1. Data Normalization

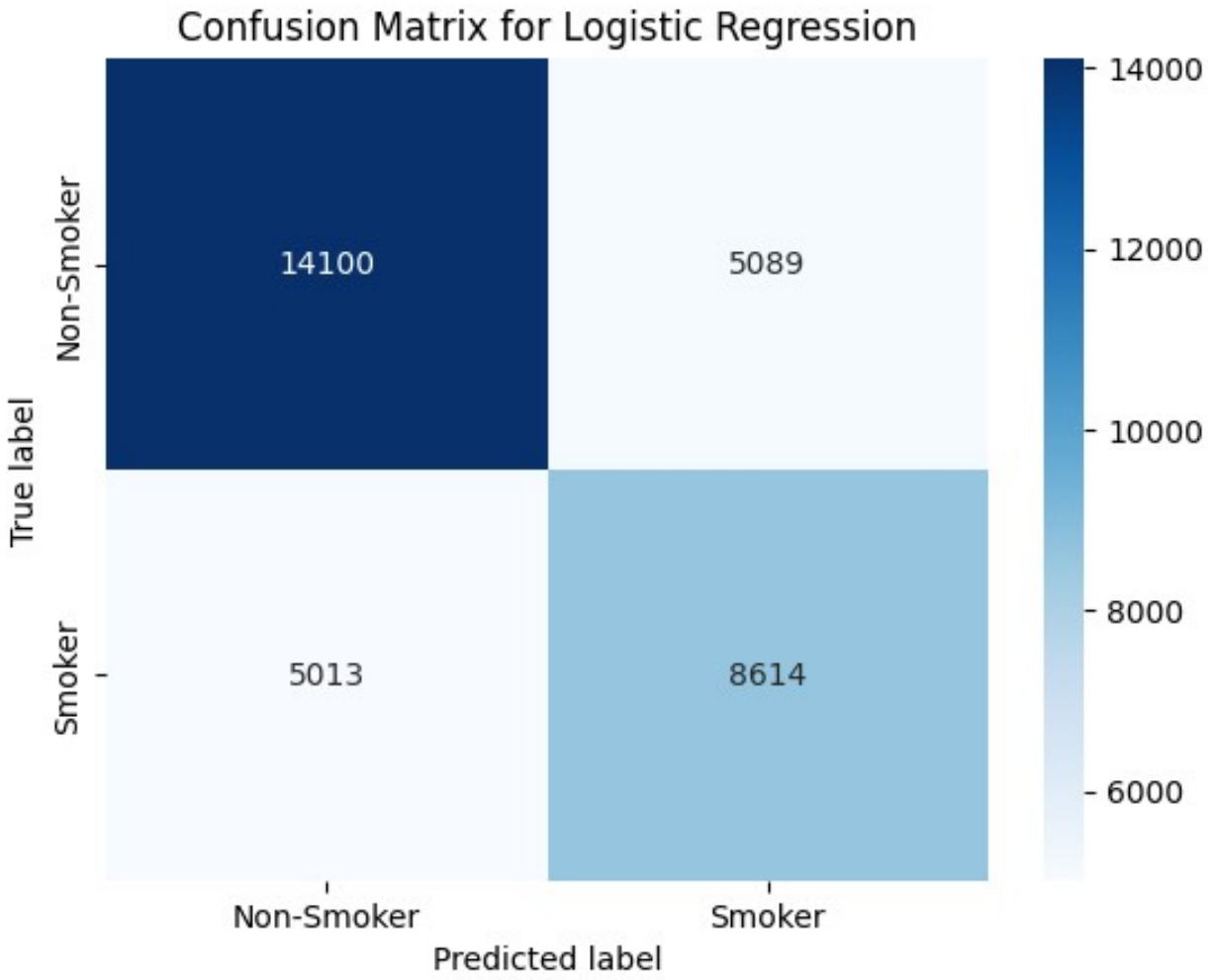
- transforms data into a standard scale without distorting differences in ranges.
- ensures all features contribute equally to the model.
- Common techniques include:
  - Min-Max Scaling: Scales features to a [0,1] range.
  - Standardization: Centers data by subtracting the mean and dividing by the standard deviation.

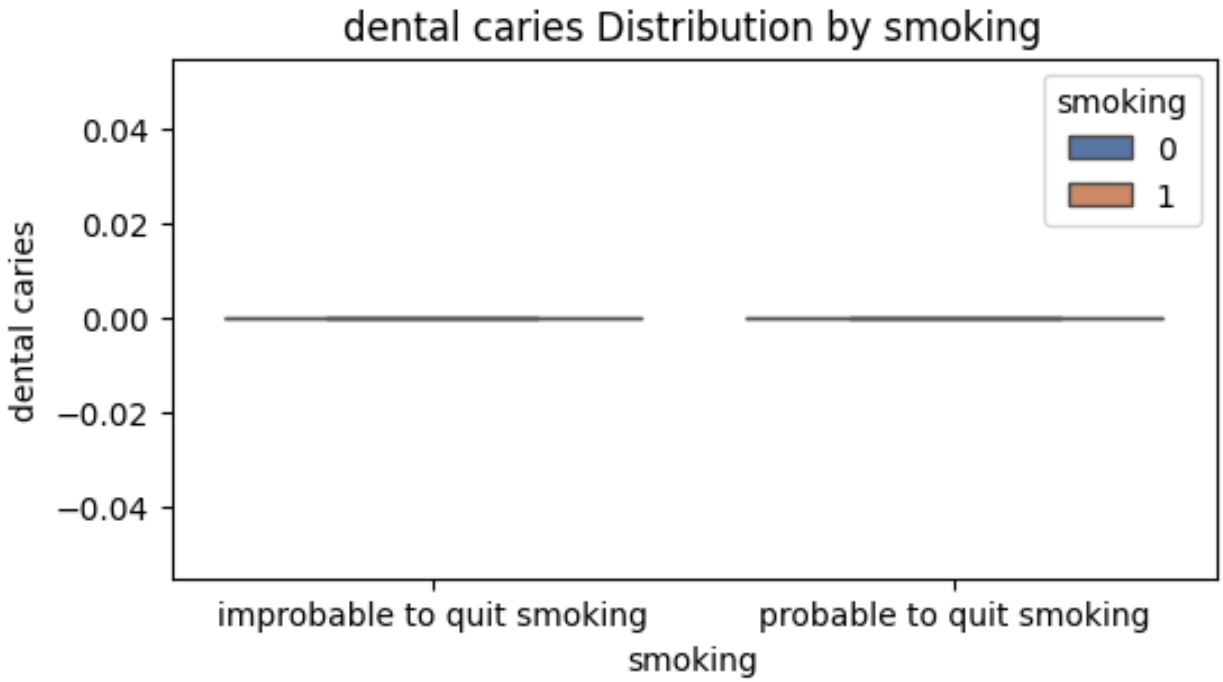
StandardScaler is used here to standardize data, making it suitable for machine learning algorithms sensitive to feature magnitudes.

2.Visualization and Data Analysis

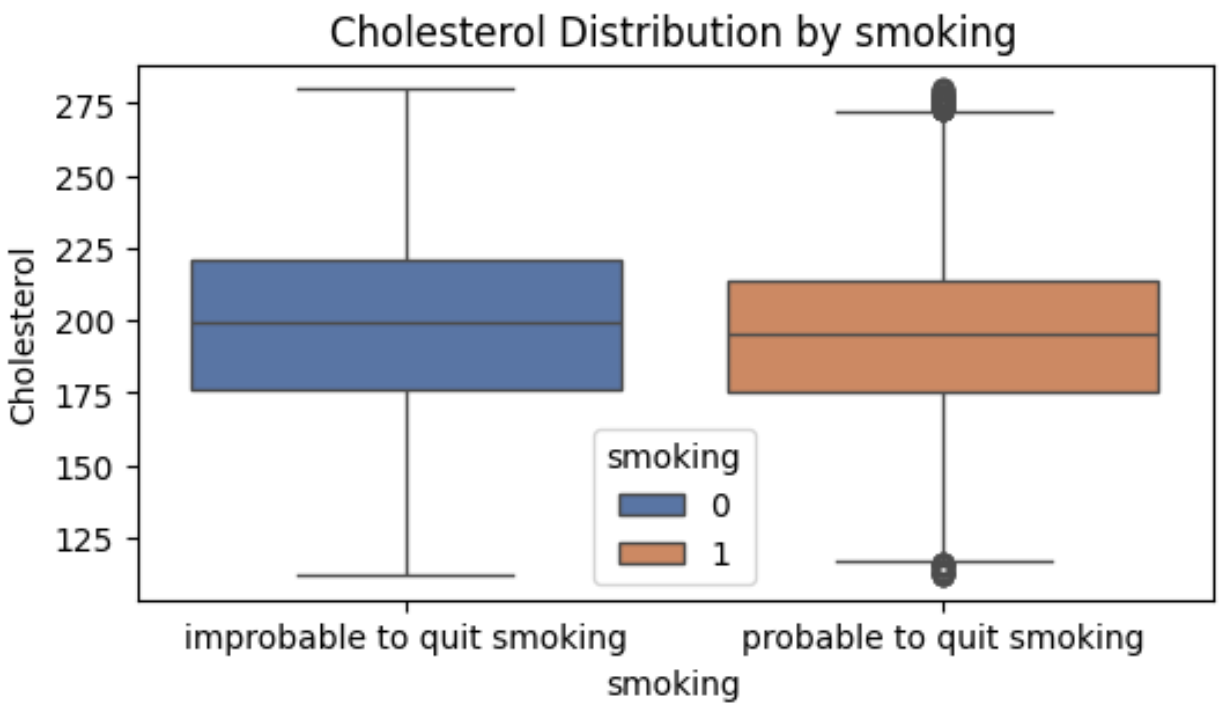
Univariate:

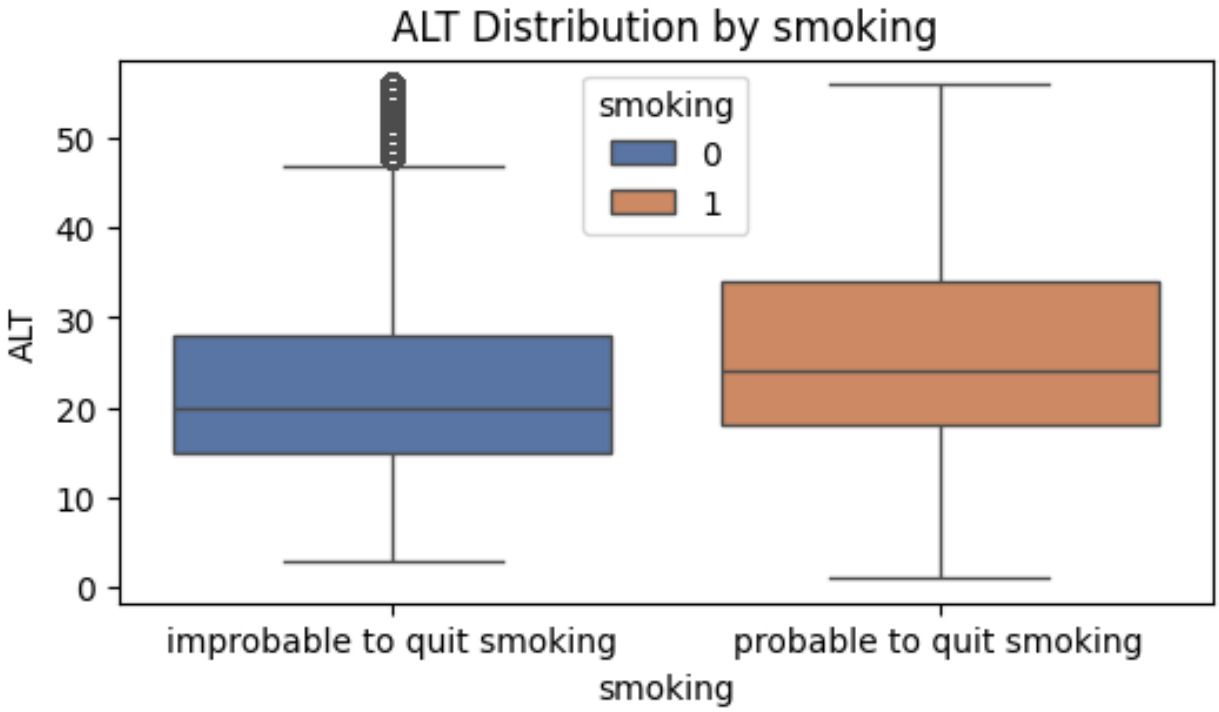

Confusion Matrix for Logistic Regression
14013
5176
4950
8677
Non-Smoker
Smoker
True label
Non-Smoker
Smoker
Predicted label
14000
12000
10000
8000
6000

