Course Project: Technical Audit of an Automated Decision System

Julie Cestaro Giancarlo Enriquez

Due: 9 May 2025

Abstract

This report audits an Automated Decision System (ADS) designed to predict smoking status from clinical and biometric data using synthetic health records. We evaluate the system's data integrity, modeling pipeline, and performance across multiple classifiers, with XGBoost showing the strongest results. Fairness analysis using age as a protected attribute reveals higher error rates for underrepresented younger individuals, raising concerns about data representativeness. Despite strong technical performance, we recommend improved age group coverage and fairness monitoring before real-world deployment.

1 Background

1.1 Real-World Implications

Smoking remains the leading cause of preventable illness and death worldwide, contributing to a wide range of chronic diseases and affecting nearly every organ in the body. According to the World Health Organization, smoking-related deaths are projected to reach 10 million annually by 2030. While treatments and counseling exist, fewer than one-third of individuals who attempt to quit smoking are successful. These challenges underscore the need for scalable, data-driven tools that can support early identification and intervention efforts. Automated decision systems (ADS) offer a promising approach by leveraging clinical and biometric data to identify individuals who may benefit most from targeted cessation support, potentially improving outcomes and optimizing healthcare resources.

1.2 Purpose and Goals

This Automated Decision System (ADS) is designed to predict whether an individual is currently a smoker, using a variety of biological and health-related indicators. Although not a direct predictor of smoking cessation success, the system supports public health and clinical goals by identifying individuals with smoking behavior, which can help target cessation resources more effectively. By accurately identifying smokers using non-invasive health metrics, the ADS could be incorporated into routine screenings to prompt timely intervention.

The system is trained on synthetic data from Kaggle's 2023 Tabular Playground Series[RC23] , which mimics realistic screening data while preserving privacy. While this enables safe experimentation and model refinement, it may limit generalizability due to the absence of social, behavioral, or environmental context—factors often correlated with smoking behavior in real-world populations.

Prior research by Pant et al. [LZW^{+}20] illustrates how machine learning can be applied to smoking behavior detection, using clinical and demographic data. Their use of residual neural networks and SHAP explainability aligns with this project's focus on combining predictive accuracy with interpretability in health settings. As with any health-related ADS, considerations of fairness, transparency, and clinical impact are essential for responsible deployment.

1.3 Trade-offs

Developing this ADS requires careful navigation of trade-offs between accuracy, interpretability, and ethical responsibility. While complex models like neural networks can deliver strong predictive performance, their opacity can undermine clinical trust and hinder auditability. In healthcare contexts, balancing sensitivity and specificity is also critical, as misclassifications may lead to adverse outcomes for patients or inefficient resource use. Furthermore, although the dataset preserves privacy through synthetic generation, it may fail to represent the full variability of real-world clinical populations, raising concerns about generalizability and potential biases. These tensions underscore the need to evaluate not just technical metrics, but also fairness, transparency, and the system's broader ethical implications.

2 Input and Output

2.1 Input Data Description

The dataset used by this Automated Decision System (ADS) is the *Smoker Status Prediction using Bio-Signals* dataset [Dut25], published as part of Kaggles's 2023 Tabular Playground Series. The data is synthetically generated to resemble realistic clinical screening data while preserving privacy. It includes 22 numerical input features and a binary target variable, `smoking`, which indicates whether the individual is a current smoker.

The full dataset contains 318,512 observations and 23 columns. All features are of type `int64` or `float64`, with no categorical or textual fields. The dataset is well-prepared for supervised learning tasks.

Each feature represents a biometric or clinical metric commonly collected during medical checkups. Features are grouped as follows:

- **Biometric**: `age`, `height(cm)`, `weight(kg)`, `wasit(cm)`
- **Sensory**: `eyesight(left)`, `eyesight(right)`, `hearing(left)`, `hearing(right)`
- **Cardiovascular and Metabolic**: `systolic`, `relaxation`, `fasting blood sugar`, `Cholesterol`, `triglyceride`, `HDL`, `LDL`, `hemoglobin`
- **Renal and Liver Function**: `serum creatinine`, `AST`, `ALT`, `Gtp`
- **Other Binary Indicators**: `urine protein`, `dental caries`

2.1.1 Missing Values and Dataset Completeness

A detailed audit confirmed 0% missingness across all 22 features and the target variable (Table 2). This eliminates the need for imputation or filtering, enabling direct modeling.

