DS202 Final Project

Tiffany Hou, Oscar Wan

April 2024

1 Background

The Automated Decision System (ADS) we selected is a solution to the Kaggle competition ”Smoker Status Prediction using Bio-Signals.” The goal of this competition is to develop a machine learning model to predict whether a person is a smoker using various bio-signals to address the problem of ineffective traditional smoking cessation methods like counseling. Providing a reliable predicting tool can facilitate targeted intervention for smokers and understanding the chance of quitting smoking for each individual smoker. This initiative responds to the challenge of high preventable morbidity and mortality rates due to smoking, aiming to improve health outcomes through better identification and support for smokers. The purpose of the ADS in response to the competition is then to accurately predict whether an individual is a smoker or non-smoker based on various bio-signal data.

2 Input and output

2.1 Data Description

The author of this ADS examines 2 datasets. One is ”Smoker Status Prediction using Bio-Signals,”. The data card of this dataset does not specify its origin. However, all the features (bio-signals) in this dataset are appropriate according to the Illinois Department of Health. The other dataset is ”Binary Prediction of Smoker Status using Bio-Signals,” which, according to the dataset description, was generated from a deep learning model trained on the former dataset. However, the model is only trained on the synthetic one and therefore, this audit will focus on this dataset.

2.2 Input Features

Table 1 shows all the features and their data types of the dataset. Figure 1 shows their distributions.

There is no missing value for any of the features. One note about the feature ”age” is that in the original data it is ordinal but not continuous. For example, a value of 25 means that the person is between 20 and 25 years old. Furthermore, although not explicitly stated, height and weight also seem to be in increment of 5 units judging from the distribution of the features. However, in the notebook we are auditing, the author uses all these variables as a continuous variable.

Data Type	**Features**
Integer	Age (5-yr gaps), Height (cm), Weight (kg), Hearing (L/R), Systolic BP, Diastolic BP, Fasting Blood Sugar, Total Cholesterol, Triglycerides, HDL Cholesterol, LDL Cholesterol, Hemoglobin, Urine Protein, Serum Creatinine, AST, ALT, GTP (Y-GTP), Dental Caries, Smoking (Target Variable)
Float	Waist Circumference (cm), Eyesight (L/R)

Table 1: Feature Names Grouped by Data Type

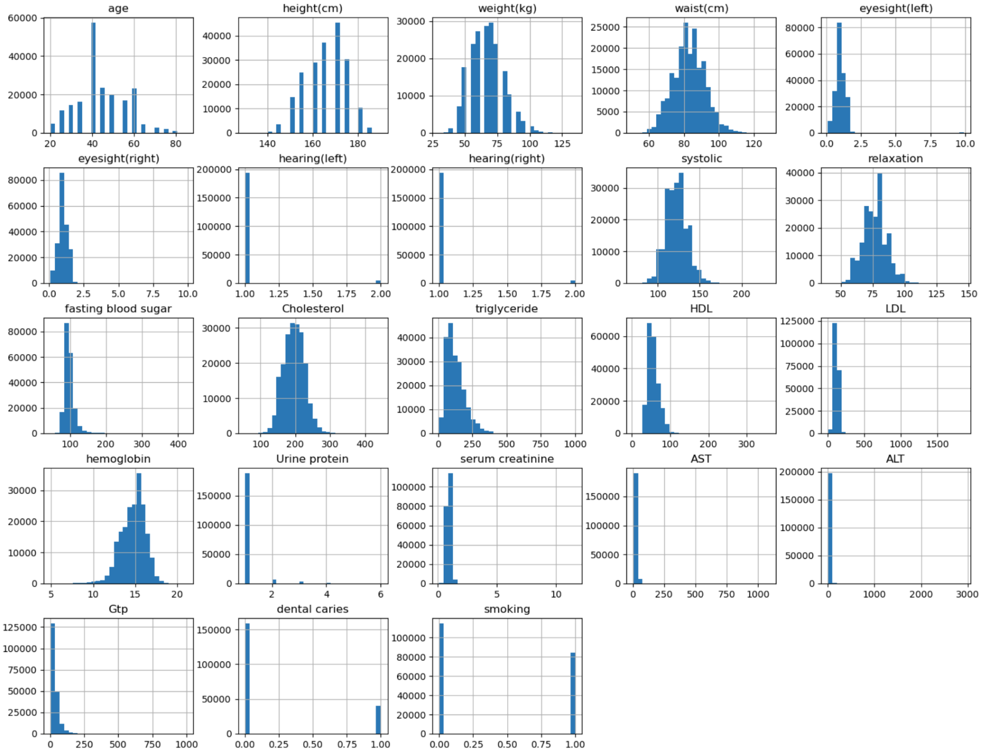

Figure 1: Histograms of All Features

Age displays a slightly right-skewed distribution, suggesting a relatively young population. Notably, the distribution peaks at 40. Both height and weight follow approximately normal distributions. Measures of eyesight and hearing are skewed towards the lower end, with a few outliers on the right, indicating good sensory functions across the dataset. Key health indicators such as blood pressure, blood sugar, and cholesterol-related parameters present a mix of distributional patterns, with some having normal curves and others showing skewness.

The distributions of serum-related features and dental caries are clustered at lower values with a few outliers with very high values. Lastly, smkoing status, the target of the ADS, is moderately unbalanced with more negatives.

Figure 2 shows all pairwise correlations between features, and table 2 shows the top 5 pairs of correlated features. From the the heatmap, we observe that many features have high correlation magnitude (above 0.4) with each other. For instance, Cholesterol and LDL, weight (kg) and waist (cm), left and right eyesight, and systolic and relaxation. Height and hemoglobin appears to have the strongest correlation with smoking status (0.44 and 0.43) while the others tend to have a weak to moderate correlation.

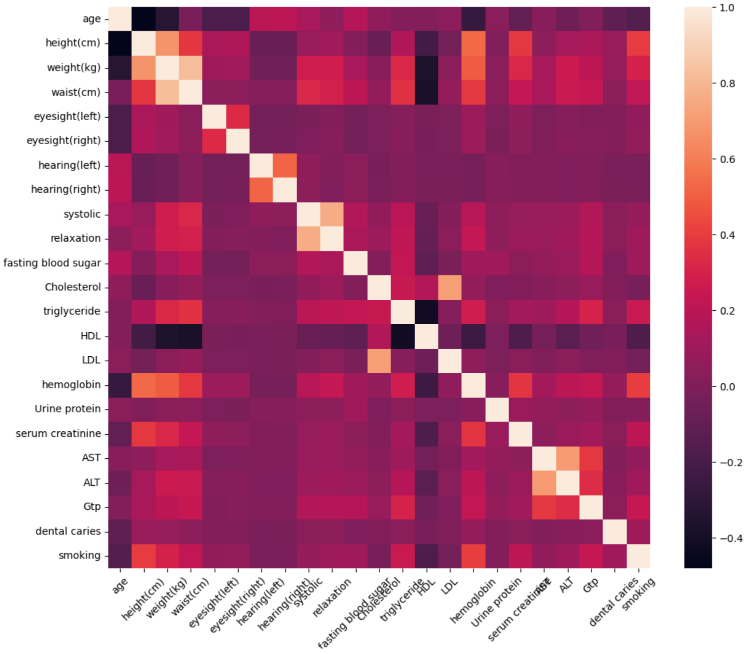

Figure 2: Correlation Matrix of the Variables

Feature 1	Feature 2	Correlation
Cholesterol	LDL	0.880968
weight(kg)	waist(cm)	0.829346
eyesight(left)	eyesight(right)	0.829092
systolic	relaxation	0.754610
AST	ALT	0.720063

Table 2: Top 5 Correlations between Features

Figure 3 shows profiling on age. We encoded age into young and old groups by 0 and 1 with median age 45 as the threshold. Those below 45 are encoded as young (0), and those above

45 are old (1). The general observation from the boxplots is that young people score better on health related attributes. With respect to the target variable ”smoking,” we observe that the proportion of smokers in the young group is much higher than that of the old group.

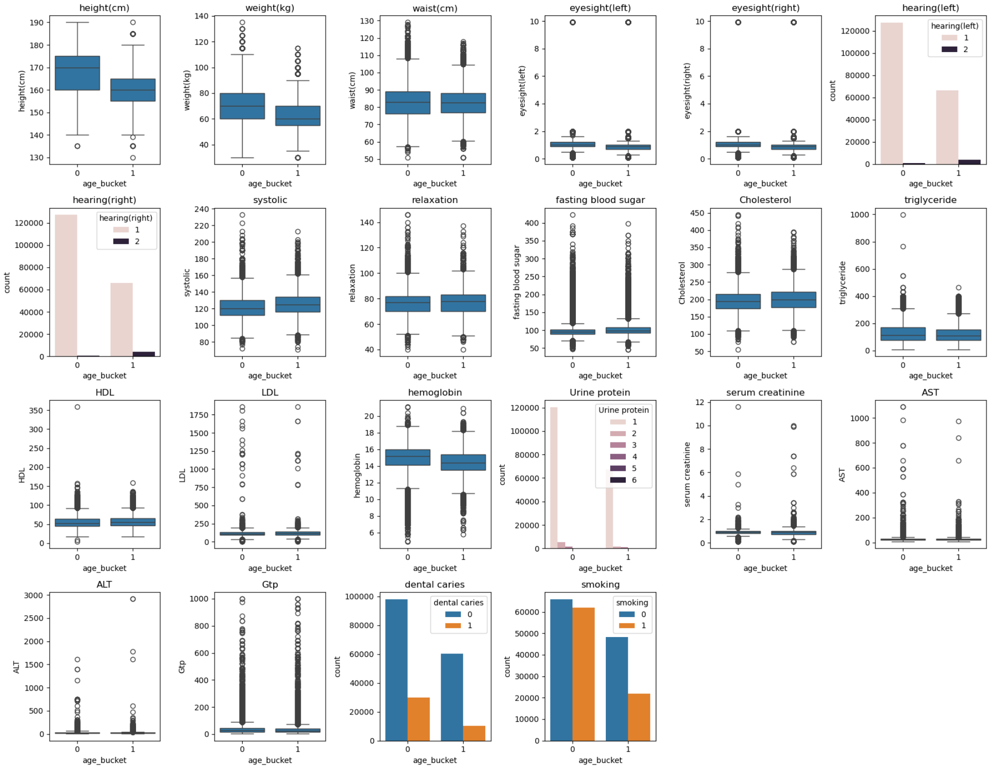

Figure 3: Profiling on Age

2.3 Output of the system

The output, ranging from 0 to 1, is the probability of an individual being a smoker. Higher values indicate a greater likelihood of smoking, while lower values imply a lower probability.

3 Implementation and Validation

3.1 Data Cleaning and Pre-processing

We found the following pre-processing to be potentially problematic:

1. **Hearing Attributes Transformation**
   - The minimum of the two values, `hearing(left)` and `hearing(right)`, is reduced by one unit and reassigned to `hearing(left)`.
   - The maximum value is similarly decreased by one unit and reassigned to `hearing(right)`.
2. **Eyesight Attributes Transformation**
   - Values for `eyesight(left)` and `eyesight(right)` that exceed 9 are reset to 0.
   - The minimum eyesight value is reassigned to `eyesight(left)`, and the maximum value is reassigned to `eyesight(right)`.

3. **Biochemical Levels Clipping**
    - Some biochemical markers are clipped within predefined medically relevant ranges. For example, GTP is clipped between 0 and 300.

For the first two transformations, the author does not explain the logic behind them, and they do not make sense but significantly alters the original data. For the third transformation, clipping, which changes the extreme values to the clipped range, may cause significant information loss, as very high or low biochemical levels can indicate critical medical conditions or unique physiological states that are clinically relevant.

