Electronic cigarette usage among Hispanic college students

Background

- The use of electronic e-cigarettes has surged in popularity over recent years
- E-cigarettes have been marketed as a safer alternative to traditional tobacco products
- The e-cigarette market is rapidly evolving with new technologies and devices

Problem Statement

“What are the risk factors that escalate the use of e-cigarettes?”

Objectives

1. Reasons Influencing E-Cigarette Usage

2. Knowledge of Health Effects Associated with E-Cigarette Usage

3. Association Between E-Cigarette and Alcohol Consumption

Methodology

- **Dataset description:** Survey at University of San Bernardino
- **Data preprocessing and EDA:** Cleaning, handling missing values, and encoding categorical variables (one-hot encoder, imputer)
- **Model Fitting:** Model development using logistic regression, decision trees, KNN and Neural Network
- **Integration Technique:** SMOTE
- **Model evaluation metrics:** Accuracy, precision, Recall

02
Data Collection

Original
Variables: 91
Observations: 931

After Cleaning
Variables: 22
Observations: 208

Data Exploration

Variables

- Gender
- First Generation
- Ever_smoked_100_cig
- Smoking_Plan
- Smoking_habit
- Social Media
- Marijuana
- Alcohol
- Alcohol_habit
- Injurious
- Reasons for e-cigarette use (feel relaxed, feel good, less stressed, socially acceptable, cool/trendy, smell like smoke, flavors)
- Health Issues (Heart Attack/stroke/coronary, Seizures, Depression, Lung disease)

Key Variables

- Gender
- First Generation
- Smoking habits
- Social Media
- Marijuana
- Alcohol
- Health issues

