DATA SCIENCE PROJECT

“Smoking Analysis Across Age Groups”

Computer Science Department
University of Sulaimani
7th Semester

Prepared by :
Bery Mansor Othman
Banaz Jaafar Mustafa
Bahra Ahmad Ali

January 2024

Contents

Abstract: 3

Introduction: 4

Methodology: 5

Data Collection: 5

Data Pre-processing: 5

Data Visualization: 5

Model Selection: 9

1. Decision Tree Classifier: 9

2. Random Forest Classifier: 9

3. Logistic Regression: 9

Differences and Complementary Application: 10

Results 10

Logistic Regression Model: 10

Decision Tree Model: 11

Random Forest Model: 11

Interpretation: 11

Conclusion: 12

Table of figures

1. **pre processing missing values** **5**
2. **smoking histogram - age histogram** **5**
3. **ciggeret and job realtion ship smoking = yes** **6**
4. **smoking across age groups** **6**
5. **ciggeret count and giving up when smoking =yes** **7**
6. **smoking and job relationship** **7**
7. **common reasons** **7**
8. **education level and job relationship** **8**
9. **education level and job realtion ship smoking = yes** **8**
10. **education level and smokingrealtionship smoking = yes** **8**

Abstract:

The pervasive habit of smoking stands as a critical public health concern with implications spanning various age groups. This study employs a comprehensive approach to analyze smoking behaviors across different age cohorts, utilizing data derived from a meticulously conducted survey and advanced analytical tools. Our aim is to unravel the intricate relationships between age, occupation, gender, education level, familial smoking habits, daily cigarette consumption, and the desire to quit.

Going beyond a mere exploration of smoking prevalence, our overarching goal involves predictive modeling to uncover the underlying factors significantly influencing smoking behaviors within diverse age demographics. By discerning these relationships, we intend to contribute valuable insights that can inform targeted interventions and policies. The predictive aspect of our analysis not only anticipates smoking behaviors but also identifies key determinants, positioning our study as a potential catalyst for evidence-based strategies, particularly tailored to specific age groups.

The methodology encompasses robust data collection, rigorous pre-processing, insightful data visualization, and the application of three machine learning models—Decision Tree Classifier, Random Forest Classifier, and Logistic Regression. These models collectively ensure a comprehensive understanding of the factors influencing smoking habits across various age groups.

Upon training and evaluating the models, all three demonstrated comparable accuracy of approximately 63.64% on the test data. Precision, recall, and F1-scores for both smoking and non-smoking classes were consistent across models. However, the moderate accuracy suggests that predicting smoking status based solely on age might have limitations, emphasizing the potential role of other factors.

In conclusion, this analysis provides insights into smoking behaviors based on age, showcasing comparable performance among the models. However, the study acknowledges the need for a more comprehensive approach, considering additional features beyond age. Recommendations for future work include exploring lifestyle, socio-economic factors, or health-related aspects and experimenting with different algorithms for improved predictive accuracy. This study serves as an initial step toward understanding smoking behaviors and highlights avenues for further exploration to enhance predictive capabilities.

Introduction:

This project delves into a meticulous analysis of smoking habits across diverse age groups, utilizing data collected through a Google Forms survey disseminated across various social media platforms. The survey encapsulates critical attributes such as age, occupation, smoking status, gender, education level, familial smoking influences, daily cigarette consumption, and the desire to quit. Leveraging Python libraries and predictive modeling, our objective is to predict the factors influencing smoking habits and propose effective solutions.

Commencing with the acquisition of data through a Google Forms survey published on social media, we meticulously pre-processed the dataset. Subsequent stages involved visualizing complex data structures and relationships between attributes like age, job, smoking status, gender, education level, familial smoking tendencies, daily cigarette consumption, and the inclination to quit.

To distill meaningful insights, three robust models – Decision Tree Classifier, Random Forest Classifier, and Logistic Regression – were employed, each selected based on its suitability for the intricacies of our dataset. Decision Tree Classifier for intricate decision boundaries, Random Forest Classifier for ensemble learning, and Logistic Regression for binary classification.

