TRƯỜNG ĐẠI HỌC SÀI GÒN

KHOA CÔNG NGHỆ THÔNG TIN

HỌC PHẦN: KHAI PHÁ DỮ LIỆU

DATA SET: MATERNAL HEALTH RISK

Họ tên: Mai Văn Thinh: 3119410413

Giáo viên hướng dẫn: Vũ Ngọc Thanh Sang

TP.HCM, Tháng 04/2023

LỜI MỞ ĐẦU

Bộ dữ liệu này cung cấp thông tin về sức khỏe của phụ nữ mang thai và các yếu tố liên quan đến rủi ro sức khỏe của mẹ và thai nhi. Việc phân tích bộ dữ liệu này có thể giúp chúng ta hiểu rõ hơn về những yếu tố ảnh hưởng đến sức khỏe của phụ nữ mang thai và những biến số quan trọng trong việc dự đoán rủi ro sức khỏe cho mẹ và thai nhi.

Bộ dữ liệu này cũng cung cấp các thông tin về chỉ số sinh học và các bệnh lý liên quan đến sức khỏe của mẹ và thai nhi, giúp cho các nhà nghiên cứu và chuyên gia y tế có thể phân tích và xây dựng các mô hình dự đoán rủi ro sức khỏe cho mẹ và thai nhi.

Tuy nhiên, bộ dữ liệu này còn tồn tại các giá trị bị khuyết, điều này có thể ảnh hưởng đến quá trình phân tích và xây dựng mô hình. Chúng ta cần thực hiện các phương pháp tiền xử lý để xử lý các giá trị bị khuyết này trước khi áp dụng các thuật toán học máy.

MỤC LỤC

I. GIỚI THIỆU:

1. Giới thiệu đề tài:

Dự án khai phá dữ liệu này dựa trên tập dữ liệu "Maternal Health Risk Data Set" được lấy từ Tổ chức Giáo dục Đại học California, Irvine (UCI) từ đường link:

https://archive.ics.uci.edu/ml/datasets/Maternal+Health+Risk+Data+Set.

Tập dữ liệu này tập trung vào vấn đề sức khỏe của phụ nữ trong thời kỳ mang thai, với những biến số được đánh giá trong khi dự báo các yếu tố rủi ro liên quan đến sức khỏe mẹ và thai nhi.

2. Mục đích:

Mục đích của dự án khai phá dữ liệu này là phân tích và đánh giá các yếu tố có ảnh hưởng đến sức khỏe của phụ nữ trong thời kỳ mang thai, từ đó đề xuất những giải pháp phòng ngừa và điều chỉnh để giảm thiểu nguy cơ và ảnh hưởng tới sức khỏe của mẹ và thai nhi. Đồng thời, thông qua việc phân tích dữ liệu, đưa ra các thông tin hữu ích giúp các chuyên gia y tế, quan chức và cả người dân trong việc đưa ra quyết định và giải pháp phù hợp với hoàn cảnh cụ thể.

3. Phạm vi:

Phạm vi của dự án khai phá dữ liệu này bao gồm việc thu thập thông tin từ tập dữ liệu đã được cung cấp, tiền xử lý dữ liệu để loại bỏ nhiễu và chuẩn hóa, phân tích mối quan hệ giữa các biến, xây dựng mô hình dự báo các yếu tố rủi ro đối với sức khỏe mẹ và thai nhi, và đánh giá hiệu quả của mô hình phát hiện rủi ro.

II. MÔ TẢ BỘ DỮ LIỆU:

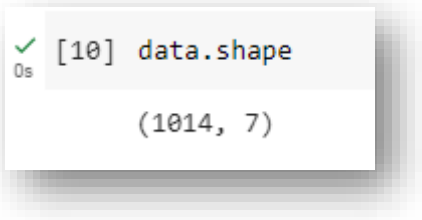
- Bộ dữ liệu Maternal Health Risk Data Set được lưu trữ tại kho lưu trữ dữ liệu của Trung tâm Hỗ trợ Máy học và Tính toán Khoa học không gian (ICS) thuộc Đại học California, Irvine (UCI). Bộ dữ liệu này được sử dụng để phân tích nguy cơ sức khỏe mẹ trong quá trình mang thai và sinh nở.

1. Nguồn gốc dữ liệu:

- Dữ liệu được thu thập từ các trung tâm giám sát và bảo vệ mẹ và trẻ, một số dữ liệu liên quan đã được biên soạn từ trước để phục vụ cho nghiên cứu theo yêu cầu của Lembaga Pengembangan Kependudukan/ Lembaga Pengembangan Kependudukan dan Keluarga Kesejahteraan (LPK/ LPK3) (1945-1992).

2. Số lượng mẫu:

- Bộ dữ liệu này chứa tổng cộng 1014 mẫu dữ liệu.



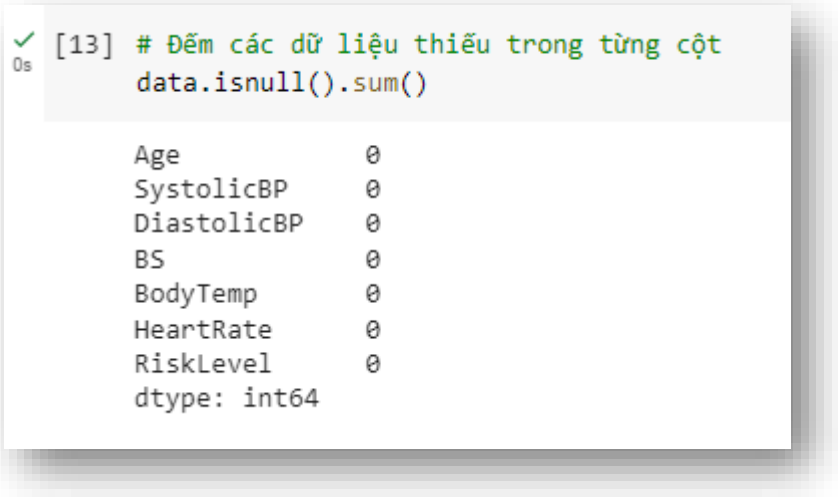
3. Số lượng thuộc tính:

- Trong bộ dữ liệu này, có 7 thuộc tính:
- Thông tin thuộc tính:
  - **Age**: Bất kỳ độ tuổi nào tính theo năm khi phụ nữ mang thai
  - **SystolicBP**: Giá trị trên của Huyết áp tính bằng mmHg, một thuộc tính quan trọng khác trong thời kỳ mang thai.
  - **DiastolicBP**: Giá trị thấp hơn của Huyết áp tính bằng mmHg, một thuộc tính quan trọng khác trong thai kỳ.
  - **BS**: Nồng độ glucose trong máu tính theo nồng độ mmol/L.
  - **BodyTemp**: Thước đo khả năng tạo và thải nhiệt của cơ thể. Đơn vị đo là Độ Fahrenheit (°F)

- **HeartRate**: Nhịp tim bình thường khi nghỉ ngơi tính bằng nhịp đập của tim mỗi phút.
- **RiskLevel**: Mức độ rủi ro được dự đoán trong thời kỳ mang thai khi xem xét thuộc tính trước đó.

4. Các giá trị bị khuyết trong dữ liệu:

- Không có



```
[13] # Đếm các dữ liệu thiếu trong từng cột
     data.isnull().sum()

Age            0
SystolicBP     0
DiastolicBP    0
BS             0
BodyTemp       0
HeartRate      0
RiskLevel      0
dtype: int64
```

III. TIỀN XỬ LÍ DỮ LIỆU:

1. Lọc dữ liệu:

Lọc dữ liệu: Loại bỏ các hàng hoặc cột không cần thiết trong tập dữ liệu, ví dụ như các cột không chứa thông tin quan trọng, hoặc các hàng chứa dữ liệu bị trùng lặp.

❖ ***Xác định các dữ liệu bị thiếu:***

```
# Xác định các dữ liệu bị thiếu
data.isnull()
```

	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
0	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...
1009	False	False	False	False	False	False	False
1010	False	False	False	False	False	False	False
1011	False	False	False	False	False	False	False
1012	False	False	False	False	False	False	False
1013	False	False	False	False	False	False	False

1014 rows × 7 columns

❖ ***Đếm các dữ liệu thiếu trong từng cột:***

```
# Đếm các dữ liệu thiếu trong từng cột
data.isnull().sum()
```

```
Age           0
SystolicBP    0
DiastolicBP   0
BS            0
BodyTemp      0
HeartRate     0
RiskLevel     0
dtype: int64
```

❖ ***Tính giá trị % dữ liệu bị thiếu trong cột:***

```
# Tính giá trị % dữ liệu bị thiếu trong cột
data[data.columns[data.isnull().any()]].isnull().sum() * 100 / data.shape[0]
```

```
Series([], dtype: float64)
```

2. *Xử lí giá trị bị khuyết:*

Xem xét dữ liệu có nên bị xóa hay thay thế:

- Nếu dữ liệu bị null nhỏ hơn 10% trong bảng dữ liệu của chúng ta, ta có thể xóa nó.
- Nếu lớn hơn 10%, ta dùng phương pháp thay thế dữ liệu thiếu, sao cho tập dữ liệu này không có dữ liệu bị thiếu.

