Predicting Smoking Behavior Using Interpretable Machine Learning on BRFSS Survey Data

Snigdha Pakala
Department of Statistics
University of Michigan
Ann Arbor, MI, USA
vpakala@umich.edu

***Abstract*—Smoking remains a leading cause of preventable death and disease worldwide. This paper presents a machine learning analysis of the 2020 BRFSS dataset to predict smoking behavior based on demographic and health features. Models including logistic regression, Random Forest, XGBoost, and a neural network were trained and interpreted using SHAP values, calibration curves, ablation studies, and t-SNE. Key predictors included general health status, diabetic condition, age category, and alcohol use. Visualization techniques and error diagnostics supported interpretability, providing insight into at-risk groups and potential for targeted health interventions.**

***Index Terms*—smoking prediction, public health, BRFSS, interpretable machine learning, neural networks, SHAP, calibration, t-SNE**

I. Introduction

Smoking remains a leading cause of preventable death worldwide, contributing to a wide range of serious health conditions including lung cancer, cardiovascular disease, and chronic respiratory illnesses. Despite longstanding awareness of its detrimental effects, smoking continues to pose a significant burden on public health systems. More recently, the increasing prevalence of electronic nicotine delivery systems, such as vaping—particularly among youth—has introduced new dimensions to this ongoing challenge, complicating both prevention efforts and epidemiological monitoring.

Understanding the sociodemographic and health-related factors associated with smoking behavior is essential for developing effective, targeted interventions. Traditional epidemiological studies have long identified links between smoking and chronic illnesses, but recent advancements in machine learning offer the opportunity to uncover more complex, nonlinear relationships within high-dimensional health data. This project leverages the 2020 Behavioral Risk Factor Surveillance System (BRFSS) dataset from the Centers for Disease Control and Prevention (CDC) to predict smoking behavior through interpretable machine learning models. The primary objective is to identify the most predictive factors and provide transparent insights into which populations are at elevated risk.

Several studies have demonstrated the utility of machine learning in public health research. For example, Marques et al. [1] reviewed emerging evidence on the effects of e-cigarette use, highlighting the complexities of modeling smoking-related behavior in the modern landscape, where traditional and electronic nicotine products coexist. Issabakhsh et al. [2] employed ensemble methods such as Random Forest and Gradient Boosting to predict smoking cessation among U.S. adults. These models demonstrated the ability to capture nonlinear associations and interaction effects, particularly those linked to mental health and comorbidities. And recent work by Lundberg et al. [3] introduced SHAP (SHapley Additive exPlanations) as a framework for interpreting complex models, enhancing both transparency and trust in predictive systems used in healthcare contexts. In addition, advances in deep learning have further enhanced model expressiveness. Neural networks such as Multilayer Perceptrons (MLPs), particularly when equipped with regularization strategies like dropout and feature masking, have shown strong generalization capabilities across large and heterogeneous survey datasets like BRFSS and NHANES. By building on this foundation, the present work contributes a comprehensive analysis that includes both classical models and neural networks, evaluated through rigorous calibration and visualization techniques including t-SNE projections, ablation studies, and SHAP-based interaction diagnostics.

II. Methods

A. Problem Formulation

Input: 17 features from BRFSS (e.g., BMI, mental health, alcohol use, demographics)

Output: Binary classification – Smoker (1) or Non-Smoker (0)

B. Methodology Walkthrough

This project employed a multi-stage methodological framework combining classical machine learning, neural networks, hyperparameter tuning, and model interpretability tools. The primary dataset used was the 2020 Behavioral Risk Factor Surveillance System (BRFSS), a comprehensive health survey conducted by the CDC, containing over 300,000 records. The prediction task was framed as a binary classification problem, with the goal of identifying whether a respondent was a current smoker based on a variety of demographic and health-related features.

Data preprocessing was a critical step to ensure model readiness. Categorical features such as race and sex were converted using one-hot encoding, while ordinal features like general health and age category were mapped to integer

scales based on meaningful orderings. The "Diabetic" feature received a custom encoding scheme to reflect varying severity levels (e.g., borderline diabetes was coded as 0.5). Binary fields with "Yes/No" responses were mapped to 1 and 0, respectively. All continuous variables were standardized using z-score normalization via the StandardScaler.