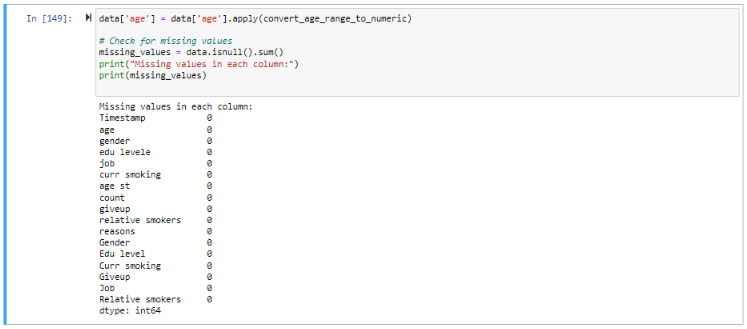
Methodology:

Data Collection:

The foundation of our analysis lies in a meticulously conducted survey that captures insights from individuals spanning diverse age groups. Collected data encompasses crucial attributes such as age, occupation, smoking status, gender, education level, familial smoking habits, daily cigarette consumption, and the aspiration to quit. This rich dataset serves as the backbone for our exploration into the factors influencing smoking habits.

Data Pre-processing:

To ensure the accuracy and reliability of our analysis, a rigorous pre-processing phase was implemented. This involved cleaning the dataset, handling missing values, and transforming categorical variables into a suitable format for machine learning algorithms.

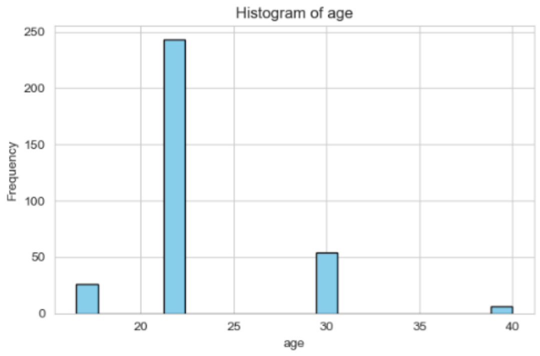


```
In [149]:  data['age'] = data['age'].apply(convert_age_range_to_numeric)

           # Check for missing values
           missing_values = data.isnull().sum()
           print("Missing values in each column:")
           print(missing_values)

           Missing values in each column:
           Timestamp           0
           age                 0
           gender              0
           edu levele          0
           job                 0
           curr smoking        0
           age st              0
           count               0
           giveup              0
           relative smokers    0
           reasons             0
           Gender              0
           Edu level           0
           Curr smoking        0
           Giveup              0
           Job                 0
           Relative smokers    0
           dtype: int64
```

1 pre processing missing values

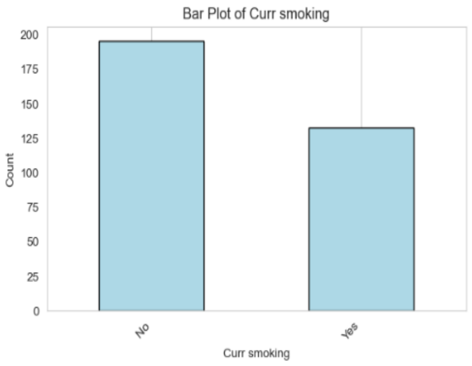
Data Visualization:

Visualizing the data structure and interrelationships among attributes was a pivotal step in our analysis. Utilizing Python libraries, we created insightful visual representations that shed light on the distribution of smoking habits across different age groups and the connections with other demographic factors.