❖ ***Loại bỏ các dữ liệu bị thiếu:***

```
# Loại bỏ các dữ liệu bị thiếu
data.dropna(inplace=True)
data
```

	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
0	25	130	80	15.0	98.0	86	high risk
1	35	140	90	13.0	98.0	70	high risk
2	29	90	70	8.0	100.0	80	high risk
3	30	140	85	7.0	98.0	70	high risk
4	35	120	60	6.1	98.0	76	low risk
...	...	...	...	...	...	...	...
1009	22	120	60	15.0	98.0	80	high risk
1010	55	120	90	18.0	98.0	60	high risk
1011	35	85	60	19.0	98.0	86	high risk
1012	43	120	90	18.0	98.0	70	high risk
1013	32	120	65	6.0	101.0	76	mid risk

1014 rows × 7 columns

❖ ***Xác định bộ giá trị trùng lặp:***

```
#Xác định bộ giá trị trùng lặp
data.duplicated()
```

```
0       False
1       False
2       False
3       False
4       False
        ...
1009     True
1010     True
1011     True
1012     True
1013     True
Length: 1014, dtype: bool
```

❖ ***Số lượng dữ liệu trùng lặp:***

```
print("Số lượng dữ liệu trùng lặp", len(data[data.duplicated()]))
data[data.duplicated()]
```

```
Số lượng dữ liệu trùng lặp 562
```

	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
67	19	120	80	7.0	98.0	70	mid risk
72	19	120	80	7.0	98.0	70	mid risk
97	19	120	80	7.0	98.0	70	mid risk
106	50	140	90	15.0	98.0	90	high risk
107	25	140	100	6.8	98.0	80	high risk
...	...	...	...	...	...	...	...
1009	22	120	60	15.0	98.0	80	high risk
1010	55	120	90	18.0	98.0	60	high risk
1011	35	85	60	19.0	98.0	86	high risk
1012	43	120	90	18.0	98.0	70	high risk
1013	32	120	65	6.0	101.0	76	mid risk

562 rows × 7 columns

❖ ***Loại bỏ các dữ liệu trùng lặp không cần thiết:***

```
print("Kích thước của Data trước khi xóa các hàng trùng lặp", data.shape)
data = data.drop_duplicates()
data.index = range(len(data))
print("Kích thước của Data sau khi xóa các hàng trùng lặp", data.shape)
```

```
Kích thước của Data trước khi xóa các hàng trùng lặp (1014, 7)
Kích thước của Data sau khi xóa các hàng trùng lặp (452, 7)
```

❖ ***Xác định và xử lí outliers của thuộc tính Age:***

```
#ngưỡng dưới, ngưỡng trên của biến Tuổi Age
min_threshold, max_threshold = data.Age.quantile([0.01, 0.99])
min_threshold, max_threshold
```

```
(12.0, 63.98000000000002)
```

```
#tập dữ liệu trên max_threshold -> outliers
data[data['Age']>max_threshold]
```

	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
236	65	90	60	6.9	98.0	70	low risk
240	66	85	60	6.9	98.0	86	low risk
246	70	85	60	6.9	102.0	70	low risk
247	65	120	90	6.9	103.0	76	low risk
304	65	130	80	15.0	98.0	86	high risk

```
#tập dữ liệu dưới min_threshold -> outliers
data[data['Age']<min_threshold]
```

	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
19	10	70	50	6.9	98.0	70	low risk
175	10	85	65	6.9	98.0	70	low risk
444	10	100	50	6.0	99.0	70	mid risk

```
#xóa outliers
data = data[(data['Age']<max_threshold)&(data['Age']>min_threshold)]
data
```

	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
0	25	130	80	15.0	98.0	86	high risk
1	35	140	90	13.0	98.0	70	high risk
2	29	90	70	8.0	100.0	80	high risk
3	30	140	85	7.0	98.0	70	high risk
4	35	120	60	6.1	98.0	76	low risk
...	...	...	...	...	...	...	...
445	15	100	49	6.8	99.0	77	low risk
446	15	100	49	6.0	99.0	77	low risk
448	15	100	60	6.0	98.0	80	low risk
449	15	100	49	7.6	98.0	77	low risk
451	21	100	50	6.8	98.0	60	low risk

423 rows × 7 columns

❖ ***Xác định và xử lí outliers của thuộc tính SystolicBP:***

```
#Huyết áp tâm thu - SystolicBP
SystolicBP_min_threshold, SystolicBP_max_threshold = data.SystolicBP.quantile([0.01, 0.99])
data = data[(data['SystolicBP']<SystolicBP_max_threshold)&(data['SystolicBP']>SystolicBP_min_threshold)]
```

❖ ***Xác định và xử outliers của thuộc tính DiastolicBP:***

```
#Huyết áp tâm trương - DiastolicBP
DiastolicBP_min_threshold, DiastolicBP_max_threshold = data.DiastolicBP.quantile([0.01, 0.99])
data = data[(data['DiastolicBP']<DiastolicBP_max_threshold)&(data['DiastolicBP']>DiastolicBP_min_threshold)]
```

❖ *Xác định và xử lí outliers của thuộc tính BS:*

```
#Đường huyết - BS
BS_min_threshold, BS_max_threshold = data.BS.quantile([0.01, 0.99])
data = data[(data['BS']<BS_max_threshold)&(data['BS']>BS_min_threshold)]
```

❖ *Xác định và xử lí outliers của thuộc tính Body Temp:*

```
#Nhiệt độ cơ thể - BodyTemp
BodyTemp_min_threshold, BodyTemp_max_threshold = data.BodyTemp.quantile([0.01, 0.99])
data = data[(data['BodyTemp']<BodyTemp_max_threshold)&(data['BodyTemp']>BodyTemp_min_threshold)]
```

❖ *Xác định và xử lí outliers của thuộc tính HeartRate:*

```
#Nhịp tim - HeartRate
HeartRate_min_threshold, HeartRate_max_threshold = data.HeartRate.quantile([0.01, 0.99])
data = data[(data['HeartRate']<HeartRate_max_threshold)&(data['HeartRate']>HeartRate_min_threshold)]
```

❖ *Kích thước tập tin sau khi xử lí outliers:*

```
data.shape
```

```
(39, 7)
```

❖ ***Xác định dữ liệu nhiễu:***

```
#Xác định dữ liệu nhiễu
Noisy_data = data[(data['SystolicBP'] < 0 ) | (data['DiastolicBP'] < 0) | (data['BS'] < 0) | (data['BodyTemp'] < 0) | (data['HeartRate'] < 0)]
Noisy_data
```

	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel

❖ ***Xử lí dữ liệu nhiễu:***

```
#Xử lý dữ liệu nhiễu
data = data[(data['SystolicBP'] > 0 ) & (data['DiastolicBP'] > 0) & (data['BS'] > 0) & (data['BodyTemp'] > 0) & (data['HeartRate'] > 0)]
data.shape
```

```
(46, 7)
```

❖ ***Xuất dữ liệu:***

```
#Xuất tập dữ liệu
data.to_csv('/content/drive/MyDrive/KPDL/GiuaKy/Output.csv', index=False)
```

❖ ***Giá trị trung bình:***

```
data.mean()
```

```
<ipython-input-56-abc01cf6c622>:1:
  data.mean()
Age             22.956522
SystolicBP     107.608696
DiastolicBP     73.782609
BS               8.108913
BodyTemp       100.965217
HeartRate       75.217391
dtype: float64
```

❖ ***Giá trị trung vị:***

```
data.median()
```

```
<ipython-input-57-135339ac59ce>:1:
  data.median()
Age             18.5
SystolicBP     115.0
DiastolicBP     75.0
BS               7.5
BodyTemp       101.0
HeartRate       76.0
dtype: float64
```

❖ ***Độ lệch chuẩn:***

```
data.std()
```

```
<ipython-input-58-a47ac8255c06>:1:
  data.std()
Age            10.090604
SystolicBP     14.327097
DiastolicBP     9.790955
BS              1.994183
BodyTemp        0.784635
HeartRate       3.943563
dtype: float64
```

❖ ***Phân vị:***

```
data.describe()
```

	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate
count	46.000000	46.000000	46.000000	46.000000	46.000000	46.000000
mean	22.956522	107.608696	73.782609	8.108913	100.965217	75.217391
std	10.090604	14.327097	9.790955	1.994183	0.784635	3.943563
min	13.000000	90.000000	60.000000	6.600000	98.400000	68.000000
25%	17.000000	90.000000	65.000000	6.900000	101.000000	70.000000
50%	18.500000	115.000000	75.000000	7.500000	101.000000	76.000000
75%	29.000000	120.000000	80.000000	7.975000	101.000000	79.500000
max	55.000000	130.000000	90.000000	16.000000	102.000000	80.000000