Nevertheless, the rest of the pre-processing steps are appropriate, including one-hot encoding categorical variables. The author also uses robust scaling, which is very suitable for this dataset because of the notable presence of extreme values for many features.

Robust Scaling: The features were scaled using a robust scaler, which is less sensitive to outliers, thereby normalizing the data and reducing the influence of extreme values.

3.2 High-level Information

The ADS implements an XGBoost model, which is capable of handling a large variety of data types, distributions, and various dimensions of input data. Tomek Links was used to address the imbalance in the dataset (smokers vs. non-smokers) by undersampling the majority class by removing those data points that are nearest to the minority class points, which enhances the decision boundary between the two classes.

3.3 Validation of the ADS

The ADS is validated by evaluating the AUC score against a separate test dataset. The original submission's official AUC score is 0.8809, which is very considerable, suggesting that this ADS meets its goal of effectively identifying smokers and non-smokers.

4 Outcomes

Does this ADS discriminate against age? We bucket age into old and young group with the median age 45 as threshold and perform the following analysis.

4.1 Accuracy Analysis

Precision: measures given predicted positive, how likely is it that the individual is actually a smoker. A high precision would mean less smoking-cessation resources are wasted on non-smokers, and thus saving more resources for those in need.
Recall: measures given an individual is a smoker, how likely will the ADS identify them. This is important as the intervention, smoking cessation sessions, is assistive to smokers, and we want as many smokers as possible to receive this treatment.
Accuracy: Provides a general measure of the model's overall effectiveness across both classes (smokers and non-smokers).

The overall metrics and metrics by age group are shown in tables 3 and 4. The confusion matrices for each group is shown in figure 4. FNR, which equals to $1-Recall$ is also shown for

Metric	Value
accuracy	0.786252
precision	0.711663
recall	0.831810
FNR	0.168190
false_negative_rate_difference	0.082884
equalized_odds_difference	0.191494
selection_rate_difference	0.240562

Table 3: Overall Model Performance Metrics

Metric	Old	Young
accuracy	0.819677	0.756079
precision	0.710379	0.712363
recall	0.780080	0.862965
FNR	0.219920	0.137035

Table 4: Performance by Age Group

convenience. The performance metrics indicate that the model classifies older individuals as smokers with higher overall accuracy compared to younger individuals. However, it is more effective at correctly identifying actual smokers in the younger group, as evidenced by a lower FNR and higher recall. Precision is nearly equivalent between the two groups, suggesting that the model's ability to correctly predict smokers as such does not vary significantly with age.

4.2 Fairness Analysis

False Negative Rate Difference (FNR Difference): Measures the disparity in FNR between subgroups (e.g., young vs. old). A higher FNR for a specific group means that the system is more likely to miss identifying actual smokers in that group.
Equalized Odds Difference: Measures discrepancies in both the FNR and FPR across groups. It ensures that the classifier's performance in predicting both positive and negative classes is consistent across groups, minimizing bias in both failing to identify smokers and incorrectly identifying non-smokers.
Selection Rate Difference: Measures differences in the rates at which different groups are predicted as smokers. It helps ensure that no group is disproportionately targeted or overlooked by the model.

The fairness metrics are shown in table 3. The FNR Difference of 0.0829 suggests that the model is more likely to misclassify older smokers. This bias could result in older smokers not being offered the opportunity for smoking cessation interventions. Moreover, the Equalized Odds Difference of 0.1915 indicates that the model's accuracy in identifying smokers and non-smokers is inconsistent across age groups. This implies that the model's performance varies depending on the age group, leading to unequal treatment and potentially affecting the effectiveness of targeted health initiatives. Additionally, the Selection Rate Difference of 0.2406 indicates that the old group are much more likely to be predicted positive than the young group, even though the base rate in the training data is the opposite (greater proportion of young smoker than olde). This discrepancy suggests a potential bias in the model's predictive behavior, which could lead to inappropriate or inefficient allocation of resources, such as preventive interventions or cessation support, potentially overlooking younger individuals who might benefit more from such efforts.

