Московский государственный технический университет

им. Н.Э. Баумана

(МГТУ им. Н.Э. Баумана)

Радиотехнический факультет (РТ)

Отчёт по лабораторной работе №3

По дисциплине

«Технологии машинного обучения»

Проверил:

Преподаватель кафедры ИУ-5

Гапанюк Ю.Е.

Подпись: _____

«___» _____ 2020 г.

Выполнил:

студент группы РТ5-61Б

Ануров Н.С.

Подпись: _____

«___» _____ 2020 г.

Москва, 2020

Задание:

1. Выбрать набор данных (датасет), содержащий категориальные признаки и пропуски в данных. Для выполнения следующих пунктов можно использовать несколько различных наборов данных (один для обработки пропусков, другой для категориальных признаков и т.д.)
2. Для выбранного датасета (датасетов) на основе материалов лекции решить следующие задачи:

- обработку пропусков в данных;
- кодирование категориальных признаков;
- масштабирование данных.

```
In [11]: df=pd.read_csv('weatherAUS.csv')

In [12]: df.head()
```

Out[12]:

MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	...	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Te
22.9	0.6	NaN	NaN	W	44.0	W	...	22.0	1007.7	1007.1	8.0	NaN	
25.1	0.0	NaN	NaN	WNW	44.0	NNW	...	25.0	1010.6	1007.8	NaN	NaN	
25.7	0.0	NaN	NaN	WSW	46.0	W	...	30.0	1007.6	1008.7	NaN	2.0	
28.0	0.0	NaN	NaN	NE	24.0	SE	...	16.0	1017.6	1012.8	NaN	NaN	
32.3	1.0	NaN	NaN	W	41.0	ENE	...	33.0	1010.8	1006.0	7.0	8.0	

```
In [13]: df.shape
Out[13]: (142193, 24)

In [14]: df.isnull().sum()
Out[14]: Date                0
         Location            0
         MinTemp           637
         MaxTemp           322
         Rainfall         1406
         Evaporation     60843
         Sunshine        67816
         WindGustDir      9330
         WindGustSpeed    9270
         WindDir9am      10013
         WindDir3pm       3778
         WindSpeed9am     1348
         WindSpeed3pm     2630
         Humidity9am      1774
         Humidity3pm      3610
         Pressure9am     14014
         Pressure3pm     13981
         Cloud9am        53657
         Cloud3pm        57094
```

In [15]:
```
df1=df.dropna(axis=1,how='any')
```

In [20]:
```
total_count=df.shape[0]
df1.shape
```

Out[20]:
```
(142193, 4)
```

In [18]:
```
df1.head()
```

Out[18]:

	Date	Location	RISK_MM	RainTomorrow
0	2008-12-01	Albury	0.0	No
1	2008-12-02	Albury	0.0	No
2	2008-12-03	Albury	0.0	No
3	2008-12-04	Albury	1.0	No
4	2008-12-05	Albury	0.2	No

In [21]:
```
# Выберем числовые колонки с пропущенными значениями
# Цикл по колонкам датасета
num_cols = []
for col in df.columns:
    # Количество пустых значений
    temp_null_count = df[df[col].isnull()].shape[0]
    dt = str(df[col].dtype)
    if temp_null_count>0 and (dt=='float64' or dt=='int64'):
        num_cols.append(col)
        temp_perc = round((temp_null_count / total_count) * 100.0, 2)
        print('Колонка {}. Тип данных {}. Количество пустых значений {}, {}%.'.format(col, dt, temp_null_count, temp_perc))
```

```
Колонка MinTemp. Тип данных float64. Количество пустых значений 637, 0.45%.
Колонка MaxTemp. Тип данных float64. Количество пустых значений 322, 0.23%.
Колонка Rainfall. Тип данных float64. Количество пустых значений 1406, 0.99%.
Колонка Evaporation. Тип данных float64. Количество пустых значений 60843, 42.79%.
Колонка Sunshine. Тип данных float64. Количество пустых значений 67816, 47.69%.
Колонка WindGustSpeed. Тип данных float64. Количество пустых значений 9270, 6.52%.
Колонка WindSpeed9am. Тип данных float64. Количество пустых значений 1348, 0.95%.
Колонка WindSpeed3pm. Тип данных float64. Количество пустых значений 2630, 1.85%.
Колонка Humidity9am. Тип данных float64. Количество пустых значений 1774, 1.25%.
Колонка Humidity3pm. Тип данных float64. Количество пустых значений 3610, 2.54%.
Колонка Pressure9am. Тип данных float64. Количество пустых значений 14014, 9.86%.
Колонка Pressure3pm. Тип данных float64. Количество пустых значений 13981, 9.83%.
Колонка Cloud9am. Тип данных float64. Количество пустых значений 53657, 37.74%.
Колонка Cloud3pm. Тип данных float64. Количество пустых значений 57094, 40.15%.
Колонка Temp9am. Тип данных float64. Количество пустых значений 904, 0.64%.
Колонка Temp3pm. Тип данных float64. Количество пустых значений 2726, 1.92%.
```