03
Exploratory Analysis

Smoking Habit-Target

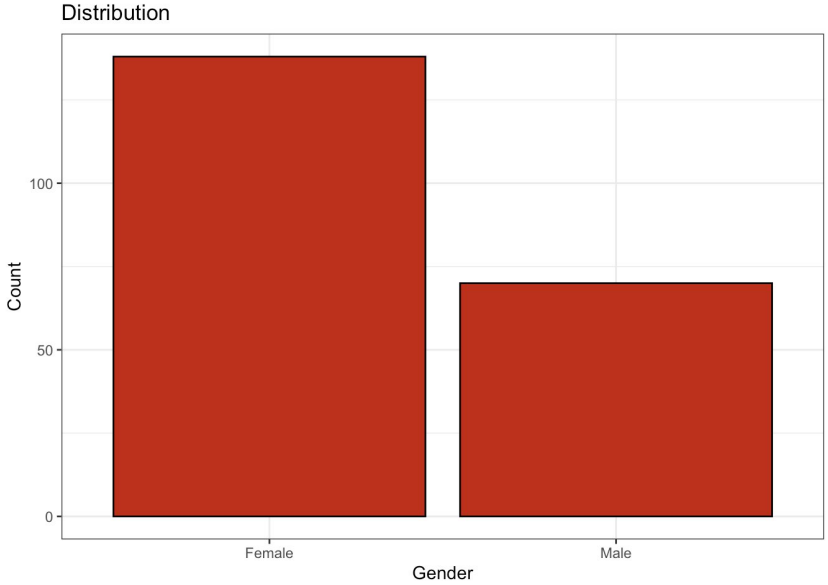
Smoke

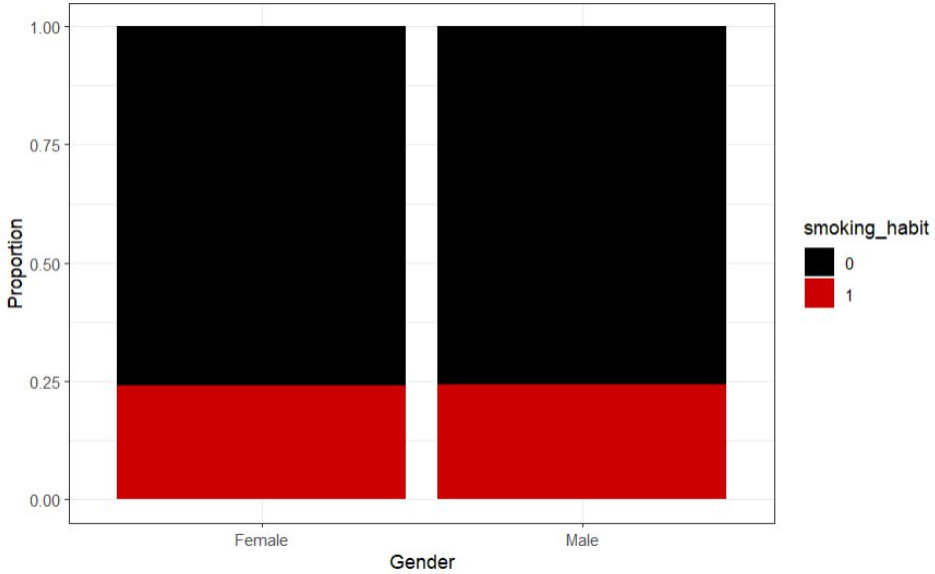
76%

Not smoke

24%

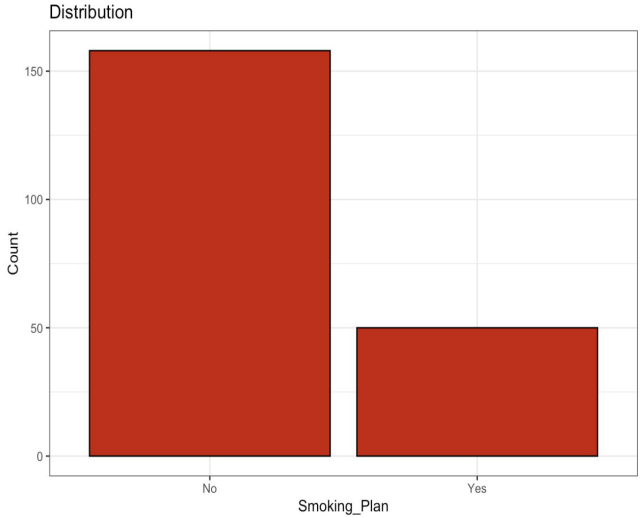
Gender

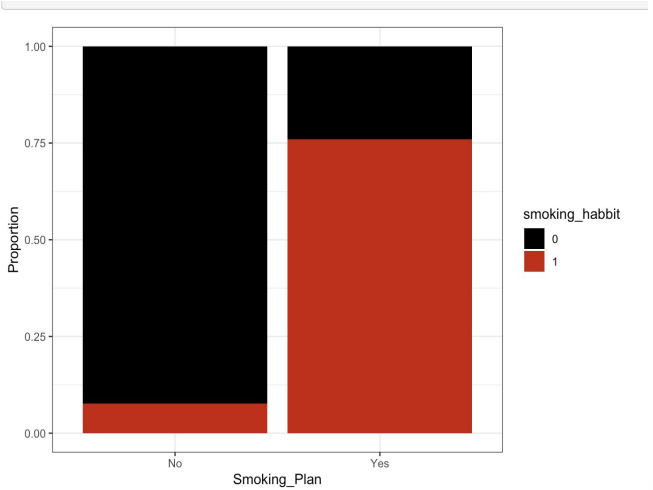




- 66.3% women, 33.7% men
- Consider potential gender influence on results

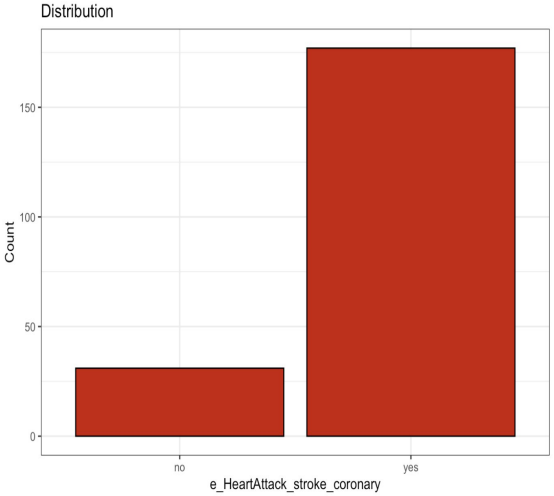
Plan to Smoke

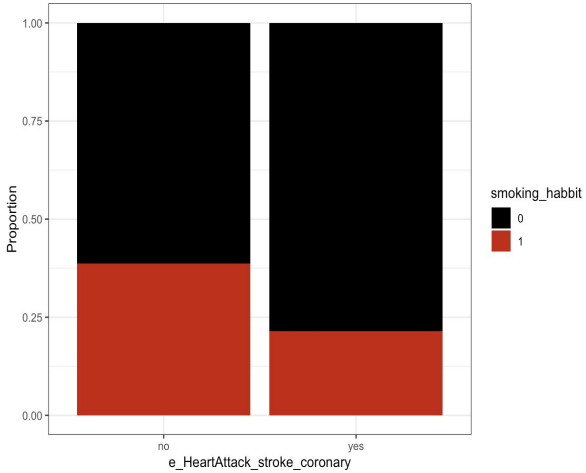




- Majority do not plan to smoke in the future (76% no, 24% yes)
- Significant correlation with smoking habits

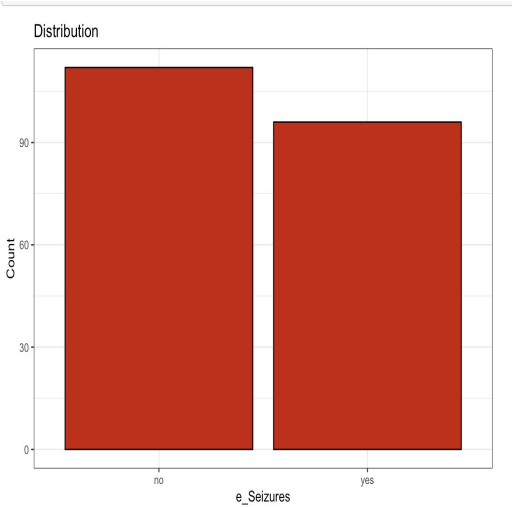
Health Issues

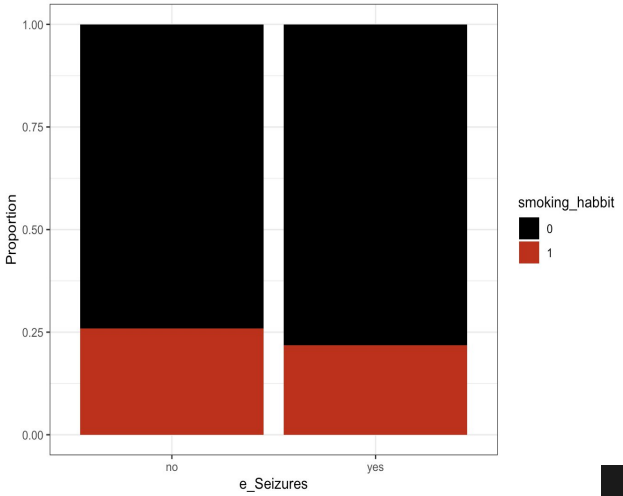




- Majority reported health issues (85.1% yes, 14.9% no)
- Potential connection between smoking and conditions like heart attacks and stroke