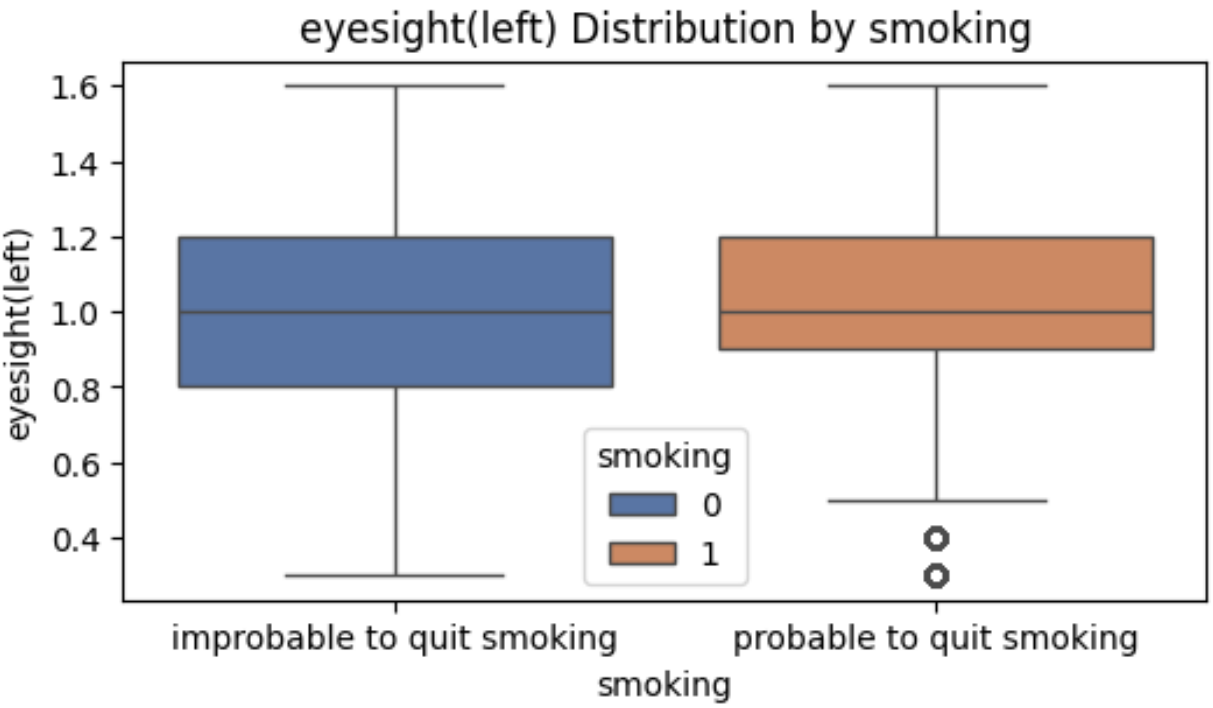

Confusion Matrix for Logistic Regression
14100
5089
5013
8614
Non-Smoker
Smoker
True label
Non-Smoker
Smoker
Predicted label
14000
12000
10000
8000
6000

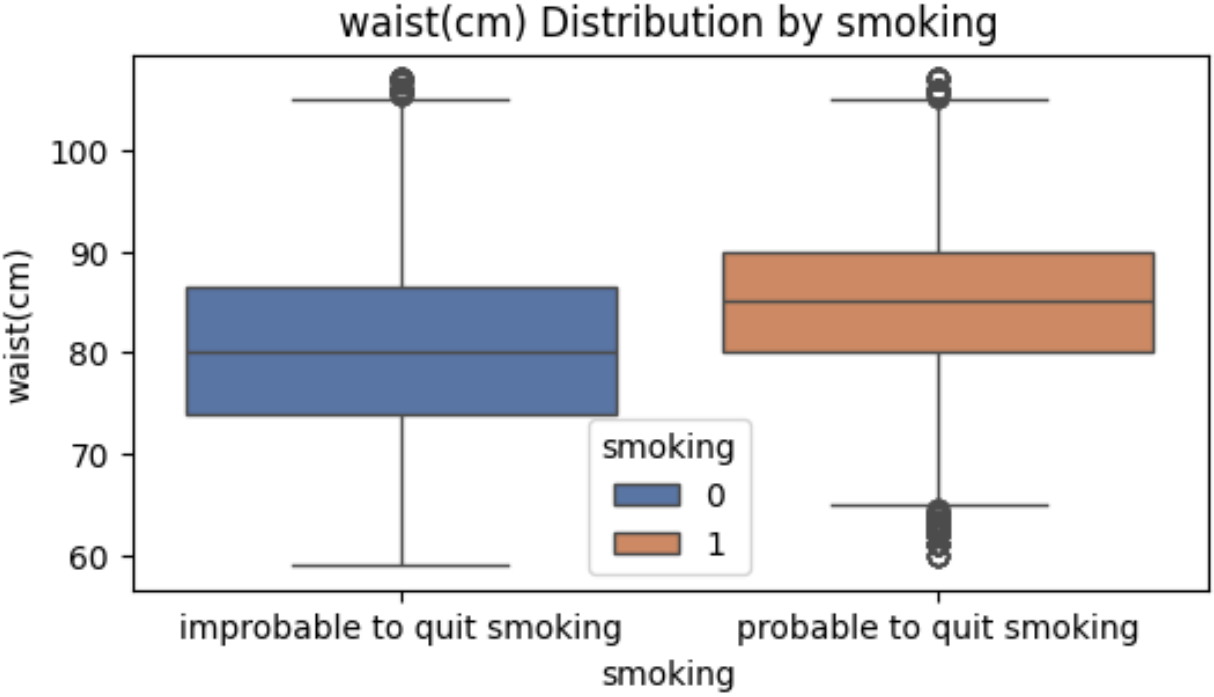


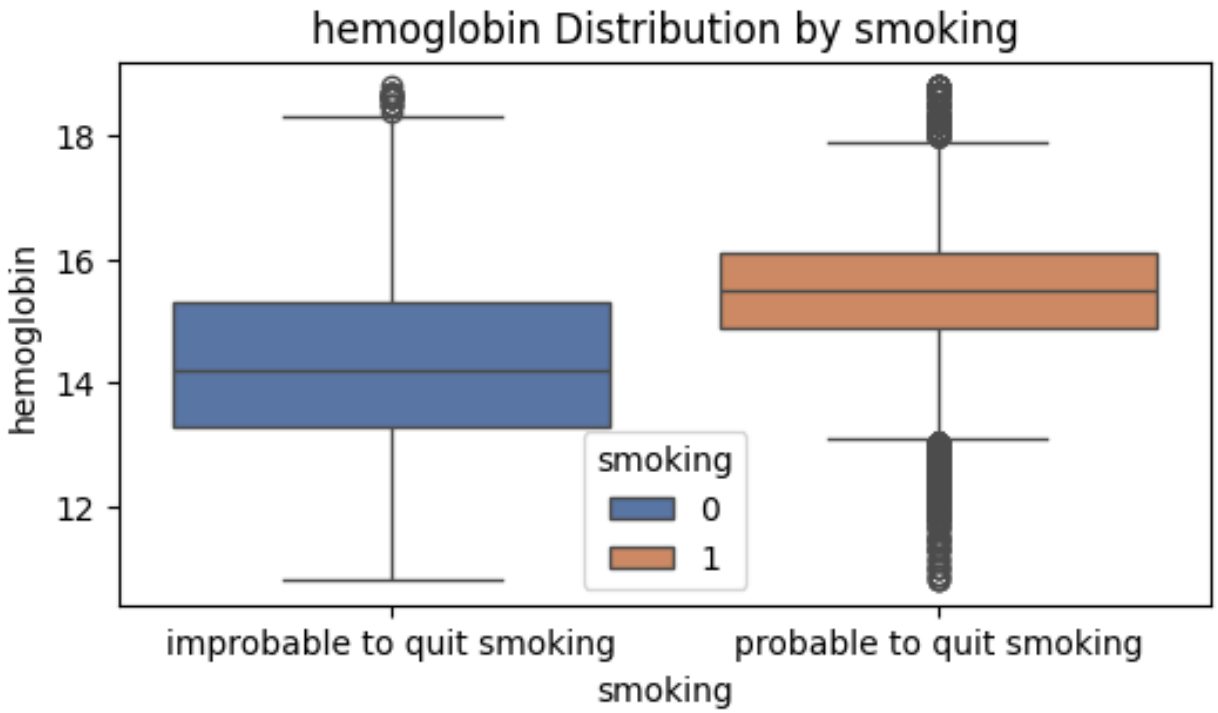
- The distribution for these 3 graphs shows a uniform spread across the range, indicating consistent representation without notable skewness or clustering.

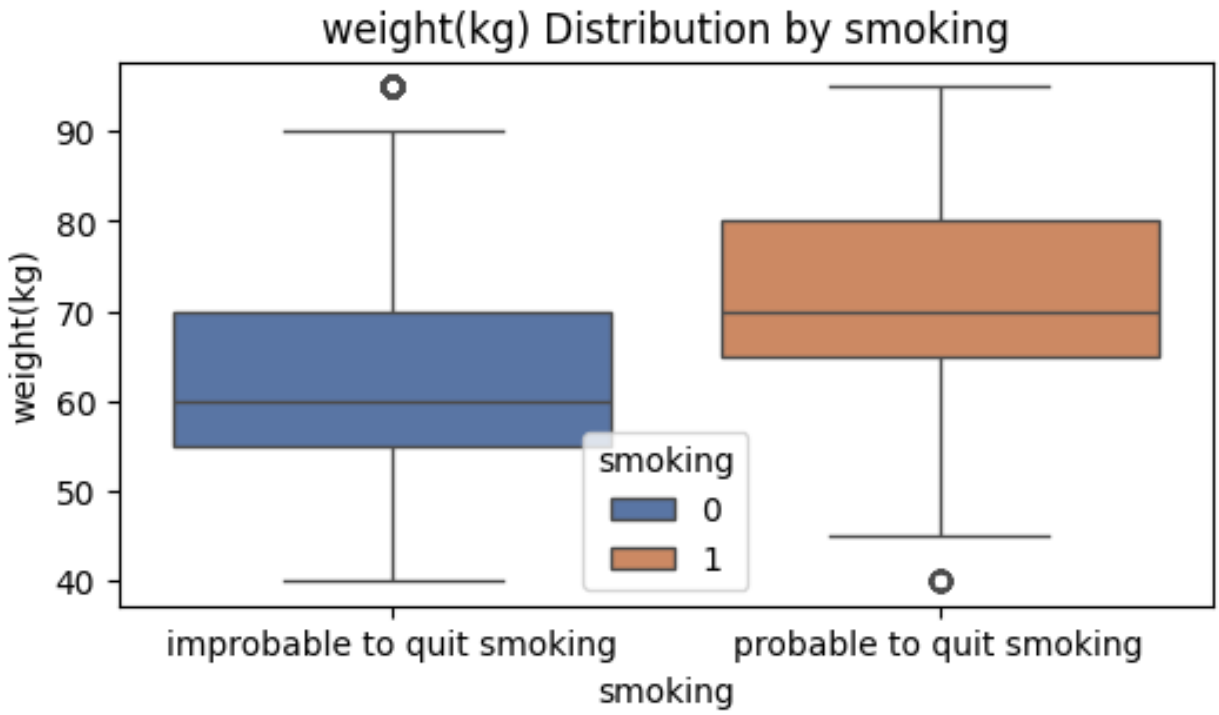

Cholesterol Distribution by smoking
Cholesterol
275
250
225
200
175
150
125
smoking
0
1
improbable to quit smoking
probable to quit smoking
smoking

