Smoking Status Prediction Project
Kevin Sullivan

Problem, Dataset Overview, Feature Engineering

According to the World Health Organization, smoking kills upwards of 8 million people per year, including 1.3 million who are harmed by second-hand smoke.[1] Furthermore, smoking can cause significant health issues that can be costly for individuals and healthcare systems alike. Helping individuals quit smoking should be a priority for the public health community. Through this project, I evaluate the efficacy of different machine learning methods in predicting the smoking status of an individual given certain bio-signals. All coding was done using R version 4.4.2. The project idea and data are from the Kaggle competition titled **Quitting Smoking - BGU2025.**[2]

The dataset used for this report was generated by Kaggle using a deep-learning model to be used in a competition setting. There are 159,256 observations and 24 features. There was no missing data or outliers present. The key outcome feature is *smoking,* a binary variable that indicates if the observation does or does not smoke. The other 23 features are either categorical or continuous bio-signals such as cholesterol, triglycerides, and blood pressure. The goal of this project was to train an effective and pragmatic model that can accurately predict an individual's smoking status given certain health characteristics. For this project, effectiveness is defined as a low misclassification error rate, and pragmatic is defined as having the lowest number of features.

Significant feature engineering was undertaken due to the computational cost and time intensity needed to run robust models with this many features. Two training and validation set pairs were created; one pair containing all of the original 23 predictors and the other systematically reduced to 9 predictors to allow for iterative combinations of models to be run. Specific feature engineering included combining the *height* and *weight* variables to form a *BMI* variable, fitting a full logistic regression model and dropping non-significant features, and conducting an in-depth literature review to pinpoint the 9 most important predictors of smoking behavior. A thorough review of these decisions can be found in the **appendix.**

Methods

Predicting the smoking status of an individual based on bio-signals is an example of a binary classification problem. Because of this, various machine learning methods were evaluated including logistic regression, k-nearest neighbors, generative classification models (QDA, LDA, Naive Bayes), and methods for model optimization such as 5-fold and 10-fold cross-validation, model subset selection, and model shrinkage methods. The methods were split into two groups: selected feature set methods and full feature set methods.

As mentioned previously, it is computationally expensive to iterate through all possible combinations of features in this dataset. This would amount to $\sum_{i=1}^{23} \binom{23}{i} = 8,388,607$ models. To ensure a level of thoroughness while remaining computationally efficient, a selected feature set method was created by systematically shrinking the number of features under consideration using both public health literature and running a full logistic regression model (see **appendix)**. The selected feature set was reduced from 23 predictors to 9. This reduced dataset was then split 70/30 into training and validation sets with the training data used to fit the models and the validation set used to test their accuracy. All possible models for the following methods were run: logistic regression, LDA, QDA, Naïve Bayes, 5-fold CV, and 10-fold CV. For the kNN method, k varied from 1 to 100. Initially, when iterating through all possible combinations, the error of too many ties was repeatedly run into. To overcome this, models with all combinations of 3 or more features were fit (see **Table 1)**. The misclassification error was computed for each of these methods and the model with the lowest error for each method was considered the best.

Other techniques did not require the same level of computational complexity. For these models, the full training set with 23 predictors was used. These methods include stepwise selection and model

shrinkage methods such as Lasso and Ridge. For model shrinkage methods, λ varied from 10^{-4} to 10^{10}. Misclassification error was then calculated for each model with the lowest error being considered the best.

The model and associated relevant parameters with the lowest misclassification error rate and the fewest number of predictors were considered the final preferred model. Misclassification error was chosen as the main unit of analysis because it is the simplest to explain and captures an important requirement of an effective machine learning model: accuracy. Furthermore, models with a low number of predictors were preferred to make implementation easier on healthcare providers and public health specialists. An accurate model that requires fewer inputs is both practical and efficient, which in turn can lead to more effective public health interventions. Lastly, smoking status was predicted using the test data for each of the "best" models identified for each ML method, as defined by the criteria outlined above. These predictions were submitted to Kaggle and the associated score (area under the ROC curve between the predicted probability and the observed target) was computed and served as a further evaluative criterion.