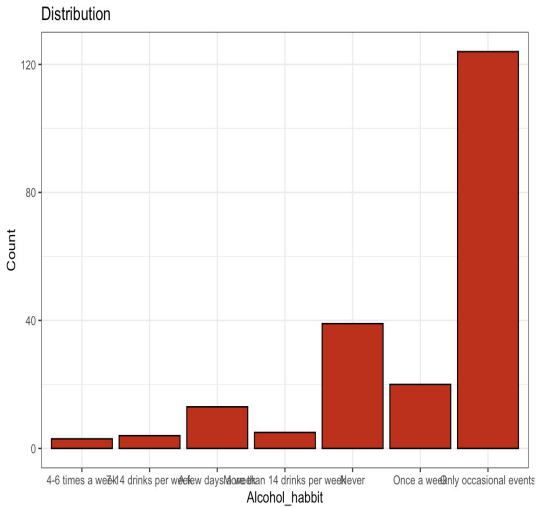
Seizures

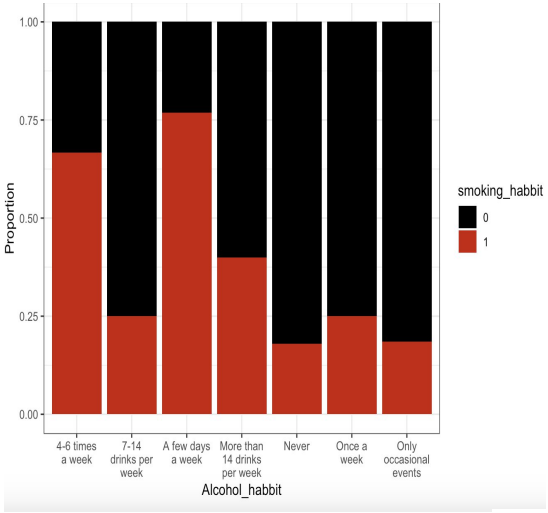




- Close split (53.8% no, 46.2% yes)
- Need to investigate potential correlation between smoking and seizures

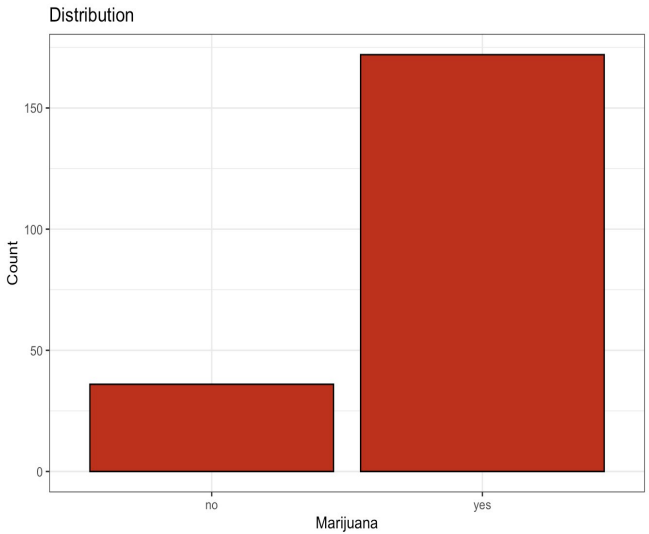
Alcohol Habit

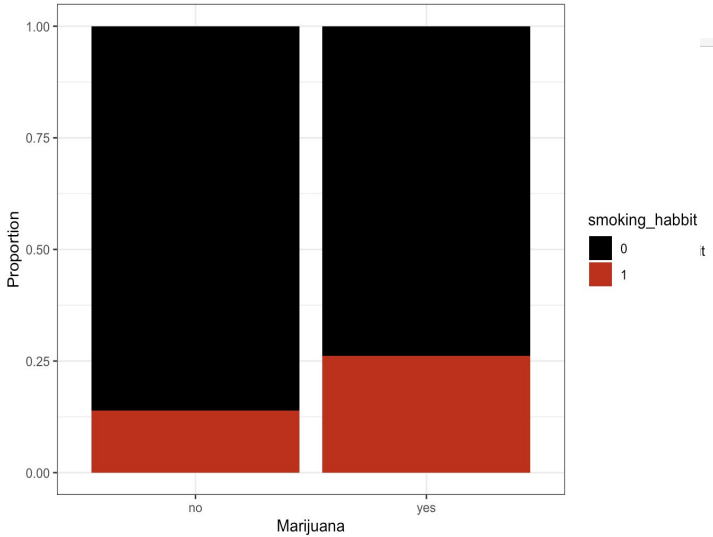




- Majority do not have a regular drinking habit
- Different drinking frequencies and habits among respondents