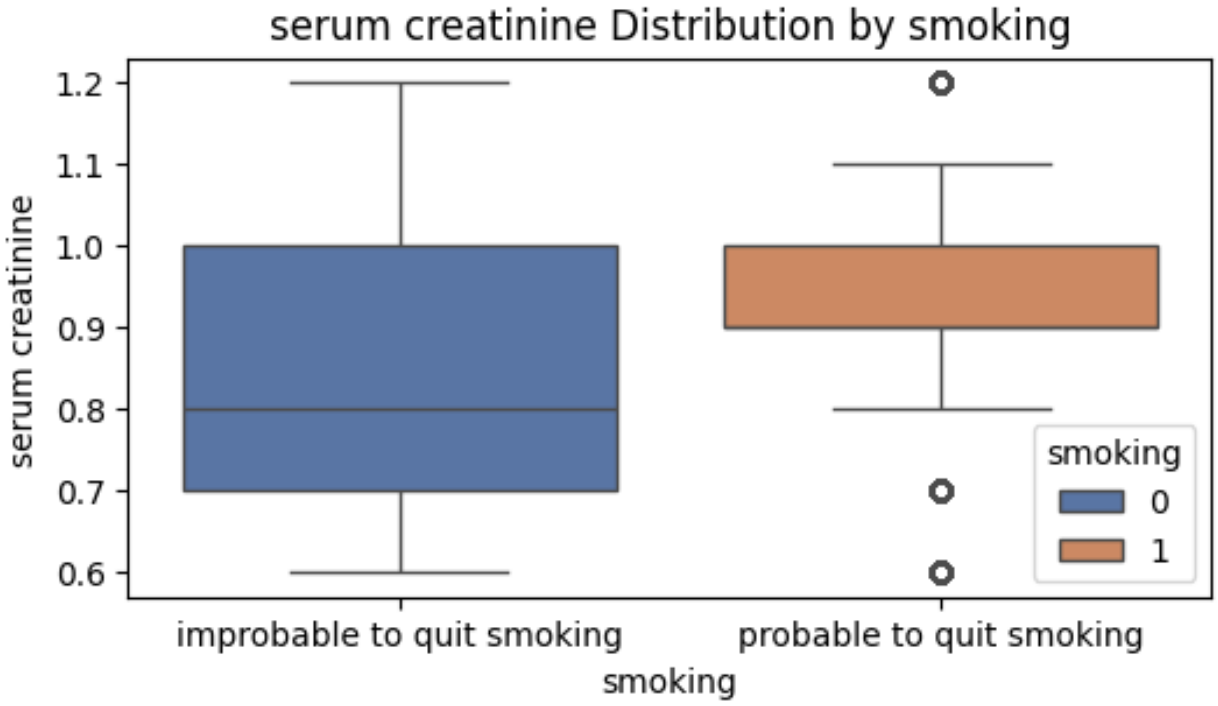

ALT Distribution by smoking
ALT
50
40
30
20
10
0
smoking
0
1
improbable to quit smoking
probable to quit smoking
smoking


eyesight(left) Distribution by smoking
eyesight(left)
1.6
1.4
1.2
1.0
0.8
0.6
0.4
smoking
0
1
improbable to quit smoking
probable to quit smoking
smoking


waist(cm) Distribution by smoking
waist(cm)
100
90
80
70
60
smoking
0
1
improbable to quit smoking
probable to quit smoking
smoking

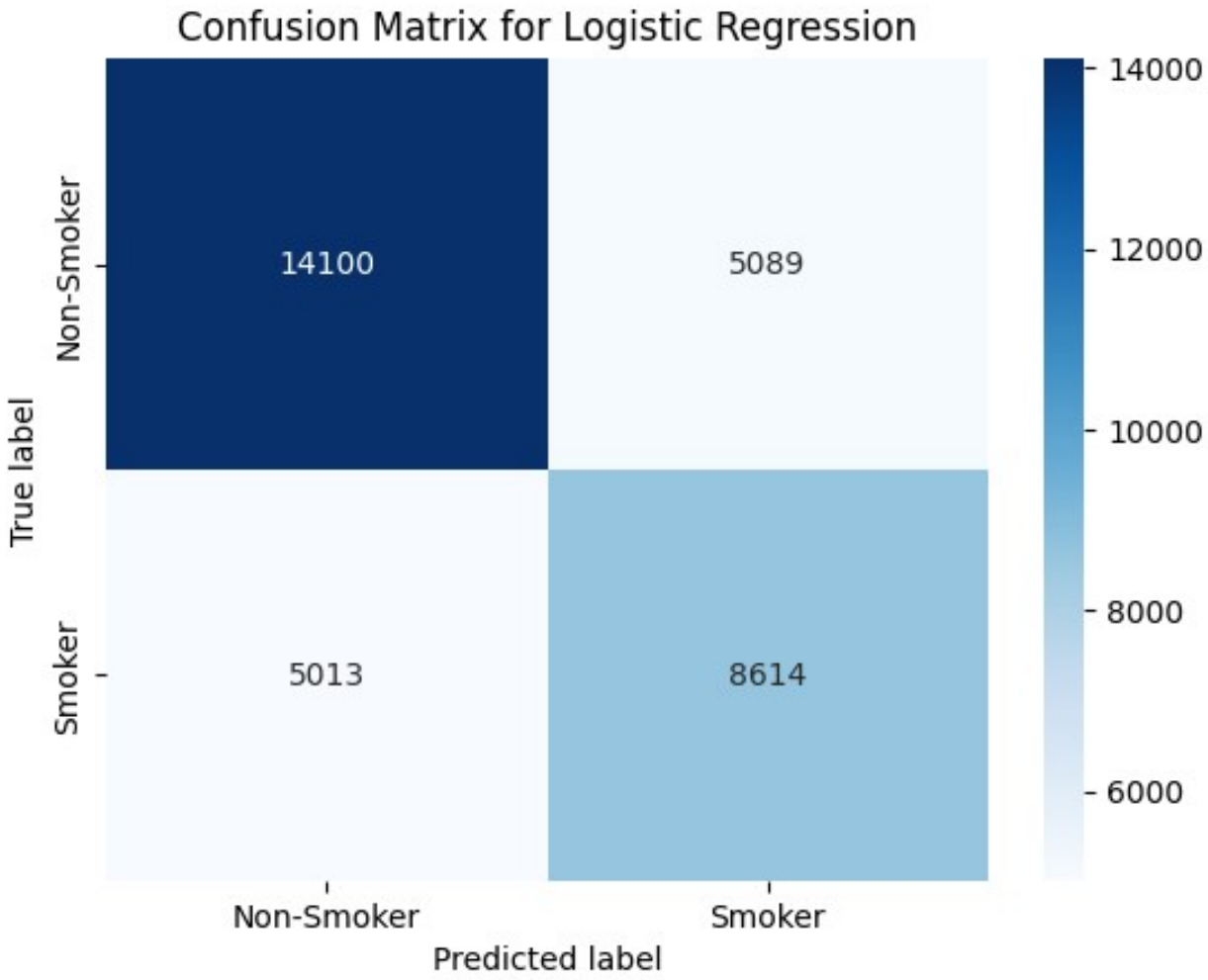

hemoglobin Distribution by smoking
hemoglobin
18
16
14
12
smoking
0
1
improbable to quit smoking
probable to quit smoking
smoking

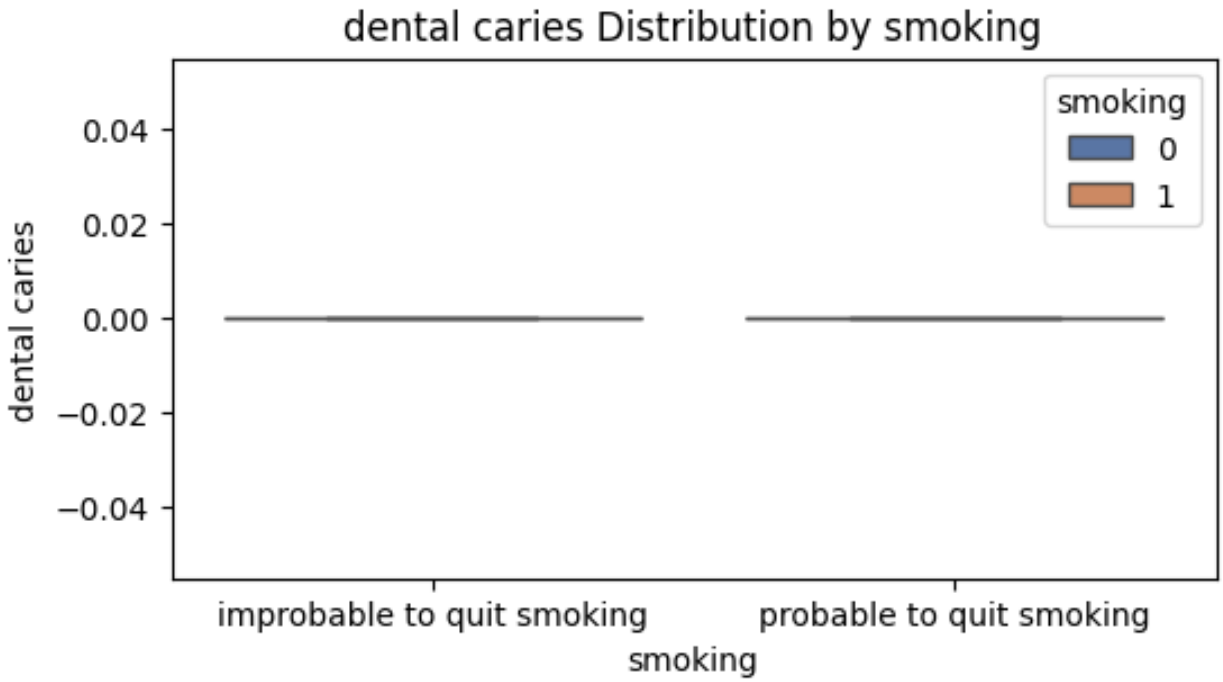




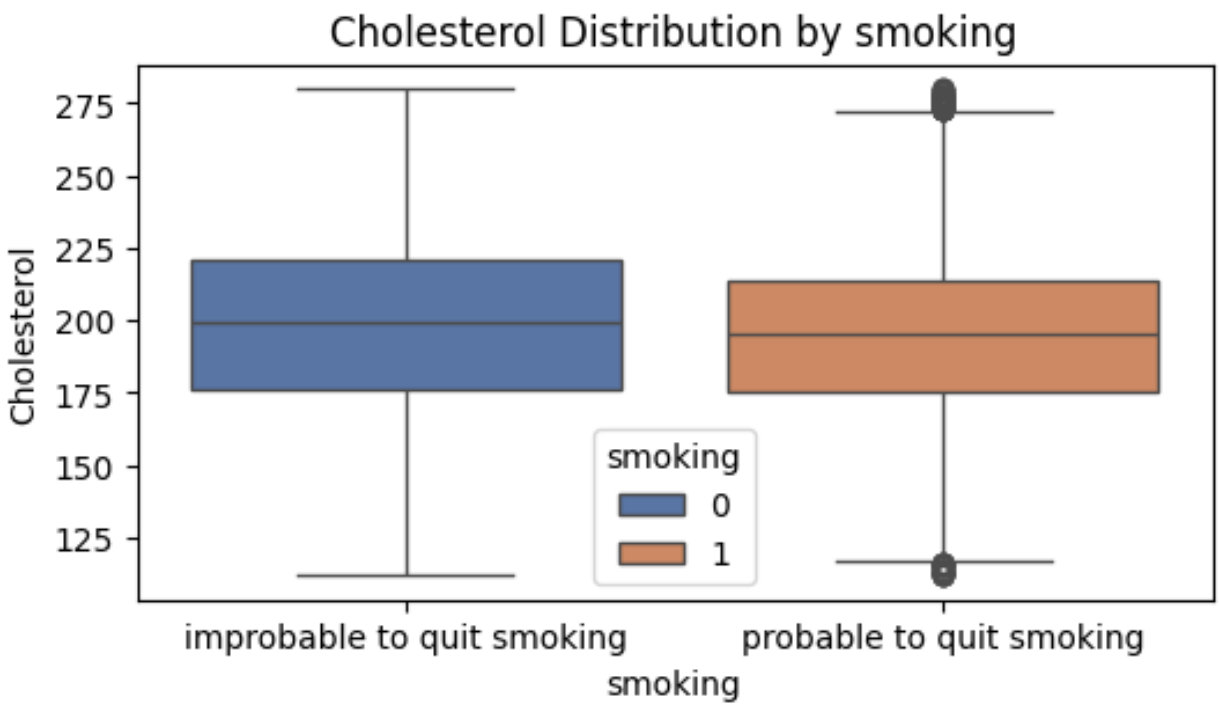
- These graphs are close to a gaussian distribution and most of them have almost zero mean.

