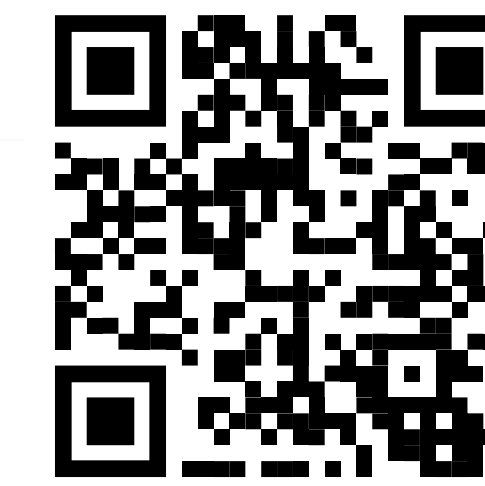


Factors of initiation to vaping products among smokers

CORESTA Congress October 13-17th, 2024

Thomas Verron¹, Xavier Cahours¹, Thomas Nahde²
¹SEITA, an Imperial Brands PLC Company, 216 Rue Raymond Losserand, 75014 Paris, France; ²Reemtsma Cigarettenfabriken GmbH, Albert-Einstein-Ring-7, D-22761, Hamburg, Germany
 Corresponding author's e-mail: thomas.verron@imperialbrands.com