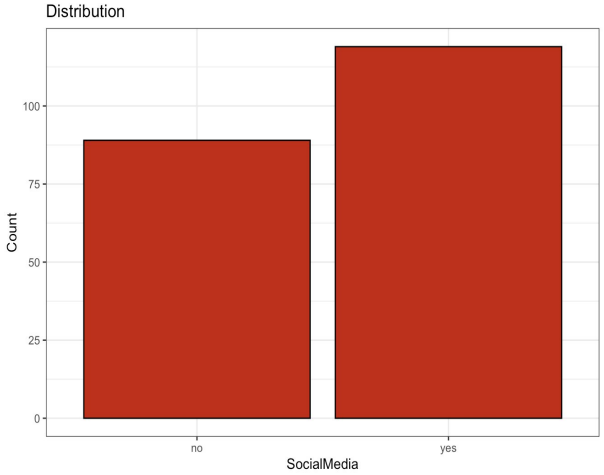
Marijuana

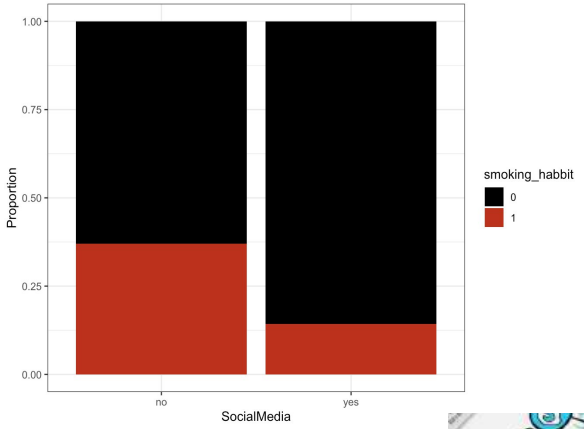




- 82.7% yes, 17.3% no
- Prevalent use of marijuana among respondents

Social Media





- 57.2% yes, 42.8% no
- Majority use social media, important for understanding smoking behavior