2.1.2 Descriptive Statistics and Skewness

Descriptive statistics for selected input features are summarized below:

Table 1: Summary Statistics for Selected Input Features

Feature	**Min**	**25%**	**Median**	**Mean**	**75%**	**Max**
Age (years)	20	40	40	44.3	55	85
Height (cm)	135	160	165	165.3	170	190
Weight (kg)	30	60	65	67.1	75	130
Waist (cm)	51	77	83	83.0	89	127
Systolic BP	77	114	121	122.5	130	213
Triglyceride	8	77	115	127.6	165	766
GTP (γ-GTP)	2	18	27	36.2	44	999
Creatinine	0.1	0.8	0.9	0.89	1.0	9.9

Several features exhibit substantial right-skew, particularly `triglyceride`, `GTP`, `ALT`, and `creatinine`, where maximum values exceed the 75th percentile by large margins. These distributions suggest heavy-tailed behavior and potential outliers. While transformations were not applied at this stage, such features may benefit from normalization techniques (e.g. log or Box-Cox) in future modeling steps.

2.1.3 Feature Distributions by Smoking Status

Violins Plots were used to visualize the distribution of each input feature plot by smoking status. Observations include:

- Features like `serum creatinine` and `waist` show visible separation between smokers and non-smokers.
- Spike-shaped distributions in `age`, `height`, and `weight` suggest binning or rounding in data generation.
- Distributions are consistent across all samples, reinforcing the dataset's internal coherence.

2.1.4 Correlation Analysis

A full Pearson correlation heatmap (Figure 1) and a focused clinical subset heatmap (Figure 2) revealed meaningful linear relationships among features, including:

- `weight` and `waist (cm)`: $r = 0.83$
- `Cholesterol` and `LDL`: $r = 0.81$
- `Systolic` and `Relaxation:` $r = 0.75$
- `ALT` and `AST`: $r = 0.62$

These associations support the dataset's biological realism and suggest that multicollinearity is manageable for most features.

2.1.5 Class Distribution

The target label `smoking` is relatively balanced, with 43.7% smokers and 56.3% non-smokers. This balance supports the use of standard performance metrics such as ROC-AUC and reduces the risk of class imbalance bias during training.

2.2 Output Description

The output of the Automated Decision System (ADS) is a probability score between 0 and 1, representing the model's confidence that a given individual is a smoker. This probability is produced by various base models (e.g., XGBoost, LightGBM, CatBoost, ANN), each trained to distinguish between smokers and non-smokers. These model outputs are then combined using a weighted ensemble, where optimal weights are determined via Optuna to maximize ROC AUC on validation data.

The final prediction for each individual is the weighted average of the predicted probabilities across all models in the ensemble. This output can be interpreted as the likelihood that the individual is a smoker, with values closer to 1 indicating higher risk. In practice, this probability can be thresholded (e.g., at 0.5 or an optimized cutoff) to yield a binary class label (smoker vs. non-smoker) if required for downstream decision-making, such as clinical intervention or targeted support.

Feature	Data Type	Train Missing %	Test Missing %
age	int64	0.0	0.0
height(cm)	int64	0.0	0.0
weight(kg)	int64	0.0	0.0
waist(cm)	float64	0.0	0.0
eyesight(left)	float64	0.0	0.0
eyesight(right)	float64	0.0	0.0
hearing(left)	int64	0.0	0.0
hearing(right)	int64	0.0	0.0
systolic	int64	0.0	0.0
relaxation	int64	0.0	0.0
fasting blood sugar	int64	0.0	0.0
Cholesterol	int64	0.0	0.0
triglyceride	int64	0.0	0.0
HDL	int64	0.0	0.0
LDL	int64	0.0	0.0
hemoglobin	float64	0.0	0.0
Urine protein	int64	0.0	0.0
serum creatinine	float64	0.0	0.0
AST	int64	0.0	0.0
ALT	int64	0.0	0.0
Gtp	int64	0.0	0.0
dental caries	int64	0.0	0.0
smoking	int64	0.0	NA

Table 2: Feature Types and Missingness

3 Implementation and Validation

3.1 Data Cleaning and Preprocessing

The data preprocessing pipeline is extensive and thoughtfully designed to prepare clinical and biometric data for modeling. Initial steps include clipping outlier values for several lab measurements (e.g., `GTP`, `HDL`, `LDL`, `ALT`, `AST`, `serum creatinine`) to mitigate the influence of extreme values.