3. *Rút trích đặc trưng và chuẩn hóa dữ liệu:*

- Dữ liệu này không cần rút trích đặc trưng

Chuẩn hóa dữ liệu:

```
from sklearn.preprocessing import MinMaxScaler
# Lấy ra các cột dữ liệu cần chuẩn hóa
columns_to_scale = ['Age', 'SystolicBP', 'DiastolicBP', 'BS', 'BodyTemp', 'HeartRate']

# Tạo đối tượng MinMaxScaler
scaler = MinMaxScaler()

# Chuẩn hóa các cột dữ liệu
data[columns_to_scale] = scaler.fit_transform(data[columns_to_scale])

# In ra dữ liệu sau khi chuẩn hóa
print(data)
```

```
           Age  SystolicBP  DiastolicBP        BS  BodyTemp  HeartRate  \
2     0.380952        0.00     0.333333  0.148936  0.444444   1.000000
13    0.166667        0.75     0.500000  0.043617  0.444444   0.166667
18    0.404762        0.75     0.666667  0.031915  0.722222   0.666667
67    0.166667        0.50     0.000000  0.042553  0.444444   0.166667
90    0.000000        0.00     0.166667  0.127660  0.722222   1.000000
111   0.404762        0.75     0.666667  0.138298  0.722222   0.666667
126   0.119048        0.75     0.666667  0.031915  1.000000   0.666667
128   0.095238        0.00     0.000000  0.031915  0.722222   0.666667
129   0.095238        0.00     0.100000  0.031915  0.722222   0.166667
130   0.285714        0.75     1.000000  0.010638  0.722222   1.000000
131   0.095238        0.75     0.666667  0.010638  1.000000   0.666667
132   0.023810        0.00     0.166667  0.042553  0.722222   0.166667
```

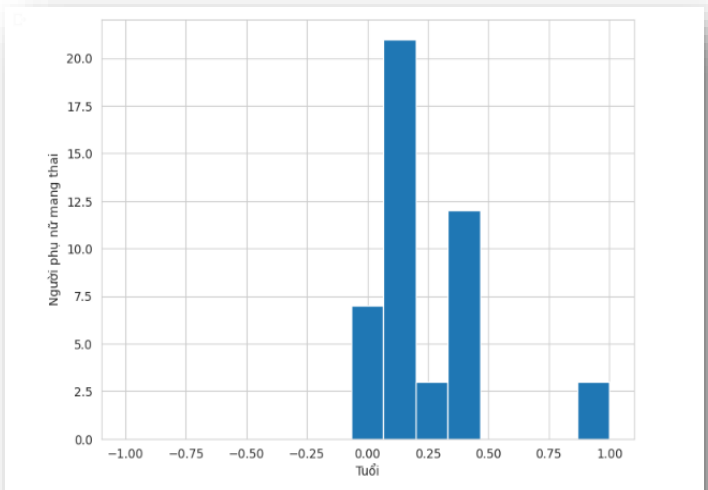
IV. PHÂN TÍCH DỮ LIỆU KHÁM PHÁ:

1. Phân tích đơn biến:

- Xem xét phân bố của từng biến trong dữ liệu bằng cách sử dụng histogram.

❖ Age (Tuổi):

```
maxAge = data['Age'].max()
minAge = data['Age'].min()
plt.hist(data['Age'], bins=15, range=(minAge-1, maxAge))
# Đặt tên cho trục x và trục y
plt.xlabel('Tuổi')
plt.ylabel('Người phụ nữ mang thai')
plt.show()
```



- Trục x: là số tuổi (Tuổi)
- Trục y: là số người phụ nữ mang thai (Người)

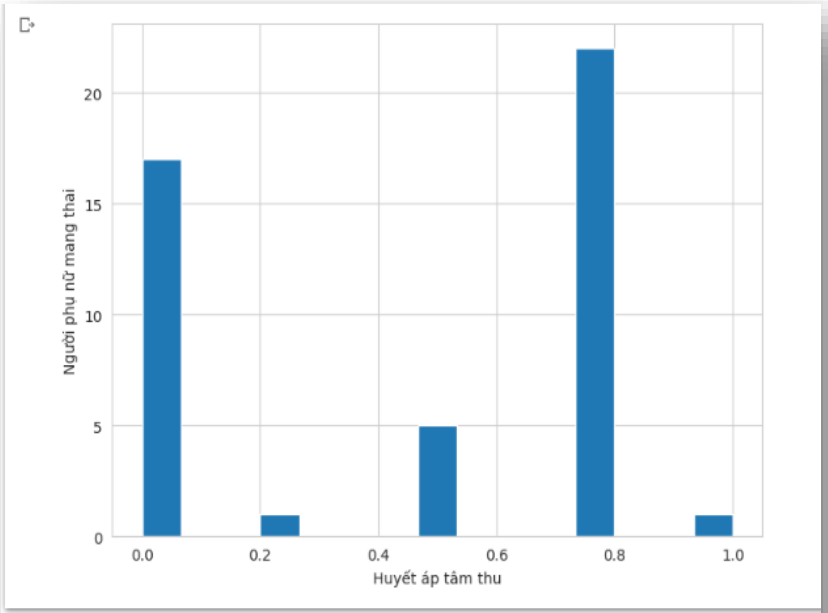
Nhận xét:

- Trong khoảng độ tuổi từ 15-18.5 tuổi là khoảng thời gian có nhiều phụ nữ mang thai nhất
- Trong khoảng từ 34 tuổi trở đi có ít phụ nữ mang thai nhất

=> Độ tuổi phụ nữ mang thai còn quá trẻ

❖ SystolicBP (Huyết áp tâm thu):

```
maxSystolicBP = data['SystolicBP'].max()
minSystolicBP = data['SystolicBP'].min()
plt.xlabel('Huyết áp tâm thu')
plt.ylabel('Người phụ nữ mang thai')
plt.hist(data['SystolicBP'], bins=15, range=(minSystolicBP, maxSystolicBP))
plt.show()
```



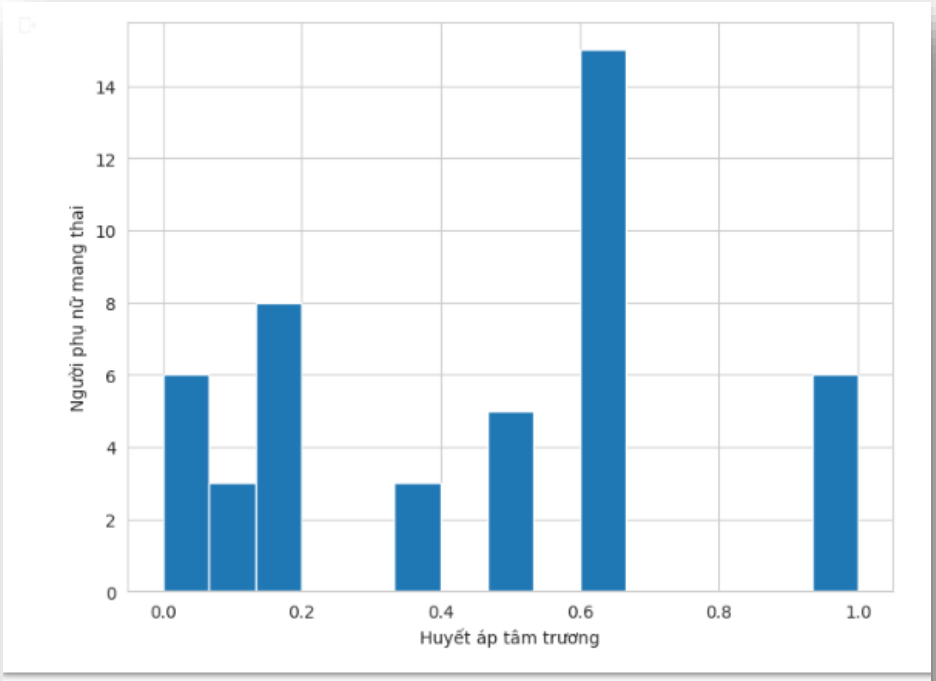
- Trục x : Số đo huyết áp tâm thu (mmHg)
- Trục y : Số người phụ nữ mang thai (Người)

Nhận xét

- Số phụ nữ mang thai có huyết áp tâm thu tập trung cao nhất chủ yếu là từ 119-122 mmHg

❖ DiastolicBP (Huyết áp tâm trương):

```
maxDiastolicBP = data['DiastolicBP'].max()
minDiastolicBP = data['DiastolicBP'].min()
plt.xlabel('Huyết áp tâm trương')
plt.ylabel('Người phụ nữ mang thai')
plt.hist(data['DiastolicBP'], bins=15, range=(minDiastolicBP, maxDiastolicBP))
plt.show()
```



