# Smoking and Its Effect on Mortality: A Case Study

Symbiosis Institute of Technology
Group Members: Ayushi Nagpure, Dhanashree Giriya

## Problem Statement

- **Brief Overview:** Smoking is a leading cause of preventable deaths globally, contributing to severe health issues like cardiovascular diseases, respiratory disorders, and cancers. This project analyzes the relationship between smoking habits and mortality rates to quantify associated risks. Using data analysis and machine learning, we uncover trends, correlations, and factors like age and smoking status that influence mortality. These insights aim to inform public health policies, targeted interventions, and awareness campaigns.

- **Key Objectives:**

1. Analyze Trends: Study smoking prevalence, mortality trends, and correlations with age and risk factors.
2. Identify High-Risk Groups: Highlight vulnerable demographics and cumulative effects of smoking.
3. Propose Recommendations: Suggest strategies, campaigns, and further research to reduce smoking risks.

**Source :** www.freepik.com/

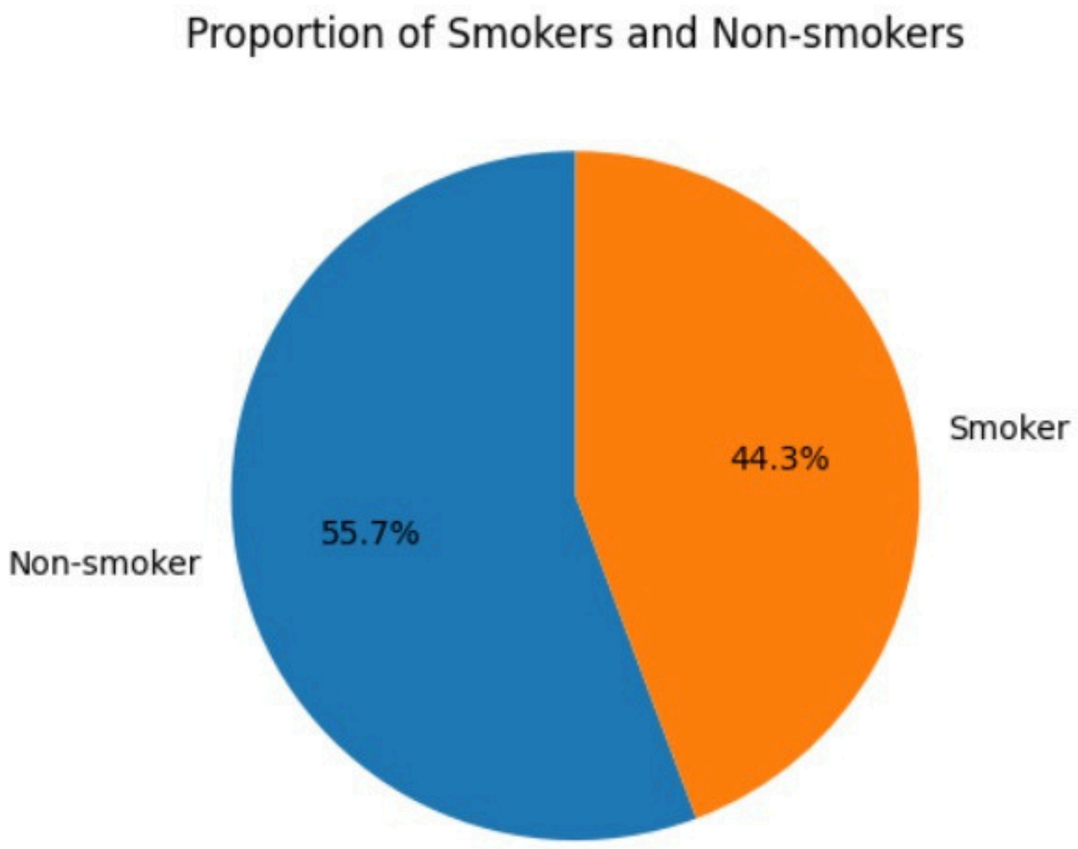
## Dataset Overview

**Dataset Description:** The dataset includes information on individuals, specifying whether they smoke, their age, and mortality status.