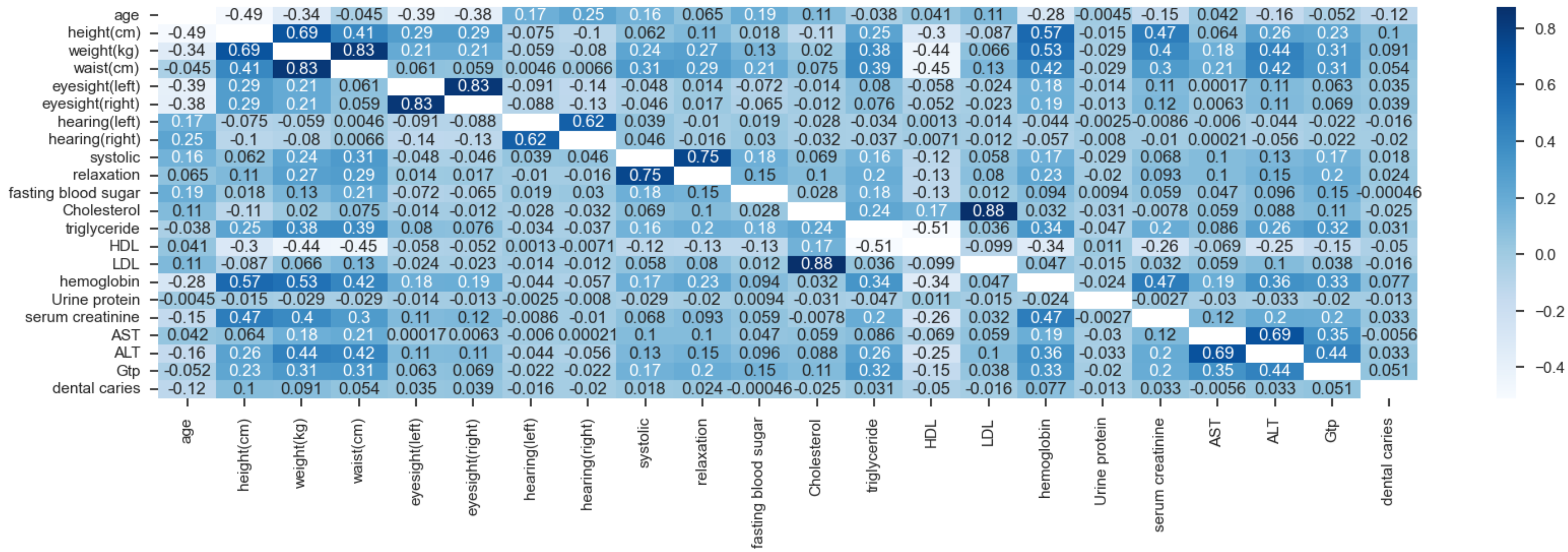

Figure 1: Pairwise Correlations for All Features

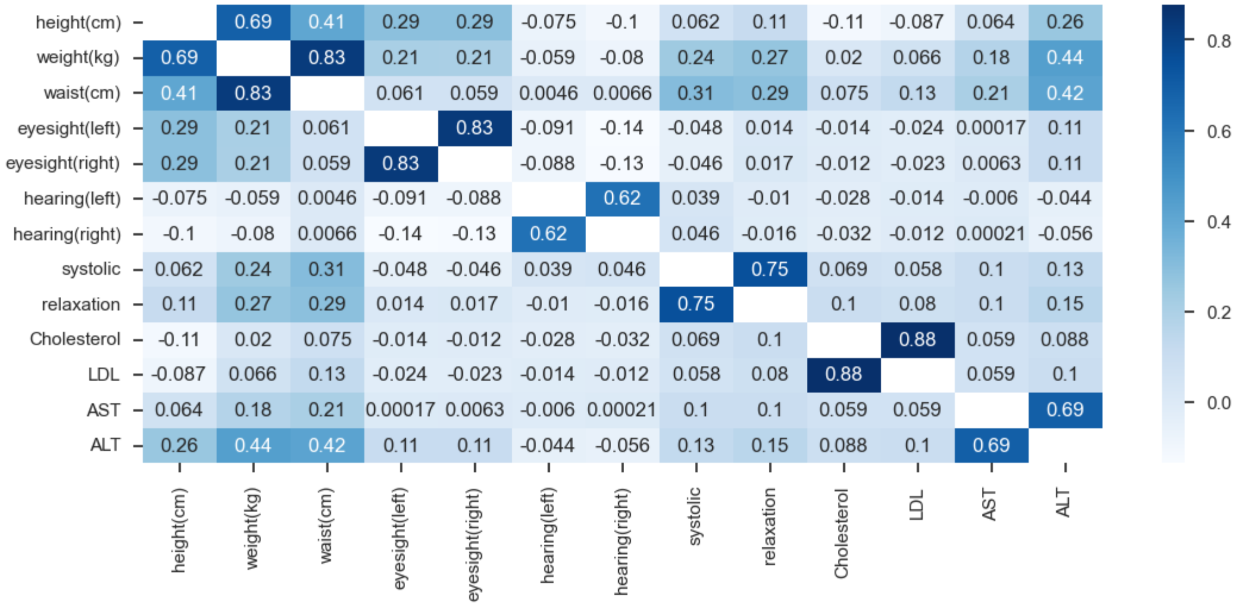

Figure 2: Pairwise Correlations for a Subset of Features

Hearing and eyesight features are reformatted to consistently represent “best” and “worst” values across both sides. Continuous variables are normalized using min-max scaling based on the combined range across training and test datasets, which ensures stability during inference.

Missing values in numerical features are addressed using a LightGBM-based iterative imputation method. This approach trains regressors for each feature with missing values, updating estimates across multiple iterations and monitoring RMSE to confirm convergence. For categorical variables, multiple encoding strategies are employed, including one-hot encoding, count encoding, target-guided mean encoding, and frequency-based labeling. Encoding methods are evaluated based on their contribution to univariate ROC AUC, and the most informative versions are retained.

The pipeline also generates new features through arithmetic combinations (e.g., `GTP/LDL`, `ALT*GTP`) and clusters less important features using k-means. Highly correlated features are consolidated using PCA, and final feature selection eliminates redundancy while preserving predictive signal. A post-processing step further removes any duplicate columns after scaling. The resulting feature matrix is both rich in signal and optimized for downstream modeling.

3.2 High-Level Implementation of the System

The ADS is implemented as a multi-model ensemble architecture, involving a diverse collection of classifiers, including tree-based models (e.g., XGBoost, CatBoost, LightGBM, Decision Trees), a logistic regression baseline, and a shallow artificial neural network (ANN) with dropout layers and leaky ReLU activation.