04
Model Fitting

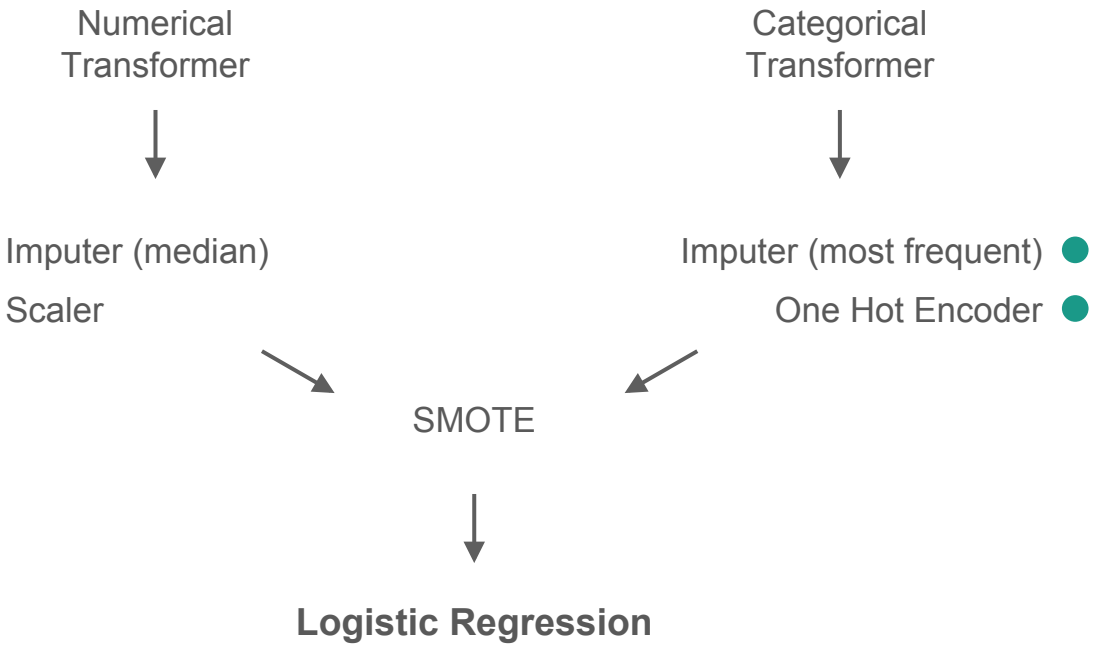
Logistic Regression

Logistic Regression

Binary Classification

Probability Estimation

GLM: Pipeline



Dataset splits

Training & Validation Set

Testing Set

Model 1: Performance

Smoking_habit
~
All Variables

86%
Accuracy

80%
Precision

57%
Recall

Statistically significant variables

Smoking_Plan

E-HearthAttack_strock_coronary

E-Seizure

Model 2: Performance

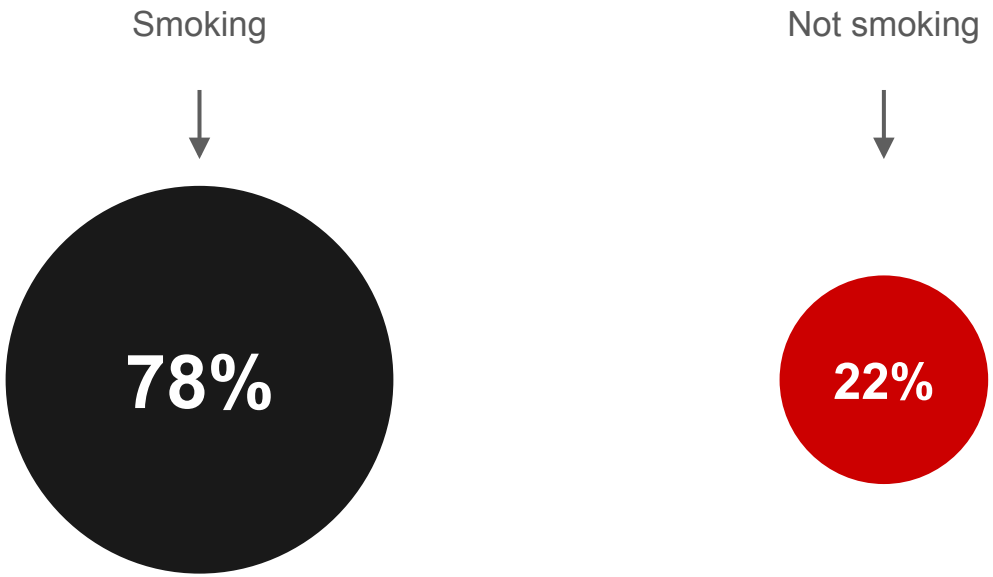
Smoking_habit
~
Backward Selection

88%
Accuracy

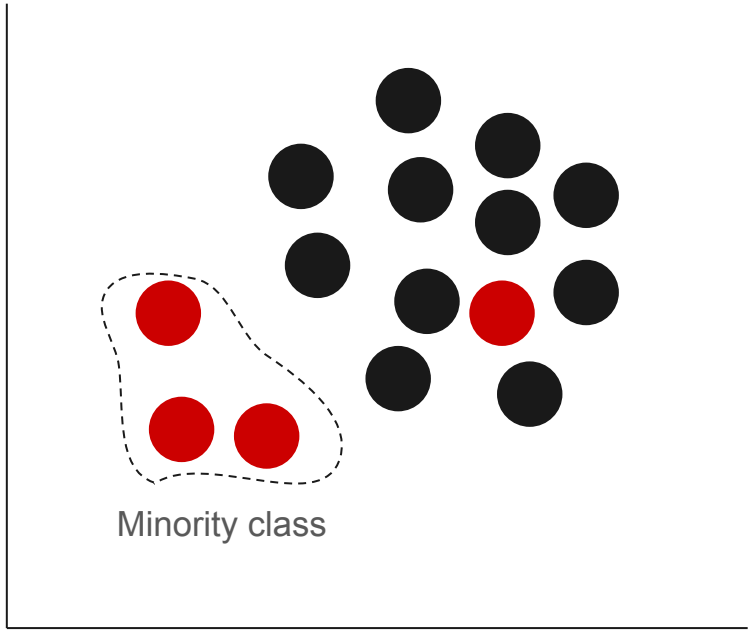
89%
Precision

57%
Recall

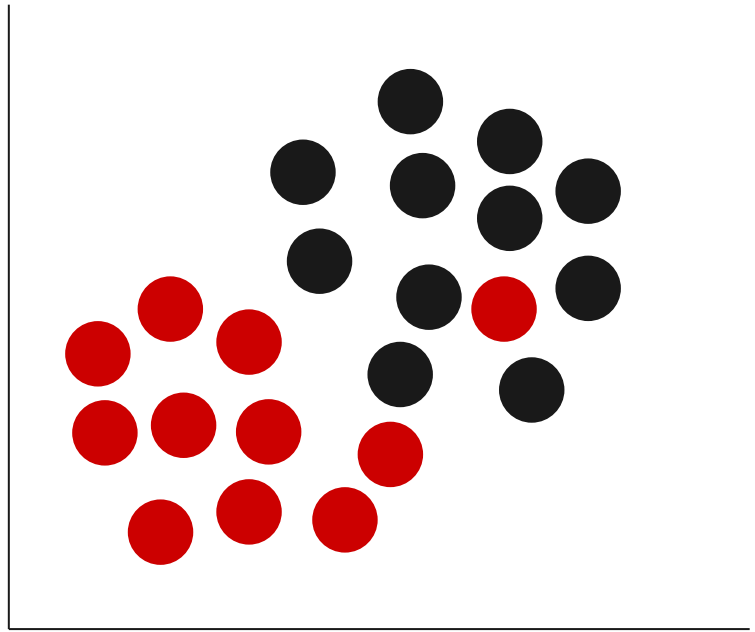
Addressing Class Imbalance



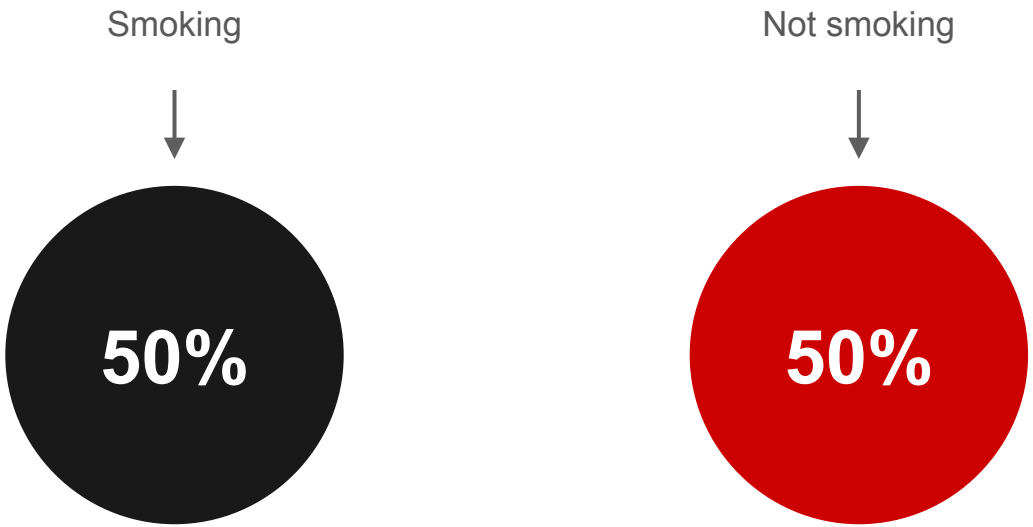
SMOTE: Identify Minority Cases



SMOTE: Synthetic Generation



Model 3: SMOTE + AIC



Model 3: SMOTE

Smoking_habit
~
Forward Selection

88%
Accuracy

89%
Precision

57%
Recall

GLM: Model Evaluation

Baseline vs. **SMOTE + AIC**

GLM: Model Evaluation

Accuracy

86%  **88%**

GLM: Model Evaluation

Precision

80% → **89%**

GLM: Model Evaluation

Recall

57% → **57%**

Statistically significant variables