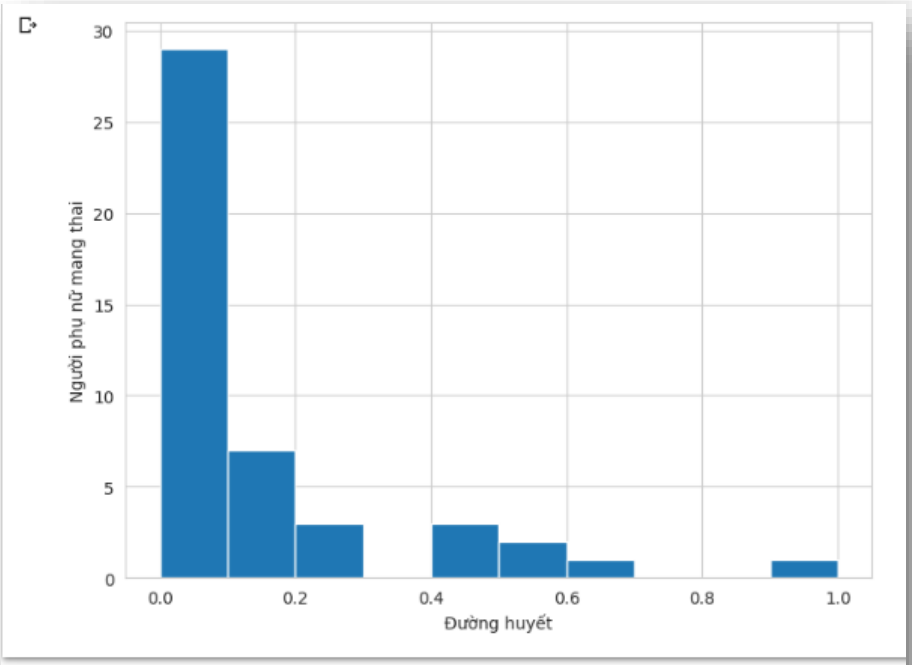
- Trục x : Số đo huyết áp tâm trương (mmHg)
- Truc y: Số người phụ nữ mang thai (Người)

Nhận xét:

- Số phụ nữ mang thai có huyết áp tâm trương cao nhất là 80-82 mmHg

❖ BS (Đường huyết):

```
maxBS = data['BS'].max()
minBS = data['BS'].min()
plt.xlabel('Đường huyết')
plt.ylabel('Người phụ nữ mang thai')
plt.hist(data['BS'], bins=10, range=(minBS, maxBS))
plt.show()
```



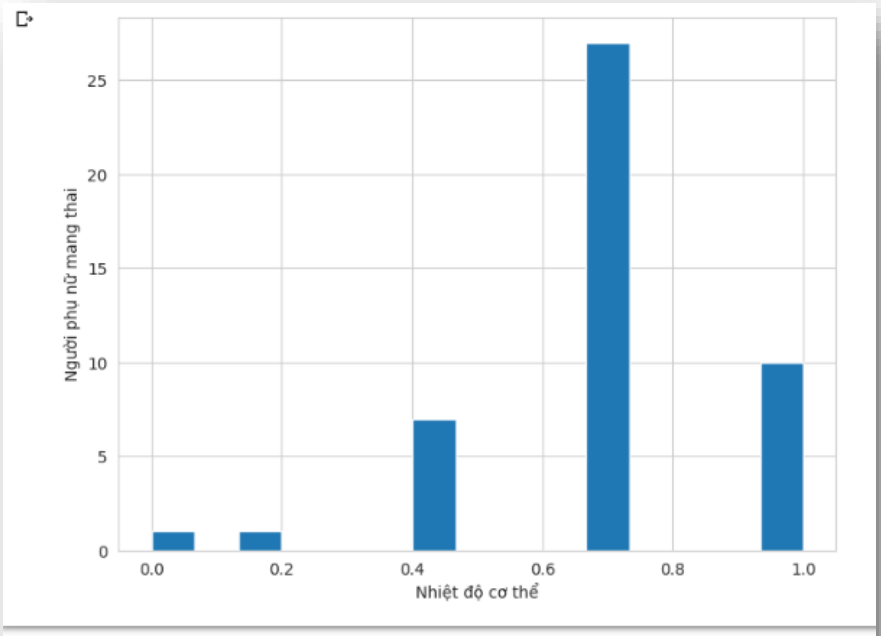
- Trục x : Số đo đường huyết (mmol/l)
- Trục y : Số người phụ nữ mang thai (Người)

Nhận xét:

- Số phụ nữ mang thai có lượng đường huyết từ 2-6 mmol/l là nhiều nhất.

❖ Body Temp (Đường huyết):

```
maxBodyTemp = data['BodyTemp'].max()
minBodyTemp = data['BodyTemp'].min()
plt.xlabel('Nhiệt độ cơ thể')
plt.ylabel('Người phụ nữ mang thai')
plt.hist(data['BodyTemp'], bins=15, range=(minBodyTemp, maxBodyTemp))
plt.show()
```



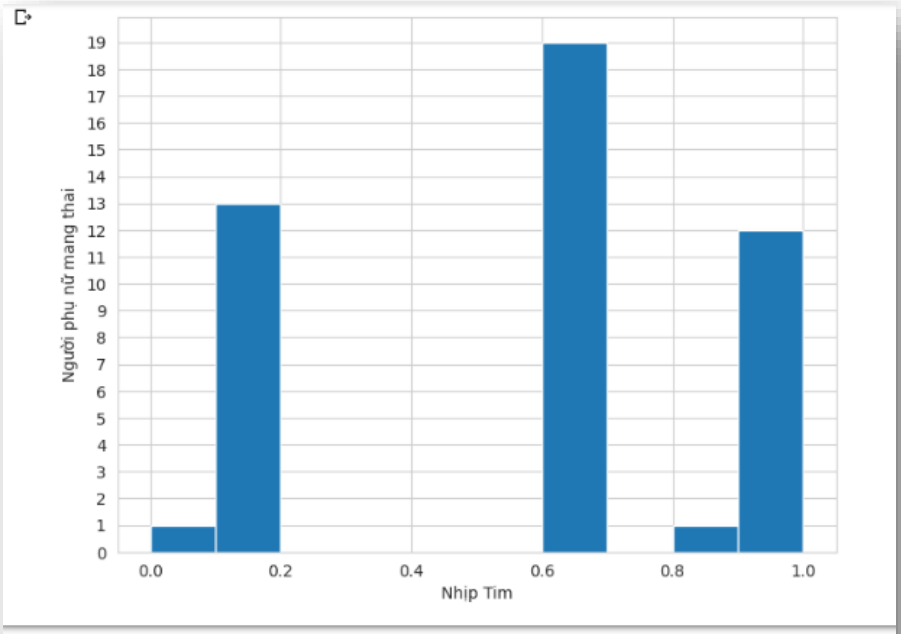
- Trục x : Nhiệt độ cơ thể (°F)
- Trục y : Số người phụ nữ mang thai (Người)

Nhận xét:

- Số phụ nữ mang thai có nhiệt độ cơ thể từ 98°F - 98.3°F là nhiều nhất

❖ Heart Rate (Nhịp tim):

```
maxHeartRate = data['HeartRate'].max()
minHeartRate = data['HeartRate'].min()
plt.xlabel('Nhịp Tim')
plt.ylabel('Người phụ nữ mang thai')
plt.yticks(range(max(Counter(data['HeartRate']).values()) + 1))
plt.hist(data['HeartRate'])
plt.show()
```



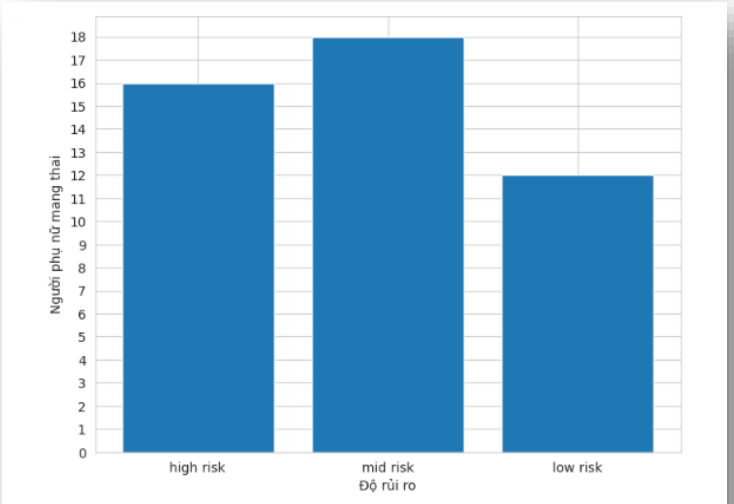
- Trục x : Nhịp tim (nhịp/phút)
- Trục y : Số người phụ nữ mang thai (người)

Nhận xét:

- Số nhụ nữ mang thai có nhịp tim trung bình từ 75-79 (nhịp/phút) là nhiều nhất

❖ Risk Level (Mức độ rủi ro):

```
riskLevel = data['RiskLevel']
# Tính số lần xuất hiện của từng giá trị
counter = Counter(riskLevel)
# Tạo biểu đồ histogram
plt.bar(counter.keys(), counter.values())
plt.xlabel('Độ rủi ro')
plt.ylabel('Người phụ nữ mang thai')
# Chỉ định giá trị cho các nhãn trên trục tung
plt.yticks(range(max(counter.values()) + 1))
# Hiển thị biểu đồ
plt.show()
```



- Trục x : Mức độ rủi ro
- Trục y : Số phụ nữ mang thai (Người)

Nhận xét:

- Phần lớn phụ nữ mang thai đều có mức rủi ro trung bình

2. Phân tích đa biến và tương quan:

- Tìm kiếm các mối quan hệ giữa các biến trong dữ liệu bằng cách sử dụng heatmap.