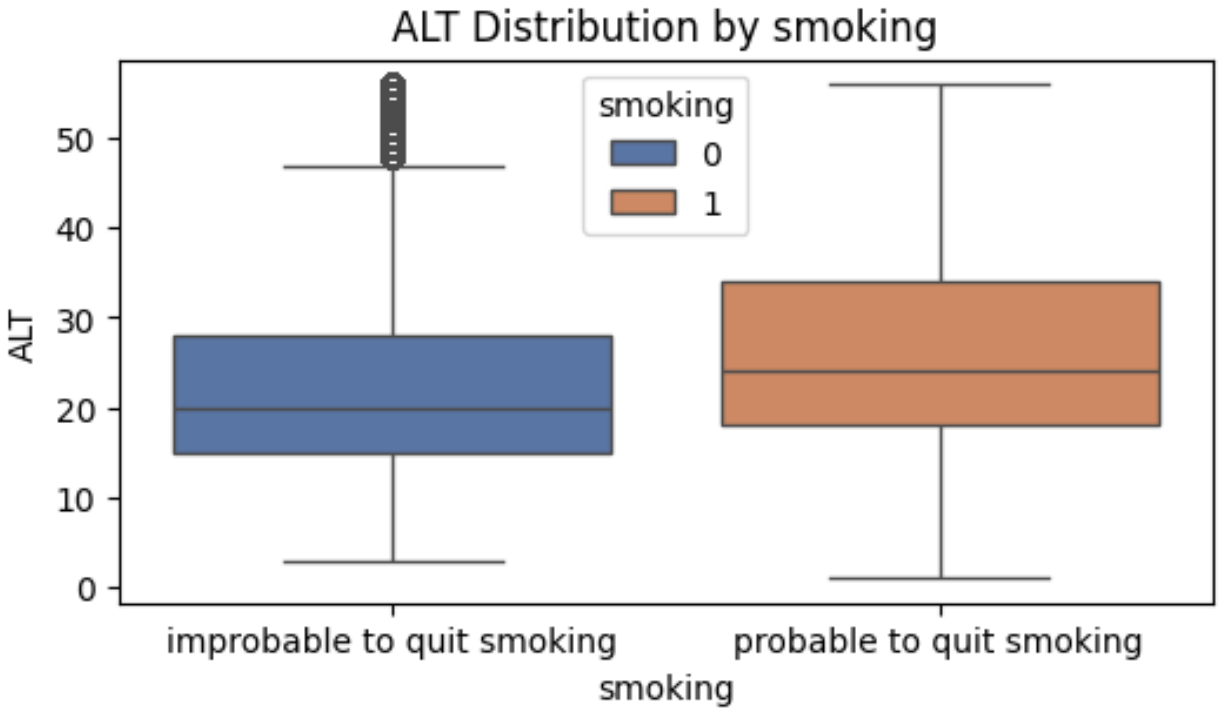

Confusion Matrix for Logistic Regression
14013
5176
4950
8677
Non-Smoker
Smoker
True label
Non-Smoker
Smoker
Predicted label
14000
12000
10000
8000
6000

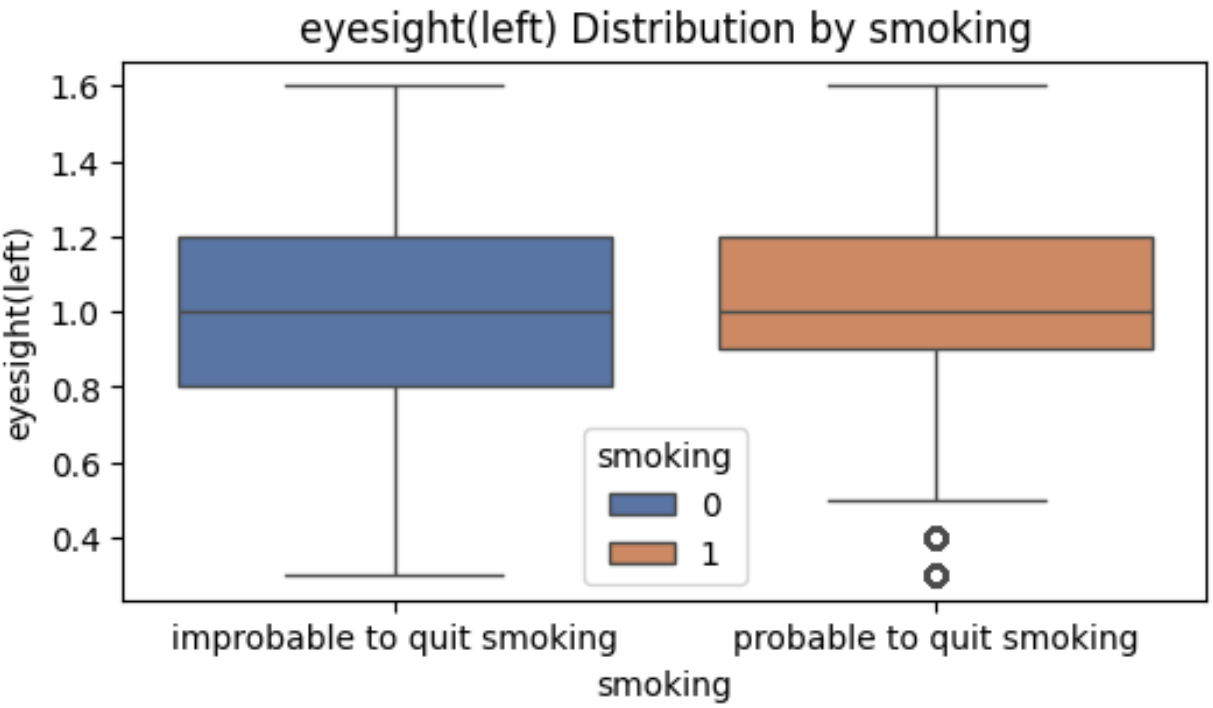

Confusion Matrix for Logistic Regression
14100
5089
5013
8614
Non-Smoker
Smoker
True label
Non-Smoker
Smoker
Predicted label
14000
12000
10000
8000
6000

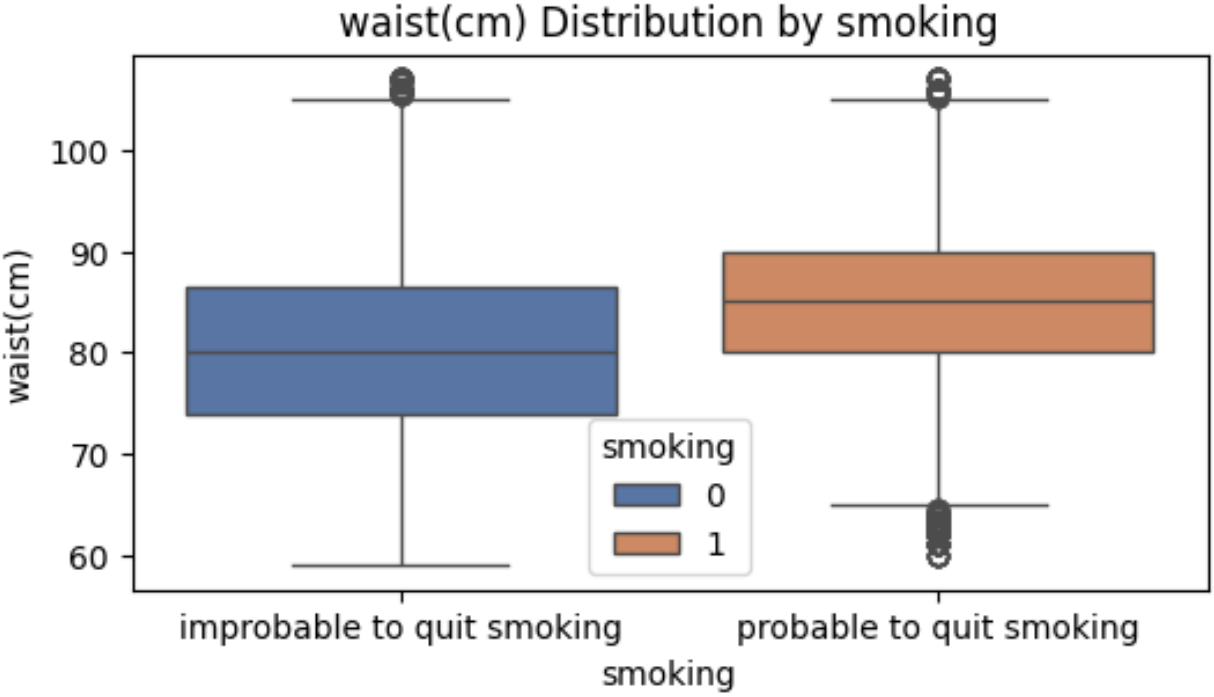


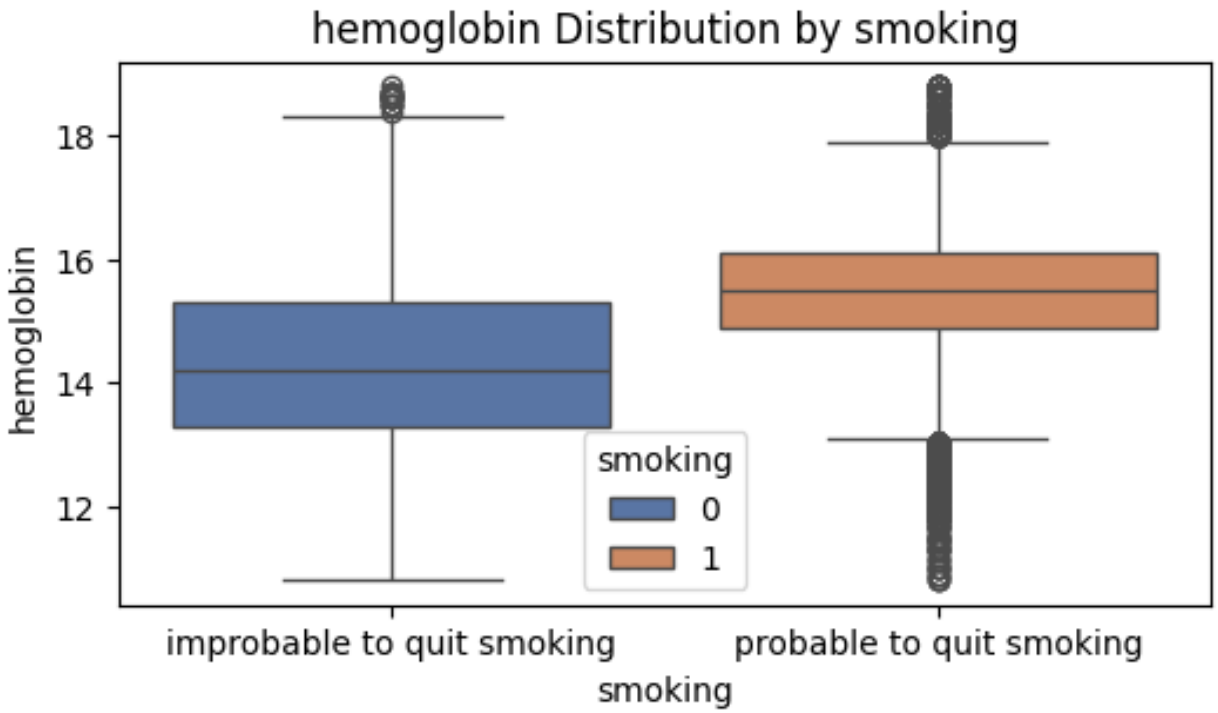
- Those graphs show uniform distribution indicating no relation to smoking