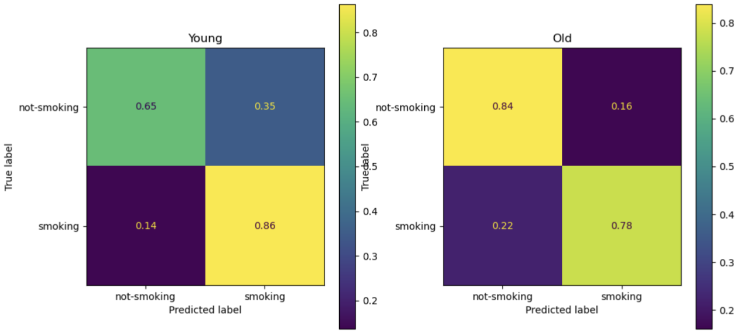

Figure 4: Confusion Matrices for Young and Old Group

4.3 Additional Performance Analysis

Model Feature Importance (Global Explainability)
Evaluating feature importance in the XGBoost model is crucial for understanding which variables most significantly impact the model's predictions, which enhances interpretability and transparency for stakeholders.

Height is by far the most important feature. GTP (γ-GTP), Hemoglobin, and Serum Creatinine follow as the next important bucket of features. Age and HDL (High-Density Lipoprotein) shows moderate influence, reinforcing the idea that smoking behaviors and their health consequences can vary across different age groups, and that smoking can alter lipid metabolism. The analysis also identifies various levels of importance for factors such as Dental Caries, LDL (Low-Density Lipoprotein), ALT (Alanine Aminotransferase), and other biochemical markers. Although these features have a lower influence, they still contribute to the model's predictions and may reflect broader health impacts related to smoking. Interestingly, some features that are commonly associated with health, such as Cholesterol, Systolic Blood Pressure, and Eyesight measurements, have lower importance scores. While these factors are relevant health indicators, they may not be as directly predictive of smoking status compared to the other high-ranking features.

Because Height is the most important feature in this model, even though the model is fairly accurate, it is not trustworthy as height and smoking shouldn't have any kind of significant relationship.

SHAP Waterfall Plots (Local Explainability)
Figure 6 shows the waterfall plots for TP, TN, FP, FN, respectively. These plots provide detailed insights into how each feature contributes to individual predictions. This transparency allows healthcare providers to validate the model's decisions based on individual patient data.

In the TP plot, Height and Hemoglobin have strong positive effects, significantly increasing the likelihood of correctly predicting an individual as a smoker. ALT (Alanine Aminotransferase) has a slight negative impact, possibly indicating health variations among smokers. The overall positive outcome suggests that when key health indicators align, the model confidently predicts smoking status.

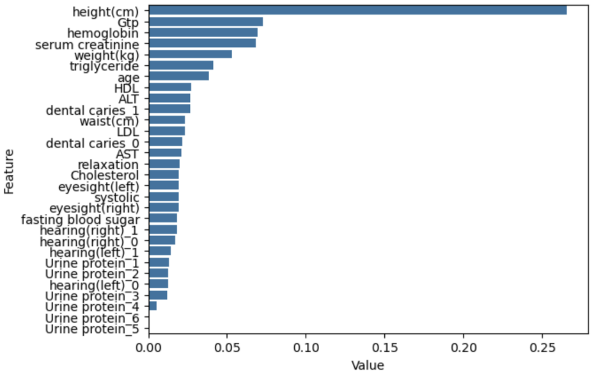

Figure 5: Feature Importance

For TN, Height has a minimal positive influence, indicating it's not a decisive factor in non-smoking predictions. Serum Creatinine and AST (Aspartate Aminotransferase) show strong negative impacts, lowering the likelihood of a smoking prediction and aligning with healthier profiles. Triglyceride levels have the strongest negative impact, driving the prediction towards non-smoking, likely reflecting healthier lifestyle choices.