Smoking_Plan

Alcohol_habit

Reason_feel_less_stressed

E-Seizure

Comparative Analysis

Alcohol_habit

Variable	Vaping status				P-value*
	None (n=29,766)	Former (n=3,351)	Current (n=2,287)		
			Intermittent (n=1,664)	Daily (n=623)	
Drinking status					
None	64.38 (0.4)	14.33 (0.6)	11.73 (0.8)	7.06 (1.3)	<0.001
Former	23.4 (0.3)	38.34 (0.9)	19.28 (1.0)	13.22 (1.6)	
Intermittent	12.11 (0.3)	46.66 (1.0)	67.42 (1.2)	64.47 (2.4)	
Daily	0.1 (0.0)	0.67 (0.1)	1.57 (0.3)	15.26 (1.8)	
Drinking quantity per intake					
None	87.78 (0.2)	52.67 (1.0)	31.01 (1.2)	20.27 (2.0)	<0.001
≤2 cups	6.26 (0.1)	9.09 (0.5)	12.33 (0.8)	6.98 (1.2)	
<1 bottles	3.14 (0.1)	11.97 (0.5)	17.13 (1.0)	14.48 (2.1)	
≥1 bottles	2.82 (0.1)	26.28 (0.9)	39.53 (1.4)	58.27 (2.5)	
Problem drinking†					
No	96.94 (0.1)	74.54 (0.8)	60.54 (1.3)	41.82 (2.4)	<0.001
Yes	3.06 (0.1)	25.46 (0.8)	39.46 (1.3)	58.18 (2.4)	

Values are presented as weighted % (standard error) by univariate analysis.
*Derived from chi-square analyses. †Measured by the CRAFFT (car, relax, alone, forget, friends, trouble) screening test.