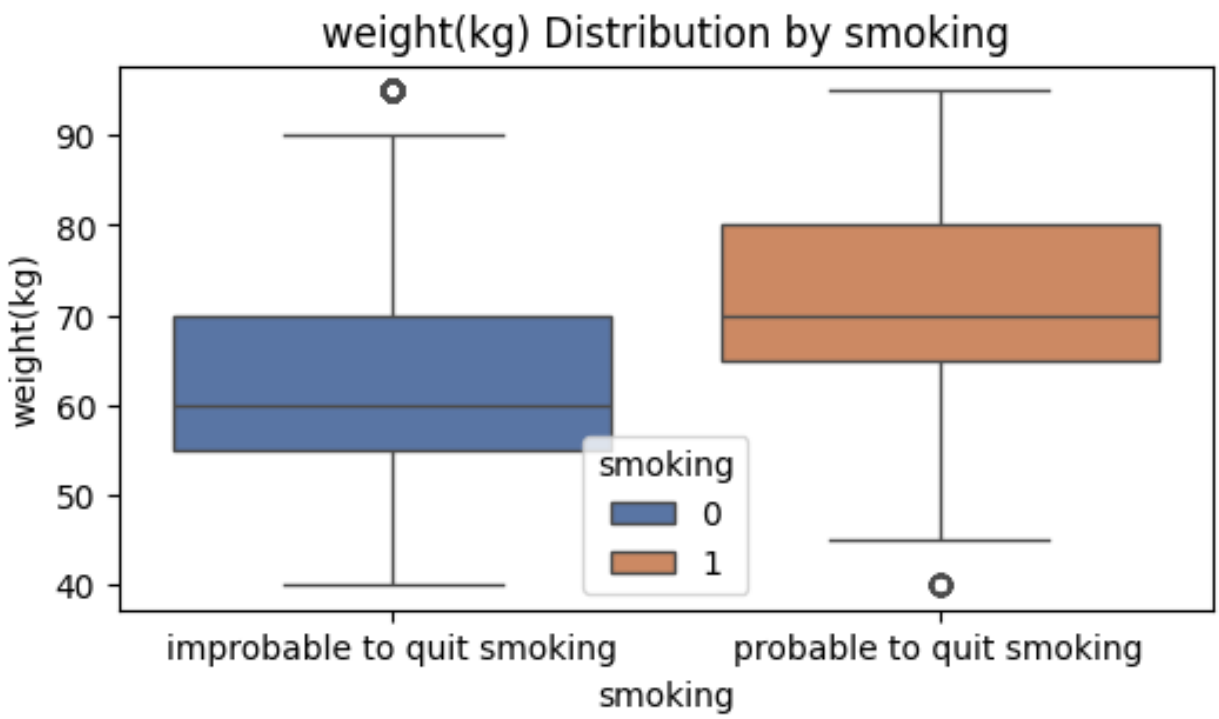

Cholesterol Distribution by smoking
Cholesterol
275
250
225
200
175
150
125
smoking
0
1
improbable to quit smoking
probable to quit smoking
smoking

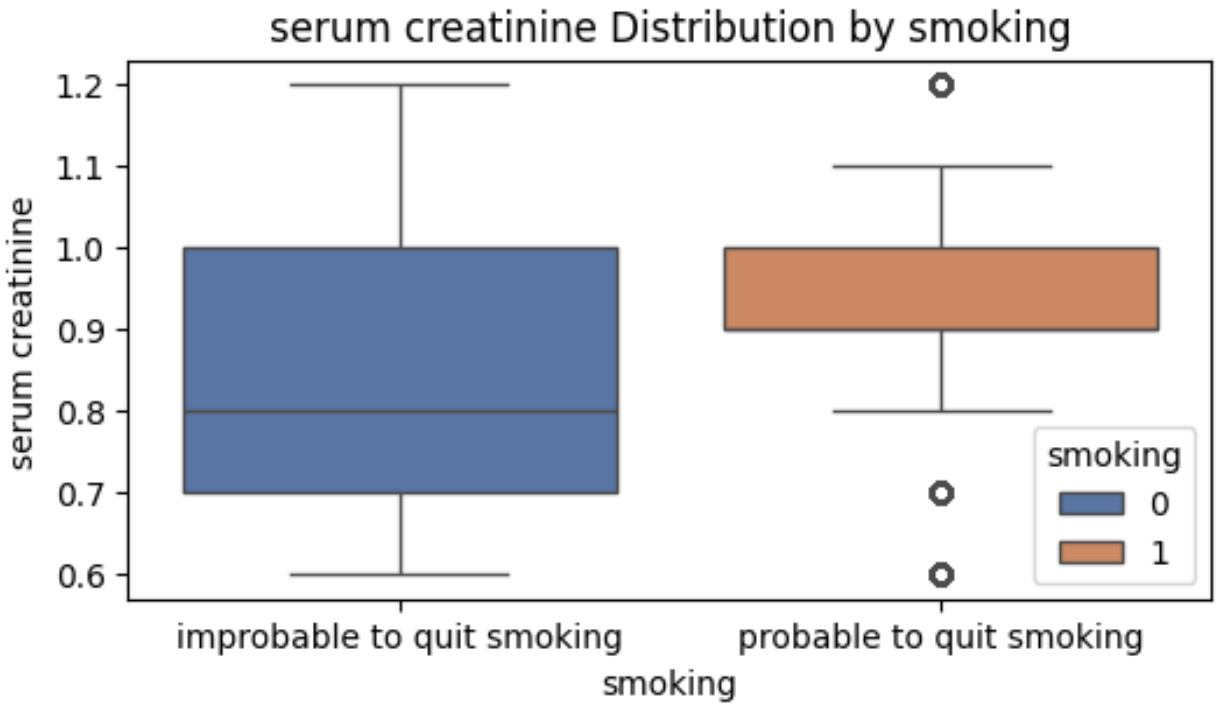

ALT Distribution by smoking
ALT
50
40
30
20
10
0
smoking
0
1
improbable to quit smoking
probable to quit smoking
smoking


eyesight(left) Distribution by smoking
eyesight(left)
1.6
1.4
1.2
1.0
0.8
0.6
0.4
smoking
0
1
improbable to quit smoking
probable to quit smoking
smoking

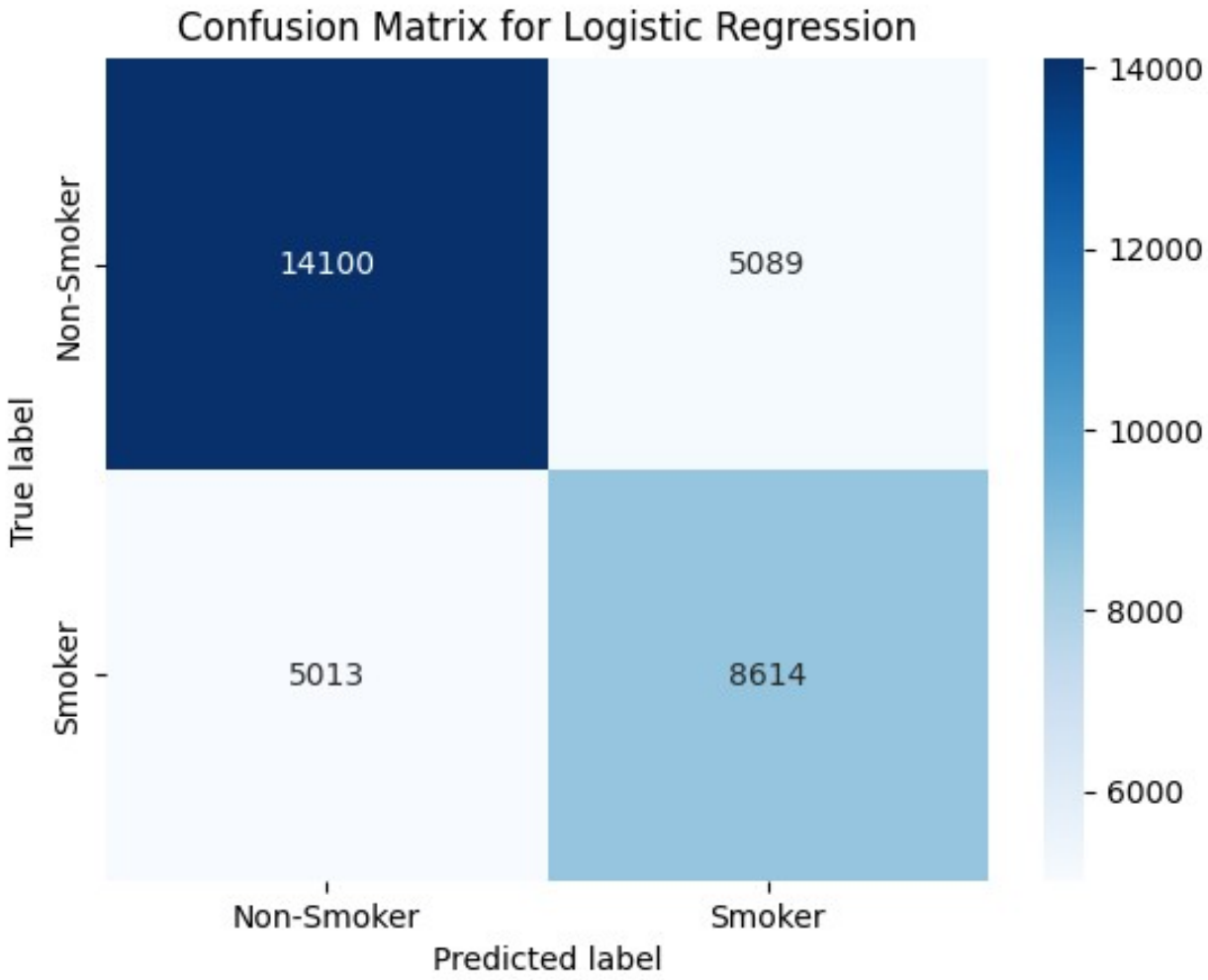

waist(cm) Distribution by smoking
waist(cm)
100
90
80
70
60
smoking
0
1
improbable to quit smoking
probable to quit smoking
smoking