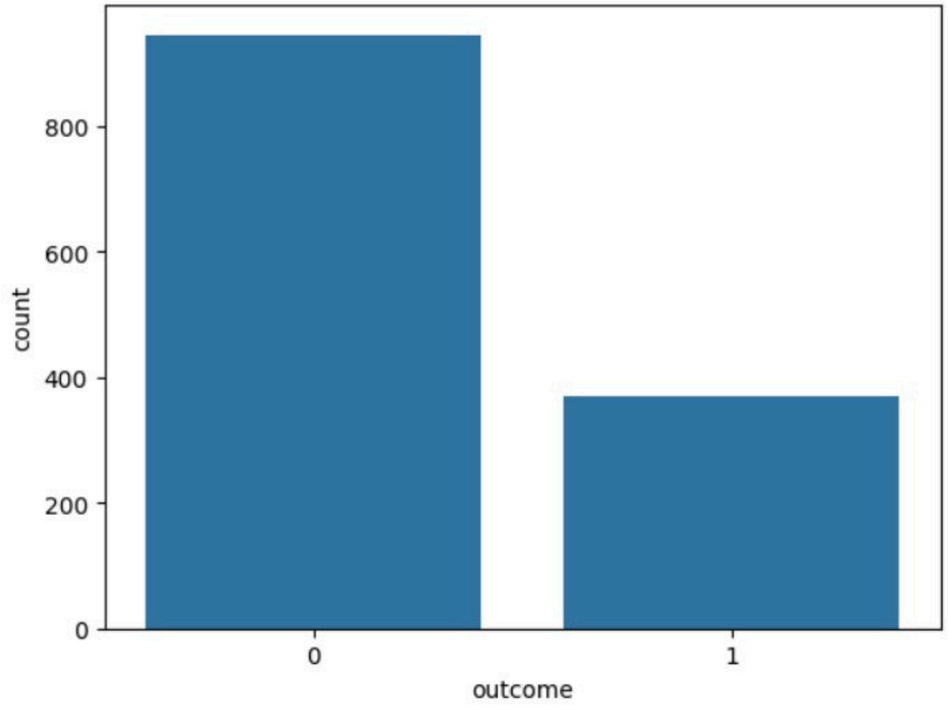
**Key features include:**

- **smoker:** A binary feature that indicates whether the individual is a smoker or not.Value:1 indicates the person is a smoker. 0 indicates the person is a non-smoker.
- **age:** The age of the individual in years. This feature is numerical and ranges from a specified minimum to maximum, representing the individual's age at the time of the study.
- **mortality:** A binary feature that reflects the death status of the individual. Value: 1 indicates the individual has died. 0 indicates the individual is alive.

**Source :** www.freepik.com/




```
outcome
0    945
1    369
Name: count, dtype: int64
```




- **Distribution of Smoking Status in the Dataset:** This pie chart visualizes the proportion of smokers versus non-smokers, allowing us to assess whether the dataset is balanced between the two groups.
- **Class Distribution of Mortality Outcome:** This count plot shows the distribution of the 'outcome' variable (alive vs. dead), allowing us to visually inspect any potential imbalance between the two classes in the dataset.

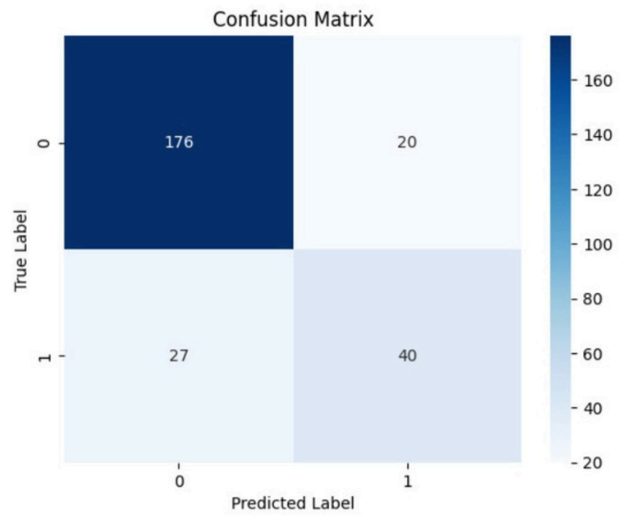
# Methodology

- **Approach:**

1. **Data Cleaning:** Handling missing values and ensuring data consistency.
2. **Exploratory Data Analysis:** Visualizing the relationship between smoking, age, and mortality.
3. **Model Building:** Using logistic regression to predict mortality based on smoking status and age.
4. **Evaluation:** Assessing model performance using feature importance and predictions.

- **Algorithms Used:**
  **Logistic Regression:** Chosen because it is well-suited for binary classification problems like predicting mortality (dead or alive) based on given factors.

**Confusion Matrix: This matrix shows the performance of the logistic regression model in predicting mortality. It highlights the number of true positives, true negatives, false positives, and false negatives.**




**Source :** www.freepik.com/

## Conclusion

- **Summary:** The logistic regression model successfully identified smoking and age as significant factors influencing mortality. The model provided reasonable accuracy in predicting mortality outcomes.

- **Future Work:** Future improvements could include collecting more data with additional features like lifestyle factors or medical history. Exploring more complex models (e.g., decision trees or random forests) could yield better performance.

**Source :** www.freepik.com/

# Thank You