Results

A total of 49,671 models across 6 different machine learning methods were run. The 10 best models, defined as having the lowest misclassification rate for each method, were then used to predict smoking status from the test set (**Table 2)**. The models ranged from 3-19 predictors with misclassification errors ranging from 0.204 to 0.385. Most models had a misclassification error of about 31%. The 5 and 10 fold CV methods yielded the models with the lowest misclassification error, while the forward stepwise model yielded the best ROC score when submitted to Kaggle.

The logistic regression model using 5-fold CV with five predictors (*age, HDL, LDL, triglyceride, dental caries)* was chosen as the preferred model. This model had a misclassification rate of 0.204, which was the lowest by far of the other methods. It also scored reasonably well on Kaggle, accruing a score of 0.75, which ranked fourth among all other models run. The forward stepwise model with 19 predictors performed the best on Kaggle but had a slightly higher error rate of 0.25 and was far more computationally complex than the preferred model.

Evaluation and Discussion

Predicting the prevalence of smoking with healthcare data to reduce smoking rates is a critical issue in the field of public health. Out of the 6 ML methods, 4 utilized a systematically reduced feature set while the other 2 used the full feature set. The logistic regression model using 5-fold CV with the predictors *age, HDL, LDL, triglyceride, and dental caries* was the most accurate model with a small number of features. These features are all commonly collected in a healthcare setting and are minimally invasive to the patient. This makes this model pragmatic. It can be run without placing a burden on healthcare providers or patients and thus has the potential to lead to accurate, effective, and actionable predictions.

The other ML methods such as kNN, LDA, QDA, Naive Bayes, Lasso, and Ridge were all minimally effective. This could be due to both the feature selection undertaken as well as the large sample size of the data. All of the models that were run had between a 20% and 35% misclassification error. This is a high error rate and may be due to the feature selection techniques implemented. While systematic and grounded in public health literature, it may not have captured the full nuances present in the data. Similarly, numerous factors influence whether or not someone smokes that go far beyond the scope of this data. Biological signals alone can only tell specialists so much, and this is reflected in the error rates found.

Future Work

Predicting the smoking status of an individual solely off of bio-signals is a complex and imperfect task. The decision to smoke is nuanced and can be affected by socioeconomic, demographic, and health factors. Because of this, future work that combines these various areas is needed to not only create more

accurate ML models but to provide a more nuanced approach that can benefit both smokers and non-smokers alike.

Appendix

Table 1: Methods of Machine Learning Evaluated

Method	Package/Function	Total Models Evaluated	Notes and Remarks
Logistic Regression	Base R, 4.4.2 *glm, family = "binomial"*	$\sum_{i=1}^{9} \binom{9}{i} = 511$	Reduced training and validation sets used
K-Nearest Neighbor	Caret *knn3*	$\sum_{i=3}^{9} \binom{9}{i} * 100 = 46,600$	Reduced training and validation sets used K varied from 1 to 100 Only models with 3 or more features were considered due to "too many ties" error with lower feature level models
Generative Classification Models: LDA, QDA, Naïve Bayes	MASS, e1071 *lda, qda, naiveBayes*	$\sum_{i=1}^{9} \binom{9}{i} * 2 = 1{,}022$	Reduced training and validation sets used
Resampling Methods: 5-Fold and 10-Fold Cross Validation	Boot *cv.glm*	$\sum_{i=1}^{9} \binom{9}{i} * 3 = 1,533$	Reduced training set used Errors obtained from cross-validation
Model Subset Selection Methods: Forward, Backward, and Forward-Backward Stepwise Selection	MASS *StepAIC*	3	Full training set used
Model Shrinkage Methods: Lasso and Ridge Regression	Glmnet *glmnet*	2	λ varied from 10^{-4} to 10^{10} Full training set used