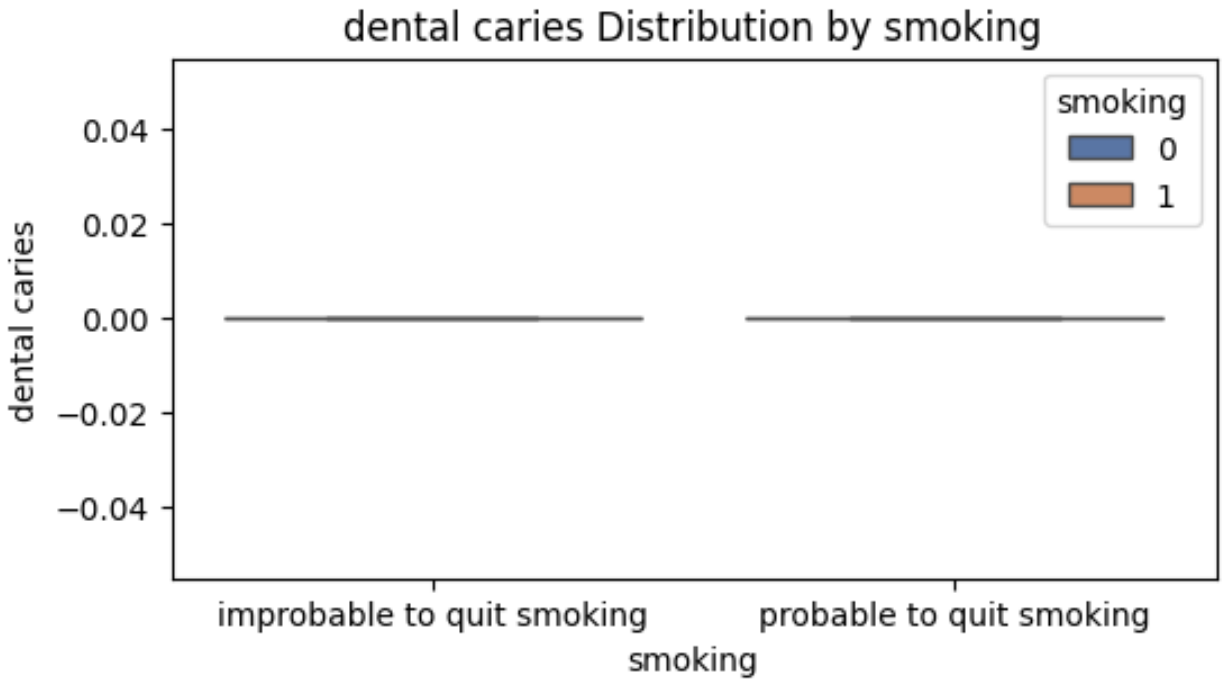

hemoglobin Distribution by smoking
hemoglobin
18
16
14
12
smoking
0
1
improbable to quit smoking
probable to quit smoking
smoking


weight(kg) Distribution by smoking
weight(kg)
90
80
70
60
50
40
smoking
0
1
improbable to quit smoking
probable to quit smoking
smoking

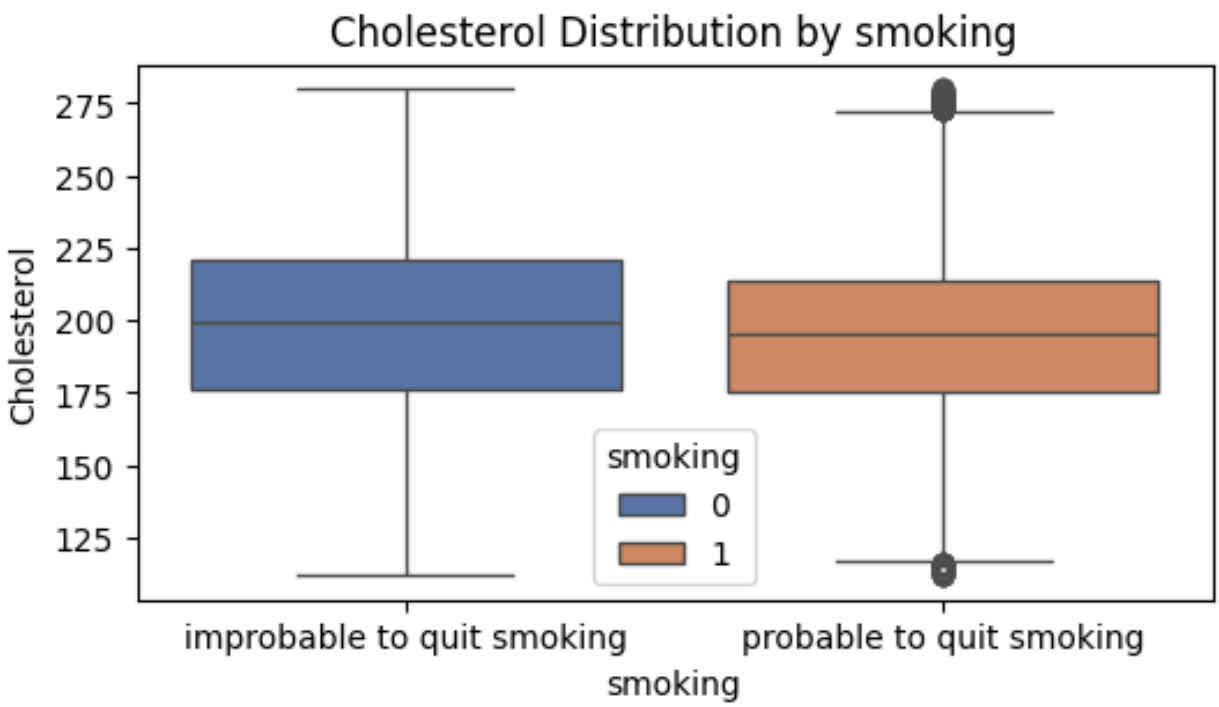

serum creatinine Distribution by smoking
serum creatinine
1.2
1.1
1.0
0.9
0.8
0.7
0.6
smoking
0
1
improbable to quit smoking
probable to quit smoking
smoking

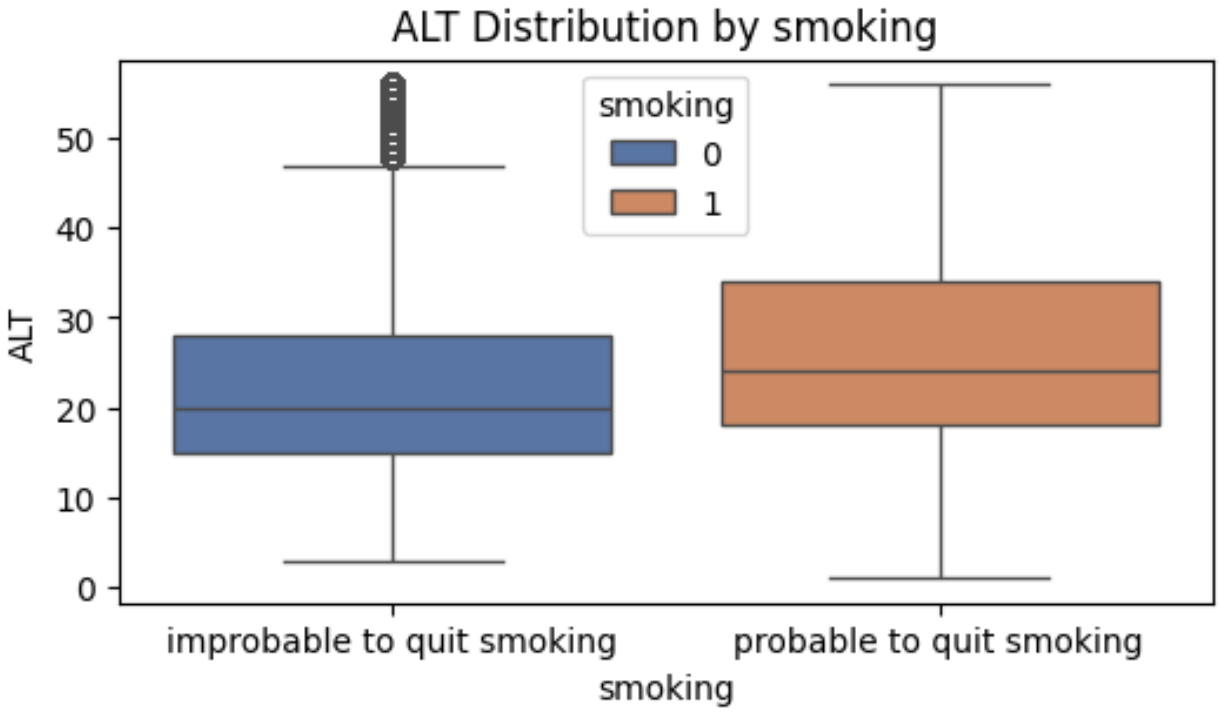

Confusion Matrix for Logistic Regression
14013
5176
4950
8677
Non-Smoker
Smoker
True label
Non-Smoker
Smoker
Predicted label
14000
12000
10000
8000
6000

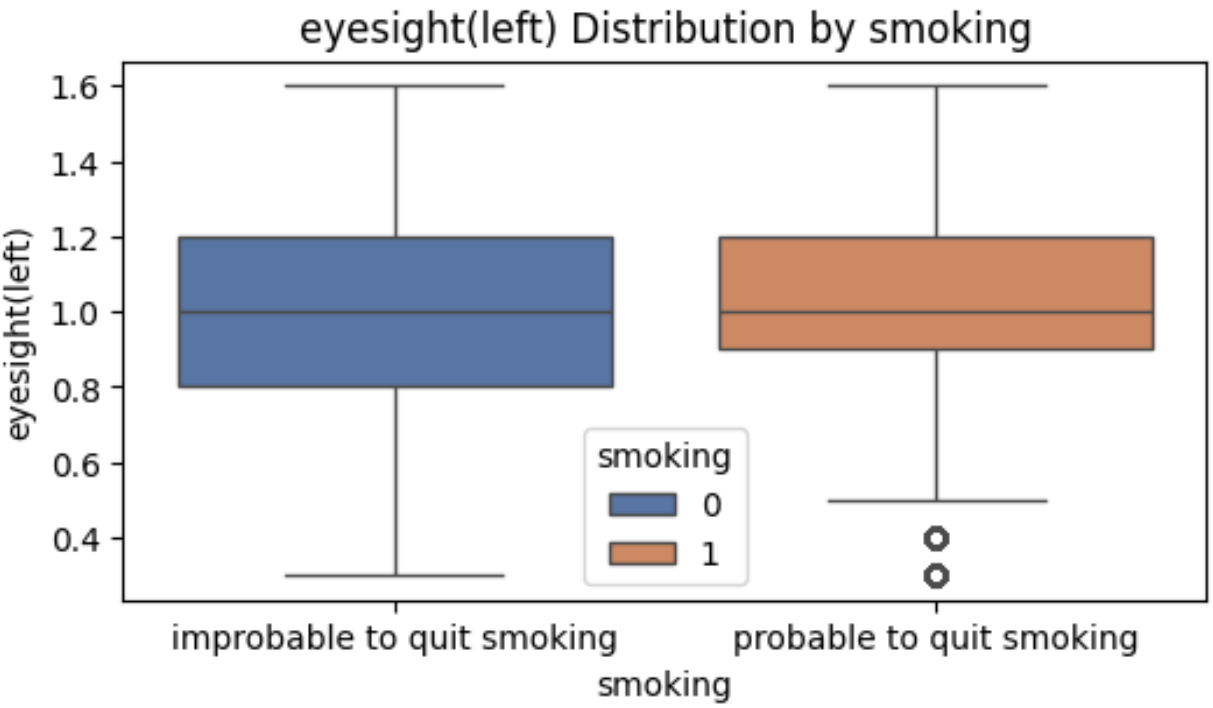

Confusion Matrix for Logistic Regression
14100
5089
5013
8614
Non-Smoker
Smoker
True label
Non-Smoker
Smoker
Predicted label
14000
12000
10000
8000
6000