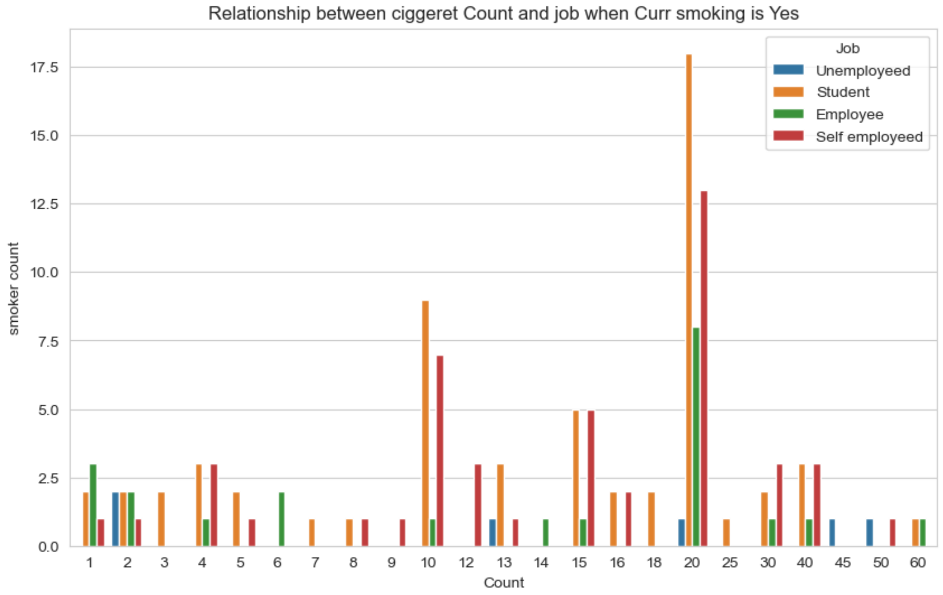
```
age
22.0    243
30.0     54
16.5     26
40.0      6
Name: count, dtype: int64
```

2smoking histogram

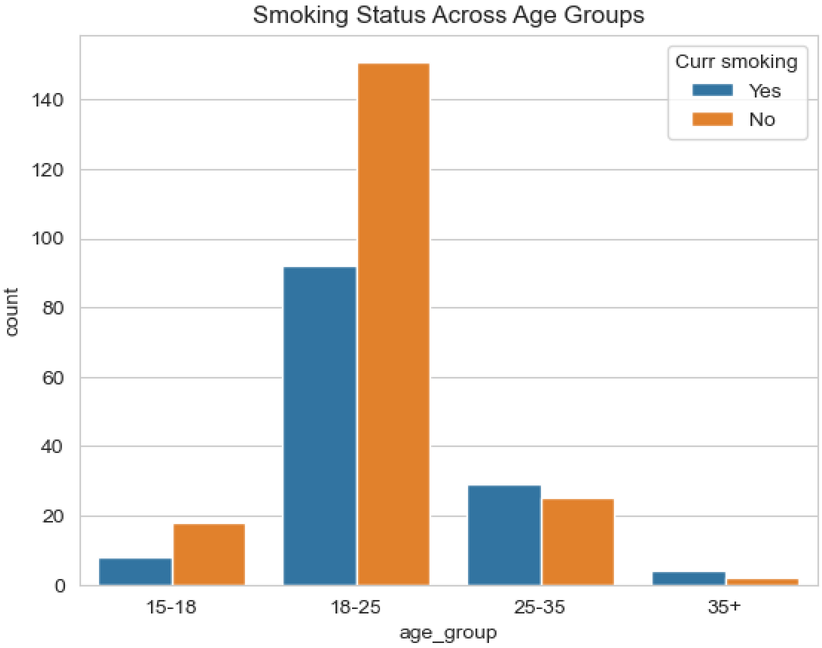


```
Curr smoking
No     196
Yes    133
Name: count, dtype: int64
```