Three classical machine learning models were trained and compared: logistic regression, Random Forest, and XGBoost. Logistic regression served as a baseline. The Random Forest model was optimized using grid search over key hyperparameters, including the number of trees, maximum depth, and minimum samples per split, evaluated using five-fold stratified cross-validation and ROC AUC as the scoring metric. XGBoost was trained with log-loss as the evaluation metric and was included due to its strong performance and compatibility with model interpretation techniques.

In parallel, a neural network model was developed using the PyTorch deep learning framework. The architecture consisted of two hidden layers with 64 and 32 neurons respectively, each followed by ReLU activation and dropout regularization (dropout rate of 0.3), and a sigmoid output layer for binary classification. The network was trained using the Adam optimizer and binary cross-entropy loss for 20 epochs with mini-batches of size 64. Dropout also functioned as an ablation mechanism to test robustness against overfitting.

Model performance was assessed using a combination of quantitative and visual evaluation techniques. These included classification reports with precision, recall, and F1-score; ROC curves and corresponding AUC values; Brier scores to assess probability calibration; and confusion matrices to examine the distribution of errors. The neural network was further evaluated using the Expected Calibration Error (ECE) to determine alignment between predicted probabilities and observed outcomes.

To enhance interpretability, SHAP values were computed for the XGBoost model, providing both global and local explanations for feature importance. Permutation importance analysis was used to confirm which features most strongly influenced predictions. An ablation study was conducted by removing top-ranked features such as general health to assess the resulting impact on model AUC. Dimensionality reduction via t-distributed stochastic neighbor embedding (t-SNE) enabled visual exploration of latent feature space, revealing distinct clusters for smokers and non-smokers. Additionally, a histogram of predicted probabilities from the neural network revealed a bimodal structure, indicating confident model decisions, and calibration curves illustrated the reliability of model outputs.

Together, these methodologies formed a robust analytical pipeline that balanced performance with interpretability, allowing for rigorous investigation of smoking-related predictors in a public health context.

III. Results

A. Classical Model Evaluation

Evaluation metrics indicated that XGBoost achieved the highest ROC AUC (0.684), followed closely by Random Forest (0.679). Logistic regression performed notably worse, suggesting the presence of nonlinear patterns within the data.

The Random Forest model produced a moderately calibrated probability distribution, with a Brier Score of 0.2207 indicating some overconfidence in its predictions. XGBoost achieved a slightly better Brier Score of 0.2186, suggesting improved probability estimation. Although the AUC scores across models were not substantially different, tree-based models consistently outperformed both logistic regression and the neural network in discriminatory power.

B. Feature Importance and Model Interpretability

SHAP summary plots indicated that "GenHealth," "Diabetic," and "AgeCategory" were among the most influential variables in predicting smoking. A SHAP dependence plot for "GenHealth" versus "AgeCategory" revealed that older adults in poor self-reported health had a higher likelihood of being smokers.

Permutation importance rankings confirmed these key features, and the ablation study further strengthened confidence: removing "GenHealth" dropped the Random Forest's ROC AUC by a noticeable margin, echoing SHAP's insights. "MentalHealth" and "Diabetic" followed close behind, suggesting both clinical and psychological factors significantly affect smoking risk.

C. Neural Network Performance

The final Multilayer Perceptron (MLP) achieved a ROC AUC of 0.674 and a Brier Score of 0.222. Though performance trailed slightly behind XGBoost and Random Forest, dropout layers minimized overfitting. The neural network was also trained with random feature masking, where a small proportion of input features were randomly set to zero during training. This approach served as a regularization technique, reducing the model's reliance on specific features and encouraging more distributed learning. The inclusion of random masking helped the MLP maintain stable performance while reducing overfitting.

The model was conservatively biased, with more false positives than false negatives, which is preferred in smoking detection, since missing actual smokers poses a greater health risk. The Expected Calibration Error (ECE) was low (0.0231), indicating high probability reliability despite the model's complexity.

D. Visualization Insights

The SHAP summary plot highlights the top predictive features affecting smoking classification. Higher age, male gender, poorer general health, and alcohol consumption significantly increased SHAP values, implying that these factors push the model toward classifying someone as a smoker.