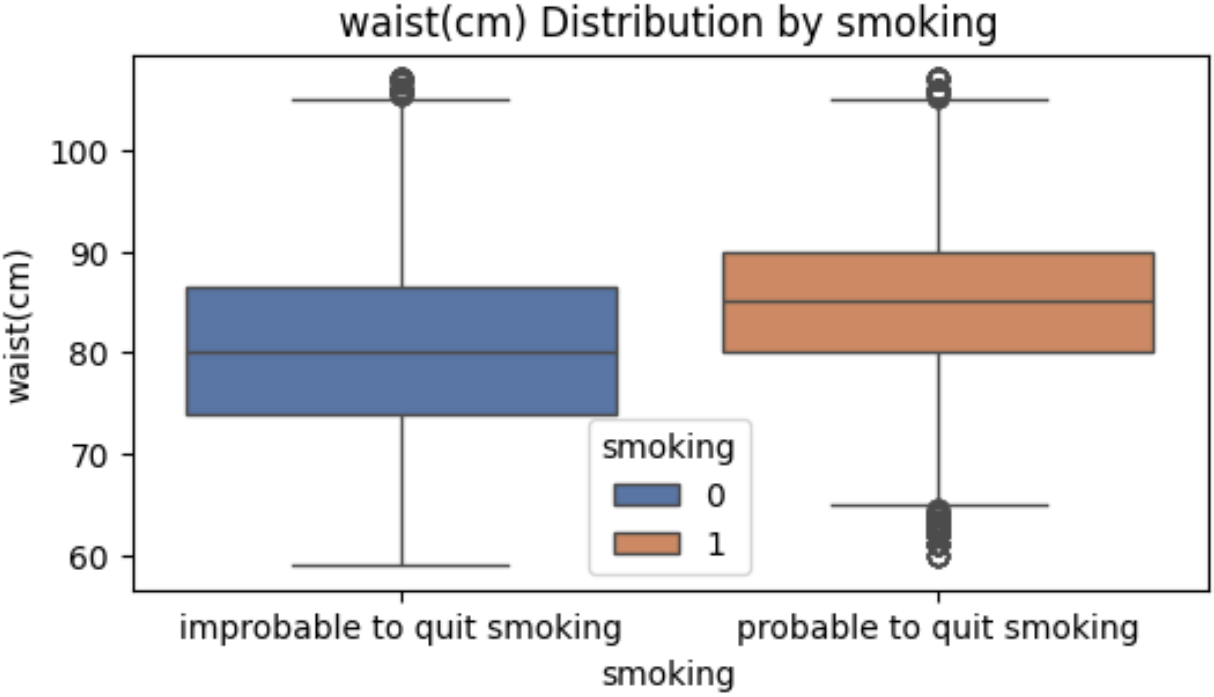


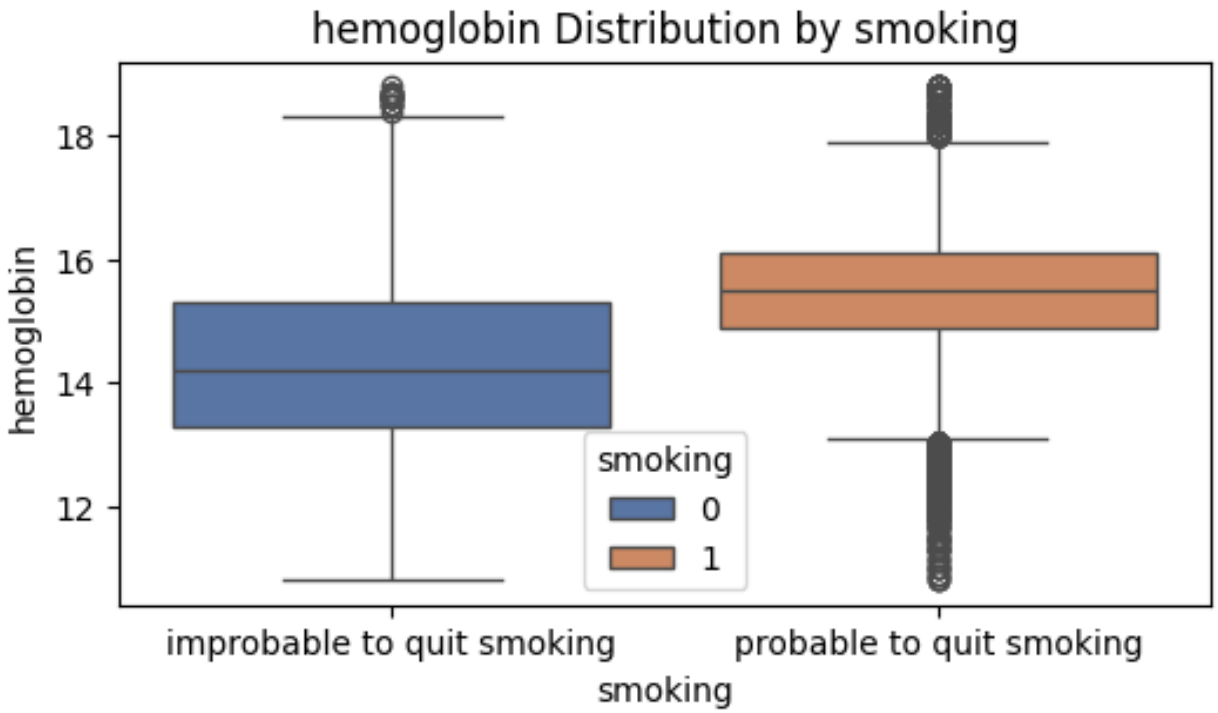
- This again shows no relation to smoking

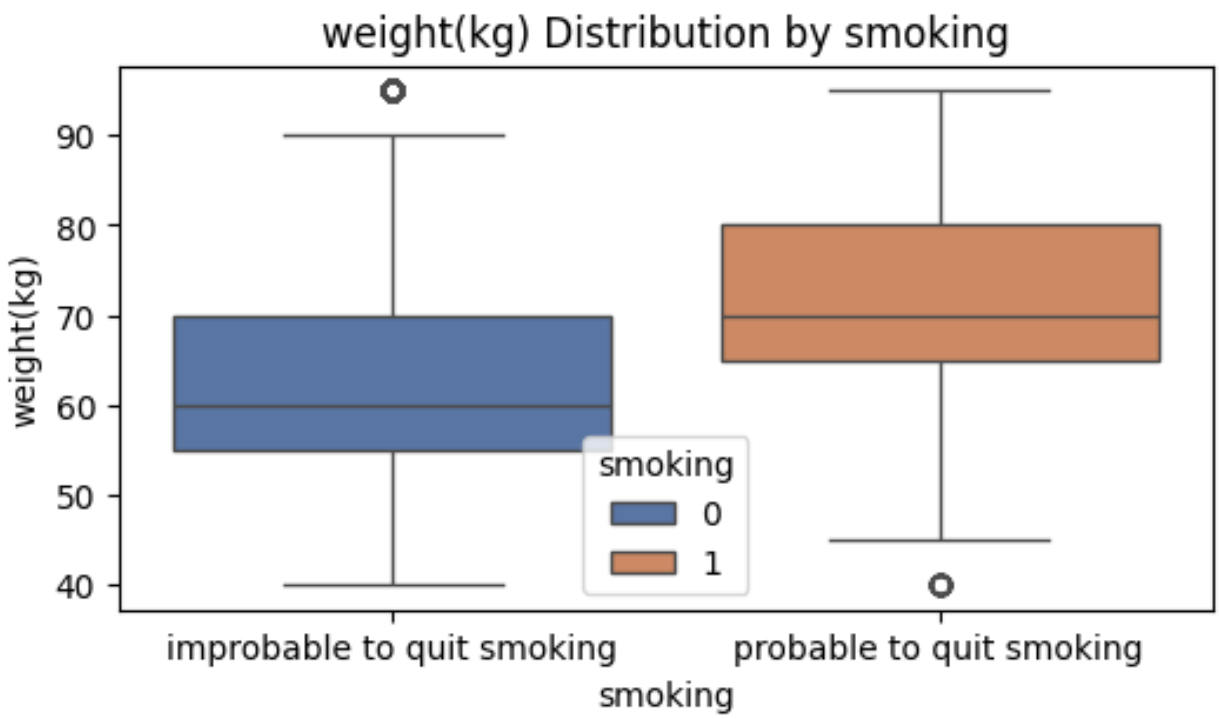

Cholesterol Distribution by smoking
Cholesterol
275
250
225
200
175
150
125
smoking
0
1
improbable to quit smoking
probable to quit smoking
smoking

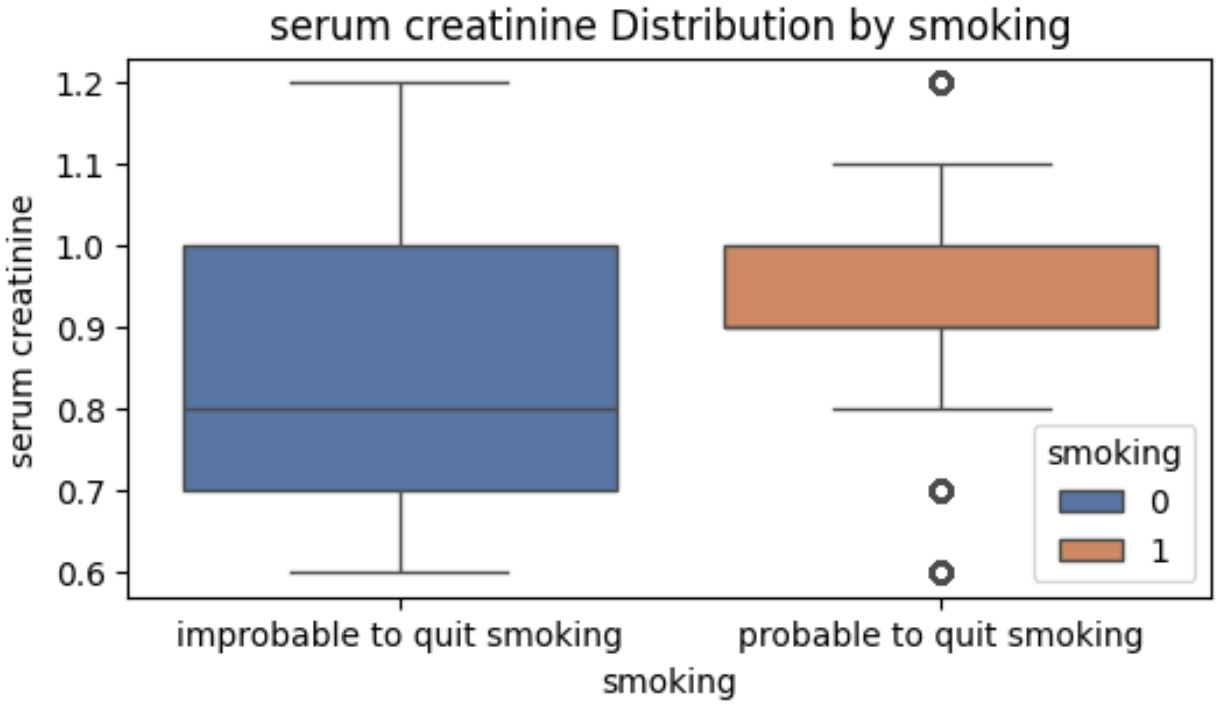

ALT Distribution by smoking
ALT
50
40
30
20
10
0
smoking
0
1
improbable to quit smoking
probable to quit smoking
smoking


eyesight(left) Distribution by smoking
eyesight(left)
1.6
1.4
1.2
1.0
0.8
0.6
0.4
smoking
0
1
improbable to quit smoking
probable to quit smoking
smoking