K-Nearest Neighbors

KNN: Model Evaluation

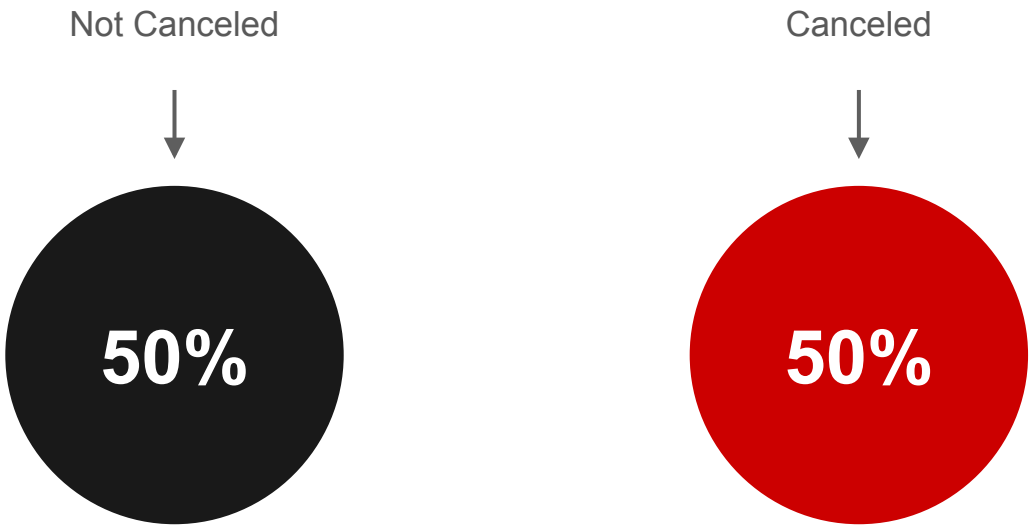
SMOTE

Dataset splits

Training & Validation Set

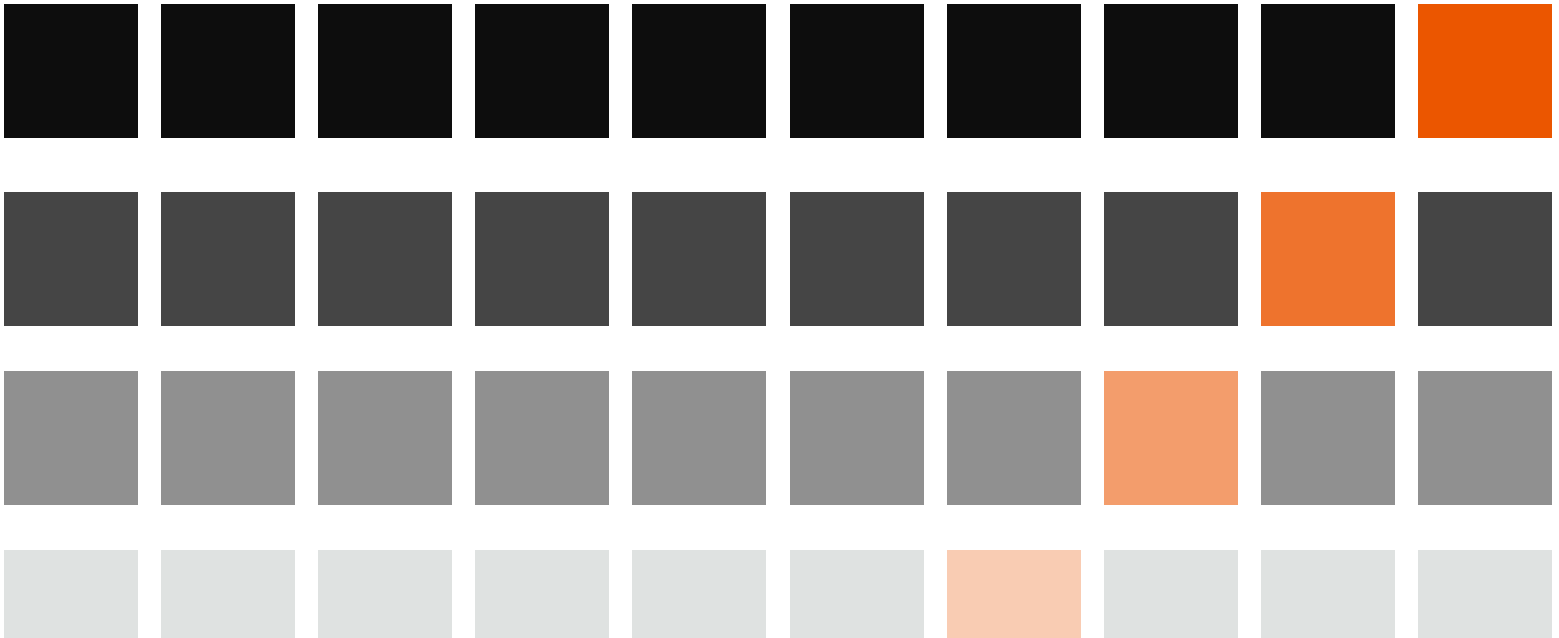
Testing Set

Model 2: SMOTE



Cross-Validation

K = 1 : 15



Model 2: Performance

Smoking_habit
~
All Variables

79%
Accuracy

56%
Precision

64%
Recall

Random Forest Classifier

Random Forest Classifiers

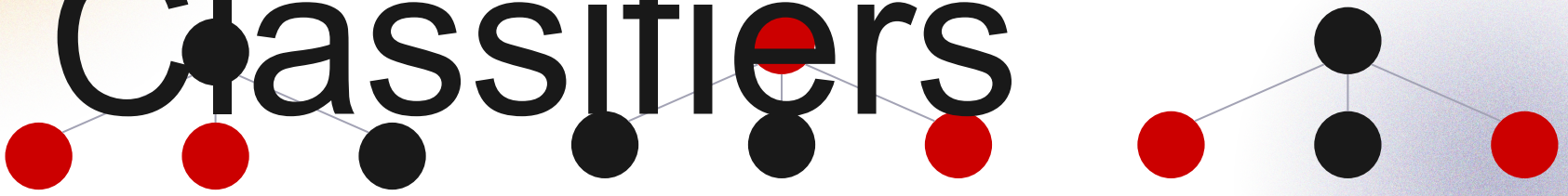
Ensemble Learning

Immune to Overfitting

Random Forest Classifiers

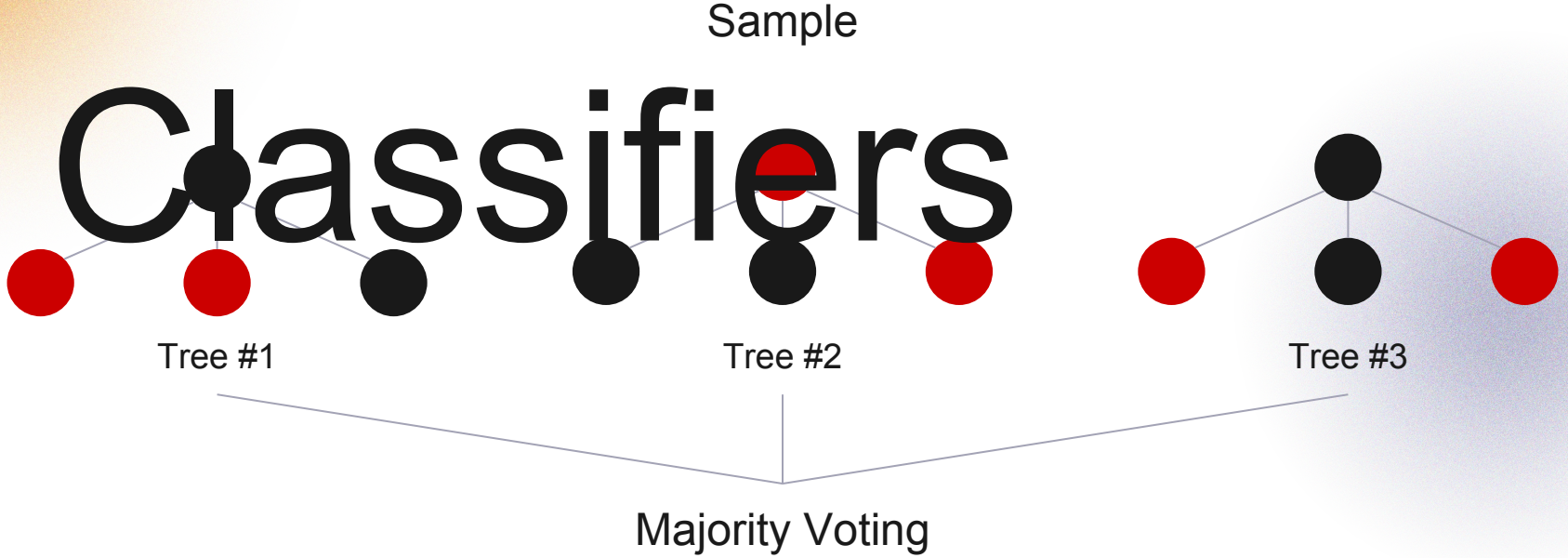
Random Forest Classifiers

Sample



Tree #1 Tree #2 Tree #3

Random Forest Classifiers



Model 1: Performance

Smoking_habit
~
All Variables

87%
Accuracy

88%
Precision

57%
Recall

Model 2: Performance

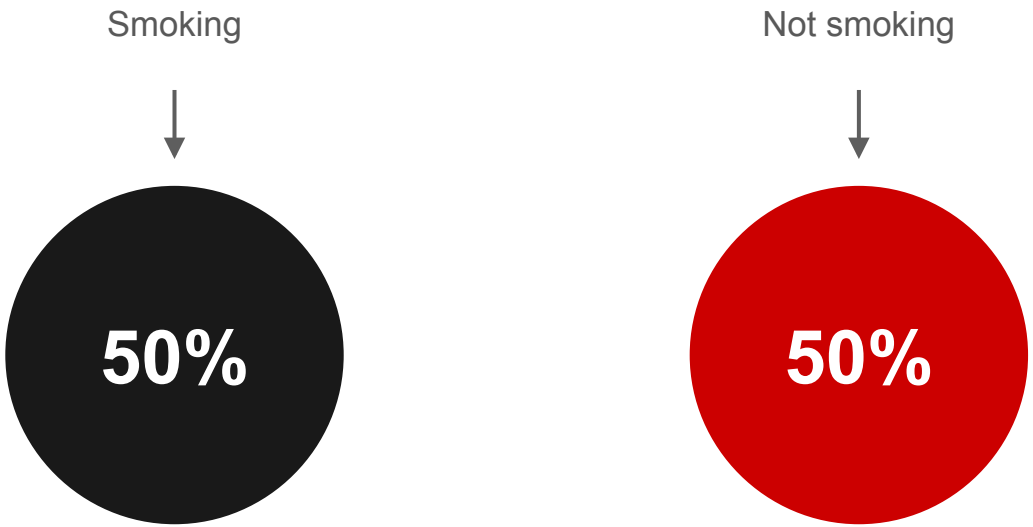
Smoking_habit
~
All Variables

88%
Accuracy

89%
Precision

57%
Recall

Model 3: SMOTE + Tuning



Model 3: Performance

88%
Accuracy

77%
Precision

71%
Recall

RF: Model Evaluation

Accuracy

88% → **88%**

RF: Model Evaluation

Precision

89% → **77%**

RF: Model Evaluation

Recall

57% → **71%**

Statistically significant variables

Smoking_PlanYes: p-value = 0.009, MeanDecreaseAccuracy = 0.170

Alcohol_habbitA_few_days_a_week: p-value = 0.009, MeanDecreaseAccuracy = 0.028

Reason_feel_relaxedyes: p-value = 0.019, MeanDecreaseAccuracy = 0.026

Reason_feel_less_stressyes: p-value = 0.009, MeanDecreaseAccuracy = 0.023

Reason_feel_goodyes: p-value = 0.009, MeanDecreaseAccuracy = 0.022

Reason_smell_like_smokeyes: p-value = 0.009, MeanDecreaseAccuracy = 0.009

ever_smoked_100_cigYes: p-value = 0.027, MeanDecreaseAccuracy = 0.007

Alcohol_habbitOnly_occasional_events: p-value = 0.059, MeanDecreaseAccuracy = 0.013

Neural Network

Data Preparation and Training

- Feature matrix extraction using **model.matrix()**
- **Batch Size:** 30 samples per batch
- Smaller batch size = more frequent updates, faster convergence, more noise
- **Epochs:** Set to 100 (complete passes through training dataset)
- **Validation Split:** 0.2 (20% of data for validation)

Model Architecture

- Constructed using **keras_model_sequential()**
- **Three dense layers** with dropout regularization to mitigate overfitting
- **Activation Functions:** ReLU and Softmax

Model : Performance

93%
Accuracy

97%
Precision

72%
Recall

05
Model Evaluation

Accuracy

Neural Network	93%
Random Forest Classifier	87%
Logistic Regression	87%
K-Nearest Neighbors	79%

Recall

Neural Network	72%
Random Forest Classifier	71%
K-Nearest Neighbors	64%
Logistic Regression	57%

Precision

Neural Network	97%
Logistic Regression	89%
Random Forest Classifier	77%
K-Nearest Neighbors	56%

Risk Factors That Escalates The Use of E-cigarettes

1. Alcohol consumption
2. Social Media
3. Stress