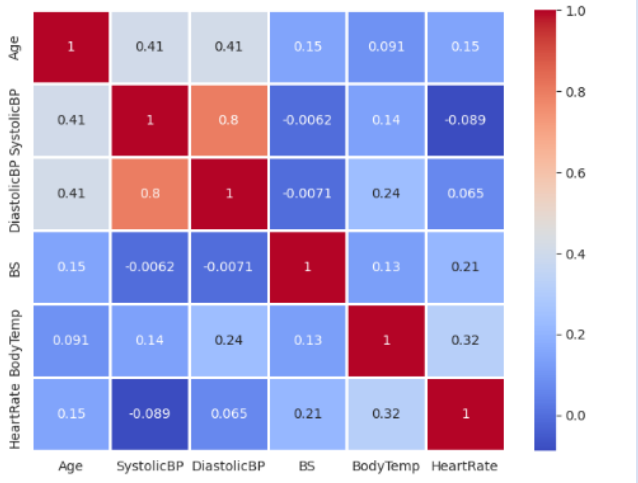
```
import seaborn as sns
data.corr()
```

```
<ipython-input-74-28f4727c3154>:2: FutureWarning: The default value of numeric
  data.corr()
```

	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate
Age	1.000000	0.406606	0.410173	0.148963	0.090743	0.146556
SystolicBP	0.406606	1.000000	0.797806	-0.006160	0.138719	-0.088923
DiastolicBP	0.410173	0.797806	1.000000	-0.007057	0.237348	0.064560
BS	0.148963	-0.006160	-0.007057	1.000000	0.128732	0.211538
BodyTemp	0.090743	0.138719	0.237348	0.128732	1.000000	0.318496
HeartRate	0.146556	-0.088923	0.064560	0.211538	0.318496	1.000000

```
sns.heatmap(data.corr(),cmap='coolwarm',annot=True, linecolor='white',linewidths=1)
```

```
<ipython-input-75-8785dfed1c17>:1: FutureWarning: The default value of numeric_only in
  sns.heatmap(data.corr(),cmap='coolwarm',annot=True, linecolor='white',linewidths=1)
<Axes: >
```



Hàm corr trả về độ tương quan giữa các cột có mối liên hệ với nhau:

- Hệ số tương quan có giá trị âm cho thấy hai biến có mối quan hệ nghịch biến hoặc tương quan âm (nghịch biến tuyệt đối khi giá trị bằng -1)
- Hệ số tương quan có giá trị dương cho thấy mối quan hệ đồng biến hoặc tương quan dương (đồng biến tuyệt đối khi giá trị bằng 1)
- Tương quan bằng 0 cho hai biến độc lập với nhau. Thông thường độ tương quan để sử dụng nằm trong khoảng 0,8 trở lên

Nhận xét:

- Các cặp biến có có hệ số tương quan bằng 1 nên có mối quan hệ đồng biến tuyệt đối như: Age và Age, SystolicBP và SystolicBP, DiastolicBP và DiastolicBP, BS và BS, BodyTemp và BodyTemp, HeartRate và HeartRate
- Các cặp biến có hệ số tương quan âm nên có mối quan hệ nghịch biến hoặc tương quan âm như: SystolicBP và BS, SystolicBP và HeartRate, DiastolicBP và BS.
- Các cặp biến còn lại có hệ số tương quan dương nên đều có mối quan hệ đồng biến hoặc tương quan dương

V. KHAI PHÁ DỮ LIỆU:

1. Câu hỏi về dữ liệu:

- Liệt kê thông tin các bà mẹ có mức độ rủi ro là "**mid risk**"?
- Hãy cho biết mức độ rủi ro của bà mẹ có **AGE** là "15", **SYSTOLICBP** là "120", **DIASTOLICBP** là "80", **BS** là "6.60", **BODYTEMP** là "99.0", **HEARTRATE** là "70"?
- Liệt kê xác suất mức độ rủi ro (**RiskLevel**) của các bà mẹ mang thai ở tuổi "17"?
- Dự đoán xác suất mức độ rủi ro (**RiskLevel**) cao nhất cho mỗi giá trị tuổi (**Age**)?

2. Đề xuất thuật toán:

- Decision Tree
- Logistic Regression

3. Trả lời câu hỏi:

❖ Với 2 câu hỏi cần sử dụng thuật toán **Decision Tree** là:

- Liệt kê thông tin các bà mẹ có mức độ rủi ro là "**mid risk**"?
- Hãy cho biết mức độ rủi ro của bà mẹ có **AGE** là "15", **SYSTOLICBP** là "120", **DIASTOLICBP** là "80", **BS** là "6.60", **BODYTEMP** là "99.0", **HEARTRATE** là "70"?

❖ Với 2 câu hỏi cần sử dụng thuật toán **Logistic Regression** là:

- Liệt kê xác suất mức độ rủi ro (**RiskLevel**) của các bà mẹ mang thai ở tuổi "17"?
- Dự đoán xác suất mức độ rủi ro (**RiskLevel**) cao nhất cho mỗi giá trị tuổi (**Age**)?

VI. ĐÁNH GIÁ VÀ CHỌN THUẬT TOÁN:

1. Đánh giá thuật toán:

❖ Thuật toán Decision Tree:

- Ưu điểm:
  - Dễ hiểu và giải thích kết quả phân loại.
  - Không yêu cầu các giả định về phân phối của dữ liệu.
  - Có thể xử lý các tính năng có giá trị bị khuyết và thể hiện mối quan hệ phi tuyến tính giữa các tính năng.
  - Có khả năng xử lý cả dữ liệu phân loại và dữ liệu liên tục.
- Nhược điểm:
  - Dễ bị quá khớp (overfitting) khi số lượng các lá (leaf) quá lớn hoặc cây quá sâu.
  - Khó xử lý các tính năng phức tạp.
  - Không thể xử lý các bài toán phân loại đa lớp (multiclass classification) một cách hiệu quả.

❖ Thuật toán Logictis Regression:

- Ưu điểm:
  - Dễ hiểu và giải thích kết quả phân loại.
  - Phù hợp với các bài toán phân loại nhị phân (binary classification).
  - Không yêu cầu các giả định về phân phối của dữ liệu.
  - Có khả năng xử lý cả dữ liệu phân loại và dữ liệu liên tục.
- Nhược điểm:
  - Không thể xử lý các bài toán phân loại đa lớp một cách hiệu quả.
  - Yêu cầu các tính năng độc lập tuyến tính.
  - Khó xử lý các tính năng phức tạp.

2. Chọn thuật toán:

❖ Sử dụng thuật toán Decision Tree để trả lời câu hỏi:

- Liệt kê thông tin các bà mẹ có mức độ rủi ro là "**mid risk**"?

Các bước thực hiện thuật toán

LIỆT KÊ THÔNG TIN CÁC BÀ MẸ CÓ MỨC ĐỘ RỦI RO LÀ "MID RISK"?

```
[ ]  import pandas as pd

     data = pd.read_csv("/content/drive/MyDrive/KPDL/GiuaKy/Output.csv")
```

import thư viện pandas và đọc file CSV bằng hàm read_csv()

```
[ ]  from sklearn.tree import DecisionTreeClassifier

     # Lấy các features và target từ dataframe
     X = data.drop('RiskLevel', axis=1)
     y = data['RiskLevel']

     # Tạo classifier và huấn luyện trên dữ liệu
     clf = DecisionTreeClassifier()
     clf.fit(X, y)

     DecisionTreeClassifier
     DecisionTreeClassifier()
```

tạo một Decision Tree classifier bằng scikit-learn, và sử dụng phương thức fit () để huấn luyện model trên dữ liệu

```
[ ]  # Dự đoán mức độ rủi ro cho toàn bộ dữ liệu
     predictions = clf.predict(X)

     # Lọc các bệnh nhân có mức độ rủi ro là "mid risk"
     mid_risk_patients = data[predictions == "mid risk"]
```

sử dụng model đã huấn luyện để dự đoán mức độ rủi ro cho các bệnh nhân.

```
[ ]  mid_risk_patients.to_csv("/content/drive/MyDrive/KPDL/GiuaKy/DecisionTree1.csv", index=False)
```

lưu các thông tin của các bệnh nhân có mức độ rủi ro là "mid risk" ra file CSV bằng phương thức to_csv () của pandas.

```
[ ]  DecisionTree1 = pd.read_csv('/content/drive/MyDrive/KPDL/GiuaKy/DecisionTree1.csv')
     DecisionTree1
```

đọc file csv để kiểm tra kết quả

	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
0	20	120	75	7.01	100.0	70	mid risk
1	30	120	80	6.90	101.0	76	mid risk
2	20	110	60	7.00	100.0	70	mid risk
3	13	90	65	7.80	101.0	80	mid risk
4	18	120	80	6.90	102.0	76	mid risk
5	17	90	60	6.90	101.0	76	mid risk
6	17	90	63	6.90	101.0	70	mid risk
7	25	120	90	6.70	101.0	80	mid risk
8	17	120	80	6.70	102.0	76	mid risk
9	13	90	65	7.90	101.0	80	mid risk

Kết quả file csv

❖ Sử dụng thuật toán Decision Tree để trả lời câu hỏi:

- Hãy cho biết mức độ rủi ro của bà mẹ có **AGE** là "15", **SYSTOLICBP** là "120", **DIASTOLICBP** là "80", **BS** là "6.60", **BODYTEMP** là "99.0", **HEARTRATE** là "70"?

