Ayushi Kate
ECE460J
October 22 2025

Binary Classification Model Report

Prediction of Smoker Status

1. Project Overview

The primary objective of this project was to analyze a dataset containing patient health indicators and build a model to accurately predict smoker status as a binary classification problem. The input data had information on features such as height, age, weight, eyesight, hearing, cholesterol, AST, ALT, GTP, and more health indicators. These features were used to develop and evaluate different binary classification models to assess their ability to predict the smoker status of people based upon their other health conditions. The smoker status was assigned as 0 for nonsmoking and 1 for smoking.

Data Preparation

The data preparation phase began with an examination of the initial dataset composition. The dataset was read using a Pandas DataFrame. The training dataset had an initial shape of (15000, 24), and the testing dataset had an initial shape of (10000, 23). The columns did not contain any null values.

To make the overall statistics easier to interpret, I combined both datasets into a single dataframe to perform univariate analysis to explore the distributions of some features. The analysis showed that there were approximately 10000 nonsmokers and 5000 smokers in the whole dataset. Using Seaborn, I also visualized the skewness distribution of features.

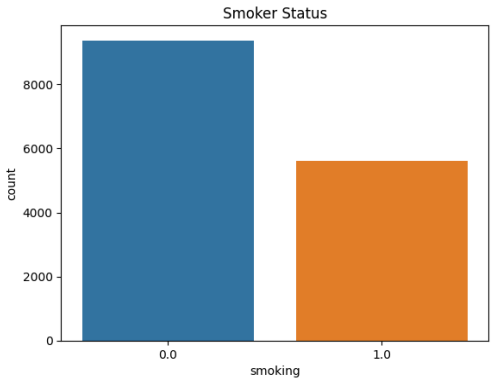


Figure 1: Distribution of smokers vs nonsmokers in the dataset.

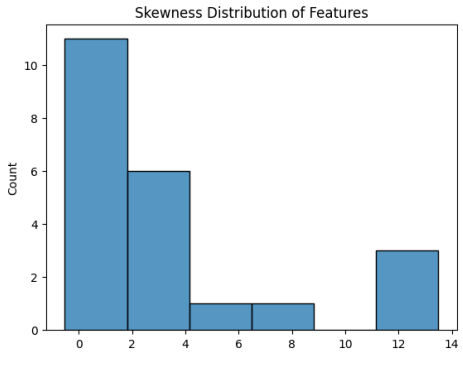
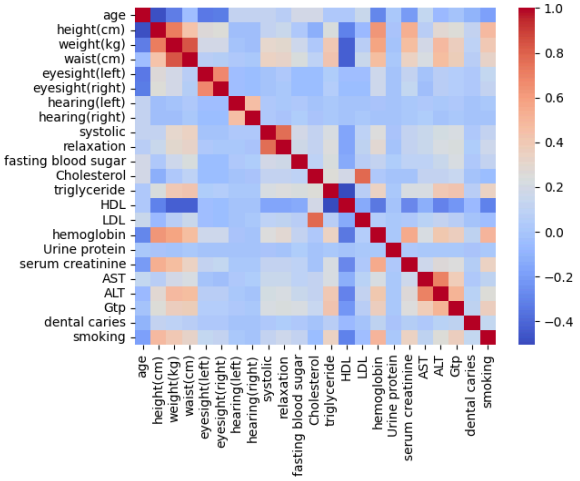


Figure 2: Distribution of skewness in the features.




```
Top 5 Correlated with Smoker
Status
hemoglobin          0.504360
height(cm)          0.476851
weight(kg)          0.409750
Gtp                 0.366762
serum creatinine    0.350503
Name: smoking, dtype: float64
```

Figure 3: Correlation matrix of all the features and their correlation with smoker status

Visualizing the correlation matrix allowed me to identify which columns had the highest effect on the smoker status. Using Seaborn, I created a correlation matrix and also displayed the top five columns that were correlated with smoker status. From this analysis, I found that hemoglobin, height, weight, Gtp, and serum creatinine had the highest correlation. Based on these results, I decided to create interaction terms with some of these columns to better analyze their combined effects.

Thus, through my feature engineering method, I created interaction terms of BMI and Cholesterol Ratio. The BMI feature was just the ratio of a person's weight and their height

in meters squared. The cholesterol ratio was their LDL level divided by their HDL level. Interaction terms are useful to capture relationships between features that might jointly influence the target variable. In many real world cases, the effect of a variable can depend on the level of another. Interaction terms help account for these combined effects and improve performance. Since both height and weight were so highly correlated with smoking status, it is possible that BMI could be a better predictor for this correlation.

2. Model Training and Evaluation

2.1 Training Procedure

Splitting the Data
The dataset was already split into the training data and testing data. I separated X_train and y_train from the train.csv. I also prepared the data by scaling it using StandardScaler on X_train and X_test.

Model Selection and Training
I trained several classification models using the cleaned dataset.

1. **Logistic Regression:** A baseline linear model
    a. L**ogistic Regression with L1 Penalty:** Added regularization, but its performance did not improve over the base model.
2. **Random Forest:** Tuned using a GridSearch to find best parameters in a param_grid defined above. This tested hyperparameters like n_estimators, max_depth, and min_samples_split.
3. **XGBoost:** Used this boosting model. Also tuned using GridSearch to find the best parameters in param_grid. Hyperparameters I selected included learning rate, n_estimators, max_depth, subsample, and colsample_bytree.
4. **Stacking**: Implemented stacking model that stacked XGB model and logistic regression. I found that it was really inaccurate when it had LR, RF, and XGBoost, so I simplified it down to only LR and XGBoost. The final estimator was the Logistic Regression model.

2.2. Model Performance

Model	Best ROC_AUC Score
Logistic Regression	0.7780
Logistic Regression (L1 Penalty)	0.7779
Random Forest	0.8840
XGBoost with using Feature Engineering Dataset	0.8917
XGBoost without using Feature Engineering Dataset	0.8923
Stacking	0.8922

2.3. Class Prediction

All of the models performed well in predicting smoker status, achieving an ROC_AUC score in the range of 0.7780 to 0.8923. The Logistic Regression models had the lowest performances, while XGBoost and Stacking achieved the highest scores. The models successfully utilized the cleaned features in the dataset to predict the outcome. To generate my final submission predictions, I ultimately chose to use XGBoost using the feature engineered dataset. This did not have the highest ROC_AUC score in comparison to the other models. However, I found that when I used Stacking or XGBoost without the FE dataset, my model's score was lower. I believe that this was due to overfitting of the model. Thus, the XGboost and incorporating feature engineering produced a more balanced model that generalized better and had higher predictive accuracy in terms of the ROC_AUC score.