In the FP plot, Height again shows a significant positive impact, potentially misleading the model to suggest smoking habits. GTP has a negative contribution, reducing the smoking likelihood and indicating possible mismatches in liver function interpretations. LDL (Low-Density Lipoprotein) and Cholesterol also have negative impacts, possibly due to variations in metabolic health not directly related to smoking.

For FN, Height has a negative impact, suggesting that certain physical characteristics might lead the model to underestimate the likelihood of smoking for some individuals. GTP has a large positive impact, indicating that despite strong indications from liver function tests, the model fails to predict the individual as a smoker, possibly due to conflicting signals from other features.

In conclusion, height consistently plays a significant role across all prediction scenarios, but its impact varies, depending on other health indicators. Biochemical markers such as Gtp, Hemoglobin, and Triglycerides also consistently influence predictions, but their effects vary as well, based on the individual's overall health profile. The differences in feature influence across various prediction scenarios suggest that while the model captures complex health patterns, there is still room for improvement, particularly in reducing false predictions by better integrating biochemical and physiological features.

5 Summary

5.1 Data Appropriateness

The data includes a comprehensive range of physiological, biochemical, and demographic features that are critical for accurately assessing smoking behavior and its effects. The inclusion of variables such as height, liver function markers (e.g., GTP), hemoglobin, lipid