Các bước thực hiện thuật toán

HÃY CHO BIẾT MỨC ĐỘ RỦI RO CỦA BÀ MẸ CÓ AGE LÀ "15" , SYSTOLICBP LÀ "120", DIASTOLICBP LÀ "80", BS LÀ "6.60", BODYTEMP LÀ "99.0", HEARTRATE LÀ "70"?

```
[ ]  import pandas as pd
     from sklearn.tree import DecisionTreeClassifier
     from sklearn.model_selection import train_test_split

     data = pd.read_csv('/content/drive/MyDrive/KPDL/GiuaKy/Output.csv')
```

import các thư viện cần thiết và đọc dữ liệu từ file csv

```
features = ['Age', 'SystolicBP', 'DiastolicBP', 'BS', 'BodyTemp', 'HeartRate']
X = data[features]
y = data['RiskLevel']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

chọn các đặc trưng để sử dụng cho mô hình Decision Tree và tạo hai tập dữ liệu riêng biệt: tập train và tập test.

```
[ ]  clf = DecisionTreeClassifier(random_state=42)
     clf.fit(X_train, y_train)

     accuracy = clf.score(X_test, y_test)
     print('Accuracy:', accuracy)

     Accuracy: 0.5
```

sử dụng tập train và đánh giá độ chính xác của mô hình trên tập test.

```
[ ]  # Tạo một DataFrame mới với thông tin của bà mẹ
     new_data = pd.DataFrame({'Age': [15], 'SystolicBP': [120], 'DiastolicBP': [80], 'BS': [6.6], 'BodyTemp': [99.0], 'HeartRate': [70]})

     # Sử dụng mô hình để dự đoán mức độ rủi ro
     risk_level = clf.predict(new_data)

     print('Risk level:', risk_level[0])

     Risk level: mid risk
```

dự đoán mức độ rủi ro của các bà mẹ có thông tin như trên.

```
[ ]  new_data['RiskLevel'] = risk_level
     new_data.to_csv('/content/drive/MyDrive/KPDL/GiuaKy/DecisionTree2.csv', index=False)
```

xuất kết quả dự đoán của mô hình vào một file csv mới.

```
[ ]  DecisionTree2 = pd.read_csv('/content/drive/MyDrive/KPDL/GiuaKy/DecisionTree2.csv')
     DecisionTree2
```

	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
0	15	120	80	6.6	99.0	70	mid risk

Kết quả từ đọc dữ liệu của file csv

❖ Sử dụng thuật toán Logistic Regression để trả lời câu hỏi:

- Liệt kê xác suất mức độ rủi ro (**Risk Level**) của các bà mẹ mang thai ở tuổi "**17**"?

Các bước thực hiện thuật toán

---

liệt kê xác suất mức độ rủi ro (RiskLevel) của các bà mẹ mang thai ở tuổi 17?

```
[ ]  import pandas as pd
     from sklearn.linear_model import LogisticRegression
```

Import thư viện

```
[ ]  df = pd.read_csv("/content/drive/MyDrive/KPDL/GiuaKy/Output.csv")
     df
```

Đọc file dữ liệu csv

```
[ ]  df_17 = df.loc[df['Age'] == 17, ['Age', 'RiskLevel']]
```

Lấy các bản ghi ở tuổi 17 và chỉ lấy cột Age và RiskLevel

```
[ ]  risk_level_map = {'low risk': 0, 'mid risk': 1, 'high risk': 2}
     df_17['RiskLevel'] = df_17['RiskLevel'].map(risk_level_map)
```

Chuyển các giá trị của cột RiskLevel thành các giá trị số

```
[ ]  model = LogisticRegression()
```

Tạo mô hình Logistic Regression

```
[ ]  X = df_17['Age'].values.reshape(-1, 1)
     y = df_17['RiskLevel'].values
```

Tạo X và y từ dataframe

```
[ ]  model.fit(X, y)
     ▾ LogisticRegression
     LogisticRegression()
```

Fit mô hình

```
[ ]  y_pred_proba = model.predict_proba(X)
```

Dự đoán xác suất mức độ rủi ro cho các bản ghi ở tuổi 17

```
[ ]  df_17['Probability of Low Risk'] = y_pred_proba[:, 0]
     df_17['Probability of Mid Risk'] = y_pred_proba[:, 1]
     df_17['Probability of High Risk'] = y_pred_proba[:, 2]
```

Thêm cột xác suất vào Data Frame

```
[ ]  df_17.to_csv('/content/drive/MyDrive/KPDL/GiuaKy/LogisticRegression1.csv', index=False)
```

Ghi kết quả vào file csv

```
LogisticRegression1 = pd.read_csv("/content/drive/MyDrive/KPDL/GiuaKy/LogisticRegression1.csv")
LogisticRegression1
```

Đọc file kết quả

	Age	RiskLevel	Probability of Low Risk	Probability of Mid Risk	Probability of High Risk
0	17	1	0.307703	0.230856	0.461442
1	17	1	0.307703	0.230856	0.461442
2	17	1	0.307703	0.230856	0.461442
3	17	2	0.307703	0.230856	0.461442
4	17	2	0.307703	0.230856	0.461442
5	17	2	0.307703	0.230856	0.461442
6	17	2	0.307703	0.230856	0.461442
7	17	2	0.307703	0.230856	0.461442
8	17	2	0.307703	0.230856	0.461442
9	17	0	0.307703	0.230856	0.461442
10	17	0	0.307703	0.230856	0.461442
11	17	0	0.307703	0.230856	0.461442
12	17	0	0.307703	0.230856	0.461442

Kết quả từ file CSV

❖ Sử dụng thuật toán Logistic Regression để trả lời câu hỏi:

- Dự đoán xác suất mức độ rủi ro (**Risk Level**) cao nhất cho mỗi giá trị tuổi (**Age**)?

Các bước thực hiện thuật toán

Dự đoán xác suất mức độ rủi ro cao nhất cho mỗi giá trị tuổi ?

```
[ ]  import pandas as pd
     import numpy as np
     from sklearn.model_selection import train_test_split
     from sklearn.linear_model import LogisticRegression
```

Import thư viện

```
[ ]  df = pd.read_csv("/content/drive/MyDrive/KPDL/GiuaKy/Output.csv")
     df
```

Đọc file dữ liệu csv

```
[ ]  X = data[['Age']]
     y = data['RiskLevel']
```

Chuẩn bị feature và label cho mô hình

```
[ ]  X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
```

Chia dữ liệu thành tập train và tập test

```
[ ]  logreg = LogisticRegression()
     logreg.fit(X_train, y_train)

     LogisticRegression
     LogisticRegression()
```

Tạo mô hình Logistic Regression và train trên tập train

```
[ ]  y_pred = logreg.predict_proba(X_test)
     y_pred_highrisk = [p[2] for p in y_pred]
```

Dự đoán xác suất mức độ rủi ro cao nhất cho mỗi giá trị tuổi trong tập test

```
[ ]  result_df = pd.DataFrame({'Age': X_test.values.ravel(), 'Highest Risk Probability': y_pred_highrisk})
     result_df.to_csv('/content/drive/MyDrive/KPDL/GiuaKy/LogisticRegression2.csv', index=False)
```

Tạo dataframe chứa kết quả và xuất ra file csv