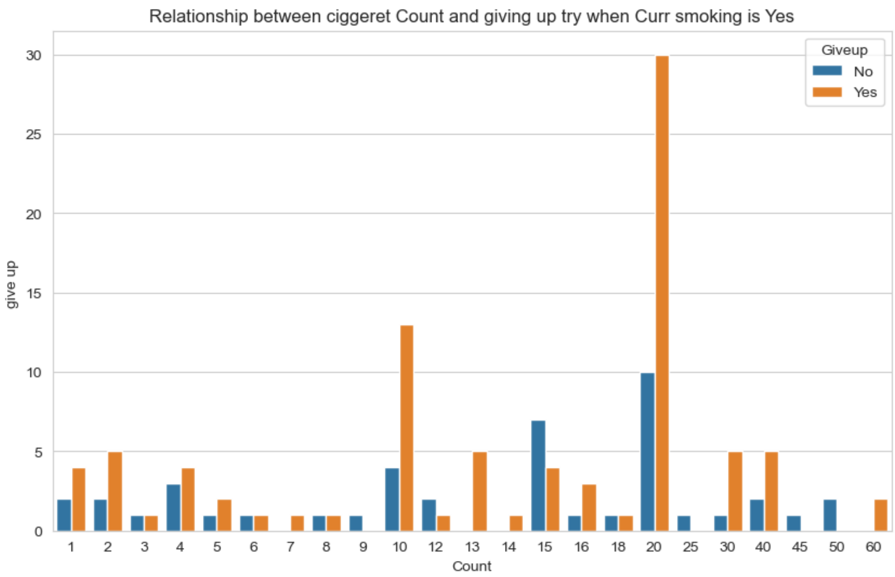
3 age histogram



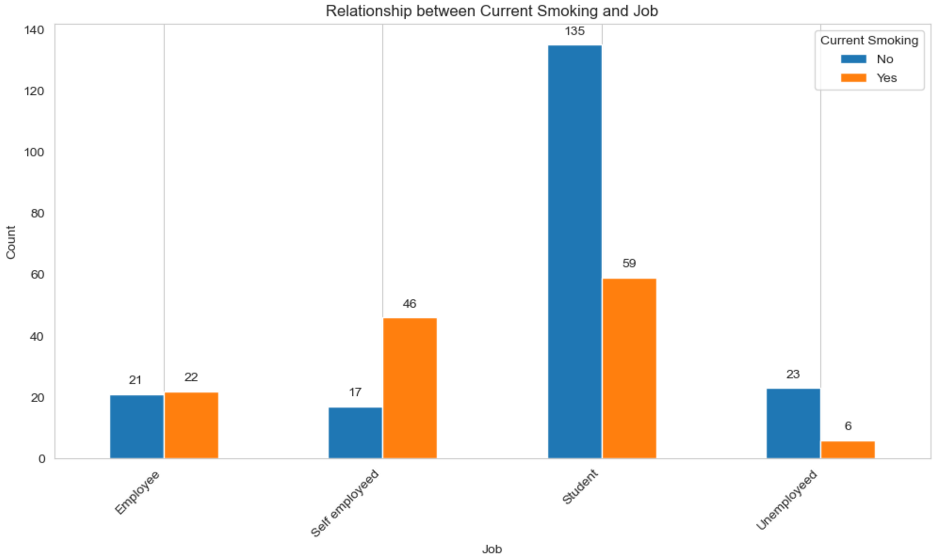
4 ciggeret and job realtion ship smoking = yes



