14/2017	Albury	24	14	57	103	7	SSE	113	54	SSE	57	83	70	24	988	1027	8	4	7	25	Yes	Yes
145481	7/15/2017	Albury	8	24	223	111	13	SSE	11	48	SSE	115	65	50	4	1014	1038	6	3	23	46	Yes	Yes
145482	7/16/2017	Albury	-3	25	69	141	7	S	68	29	S	127	86	21	21	1005	1022	2	4	23	18	Yes	Yes
145483	7/17/2017	Albury	13	31	369	121	9	NE	5	1	NE	45	24	43	71	1029	1001	2	3	29	31	Yes	Yes
145484	7/18/2017	Albury	0	23	199	10	13	WNW	106	74	WNW	35	69	32	54	1013	995	5	8	10	13	Yes	Yes
145485	7/19/2017	Albury	20	4	252	41	9	N	69	29	N	126	37	99	95	1037	997	3	9	-4	42	Yes	Yes
145486	7/20/2017	Albury	17	34	44	117	9	W	108	37	W	56	14	99	56	1007	986	2	8	12	12	Yes	Yes
145487	7/21/2017	Albury	9	34	9	142	10	WSW	24	29	WSW	0	6	66	10	991	1009	6	8	32	12	Yes	Yes

Vậy ta sẽ có dataset mới bao gồm được thu thập từ 11 năm (1/12/2008 đến 25/6/2018) với 161154 dòng và 23 thuộc tính.

Thêm dòng giá trị mới và so sánh độ tương thích dữ liệu so với dataset cũ:

- Date: ngày tháng năm thu thập dòng dữ liệu (mm/dd/yyyy) sẽ được thêm 6/26/2017 đến 6/25/2018 (321 dòng).
- Location: địa điểm (name) được thêm 49 địa điểm ở nước Úc tương ứng với 6/26/2017 đến 6/25/2018 (mỗi địa điểm sẽ được thêm 321 dòng dữ liệu mới).
- MinTemp: nhiệt độ tối thiểu (oC) được dựa trên giá trị của dataset có sẵn giao động từ -8.5 đến 33.9 oC và được lấy tương thích nếu nhiệt độ cao gây ra sự bốc hơi thì khả năng mưa ngày hôm nay và ngày mai sẽ cao.
- MaxTemp: nhiệt độ tối đa (oC) được dựa trên giá trị của dataset có sẵn giao động từ -4.8 đến 44.1 oC và được lấy tương thích nếu nhiệt độ cao gây ra sự bốc hơi thì khả năng mưa ngày hôm nay và ngày mai sẽ cao.

- Rainfall: lượng mưa được ghi nhận trong ngày (mm) được thêm vào dựa trên giá trị cũ 0 đến 371 mm, nếu ngày hôm đó mưa sẽ được lấy lượng mưa rơi trong ngày và dự đoán cho ngày tiếp theo.
- Evaporation: sự bốc hơi được ghi nhận từ 24 giờ đêm đến 9h sáng (mm) được lấy từ 0 đến 145 mm dự trên data cũ nếu ngày hôm đó có nhiệt độ cao khả năng bốc hơi cũng sẽ cao.
- Sunshine: số giờ nắng trong ngày (hour) được lấy từ 0 đến 14.5hour dựa trên data cũ sẽ phụ thuộc vào nhiệt độ trong ngày và sự bốc hơi.
- WindGustDir: hướng gió (N-E-W-S).
- WindGustSpeed: vận tốc gió (km/h) được lấy từ 0 đến 135 km/h của data cũ.
- WindDir9am: hướng gió lúc 9 giờ sáng (N-E-W-S).
- WindDir3pm: hướng gió lúc 3 giờ chiều (N-E-W-S).
- WindSpeed9am: vận tốc gió lúc 9 giờ sáng (km/h) được lấy 0 đến 130km/h của data cũ dựa trên hướng gió 9h sáng và 3h chiều.
- WindSpeed3pm: vận tốc gió lúc 3 giờ chiều (km/h) được lấy 0 đến 87km/h của data cũ dựa trên hướng gió 9h sáng và 3h chiều.
- Humidity9am: độ ẩm không khí lúc 9 giờ sáng (g/m3) được lấy từ 0 đến 100 g/m3 của data cũ dựa trên nhiệt độ tối đa và tối thiểu trong ngày, sự bốc hơi của ngày hôm đó.
- Humidity3pm: độ ẩm không khí lúc 3 giờ chiều (g/m3) được lấy từ 0 đến 100 g/m3 của data cũ dựa trên nhiệt độ tối đa và tối thiểu trong ngày, sự bốc hơi của ngày hôm đó.
- Pressure9am: áp suất không khí lúc 9 giờ sáng (N/m2) được lấy từ 980.5 1041 dựa trên data cũ áp suất không khí cao thì ghi nhận mưa trong ngày là giá trị No, ấp suất không khí thấp thì hôm đó ghi nhận được trời sẽ đổ mưa trong ngày dựa trên các trị nhiệt độ tối đa tối thiểu và độ ẩm hướng gió vận tốc gió.
- Pressure3pm: áp suất không khí lúc 3 giờ chiều (N/m2) được lấy từ 977.1 1039.6 dựa trên data cũ áp suất không khí cao thì ghi nhận mưa trong ngày là giá trị No, ấp suất không khí thấp thì hôm đó ghi nhận được trời sẽ đổ mưa trong ngày dựa trên các trị nhiệt độ tối đa tối thiểu và độ ẩm hướng gió vận tốc gió
- Cloud9am: mật độ mây theo các mức độ lúc 9 giờ sáng được lấy từ (0-9).