```
LogisticRegression2 = pd.read_csv("/content/drive/MyDrive/KPDL/GiuaKy/LogisticRegression2.csv")
LogisticRegression2
```

Đọc file CSV

	Age	Highest Risk Probability
0	50	0.631517
1	13	0.266961
2	30	0.426371
3	13	0.266961
4	17	0.301124
5	17	0.301124
6	31	0.436642
7	14	0.275271
8	19	0.319092
9	18	0.310037

Kết quả từ file CSV

VII. KẾT QUẢ VÀ THẢO LUẬN:

1. Câu hỏi “Liệt kê thông tin các bà mẹ có mức độ rủi ro là "mid risk"?

❖ Kết quả:

	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
0	20	120	75	7.01	100.0	70	mid risk
1	30	120	80	6.90	101.0	76	mid risk
2	20	110	60	7.00	100.0	70	mid risk
3	13	90	65	7.80	101.0	80	mid risk
4	18	120	80	6.90	102.0	76	mid risk
5	17	90	60	6.90	101.0	76	mid risk
6	17	90	63	6.90	101.0	70	mid risk
7	25	120	90	6.70	101.0	80	mid risk
8	17	120	80	6.70	102.0	76	mid risk
9	13	90	65	7.90	101.0	80	mid risk
10	50	130	80	16.00	102.0	76	mid risk
11	27	120	90	6.80	102.0	68	mid risk
12	55	100	70	6.80	101.0	80	mid risk
13	31	110	90	6.80	100.0	70	mid risk
14	18	120	80	7.90	102.0	76	mid risk
15	30	120	80	7.50	101.0	76	mid risk
16	32	120	90	7.00	100.0	70	mid risk
17	30	120	80	9.00	101.0	76	mid risk

❖ Điểm mạnh:

- Kết quả cung cấp thông tin chi tiết về các bà mẹ có mức độ rủi ro "mid risk" trong bảng dữ liệu, bao gồm các thông số về tuổi (Age), huyết áp tâm thu (SystolicBP), huyết áp tâm trương (DiastolicBP), chỉ số đường huyết (BS), nhiệt độ cơ thể (BodyTemp), nhịp tim (HeartRate) và mức độ rủi ro (RiskLevel).
- Kết quả giúp người đọc dễ dàng nhận thấy phân bố của dữ liệu và xu hướng của từng thông số trong dữ liệu.

❖ Điểm yếu:

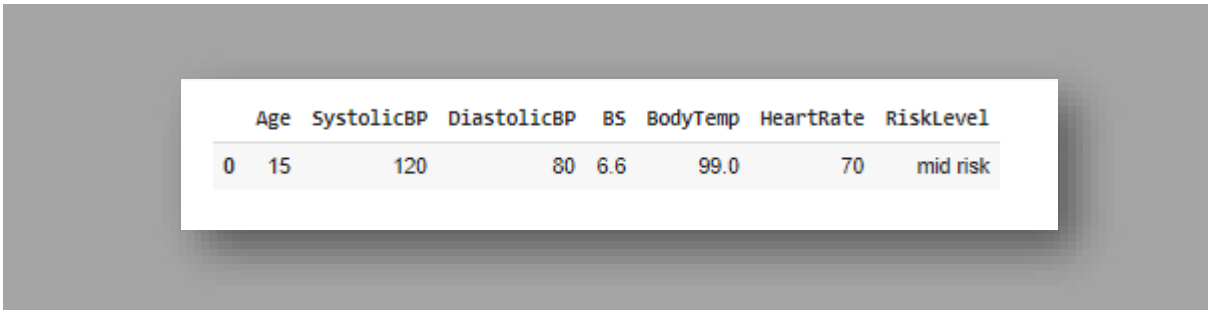
- Dữ liệu chưa được phân tích sâu về mối quan hệ giữa từng thông số và mức độ rủi ro. Ví dụ, liệu có mối tương quan giữa tuổi với mức độ rủi ro, hay giữa huyết áp và rủi ro, hoặc giữa chỉ số đường huyết và mức độ rủi ro không?
- Chưa rõ ràng về nguyên nhân dẫn đến mức độ rủi ro "mid risk" trong dữ liệu. Điều này đòi hỏi phải tìm hiểu thêm về nguyên nhân đằng sau số liệu để có cái nhìn đầy đủ hơn về dữ liệu.
- Thiếu thông tin về những yếu tố ngoài, như guồng máu chảy qua tai, tổn thương, hạ mức oái ẩm, bệnh lý tiền đợi, môi trường sống, thói quen sinh hoạt, chế độ dinh dưỡng của các bà mẹ, vv., những yếu tố này có thể ảnh hưởng đến mức độ rủi ro.
- Dữ liệu chỉ cung cấp thông tin về số lượng bà mẹ có mức độ rủi ro "mid risk" mà không cho biết tỷ lệ so với tổng số bà mẹ trong dữ liệu. Điều này sẽ giúp người đọc đánh giá được mức độ phổ biến của nhóm "mid risk" trong tổng thể.

❖ Kết luận:

- Nhìn chung, kết quả trên có thể là một nguồn thông tin hữu ích cho việc nghiên cứu về mức độ rủi ro của các bà mẹ, nhưng cần phải tìm hiểu thêm và phân tích sâu hơn để có những kết luận chính xác và đúng đắn.

2. **Câu hỏi "Hãy cho biết mức độ rủi ro (**Risk Level**) của bà mẹ có Age là** "15", **SYSTOLICBP là** "120", **DIASTOLICBP là** "80", **BS là** "6.60", **BODYTEMP là** "99.0", **HEARTRATE là** "70"**?"**

❖ Kết quả:



	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
0	15	120	80	6.6	99.0	70	mid risk

- Kết quả trên cho thấy mức độ rủi ro (Risk Level) của bà mẹ có độ tuổi 15, huyết áp tâm thu (SystolicBP) là 120, huyết áp tâm trương (DiastolicBP) là 80, đường huyết (Blood Sugar - BS) là 6.6, nhiệt độ cơ thể (BodyTemp) là 99.0 và nhịp tim (HeartRate) là 70. Mức độ rủi ro được xác định là "mid risk".
- Những khía cạnh quan trọng của kết quả trên bao gồm:
  - ✓ Độ tuổi: Bà mẹ là một thiếu nữ 15 tuổi, độ tuổi này thường chưa đủ trưởng thành để mang thai và sinh con. Việc mang thai ở độ tuổi này có thể dẫn đến những rủi ro liên quan đến sức khỏe của cả mẹ và bé.
  - ✓ Huyết áp: Huyết áp tâm thu 120 và huyết áp tâm trương 80 được coi là bình thường. Tuy nhiên, tùy thuộc vào tình trạng sức khỏe của bà mẹ trẻ, các chỉ số này có thể thay đổi trong suốt thai kỳ.
  - ✓ Đường huyết: Chỉ số đường huyết 6.6 mmol/L cho thấy bà mẹ không có nguy cơ tiểu đường thai kỳ. Mức này nằm trong khoảng bình thường cho phụ nữ mang thai.
  - ✓ Nhiệt độ cơ thể: Nhiệt độ cơ thể 99.0 độ F (37.2 độ C) hơi cao so với nhiệt độ cơ thể bình thường là 98.6 độ F (37 độ C). Nếu nhiệt độ cơ thể tiếp tục tăng, điều này có thể gây lo ngại.
  - ✓ Nhịp tim: Nhịp tim ở mức 70 lần/phút là bình thường.

- ✓ Dựa trên các chỉ số trên, mức độ rủi ro được xác định là "mid risk".

❖ Điểm mạnh:

- Phân loại rủi ro giúp bác sĩ và gia đình nắm bắt tình trạng sức khỏe của bà mẹ và đưa ra quyết định phù hợp cho chăm sóc thai kỳ.
- Các chỉ số huyết áp, đường huyết và nhịp tim đều nằm trong khoảng bình thường, giảm bớt một số nguy cơ liên quan đến sức khỏe.

❖ Điểm yếu:

- Việc phân loại rủi ro chỉ dựa trên một số ít chỉ số nên không đủ để đánh giá toàn diện tình trạng sức khỏe của bà mẹ.
- Kết quả không đề cập đến những yếu tố khác liên quan đến sức khỏe, chẳng hạn như lịch sử gia đình, lối sống, dinh dưỡng, tình trạng tâm lý, cũng như các khó khăn trong thai kỳ.

❖ Kết luận:

- Nhìn chung, kết quả trên cung cấp một cái nhìn tổng quát về mức độ rủi ro của bà mẹ 15 tuổi nhưng không thể thay thế sự đánh giá toàn diện từ các chuyên gia y tế.

3. **Câu hỏi "Hãy liệt kê xác suất mức độ rủi ro (**RiskLevel**) của các bà mẹ mang thai ở tuổi "**17**"?"**

❖ Kết quả:

	Age	RiskLevel	Probability of Low Risk	Probability of Mid Risk	Probability of High Risk