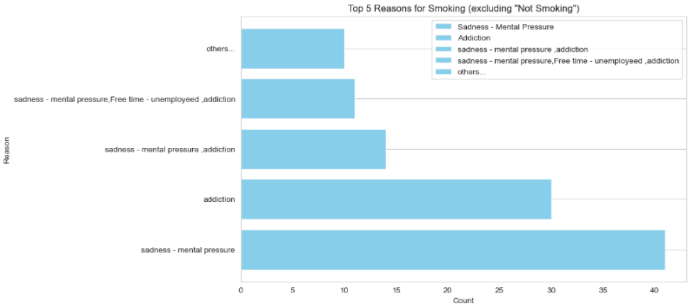
5 smoking across age groups



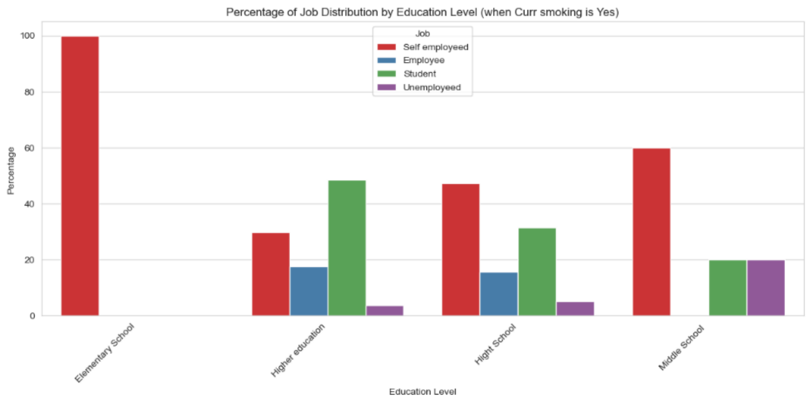
6 ciggeret count and giving up when smoking =yes



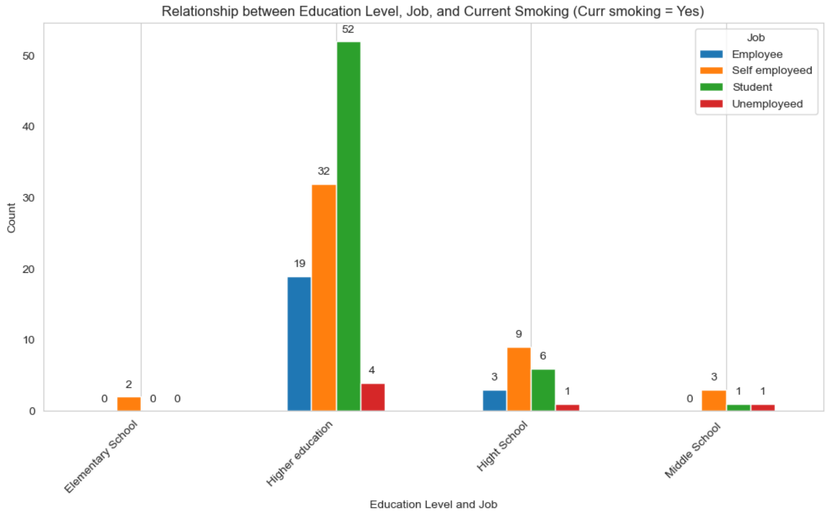
7 smoking and job relationship



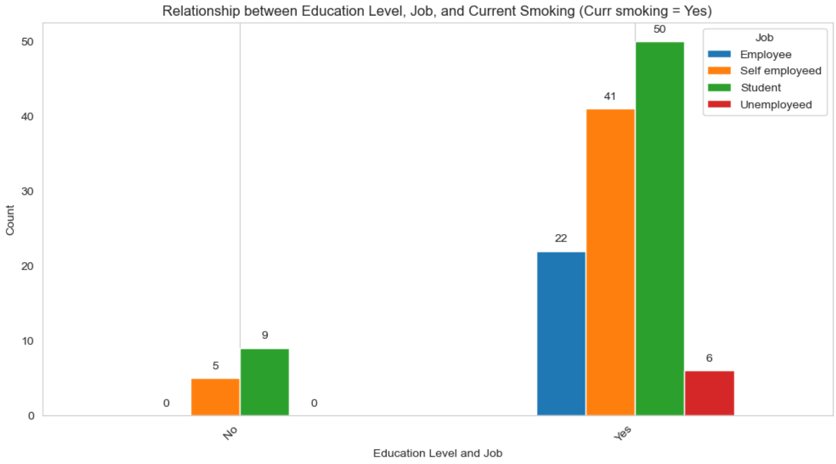
8 common reasons



9education level and job relationship



10education level and job realtion ship smoking = yes



11education level and smokingrealtionship smoking = yes

Model Selection:

Three distinct machine learning models—Decision Tree Classifier, Random Forest Classifier, and Logistic Regression—were carefully chosen for our predictive analysis, with the selection rationale deeply rooted in the unique characteristics of our dataset and the specific goals of our study.