Table 2: Final Model Performance Evaluation

Algorithm	Model	Remarks	Error*	Kaggle Score**
Logistic	*Smoking* ~ Age + Relaxation + Systolic + HDL + Cholesterol + Triglyceride + Dental Caries		0.3190	0.7540
kNN	*Smoking* ~ Age + BMI + Relaxation	k = 1	0.3140	0.7350
kNN	*Smoking* ~ Age + BMI + HDL	k = 1	0.3160	0.7230
LDA	*Smoking* ~ Age + Relaxation + Systolic + HDL + LDL + Triglyceride + Dental Caries		0.3190	0.7550
QDA	*Smoking* ~ Age + LDL + Cholesterol + Triglyceride		0.3230	0.7370
Naive Bayes	*Smoking* ~ Age + Relaxation + Systolic + HDL + Cholesterol + Triglyceride		0.3220	0.7380
5-Fold Cross-Validation†	*Smoking* ~ Age + HDL + LDL + Triglyceride + Dental Caries	All models with 5-fold CV performed better than all other methods. No 5-fold CV model had an error rate > 25%	0.2040	0.7530
10-Fold Cross-Validation	*Smoking* ~ Age + Cholesterol + Triglyceride + Dental Caries		0.2040	0.7507
Forward Select	*Smoking* ~ Hemoglobin + Height (cm) + Gtp + Triglyceride + Cholesterol + ALT + Dental Caries + Weight (kg) + BMI + Systolic + Fasting Blood Sugar + AST + HDL + Relaxation + Urine Protein + Eyesight (Right Eye) + Waist Size (cm) + LDL + Hearing (Right Ear)	Stepwise	0.2508	0.8380
Lasso	*Smoking* ~ Age + BMI + Cholesterol + Dental Caries + LDL + HDL + Triglyceride + Systolic + Relaxation	λ = 0.000438	0.3850	0.6200

[a] *Misclassification error*
**ROC score (Higher score = better)
†Preferred final model

Systematic Feature Reduction Methodology

There is often a trade-off between thoroughness and computing power when fitting machine learning models. I ran into this problem firsthand through this project. The dataset provided by Kaggle had 24 features. Ideally, I wanted to iterate through all possible combinations of features and fit each model using different machine learning methods. This would allow for a thorough analysis of all possible predictors and hopefully lead to the best model. However, this would involve fitting over 8 million models, a computationally impossible task due to my personal computing and time constraints.

This led me to seek out a systematic way to reduce the number of features in the dataset to a manageable number without losing predictive power. I settled on two approaches to accomplish this. First, I fit a full logistic regression model to see which variables were not statistically significant in predicting *smoking*. This led me to drop the following predictors: *id, hearing.left, hearing.right, eyesight.left, serum.creatinine* leaving me with 18. I then conducted a literature review to determine what public health experts consider the best biological predictors of smoking behavior. A condensed summary of this literature review can be found in **Table 3.** Although *age* was found to be statistically insignificant in the logistic regression model, the literature indicated it should be kept so I did not remove it. Lastly, I combined the *height* and *weight* variables to form a *BMI* variable. This left me with 9 features to predict *smoking*.