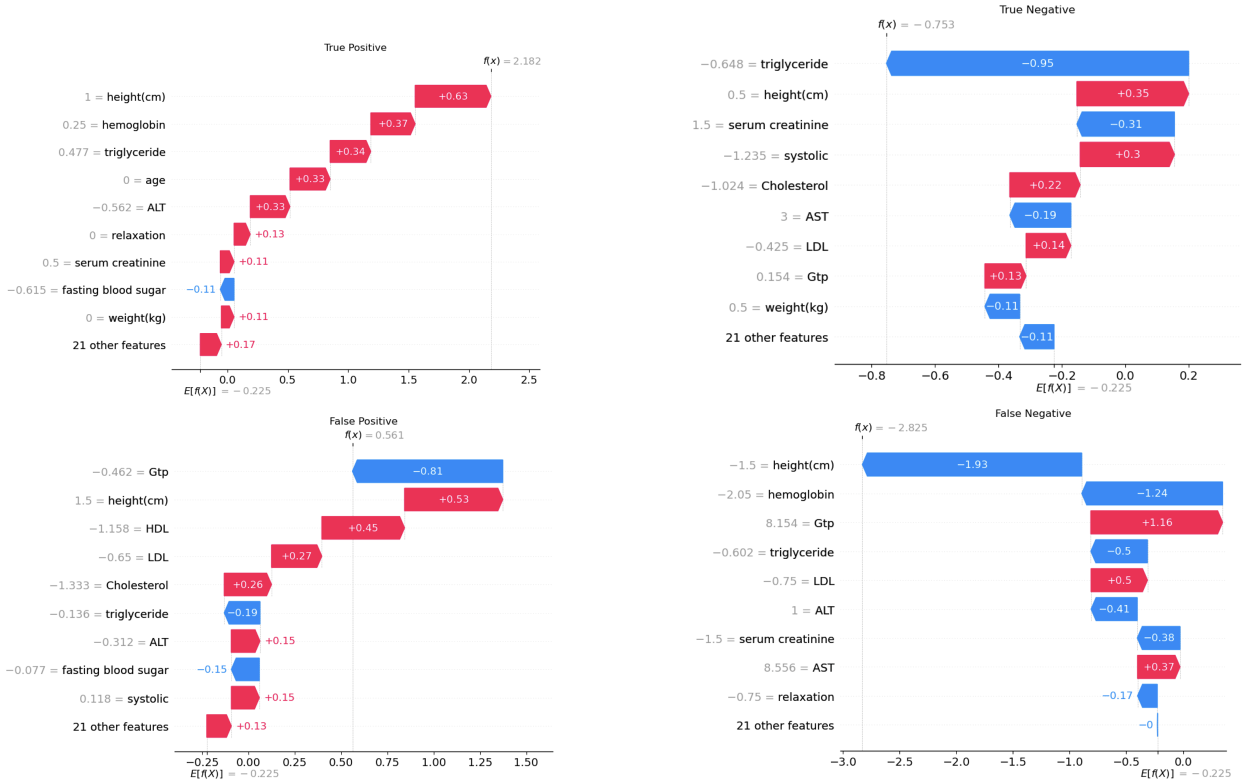

Figure 6: SHAP Waterfall Plots for TP, TN, FP, FN

profiles, and age provides valuable information about an individual's health status and potential smoking habits. The use of such extensive data is appropriate, as it is consistent with clinical evidence that connects these factors to smoking. Nevertheless, the varying influence of certain features across prediction scenarios could be further examined, improving the model's precision and reliability by identifying potential biases or confounding factors affecting predictions.

5.2 Robustness, Accuracy, and Fairness

Robustness: The ADS's implementation demonstrates robustness in handling diverse health-related features to predict smoking status. The use of machine learning techniques and the incorporation of a wide range of health indicators suggest a strong ability to capture complex patterns associated with smoking behaviors. However, the SHAP analysis reveals variations in feature influence across different scenarios (TP, TN, FP, FN), indicating further refinement to more effectively handle anomalies or conflicting signals.

Accuracy: The ADS exhibits a good level of accuracy, as shown by the performance metrics. The use of accuracy, precision, and recall provides a comprehensive view of the model's predictive performance. High recall rates are particularly crucial in this case as failing to identify a smoker could result in missed opportunities for timely interventions. The presence of false positives and false negatives underscores the need for further improvement.

Fairness: The model's fairness was evaluated using FNR Difference, Equalized Odds Difference, and Selection Rate Difference. These metrics are critical for ensuring that the ADS does not systematically disadvantage any subgroup (age group in our case). The results revealed some disparities in predictions between different age groups, which requires further adjustments to enhance fairness.

Healthcare Providers would prioritize accuracy and recall measures, as they directly impact clinical decisions and patient care. High recall ensures that fewer smokers are missed, which is vital for effectively targeting interventions.
Patients and advocacy groups would be particularly interested in fairness measures, ensuring that the ADS provides equitable health assessments across all demographics.
Healthcare Policymakers would likely focus on the overall effectiveness and fairness of the ADS to inform public health strategies and resource allocation.

5.3 Comfort in Deployment

Deploying this ADS is not recommended at this stage without further enhancements. As discussed earlier, this ADS is not trustworthy as it prioritizes a feature that is unrelated to smoking status. The other concerns is fairness: large disparities in performance across different age groups and high false negative rate. This issue is critical because they can result in unequal opportunity for intervention, which are especially problematic in public health contexts where trust and equity are of utmost importance.

5.4 Recommended Improvements

In terms of data collection, there should be a data card that informs the users about the original source of the data and a clear explanation of all the features. However, this is not in the author's control. As we pointed out in section 3.1, some of the feature transformations do not make sense. Perhaps the actual reasoning is valid, but the author should explicitly state the reasoning. Although the model is decently accurate, it may be further improved with more appropriate feature transformation, selection and engineering. For example, from the correlation heatmap, we see clusters of highly correlated features and features that are completely unrelated to the target variable smoking, perhaps the author could perform dimension reduction and feature selection to reduce the number of features and thus reduce overfitting. Even though XGBoost is fairly robust to this issue, these steps may be worth taking to further enhance model performance and more importantly, derive deeper insights into the dataset.

This ADS suffers in fairness, and thus we suggest that it needs to be calibrated. The large disparity in the base rate of the original data and the selection rate suggests a very strong technical bias. We recommend the author to consider performing some bias reducing techniques such as Correlation Remover, as we see in Figure 2 that age has notable correlations with many features, which could result in age affecting model's decisions.

6 Contribution

Equal Contribution on code and report.