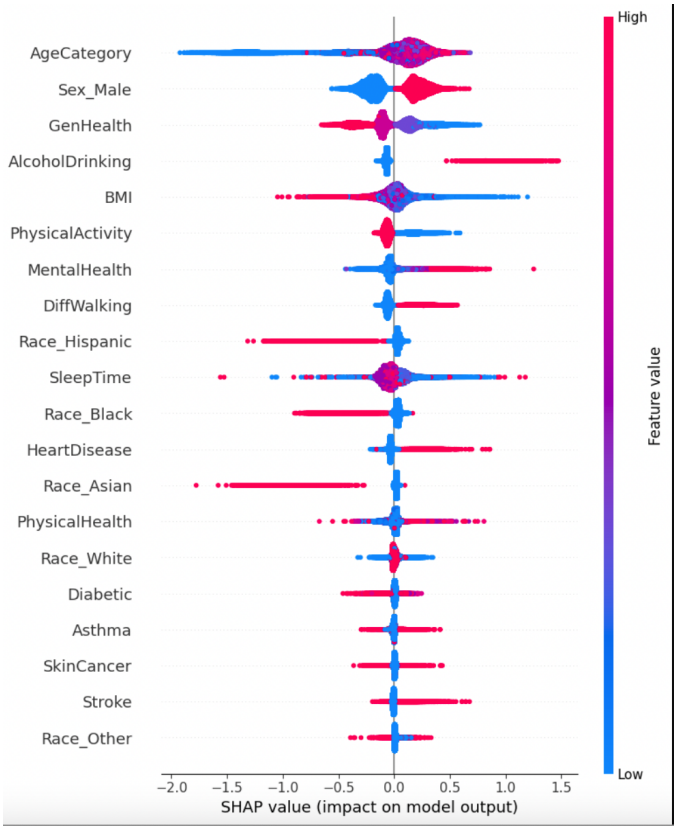

Fig. 1. SHAP summary plot highlighting top predictive features by importance and direction of effect.

The histogram of predicted probabilities from the Multilayer Perceptron (MLP) shows that most values fall between 0.3 and 0.5, indicating moderate confidence in classification. This cautious prediction pattern suggests the model avoids overconfidence but may struggle to sharply distinguish between smokers and non-smokers. The distribution aligns with the model's balanced calibration performance observed during evaluation.

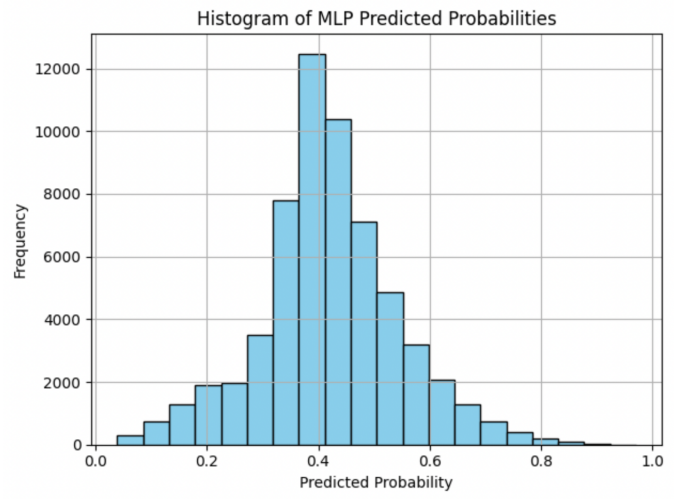

Fig. 2. Histogram of MLP predicted probabilities. Most values cluster around 0.4, reflecting moderate confidence and conservative prediction boundaries.

The feature ablation plot shows the change in ROC AUC when individual features are removed. `AgeCategory` had the largest impact, reducing AUC by approximately 0.02 when omitted, followed by `Sex_Male`, `GenHealth`, and `AlcoholDrinking`. These findings affirm the predictive strength of demographic and self-reported health variables. In contrast, several clinical features such as `Stroke` and `KidneyDisease` had minimal or even negative impact on model performance, suggesting redundancy or limited standalone predictive value.