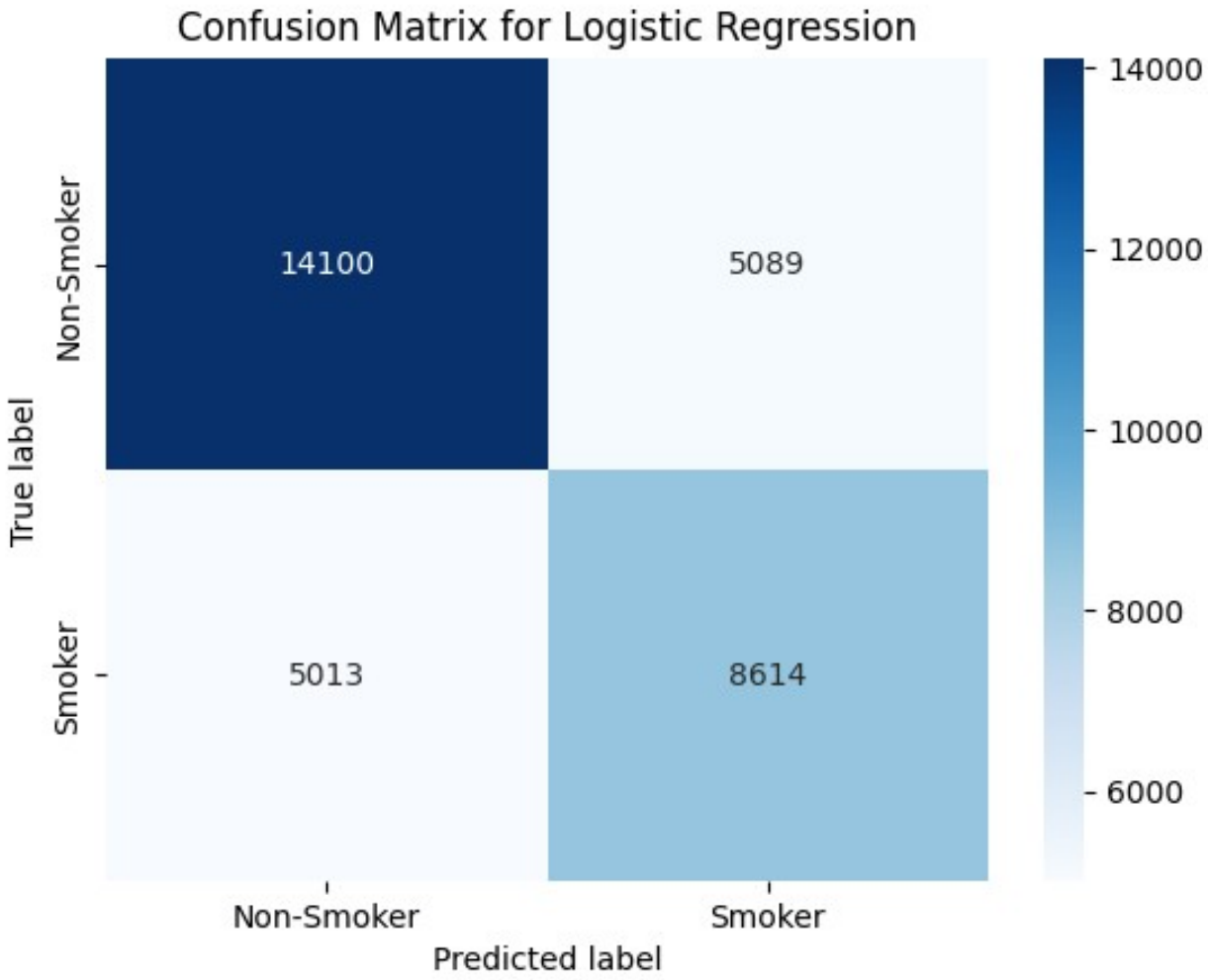

waist(cm) Distribution by smoking
waist(cm)
100
90
80
70
60
smoking
0
1
improbable to quit smoking
probable to quit smoking
smoking


hemoglobin Distribution by smoking
hemoglobin
18
16
14
12
smoking
0
1
improbable to quit smoking
probable to quit smoking
smoking


weight(kg) Distribution by smoking
weight(kg)
90
80
70
60
50
40
smoking
0
1
improbable to quit smoking
probable to quit smoking
smoking

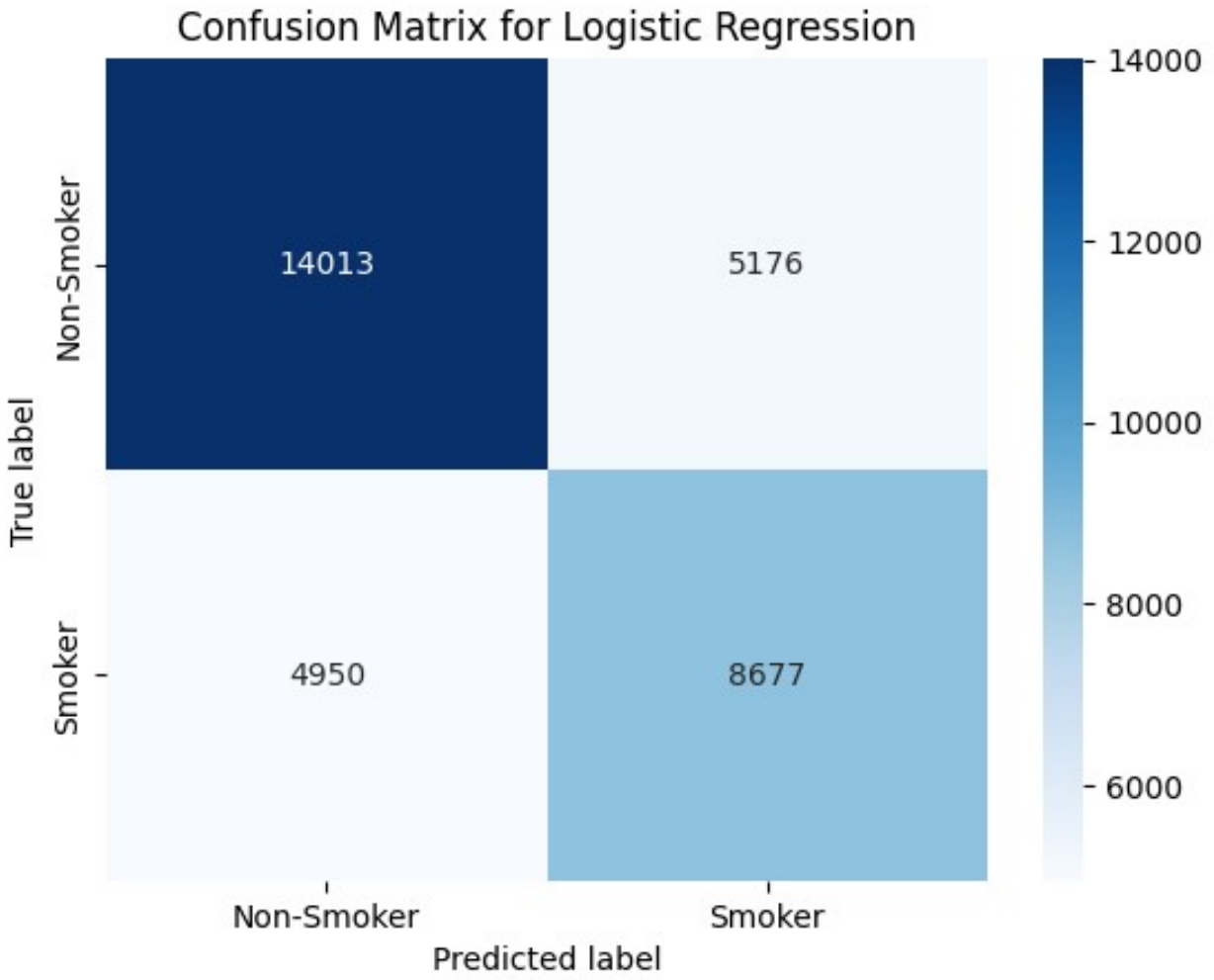

serum creatinine Distribution by smoking
serum creatinine
1.2
1.1
1.0
0.9
0.8
0.7
0.6
smoking
0
1
improbable to quit smoking
probable to quit smoking
smoking

Bivariate:



from analyzing the data relative to the label 'smoking' we were able to identify certain faetures that correlate the most to the label.

- we note that some features have somewhat postive correlations with one another.
- positive correlations with smoking: [waist, hemoglobin, weight, serum creatinine]
- paris with postive correlations: (ALT,weight) (ALT,hemoglobin) (ALT,waist)
- (waist,serum creatinine) (waist,wieght) (waist,hemoglobin)
- (weight,serum creatinine) (weight,hemoglobin)



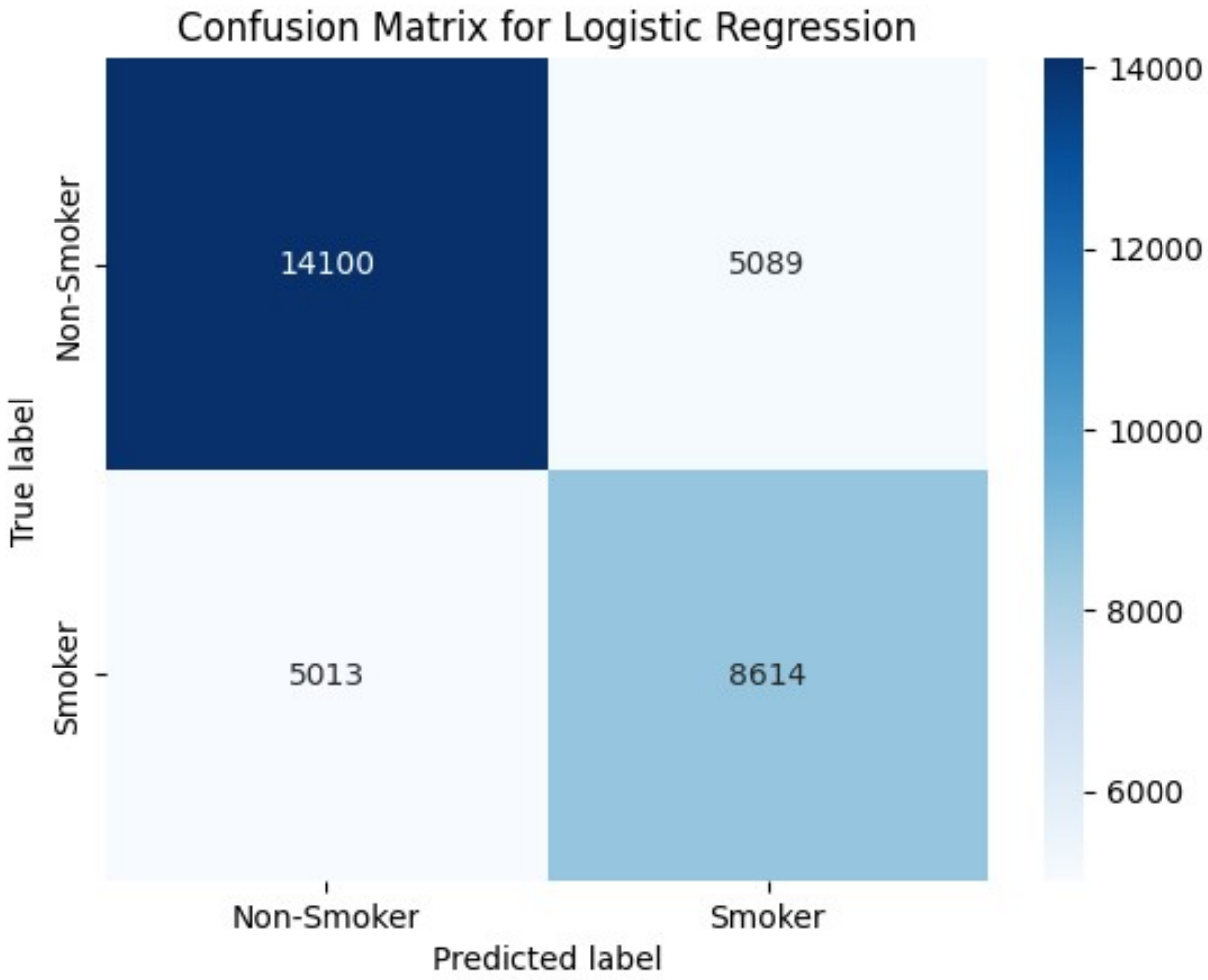
- Here is a collective graph of the relationships between each two features

Multivariate: Using linear regression (label is binary)

Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional space while preserving as much variance as possible, making it useful for reducing complexity and mitigating multicollinearity.

multivariate analysis showed that hearing left, hearing right , and dental carries contribute very little. we might drop them for dimensionality reduction.

note that hearing left, hearing right , and dental carries contribute very little. we might drop them for dimensionality reduction. now, we attempt to retrain the model using only seven features(excluding hearing left, hearing right , and dental carries) with high correlation to smoking.



2. Feature Engineering

we concluded that we could reduce the dimensionality only to four features while maintaing the same information with regard to the label. we then engineered a new feature that slightly improved the model accuracy.

3.Ensemble Methods

we created a class for each of the ensemble methods

4. Hyperparametr tuning

Bagging

```
Bagging, the short form for bootstrap aggregating, is mainly applied
in classification and regression. It increases the accuracy of models
through decision trees, which reduces variance to a large extent. The
```

```
reduction of variance increases accuracy, eliminating overfitting,
which is a challenge to many predictive models.
```

Boosting

Boosting is an ensemble technique that learns from previous predictor mistakes to make better predictions in the future. The technique combines several weak base learners to form one strong learner, thus significantly improving the predictability of models. Boosting works by arranging weak learners in a sequence, such that weak learners learn from the next learner in the sequence to create better predictive models.