FIND OUT MORE

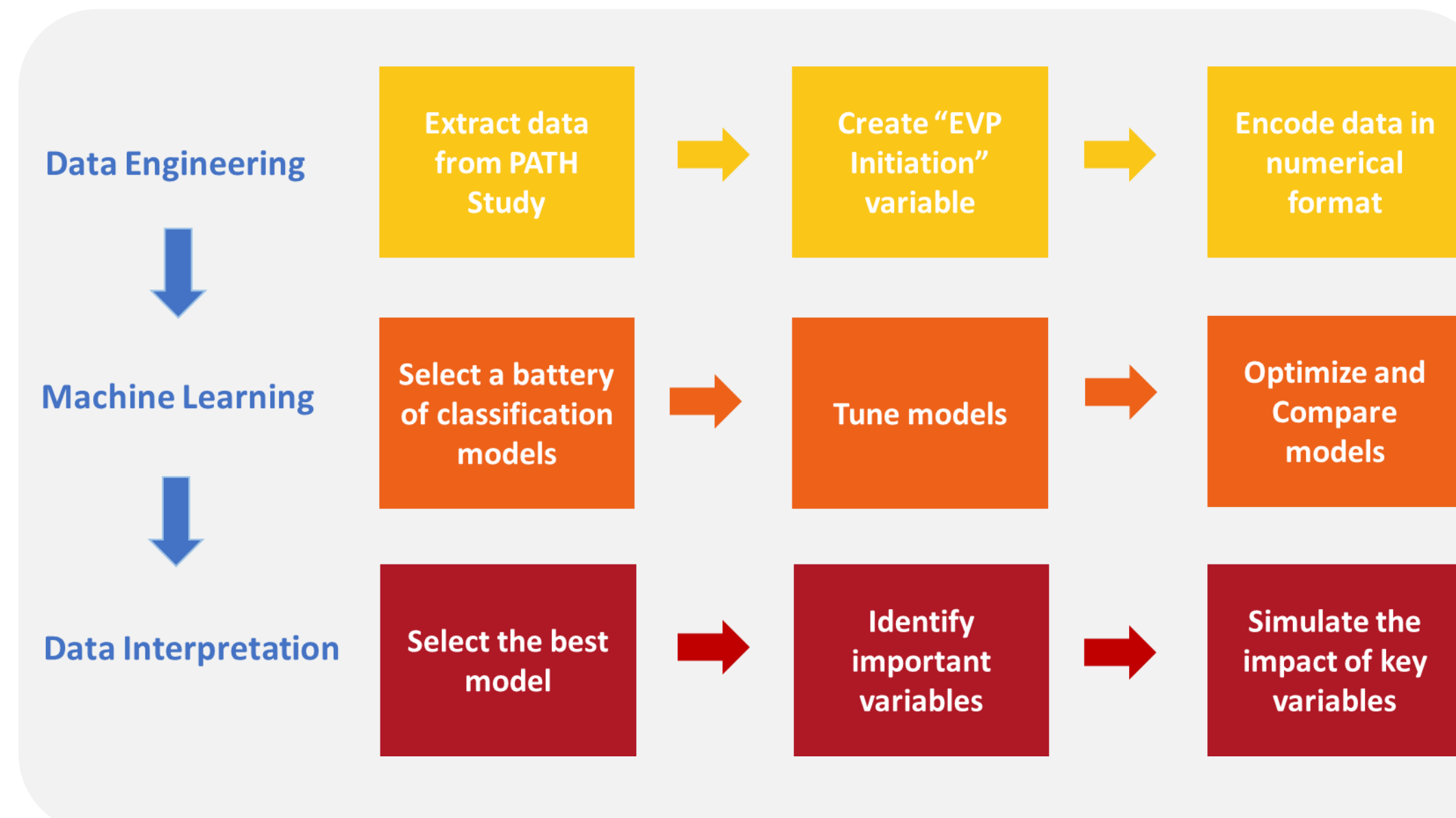


INTRODUCTION

The objective of this study was to determine the relevant factors related to the adoption of electronic vapor products (EVP) among adult smokers. To identify predictive factors of EVP initiation among adult smokers, we applied machine learning approaches to data from the Population Assessment of Tobacco and Health (PATH) study [1]. This longitudinal study, conducted in the United States, assesses smoking-related behaviors, tobacco consumption, nicotine exposure, and health effects on individuals.

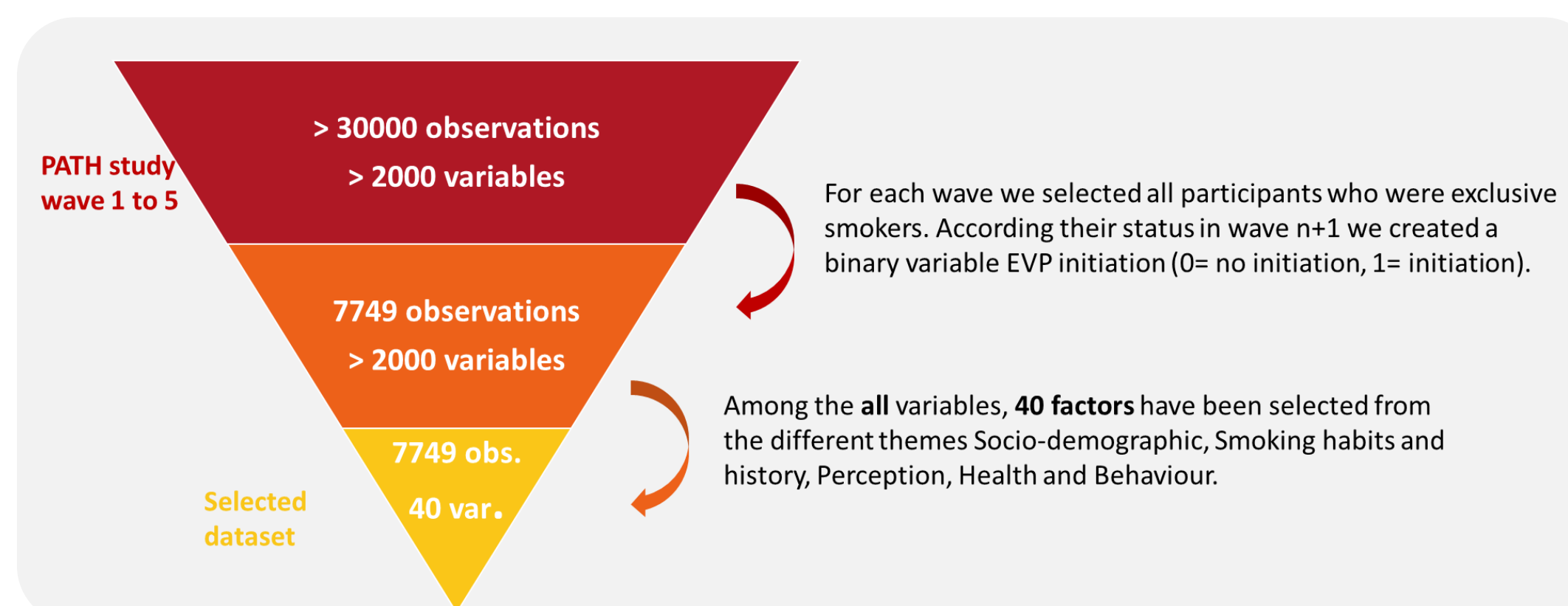
1. APPROACH

Our overall approach is outlined below, starting with data engineering, followed by machine learning and data interpretation. The processes include data collection and reprocessing, creation of training models, evaluating performance of models, selecting the most appropriate model and data interpretation.