- Cloud3pm: mật độ mây theo các mức độ lúc 3 giờ chiều được lấy từ (0-9).
- Temp9am: nhiệt độ ghi nhận lúc 9 giờ sáng (oC) được lấy từ -7.2 đến 40.2 oC ảnh hưởng đến sự bốc hơi và độ ẩm trong không khí dẫn đến tích tụ nước và gây mưa
- Temp3pm: nhiệt độ ghi nhận lúc 3 giờ chiều (oC) được lấy từ -5.4 đến 46.7 oC ảnh hưởng đến sự bốc hơi và độ ẩm trong không khí dẫn đến tích tụ nước và gây mưa
- RainToday: thuộc tính ghi nhận mưa trong ngày (yes/no).
- RainTomorrow: thuộc tính dự đoán mưa trong ngày tiếp theo (yes/no).

Điều chỉnh 2: Sử dụng dataset mới (đã thêm dữ liệu mới vào dataset cũ)

Với các dữ liệu bị thiếu trong dataset weatherAUS.csv sẽ được xử lý missing value như sau:

- Bước 1: Loading dataset
- Bước 2: Phân tích các thuộc tính thành 2 loại (kiểu số và kiểu kí tự).
- Bước 3: Có nhiều cách xử lý dữ liệu bị thiếu: Sử dụng trung vị hoặc giá trị trung bình, chọn giá trị ngẫu nhiên để đưa vào hoặc giá trị có tần suất xuất hiện nhiều nhất của mỗi thuộc tính. Ở trường hợp dataset này sử dụng điền các giá trị còn thiếu bằng giá trị có tần suất xuất hiện nhiều nhất của mỗi thuộc tính. Sử dụng hàm fillna() để tìm dữ liệu xuất hiện nhiều nhất lấp vào chỗ trống.

Về độ tương thích của dữ liệu sao khi missing value: sẽ được so sánh bằng cách sử dụng điền các giá trị còn thiếu bằng giá trị có tần suất xuất hiện nhiều nhất của mỗi thuộc tính. Bằng cách dự đoán và kiểm tra kết quả đánh giá theo thuật toán Sử dụng thuật toán Logistic Regression.

- Trước khi missing value

```
y_pred_test = logreg.predict(X_test)

y_pred_test
```

```
array(['No', 'No', 'No', ..., 'No', 'No', 'Yes'], dtype=object)
```

Kiểm tra đánh giá độ chính xác

```
[35]: from sklearn.metrics import accuracy_score

print('Model accuracy score: {0:0.4f}'. format(accuracy_score(y_test, y_pred_test)))
```

```
Model accuracy score: 0.8498
```

- Sau khi missing value:

```
model = LogisticRegression(max_iter=500)
model.fit(x_train, y_train)
predicted=model.predict(x_test)
conf = confusion_matrix(y_test, predicted)
print ("The accuracy of Logistic Regression is : ", accuracy_score(y_test, predicted)*100, "%")
```

```
The accuracy of Logistic Regression is :  84.00494267865723 %
```

Kết luận: vậy giá trị điền vào NaN trống của data về độ tương thích là không chênh lệch kết quả quá nhiều so với trước. Khi dự đoán xác suất của trước khi missing là 84.98% còn sau khi missing dự đoán xác suất là 84%.

Điều chỉnh 3: Đưa ra đánh giá với từng thuật toán sử dụng trong bài.

Với đề tài dự báo thời tiết ở nước Úc, quá trình phân tích train và test tập dữ liệu. Với thuật toán Logistic Regression cho ra độ chính xác 84%, thuật toán XGBoost 84.75%, thuật toán Gaussian Naive Bayes 80.1%, thuật toán Bernoulli Naive Bayes 76.67% và thuật toán Random Forest Model cho độ chính xác 85.48%.

Vậy với chủ đề dự báo thời tiết ở nước Úc thuật toán Random Forest Model cho ra dự đoán chính xác cao nhất. Ngược lại thuật toán Bernoulli Naive Bayes cho ra dự đoán thấp nhất.

Thuật toán Random Forest Model cho ra kết quả chính xác nhất tuy nhiên tốn nhiều thời gian hơn so với các thuật toán còn lại.

Thuật toán XGBoost cho ra kết quả dự đoán thời gian ngắn nhất mà độ chính xác so với Thuật toán Random Forest Model cũng không chênh lệch nhiều.

Thuật toán Logistic Regression cho ra độ chính xác cao đứng thứ 3 tuy nhiên thời gian dự đoán nhanh mà độ chính xác so với Thuật toán Random Forest Model cũng không chênh lệch nhiều.

Thuật toán Gaussian Naive Bayes cùng với thuật toán Bernoulli Naive Bayes cho ra độ chính xác đứng thứ 4 và thứ 5 thời gian dự đoán ngắn tuy nhiên kết quả dự đoán chênh lệch cao.

Kết luận: đối với dataset dự báo thời tiết ở nước Úc, chúng em sẽ cho thuật toán XGBoost là phù hợp nhất vì thời gian dự đoán nhanh và cho ra kết quả chính xác không quá chênh lệch với Random Forest Model. Random Forest Model cho ra kết quả chính xác cao nhất nhưng thời gian dự đoán khá lâu nên chúng em sẽ không chọn thuật toán này là tốt nhất.