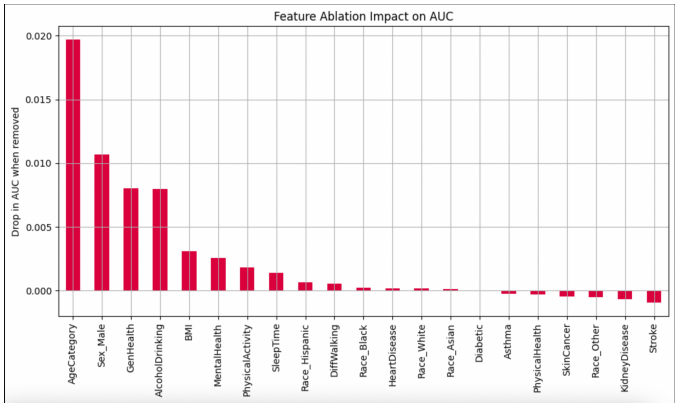

Fig. 3. Feature Ablation Study: Drop in ROC AUC when key features are removed individually.

The t-SNE visualization shows that the feature space learned by the model captures distinct patterns associated with smoking status. Clusters of smokers and non-smokers emerge in the ground truth plot, and the MLP approximates these clusters reasonably well. However, the less distinct separation in the prediction plot suggests that the model has limited confidence around decision boundaries, consistent with the previously observed probability histogram.

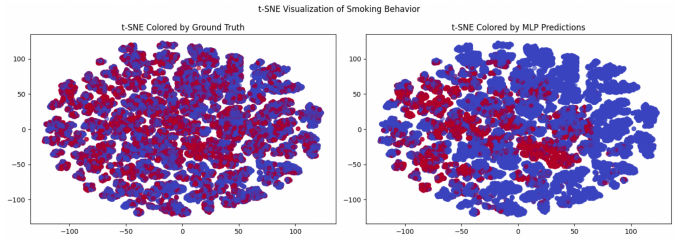

Fig. 4. t-SNE projection of the test set. The left plot shows clustering of smokers (red) and non-smokers (blue) based on ground truth labels. The right plot shows corresponding predictions from the MLP model, highlighting learned structure and areas of uncertainty near class boundaries.

IV. Conclusion

Robust models were constructed to predict smoking behavior using the BRFSS dataset. Analysis revealed that general health status, alcohol use, age, and sex were among the most predictive features. The use of SHAP values, permutation importance, and ablation studies enabled transparent interpretation of both classical and neural models.

Visualizations such as t-SNE projections, SHAP dependence plots, and predicted probability histograms added significant analytical depth. The t-SNE projection revealed partially separable clusters of smokers and non-smokers, indicating that the learned representations captured meaningful behavioral structure. SHAP dependence plots demonstrated how combinations of variables, such as poor general health and increasing age, jointly increased smoking risk. While the MLP model exhibited moderate confidence in its classifications, the low Expected Calibration Error (ECE) indicated that it was well-calibrated for probabilistic decision-making.

These insights carry important public health implications. By identifying features such as alcohol consumption, self-reported health status, and age as high-impact predictors, the model highlights key behavioral and demographic profiles where smoking intervention programs could be most effectively deployed. For example, older adults reporting poor general health or high alcohol intake may benefit from tailored cessation resources. The ability to segment populations based on risk also supports the allocation of prevention efforts and health education campaigns more efficiently.

Future directions may include integrating temporal BRFSS data to study evolving behavioral trends or expanding the feature space to include socioeconomic indicators such as income, education, or insurance access. More broadly, this work underscores the practical and ethical value of interpretable machine learning in designing targeted, data-driven strategies for population health management.

REFERENCES

[1] P. Marques, L. Piqueras, and M.-J. Sanz, "An updated overview of e-cigarette impact on human health," *Respir. Res.*, vol. 22, no. 151, May 2021. [Online]. Available: https://doi.org/10.1186/s12931-021-01737-5

[2] M. Issabakhsh, L. M. Sánchez-Romero, T. T. T. Le, A. C. Liber, J. Tan, Y. Li, R. Meza, D. Mendez, and D. T. Levy, "Machine learning application for predicting smoking cessation among US adults: An analysis of waves 1–3 of the PATH study," *PLOS ONE*, vol. 18, no. 6, p. e0286883, Jun. 2023. [Online]. Available: https://doi.org/10.1371/journal.pone.0286883

[3] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, Jan. 2020. [Online]. Available: https://doi.org/10.1038/s42256-019-0138-9