0	17	1	0.307703	0.230856	0.461442
1	17	1	0.307703	0.230856	0.461442
2	17	1	0.307703	0.230856	0.461442
3	17	2	0.307703	0.230856	0.461442
4	17	2	0.307703	0.230856	0.461442
5	17	2	0.307703	0.230856	0.461442
6	17	2	0.307703	0.230856	0.461442
7	17	2	0.307703	0.230856	0.461442
8	17	2	0.307703	0.230856	0.461442
9	17	0	0.307703	0.230856	0.461442
10	17	0	0.307703	0.230856	0.461442
11	17	0	0.307703	0.230856	0.461442
12	17	0	0.307703	0.230856	0.461442

- Mức độ rủi ro (**RiskLevel**) của các bà mẹ mang thai ở tuổi 17 được phân thành 3 nhóm: Thấp (**Low Risk**), Trung bình (**Mid Risk**) và Cao (**High Risk**) với các xác suất tương ứng là 0.307703, 0.230856 và 0.461442.
- Phân bố xác suất mức độ rủi ro của các bà mẹ mang thai ở tuổi 17 cho thấy rằng có tỷ lệ cao nhất là 46,14% thuộc nhóm rủi ro cao, tỷ lệ thấp nhất là 23.09% thuộc nhóm rủi ro trung bình, và 30.77% thuộc nhóm rủi ro thấp.

❖ Điểm mạnh:

- Việc phân loại rõ ràng mức độ rủi ro, giúp các chuyên gia y tế, bà mẹ mang thai và gia đình có thể nắm bắt được tình hình sức khỏe của người mẹ và thai nhi, từ đó đưa ra các biện pháp phòng ngừa, chăm sóc và theo dõi sức khỏe tốt hơn.

❖ Điểm yếu:

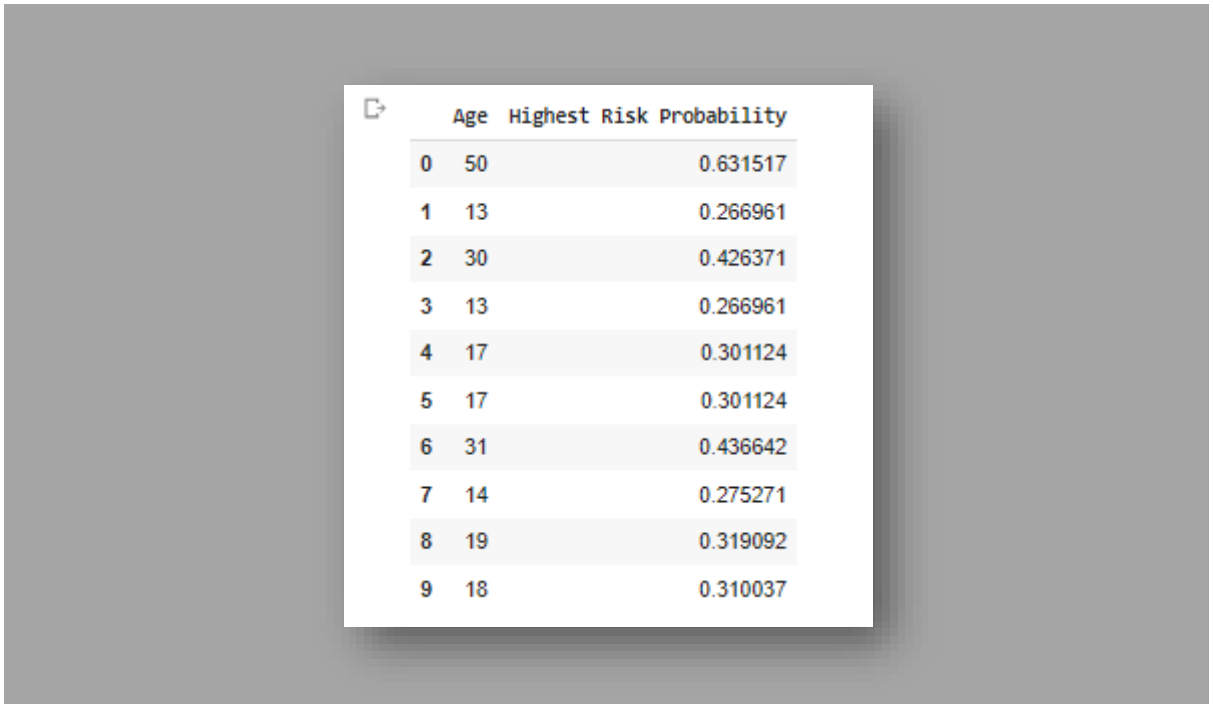
- Không đưa ra nguyên nhân dẫn đến mức độ rủi ro cao, trung bình hay thấp. Điều này làm cho việc đưa ra giải pháp cải thiện tình hình sức khỏe của bà mẹ và thai nhi trở nên khó khăn.
- Ngoài ra, kết quả trên chỉ xét một yếu tố duy nhất là tuổi của bà mẹ mang thai, trong khi đó có nhiều yếu tố khác cũng ảnh hưởng đến mức độ rủi ro trong quá trình mang thai như chế độ dinh dưỡng, lịch sử sức khỏe, môi trường sống, tâm lý, … Chúng ta cần xem xét đánh giá rủi ro dựa trên nhiều yếu tố hơn để có cái nhìn tổng quát hơn.

❖ Kết luận:

- Để đánh giá và giải quyết các điểm yếu nêu trên, chúng ta cần tiếp tục nghiên cứu và phân tích các yếu tố ảnh hưởng đến mức độ rủi ro của bà mẹ mang thai, đồng thời xây dựng các mô hình dự đoán và đề xuất các giải pháp dựa trên kết quả nghiên cứu.

4. **Câu hỏi "Dự đoán xác suất mức độ rủi ro (RiskLevel) cao nhất cho mỗi giá trị tuổi (Age)?"**

❖ Kết quả:



	Age	Highest Risk Probability
0	50	0.631517
1	13	0.266961
2	30	0.426371
3	13	0.266961
4	17	0.301124
5	17	0.301124
6	31	0.436642
7	14	0.275271
8	19	0.319092
9	18	0.310037

- Kết quả cho thấy mức độ rủi ro cao nhất (Highest Risk Probability) rất khác nhau đối với các độ tuổi. Người có độ tuổi 50 có xác suất rủi ro cao nhất là 0.631517, trong khi đó, người có độ tuổi 13 chỉ có xác suất rủi ro 0.266961. Đây là một thông tin quan trọng cho chúng ta để nhận thức được sự khác biệt về mức độ rủi ro đối với các độ tuổi.

❖ Điểm mạnh:

- Kết quả này giúp chúng ta hiểu rõ hơn về mối quan hệ giữa độ tuổi và mức độ rủi ro cao nhất. Điều này rất hữu ích khi đưa ra các biện pháp quản lý rủi ro hoặc đưa ra các chiến lược phòng ngừa, như đầu tư vào công tác huấn luyện, giáo dục, chăm sóc sức khỏe phù hợp với từng độ tuổi.
- Kết quả dễ hiểu và có thể trực quan hóa đơn giản giúp người dùng dễ dàng tiếp cận và phân tích kết quả nghiên cứu.

❖ Điểm yếu:

- Một số dữ liệu trùng lặp về độ tuổi (ví dụ, độ tuổi 13 và 17 xuất hiện 2 lần) có thể gây nhầm lẫn. Để giải quyết vấn đề này, chúng ta có thể tổng hợp và thống kê dữ liệu về tính chất phân bố của độ tuổi.
- Kết quả chỉ đưa ra mối quan hệ giữa độ tuổi và xác suất rủi ro cao nhất, nhưng chưa giải thích về các nguyên nhân và yếu tố gây ảnh hưởng đến xác suất này. Để có cái nhìn toàn diện hơn, chúng ta cần tiếp tục phân tích và tìm hiểu thêm các yếu tố liên quan đến mức độ rủi ro.
- Kết quả có thể không đại diện cho toàn bộ quần thể nếu như mẫu dữ liệu không đủ đa dạng hoặc lớn. Chúng ta cần phải xem xét vấn đề này để đánh giá tính chính xác và khả năng áp dụng kết quả này vào thực tế.

❖ Kết luận:

- Tóm lại, kết quả trên cung cấp cho chúng ta một cái nhìn về mối quan hệ giữa độ tuổi và xác suất rủi ro cao nhất. Chúng ta đã nêu ra một số điểm mạnh và điểm yếu của kết quả này, giúp chúng ta có những hướng điều chỉnh và cải tiến trong các nghiên cứu tiếp theo.

VIII. KẾT LUẬN:

Sau khi tiến hành khai phá tập dữ liệu Maternal Health Risk Data Set ta có kết luận như sau:

❖ Số phụ nữ mang thai trong độ tuổi 15-18 cao và ở độ tuổi 34 trở lên có rất ít. Huyết áp tâm thu của họ chủ yếu dao động từ 119-122 mmHg và huyết áp tâm trương dao động chủ yếu 80-82 mmHg. Phụ nữ mang thai có lượng đường huyết trong người khoảng từ 2-6 mmol/l và nhiệt độ cơ thể của họ luôn từ 98°F - 98.3°F. Nhịp tim trung bình của phụ nữ mang thai từ 75-79 (nhịp/phút).

❖ Những người mang thai ở dưới tuổi vị thành niên thì độ rủi ro sẽ rất cao ảnh hưởng không tốt đến sức khỏe và nguy cơ tử vong mẹ vẫn còn cao so với các bà mẹ sinh con ở tuổi trưởng thành.

❖ Tuy nhiên, những phụ nữ mang thai sau độ tuổi từ 35 cũng thuộc trường hợp rủi ro cao về sức khỏe như huyết áp tăng cao chỉ số huyết áp tâm thu và tâm trương có thể lớn hơn hoặc bằng 140/90mmHg, nguy cơ mắc các bệnh di truyền như Down, cũng như nguy cơ sẩy thai hay gặp phải các vấn đề sản khoa tăng lên theo tuổi của mẹ.

TÀI LIỆU THAM KHẢO

- Dey, I. (n.d.). Titanic Survival. Retrieved from Kaggle: **https://www.kaggle.com/code/indraneeldey/titanic-survival**
- Mirjalili, S. R. (n.d.). *Python Machine Learning.*
- P.Murphy, K. (n.d.). *Machine Learning: A Probabilistic Perspective.*