2. DATA ENGINEERING

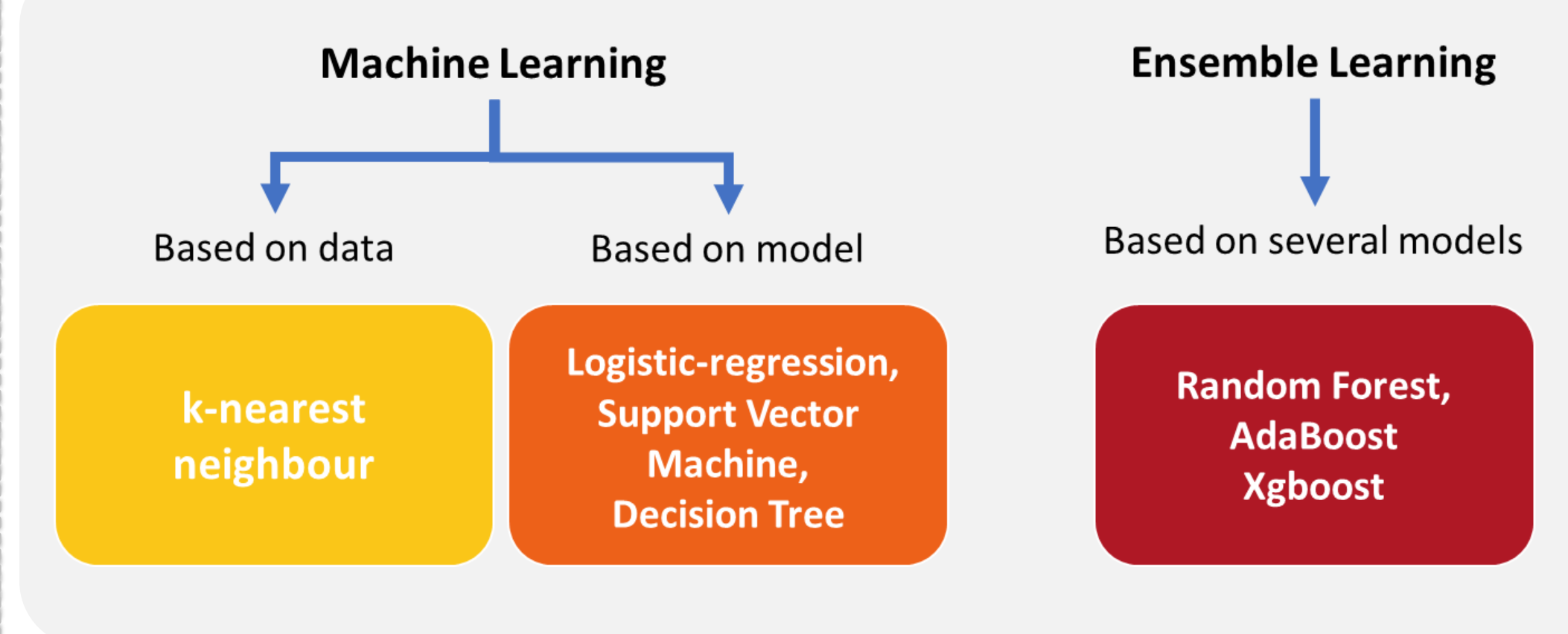
Based on the data from PATH wave1 to wave 5, we extracted data for 7,749 exclusive smokers and identified 40 variables.



To ensure that the machine learning algorithms can interpret and learn from the categorical data effectively, one-hot encoding has been used to transform categorical variables into a numerical format.

3. MACHINE LEARNING

Machine learning algorithms learn from data by identifying patterns, relationships, and structures within the data. A variety of seven Machine Learning algorithms were tested using Python and its Pycaret library [2]. These seven algorithms can be categorized as machine learning on based data or model, and ensemble learning based several models.



Fine tuning: each model was optimized using fine tuning to select their best hyperparameters, and then the models were compared.