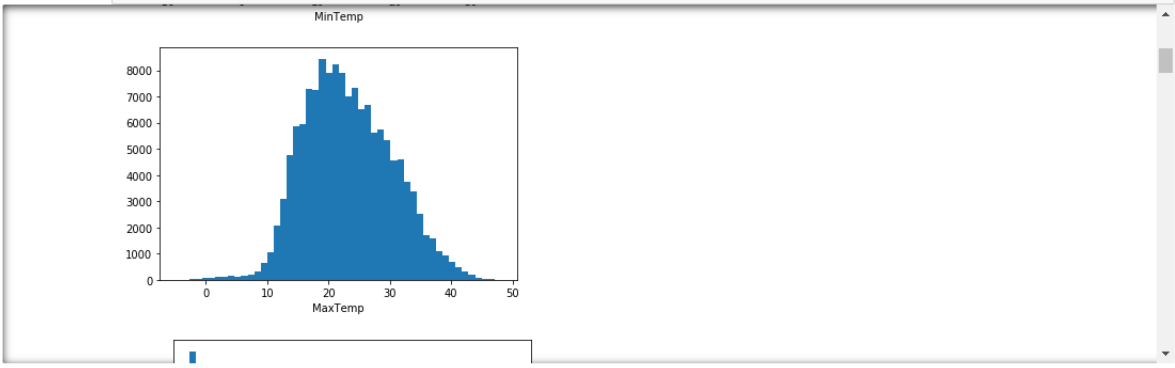
In [24]:
```
data_num=df[num_cols]
data_num
```

Out[24]:

	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustSpeed	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pre
0	13.4	22.9	0.6	NaN	NaN	44.0	20.0	24.0	71.0	22.0	1007.7	
1	7.4	25.1	0.0	NaN	NaN	44.0	4.0	22.0	44.0	25.0	1010.6	
2	12.9	25.7	0.0	NaN	NaN	46.0	19.0	26.0	38.0	30.0	1007.6	
3	9.2	28.0	0.0	NaN	NaN	24.0	11.0	9.0	45.0	16.0	1017.6	
4	17.5	32.3	1.0	NaN	NaN	41.0	7.0	20.0	82.0	33.0	1010.8	
...	...	...	...	...	...	...	...	...	...	...	...	...
142188	3.5	21.8	0.0	NaN	NaN	31.0	15.0	13.0	59.0	27.0	1024.7	
142189	2.8	23.4	0.0	NaN	NaN	31.0	13.0	11.0	51.0	24.0	1024.6	
142190	3.6	25.3	0.0	NaN	NaN	22.0	13.0	9.0	56.0	21.0	1023.5	
142191	5.4	26.9	0.0	NaN	NaN	37.0	9.0	9.0	53.0	24.0	1021.0	
142192	7.8	27.0	0.0	NaN	NaN	28.0	13.0	7.0	51.0	24.0	1019.4	

142193 rows × 16 columns

In [26]:
```
for col in data_num:
    plt.hist(df[col], 50)
    plt.xlabel(col)
    plt.show()
```



In [ ]: