Maternal Health Risks Logistic Regression

By: Michelle Garcia

Step 1: Data Understanding

Independent Variables

- Age
- SystolicBP(Systolic Blood Pressure)
- DiastolicBP(Diastolic Blood Pressure)
- BS(Blood Sugar)
- BodyTemp
- HeartRate

Target/Dependent Variable

- Risk Level

Step 2: Data Preparation Cont...

Data has been collected from different hospitals, community clinics, maternal health cares from the rural areas of Bangladesh through the IoT based risk monitoring system.

	A	B	C	D	E	F	G	H
1	Age	SystolicBP	DiastolicB	BS	BodyTemp	HeartRate	RiskLevel	
2	25	130	80	15	98	86	high risk	
3	35	140	90	13	98	70	high risk	
4	29	90	70	8	100	80	high risk	
5	30	140	85	7	98	70	high risk	
6	35	120	60	6.1	98	76	low risk	
7	23	140	80	7.01	98	70	high risk	
8	23	130	70	7.01	98	78	mid risk	
9	35	85	60	11	102	86	high risk	
10	32	120	90	6.9	98	70	mid risk	
11	42	130	80	18	98	70	high risk	
12	23	90	60	7.01	98	76	low risk	
13	19	120	80	7	98	70	mid risk	
14	25	110	89	7.01	98	77	low risk	
15	20	120	75	7.01	100	70	mid risk	
16	48	120	80	11	98	88	mid risk	
17	15	120	80	7.01	98	70	low risk	
18	50	140	90	15	98	90	high risk	
19	25	140	100	7.01	98	80	high risk	
20	30	120	80	6.9	101	76	mid risk	
21	10	70	50	6.9	98	70	low risk	
22	40	140	100	18	98	90	high risk	
23	50	140	80	6.7	98	70	mid risk	
24	21	90	65	7.5	98	76	low risk	
25	18	90	60	7.5	98	70	low risk	
26	21	120	80	7.5	98	76	low risk	
27	16	100	70	7.2	98	80	low risk	

Step 2: Data Preparation Cont...

1015 observations, 7 variables

Step 2: Data Prep. Cont...

1. Make sure there no incorrect values

2. Check for outliers

3. Check for missing values

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

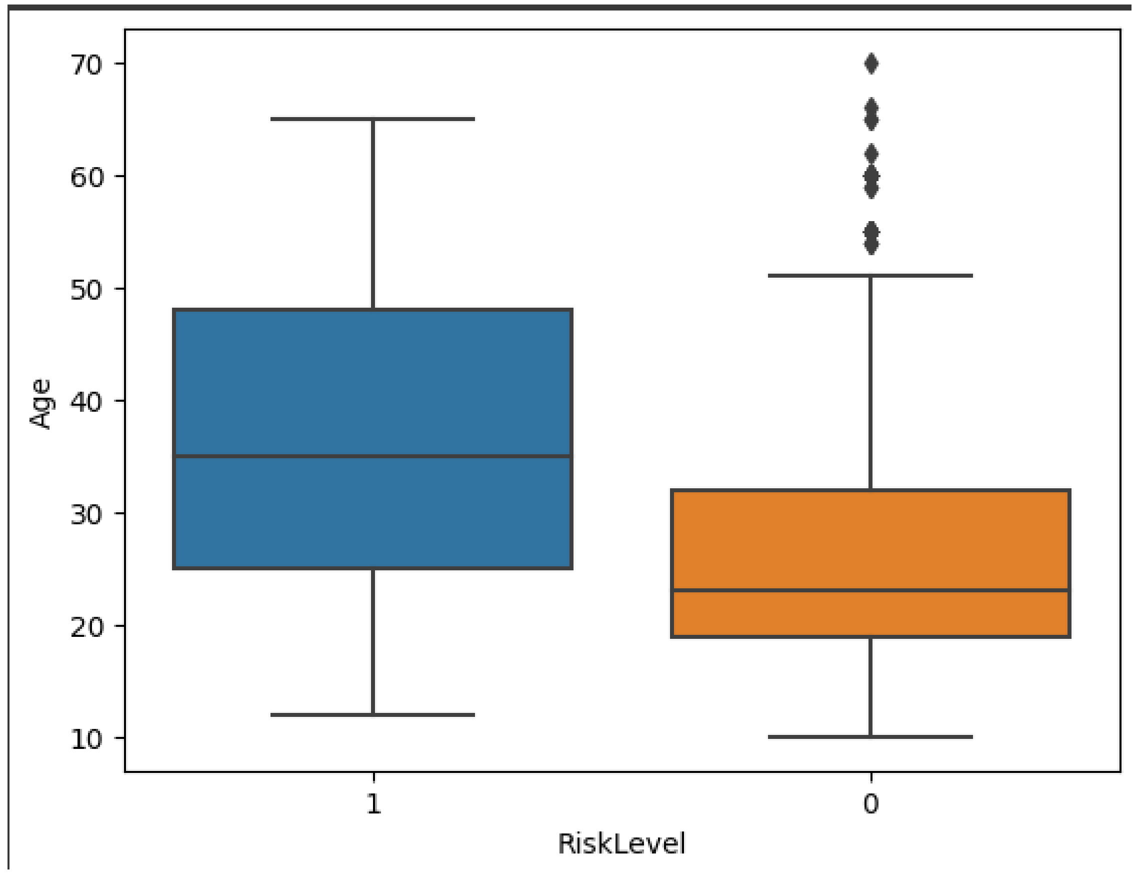
	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate
count	1014.000000	1014.000000	1014.000000	1014.000000	1014.000000	1014.000000
mean	29.871795	113.198225	76.460552	8.725986	98.665089	74.301775
std	13.474386	18.403913	13.885796	3.293532	1.371384	8.088702
min	10.000000	70.000000	49.000000	6.000000	98.000000	7.000000
25%	19.000000	100.000000	65.000000	6.900000	98.000000	70.000000
50%	26.000000	120.000000	80.000000	7.500000	98.000000	76.000000
75%	39.000000	120.000000	90.000000	8.000000	98.000000	80.000000
max	70.000000	160.000000	100.000000	19.000000	103.000000	90.000000

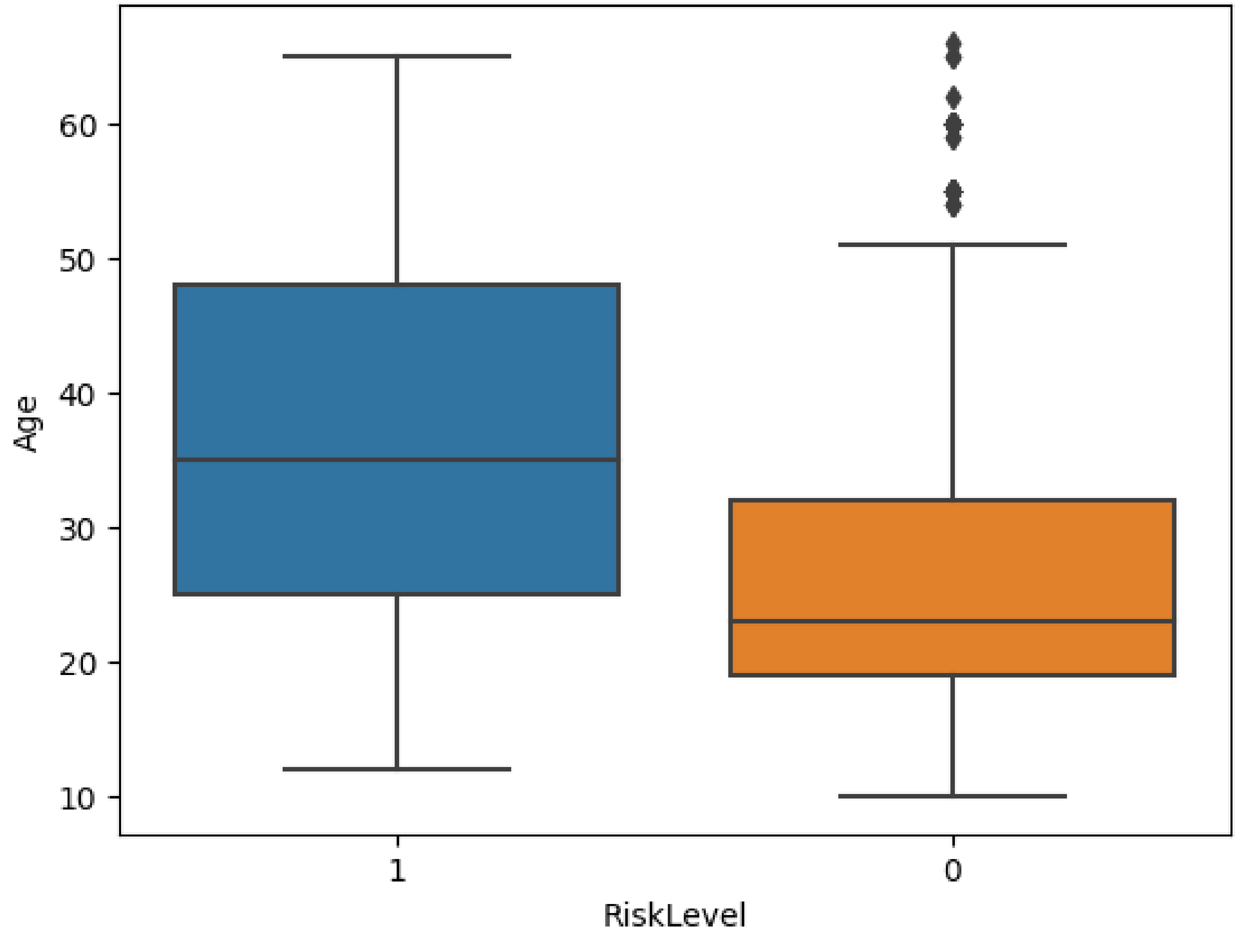
Step 2: Data Preparation Cont...

Fixing Outliers

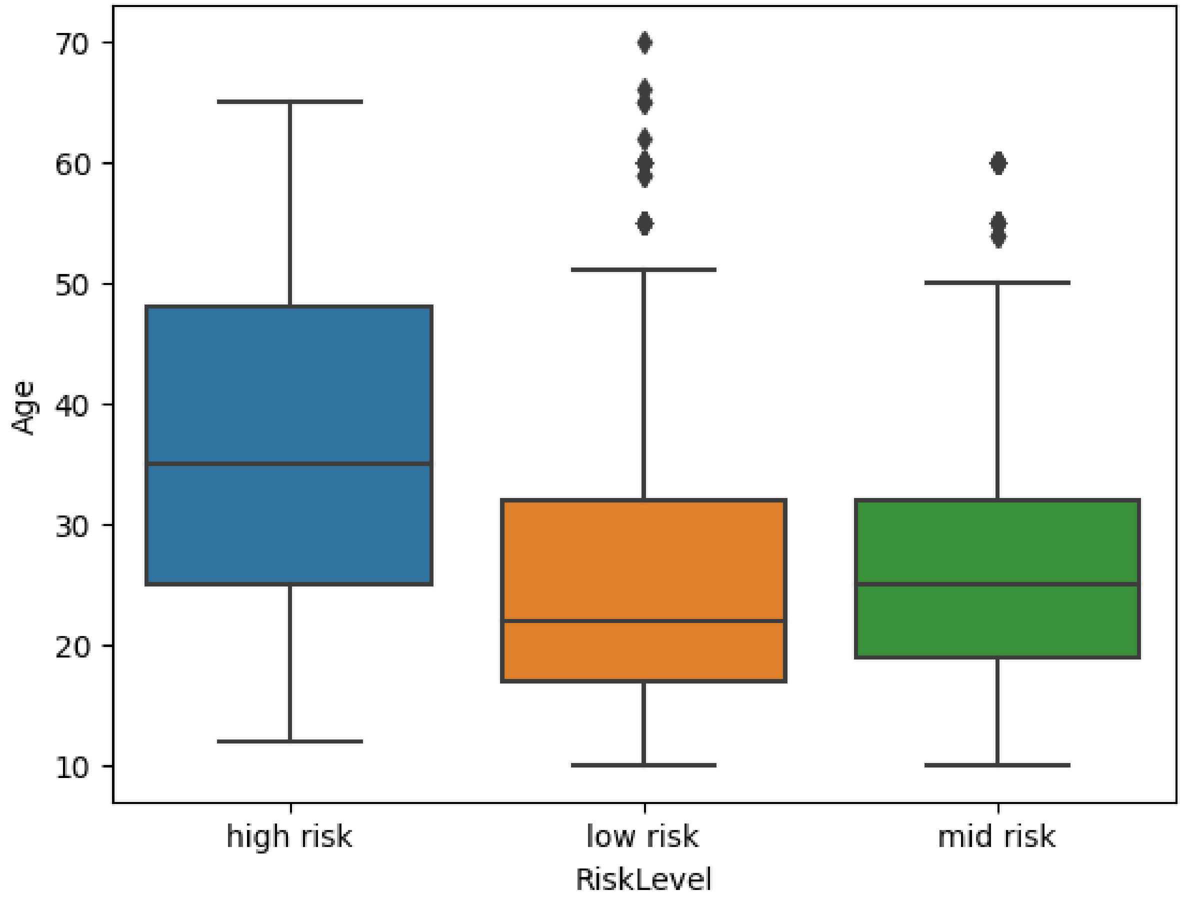
- After doing IQR there are still outliers

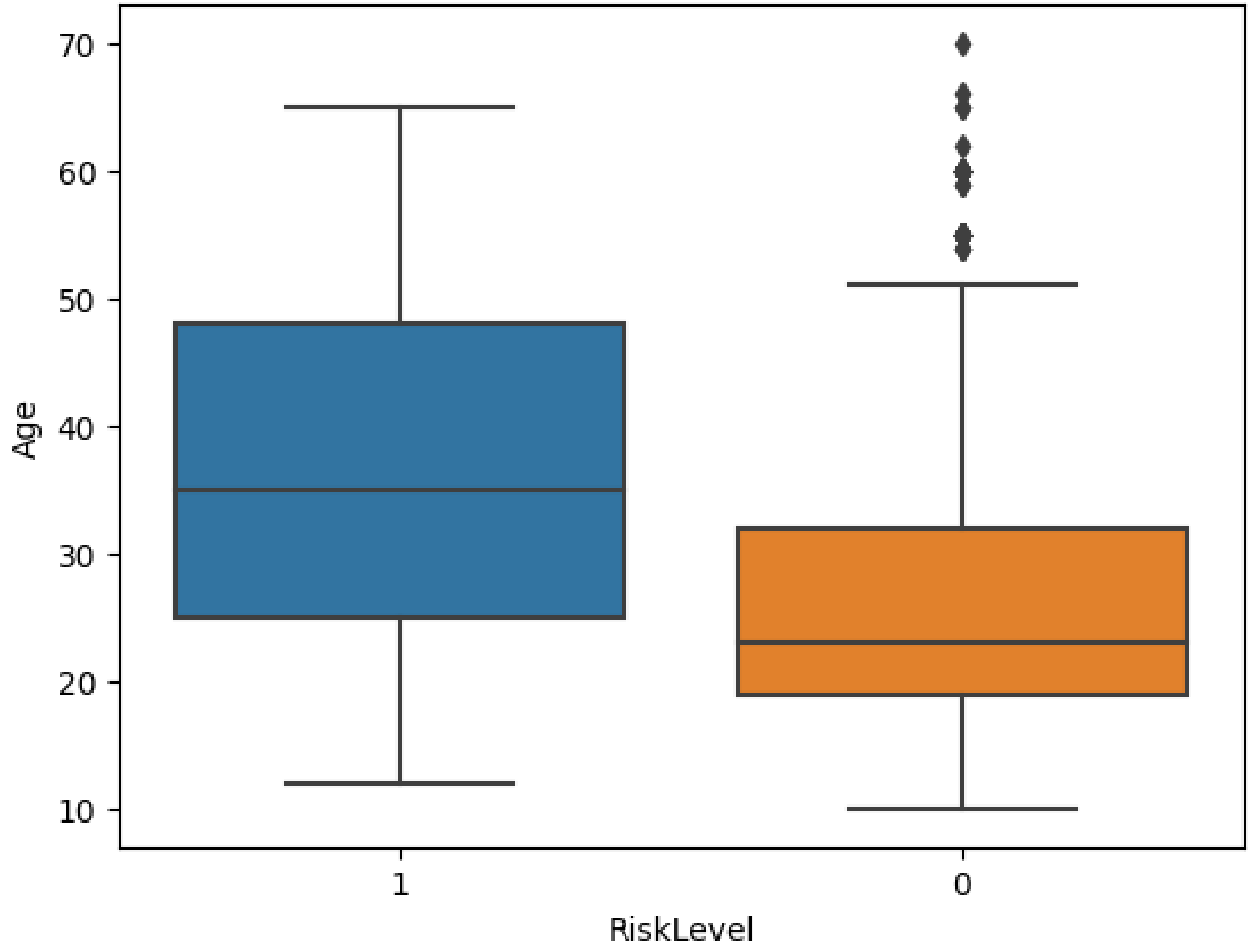
```
# Calculate the interquartile range (IQR)
q1 = MaternalRisk['Age'].quantile(0.25)
q3 = MaternalRisk['Age'].quantile(0.75)
iqr = q3 - q1
# Set a threshold for outliers (e.g., 1.5 times the IQR)
threshold = 1.5
lower_fence = q1 - threshold * iqr
upper_fence = q3 + threshold * iqr
MaternalRisk = MaternalRisk[(MaternalRisk['Age']>lower_fence)&(MaternalRisk['Age']<upper_fence)]
```





Step 2: Data Preparation Cont...





Fixing the Target Variable

- Combined low risk and mid risk data to low risk
- Converted high risk to 1 and low risk to 0

Step 3: Modeling

1. Decision Tree
2. Logistic Regression
3. Linear Regression

Step 3 : Modeling

Building the Model 1

```
# Building first model
MaternalRisk_mod1 = MaternalRisk[['Age', 'SystolicBP', 'DiastolicBP', 'BS', 'BodyTemp', 'RiskLevel']]
X = MaternalRisk_mod1.drop('RiskLevel', axis=1)
#Standardize the feature
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
y = MaternalRisk_mod1['RiskLevel']
# Split the data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.3,random_state=42)
# Initialize the logistic regression model
logreg_mod1 = LogisticRegression()
# Fit the model to the training data
logreg_mod1.fit(X_train, y_train)
```

```
▾ LogisticRegression
LogisticRegression()
```

The Coefficients

```
coef_df = pd.DataFrame({'Features': X.columns, 'Coefficient': logreg_mod1.coef_[0]})
print(coef_df)
```

```
      Features  Coefficient
0          Age    -0.010325
1   SystolicBP     0.867847
2  DiastolicBP     0.577936
3           BS     0.635254
4     BodyTemp     1.005752
```

AUROC & Accuracy

```
# Predictions on the test data
y_pred = logreg_mod1.predict(X_test)
# Calculating accuracy
accuracy = accuracy_score(y_test, y_pred)
print(f'Accuracy: {accuracy}')
# Calculating area under receiver operating characteristic curve
y_pred_prob = logreg_mod1.predict_proba(X_test)[:, 1]
auroc = roc_auc_score(y_test, y_pred_prob)
print(f'Area under receiver operating characteric curve (AUROC): {auroc}')
```

```
Accuracy: 0.8953974895397490
Area under receiver operating characteric curve (AUROC): 0.827775897088693
```

Step 3 : Modeling

Building the Model 2

```
# Building second model
MaternalRisk_mod2 = MaternalRisk[['Age', 'SystolicBP', 'DiastolicBP', 'BS', 'RiskLevel']]
X = MaternalRisk_mod2.drop('RiskLevel', axis=1)
#Standardize the feature
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
y = MaternalRisk_mod2['RiskLevel']
# Split the data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.3,random_state=42)
# Initialize the logistic regression model
logreg_mod2 = LogisticRegression()
# Fit the model to the training data
logreg_mod2.fit(X_train, y_train)
```

```
▾ LogisticRegression
LogisticRegression()
```

The Coefficients

```
coef_df = pd.DataFrame({'Features': X.columns, 'Coefficient': logreg_mod2.coef_[0]})
print(coef_df)
```

```
      Features  Coefficient
0          Age    -0.084171
1   SystolicBP     0.573960
2  DiastolicBP     0.298609
3           BS     0.669433
```

AUROC & Accuracy

```
# Predictions on the test data
y_pred = logreg_mod2.predict(X_test)
# Calculating accuracy
accuracy = accuracy_score(y_test, y_pred)
print(f'Accuracy: {accuracy}')
# Calculating area under receiver operating characteristic curve
y_pred_prob = logreg_mod2.predict_proba(X_test)[:, 1]
auroc = roc_auc_score(y_test, y_pred_prob)
print(f'Area under receiver operating characteric curve (AUROC): {auroc}')
```

```
Accuracy: 0.899581589958159
Area under receiver operating characteric curve (AUROC): 0.7219871360866621
```

Step 3: Modeling

Logit(P(RiskLevel))=intercept+(.86*SystolicBP)+(.57*DiastolicBP)+(.63*BS)+(1*BodyTemp)

P (RiskLevel) = (e^(RiskLevel)/1 + e^(RiskLevel))

Step 4: Data Insights

Model 1

- Model 1 indicates that higher SystolicBP, DiastolicBP, BS, and BodyTemp are positively associated, while higher Age is negatively associated, resulting in a model with an accuracy of 89.54% and an AUROC of 82.78%.

Model 2

- Model 2 indictaes that higher SystolicBP, DiastolicBP, and BS are positively associated, while higher Age is negatively associated, resulting in a model with an accuracy of 89.96% and an AUROC of 72.20%.

Thank you!