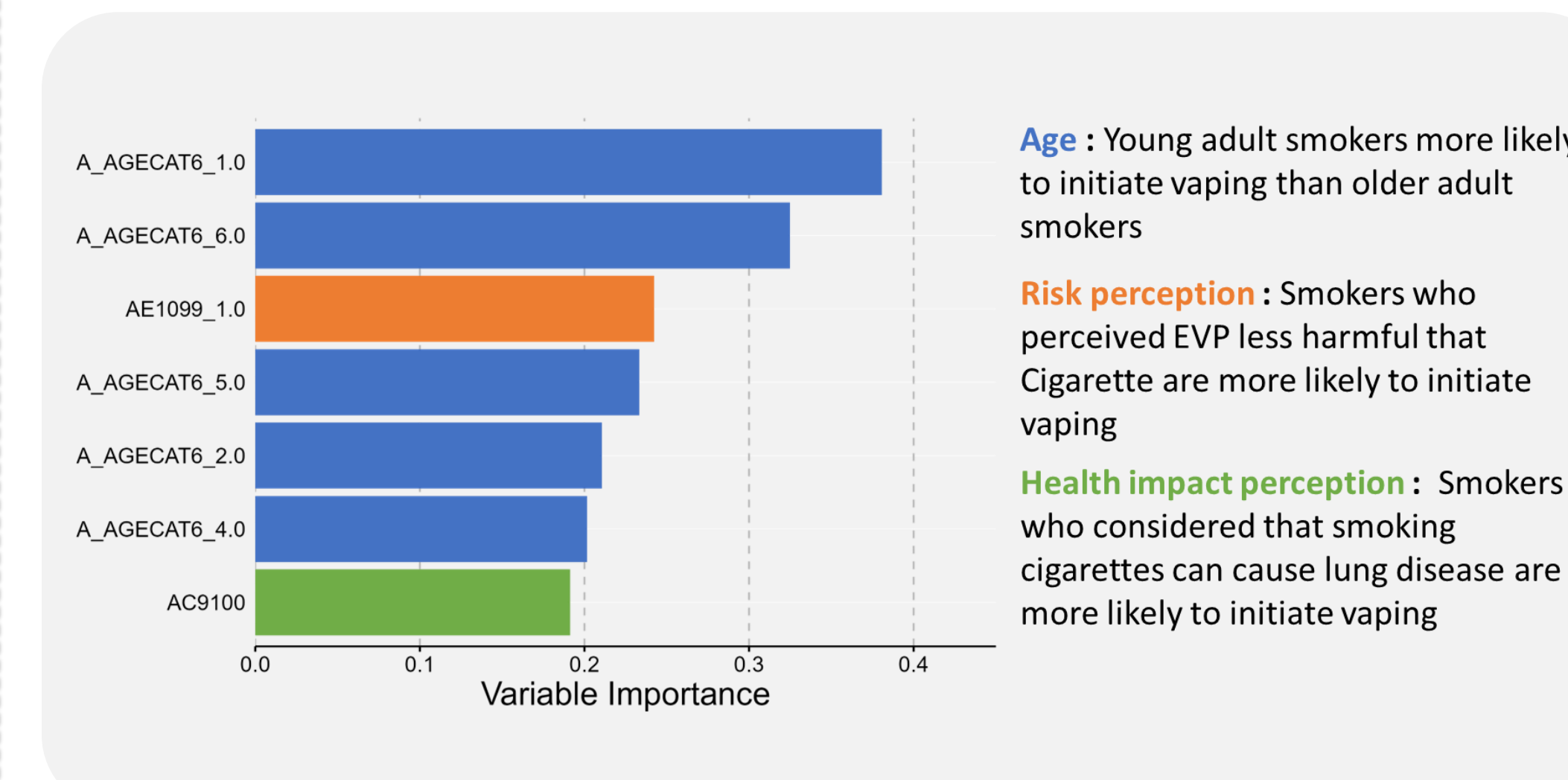
Performance assessment: The performance of the seven classification methods were sorted by "F1-score". F1-score is a metric to evaluate the performance of classification model based on a balance between precision and recall.

Model	Accuracy	AUC	Recall	Prec.	F1
Logistic Regression	69%	69%	54%	24%	33%
SVM - Linear Kernel	60%	68%	70%	22%	33%
Random Forest Classifier	68%	66%	50%	22%	31%
Extreme Gradient Boosting	53%	66%	74%	20%	31%
K Neighbors Classifier	53%	57%	60%	17%	27%
Decision Tree Classifier	14%	55%	100%	14%	25%
Ada Boost Classifier	85%	69%	3%	33%	5%

Model selected: Based on F1-score, **Logistic regression** model has been selected

