Interpreting the result

Before training, we utilized label encoding and ordinal encoding to preprocess the response variable and categorical variables. During EDA, we don't find any correlation between pairs feature variables.

To predict the smoking status of the participant, we utilized different machine learning algorithms, including logistic regression, random forest, and support vector machine. Specifically, we use logistic regression as the base model. For other algorithms, since our dataset is not big enough, we performed k-folds cross validation to select the best model. For random forest, we utilized grid search to explore different combinations of hyperparameters (n_estimators, max_depth, max_features, min_samples_split, min_samples_leaf, bootstrap, criterion). And we used 5 folds in this case. For SVM, we first used randomized search to narrow down the search space by exploring different combinations of kernels and C. We also use 5 folds in this case. Then we used Grid search to explore the different combination of hyperparameters particularly for rbf kernels but got similar best performance as before using randomized search. Comparing the metrics (precision, recall, accuracy) of SVM and random forest, SVM after tuning is the better model for prediction with a 0.94 accuracy.

Limitations

At first, after processing our data, we find that a lot of data is missing, and after processing these missing data, nearly all data relate to Scotland and seldom people don't smoke. And, The variables selected in the data do not necessarily have a direct and decisive impact on predicting whether to smoke or not. So many categorical variables that we can't set up our function precisely. So, due to all these limitations above, our predictions will be affected. In this way, we need more data that contain people who don't smoke and different kinds of people that come from different countries.

Conclusion

In this analysis, we utilized machine learning techniques, specifically the Random Forest Classifier and SVC, to predict smoking behavior based on various features. We employed hyperparameter tuning using Randomized Search and Grid Search to find the optimal parameters for each model.The results of this analysis provide evidence that machine learning models can effectively predict smoking behavior based on the provided features. The importance of certain features highlights their relevance in understanding smoking habits.In the context of related research, this work contributes to the existing literature on predicting smoking behavior using machine learning. It reinforces the idea that machine learning models can be

valuable tools in public health research aimed at understanding and addressing smoking-related issues.