Figure 1: Full Logistic Regression Results Used for Feature Reduction

```
glm(formula = smoking ~ ., family = "binomial", data = train_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.5553  -0.7516  -0.2798   0.8641   6.5616

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -3.752e+01  1.038e+00 -36.135  < 2e-16 ***
BMI                  3.535e-01  2.096e-02  16.866  < 2e-16 ***
id                  -8.060e-08  1.351e-07  -0.597 0.550838
age                 -3.520e-04  6.876e-04  -0.512 0.608665
height.cm.           1.912e-01  6.226e-03  30.713  < 2e-16 ***
weight.kg.          -1.450e-01  7.364e-03 -19.685  < 2e-16 ***
waist.cm.            4.498e-03  1.545e-03   2.912 0.003594 **
eyesight.left.       6.128e-03  1.724e-02   0.355 0.722300
eyesight.right.      6.470e-02  1.777e-02   3.640 0.000272 ***
hearing.left.        3.372e-02  5.077e-02   0.664 0.506549
hearing.right.       7.052e-02  5.160e-02   1.367 0.171698
systolic            -1.261e-02  7.812e-04 -16.147  < 2e-16 ***
relaxation           6.720e-03  1.076e-03   6.245 4.25e-10 ***
fasting.blood.sugar  6.967e-03  4.391e-04  15.866  < 2e-16 ***
Cholesterol         -6.765e-03  7.650e-04  -8.843  < 2e-16 ***
triglyceride         6.670e-03  1.903e-04  35.041  < 2e-16 ***
HDL                 -6.484e-03  9.494e-04  -6.829 8.52e-12 ***
LDL                 -1.890e-03  7.313e-04  -2.584 0.009767 **
hemoglobin           4.807e-01  6.635e-03  72.453  < 2e-16 ***
Urine.protein       -1.068e-01  1.860e-02  -5.743 9.30e-09 ***
serum.creatinine     2.616e-02  4.148e-02   0.631 0.528274
AST                 -8.397e-03  1.016e-03  -8.264  < 2e-16 ***
ALT                 -7.002e-03  6.370e-04 -10.992  < 2e-16 ***
Gtp                  1.999e-02  3.151e-04  63.454  < 2e-16 ***
dental.caries        3.660e-01  1.548e-02  23.644  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 218270  on 159255  degrees of freedom
Residual deviance: 156190  on 159231  degrees of freedom
AIC: 156240
```

Table 3: Key Findings from Literature Review – Features in dataset identified as best predictors of smoking behavior

Variable	Rationale	Source
Age	Smoking behavior is highly age dependent.	Identifying Predictors of Smoking Switching Behaviours Among Adult Smokers in the United States: A Machine Learning Approach: Cao et al.
BMI	Smokers often have lower BMI, making this a relevant predictor.	Identifying Predictors of Smoking Switching Behaviours Among Adult Smokers in the United States: A Machine Learning Approach: Cao et al.
Systolic Blood Pressure	Smoking is a key contributor to higher blood pressure.	Predictors of 7-Year Changes in Exercise Blood Pressure, Effects of Smoking, Physical Fitness and Coronary Function: Mundal et al.
Diastolic Blood Pressure	Smoking is a key contributor to higher blood pressure.	Smoking status and its effect on blood pressure A study on medical students: Jena & Purohit
Cholesterol	Smoking can significantly impact lipid profiles.	Smoking Prediction Using Bio-Signals: Alquran et al.
HDL	Smoking can significantly impact lipid profiles.	Smoking Prediction Using Bio-Signals: Alquran et al.
LDL	Smoking can significantly impact lipid profiles.	Smoking Prediction Using Bio-Signals: Alquran et al.
Triglycerides	Smoking is associated with higher triglycerides.	Meta-analysis of the effects of smoking and smoking cessation on triglyceride levels: Van der Plas et al.
Dental Caries	Smokers are more likely to have dental issues.	Correlation between tobacco smoking and dental caries: A systematic review and meta-analysis: Jiang et al.

References

Report

1. https://www.who.int/news-room/fact-sheets/detail/tobacco
2. https://www.kaggle.com/competitions/quitting-smoking-bgu-2025/overview

Appendix

- Smoking Prediction Using Bio-Signals: Alquran et al.
- Identifying Predictors of Smoking Switching Behaviours Among Adult Smokers in the United States: A Machine Learning Approach: Cao et al.
- Predictors of 7-Year Changes in Exercise Blood Pressure, Effects of Smoking, Physical Fitness and Coronary Function: Mundal et al.
- Smoking status and its effect on blood pressure A study on medical students: Jena & Purohit
- Meta-analysis of the effects of smoking and smoking cessation on triglyceride levels: Van der Plas et al.