4. DATA INTERPRETATION

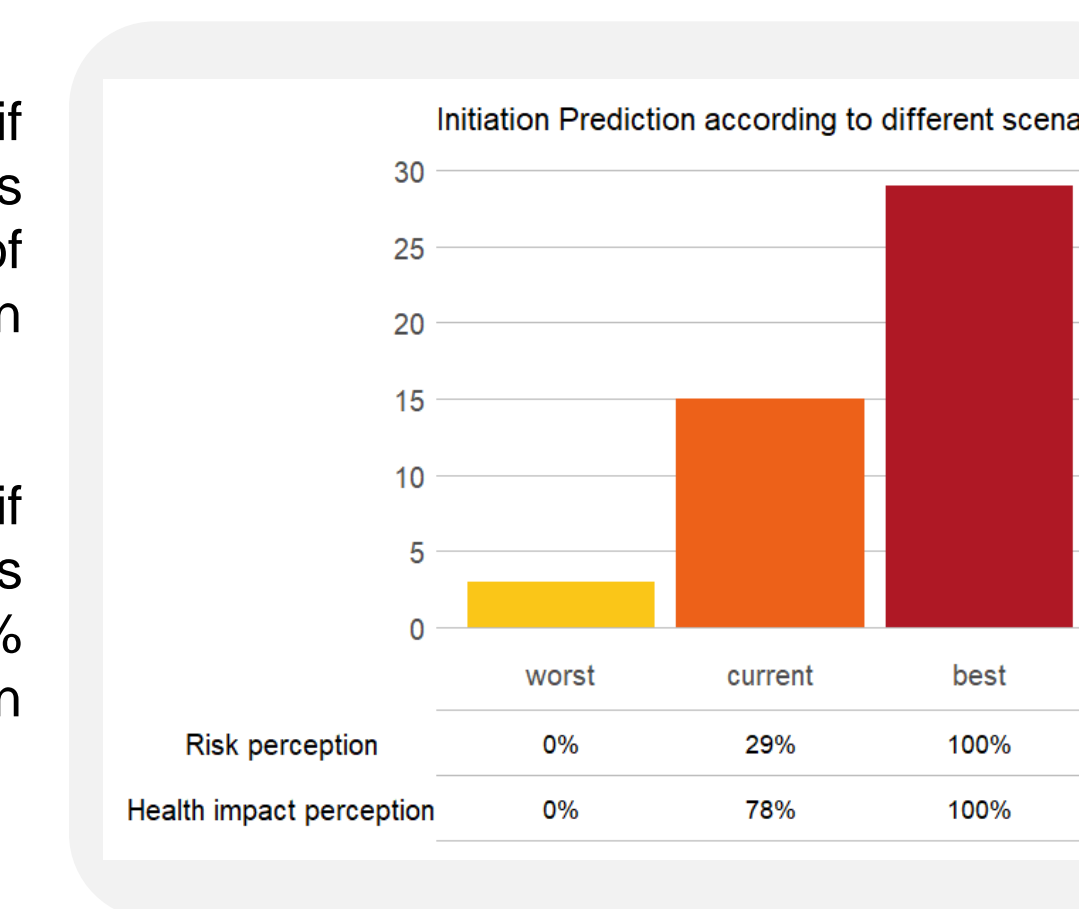
'Variable Importance' is a valuable tool to interpret machine learning models and understand which factors have the most significant impact on the model's predictions.



5. SIMULATION

From current PATH data, 15% of smokers used EVP. In our model, we predicted how risk perception and perception of health impact may influence how many smokers use EVP.

- The initiation of EVP would be low if 0% of smokers perceived EVP as less harmful than cigarette, and 0% of smokers considered cigarettes can cause lung disease.
- The initiation of EVP would be high if 100% of smokers perceived EVP as less harmful than cigarette, and 100% of smokers considered cigarettes can cause lung disease



A change in risk perception and health impact perception could potentially increase the EVP initiation rate of smokers from 3% to 30%

6. LIMITATIONS

The limited predictive power, indicates that the available factors do not fully explain the initiation of EVP. Additional factors might be necessary to gain a complete understanding of the initiation of EVP by smokers.

People's risk perception may change from one year to another, which is not accounted for in the model. As a result, the model may incorrectly classify some individuals as having initiated vaping than they actually not have (false positives) due to an increase in their relative risk perception. Conversely, it may also incorrectly classify others as not having initiated vaping when they actually have (false negatives) due to a decrease in their relative risk perception.

7. CONCLUSIONS

These findings provide valuable insights for public health strategies aiming to reduce smoking prevalence, particularly among younger adult individuals. Targeted education and awareness campaigns that address risk perception and health impact perception could be effective in helping adult smokers to initiate EVP use and thereby transitioning away from smoking.

REFERENCES

- United States Department of Health and Human Services. National Institutes of Health. National Institute on Drug Abuse, and United States Department of Health and Human Services. Food and Drug Administration. Center for Tobacco Products. Population Assessment of Tobacco and Health (PATH) Study [United States] Public-Use Files. Inter-university Consortium for Political and Social Research [distributor], 2024-06-14. <https://doi.org/10.3886/ICPSR36498.v20>
- Moez Ali (2020), PyCaret: An open source, low-code machine learning library in Python, PyCaret version 1.0.0, <https://www.pycaret.org>