1. Decision Tree Classifier:

Dependence: Decision Tree Classifier excels in capturing complex decision boundaries, making it particularly effective in scenarios where certain features strongly influence the outcome. This model relies on constructing a binary tree structure based on the features in the dataset, recursively splitting the data based on the feature that provides the best information gain at each step.

Strengths:

Excellent at capturing intricate relationships between features.

Easily interpretable, providing a clear visualization of the decision-making process.

Requires minimal data preparation and is less sensitive to outliers.

2. Random Forest Classifier:

Dependence: Random Forest Classifier leverages an ensemble of decision trees to enhance predictive accuracy. By averaging predictions from multiple models and reducing overfitting, it provides robustness against noise in the dataset. Each tree is trained on a random subset of the data with replacement (bootstrap sampling), and the final prediction is obtained through averaging or voting.

Strengths:

Enhances predictive accuracy and robustness through ensemble learning.

Can handle a large number of features and capture complex relationships in the data.

Reduces overfitting by aggregating predictions from diverse trees.

3. Logistic Regression:

Dependence: Logistic Regression models the relationship between the dependent binary variable (smoking or non-smoking) and the independent variables, estimating the probability of an instance belonging to a particular class. It is widely used in binary classification tasks, especially when exploring the likelihood of binary outcomes.

Strengths:

Well-suited for binary classification tasks, providing probabilities for class membership.

Simplicity and interpretability make it easy to understand the impact of each feature on the prediction.

Less prone to overfitting, especially with a relatively small number of features.

Differences and Complementary Application:

Decision Tree Classifier captures local relationships and decision boundaries but may lack generalization.

Random Forest Classifier mitigates the overfitting of decision trees and provides robust predictions through ensemble learning.

Logistic Regression models the probability of binary outcomes with simplicity and interpretability. The collective application of these models ensures a comprehensive understanding of the factors influencing smoking habits across various age groups. Decision trees capture local nuances, random forests enhance predictive accuracy and robustness, while logistic regression provides insights into the likelihood of binary outcomes. This strategic combination aims not only to predict smoking behavior but also to uncover significant features contributing to these predictions. The results of this analysis have the potential to inform targeted interventions and policies addressing smoking-related issues within different age demographics.

Results

training three different classification models (Logistic Regression, Decision Tree, and Random Forest) to predict the 'Curr smoking' status based on the 'age' feature. Let's go through the results of each model:

Logistic Regression Model:

Accuracy: 0.6364 (or 63.64%)

Precision (No): 0.65

Precision (Yes): 0.55

Recall (No): 0.88

Recall (Yes): 0.24

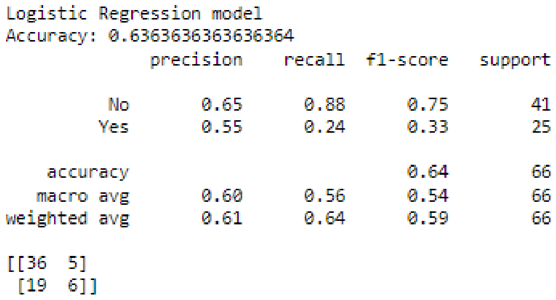
F1-Score (No): 0.75

F1-Score (Yes): 0.33

Confusion Matrix:

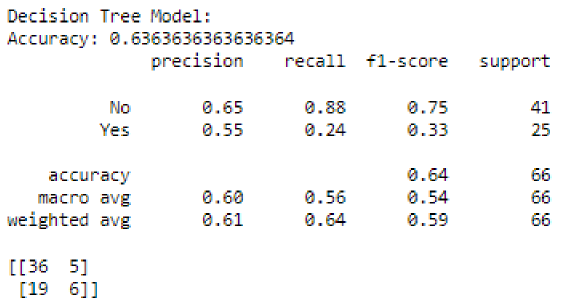
[[36 5]

 [19 6]]



```
Logistic Regression model
Accuracy: 0.6363636363636364
              precision    recall  f1-score   support

          No       0.65      0.88      0.75        41
         Yes       0.55      0.24      0.33        25

    accuracy                           0.64        66
   macro avg       0.60      0.56      0.54        66
weighted avg       0.61      0.64      0.59        66

[[36  5]
 [19  6]]
```

Decision Tree Model:



```
Decision Tree Model:
Accuracy: 0.6363636363636364
              precision    recall  f1-score   support

          No       0.65      0.88      0.75        41
         Yes       0.55      0.24      0.33        25

    accuracy                           0.64        66
   macro avg       0.60      0.56      0.54        66
weighted avg       0.61      0.64      0.59        66

[[36  5]
 [19  6]]
```

Accuracy: 0.6364 (or 63.64%)

Precision (No): 0.65

Precision (Yes): 0.55

Recall (No): 0.88

Recall (Yes): 0.24

F1-Score (No): 0.75

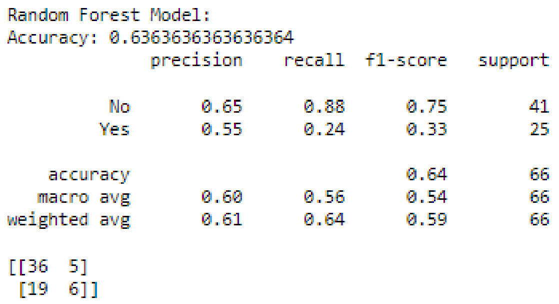
F1-Score (Yes): 0.33

Confusion Matrix:

[[36 5]

 [19 6]]

Random Forest Model:



```
Random Forest Model:
Accuracy: 0.6363636363636364
              precision    recall  f1-score   support

          No       0.65      0.88      0.75        41
         Yes       0.55      0.24      0.33        25

    accuracy                           0.64        66
   macro avg       0.60      0.56      0.54        66
weighted avg       0.61      0.64      0.59        66

[[36  5]
 [19  6]]
```

Accuracy: 0.6364 (or 63.64%)

Precision (No): 0.65

Precision (Yes): 0.55

Recall (No): 0.88

Recall (Yes): 0.24

F1-Score (No): 0.75

F1-Score (Yes): 0.33

Confusion Matrix:

[[36 5]

 [19 6]]

Interpretation:

All three models (Logistic Regression, Decision Tree, and Random Forest) achieved the same accuracy of approximately 63.64% on the test data.

The precision, recall, and F1-scores for both classes ('No' and 'Yes') are consistent across the models.

The confusion matrices show the distribution of true positives, true negatives, false positives, and false negatives for each model.

Conclusion:

In this specific analysis, the models, regardless of the algorithm used, showed similar performance in predicting smoking status based on age.

The accuracy achieved is moderate, indicating that the 'age' feature alone may not be sufficient for highly accurate predictions.

Further exploration and feature engineering might be needed to improve model performance. Additionally, considering more features in the modeling process could enhance predictive capabilities.

In conclusion, the analysis aimed to predict the smoking status ('Curr smoking') based on the individual's age using three different classification models: Logistic Regression, Decision Tree, and Random Forest. The results indicated that all three models achieved comparable performance on the test data, with an accuracy of approximately 63.64%.

While the models demonstrated consistent precision, recall, and F1-scores for both smoking and non-smoking classes, it's important to note that the prediction based solely on age might have limitations. The moderate accuracy suggests that other factors beyond age may play a crucial role in determining an individual's smoking status.

The confusion matrices provided insights into the distribution of true positives, true negatives, false positives, and false negatives. However, achieving higher accuracy might require considering additional features and conducting further analysis.