For simplicity, we focused our audit on the independent implementation of three models as binary classifiers: the Logistic Regression model, Decision Tree model, and XGBoost model. Each model outputs a probability that each individual in the dataset is a smoker. We thresholded these probabilities at 0.5 to create a set of binary predictions for simpler analysis.

3.3 Validation and Goal Alignment

Final models are validated using ROC AUC, a suitable metric for imbalanced binary classification, and for this particular prediction task. Because correctly identifying smokers will directly impact the future medical care and future medical costs of each individual, it is important to ensure a high accuracy in these predictions. This ensures that model behavior is aligned with stakeholder expectations and real-world decision-making needs.

4 Outcomes

4.1 Defining the Protected Attribute

For this analysis, we define the protected attribute to be age to ensure that undue burden is not placed on an individual based on an immutable characteristic of that individual. The individuals in the dataset are divided into three categories: 'young', 'middle aged', and 'elderly'. We define these categories relative to the age characteristics in the dataset. The threshold between 'young' and 'middle aged' is the mean age minus one standard deviation. Similarly, the threshold between 'middle aged' and 'elderly' is the mean age plus one standard deviation. The approximate value count breakdown is as follows:

Group Name	Count	Percent of Total
elderly	150100	47.125%
middle aged	119544	37.532%
young	48868	15.342%

4.2 Performance Across Accuracy Metrics

We primarily looked at four accuracy metrics in our analysis: accuracy, false positive rate (FPR), false negative rate (FNR), and selection rate. We primarily focused on error rate metrics in our analysis as both false positives and false negatives would burden the individual subject to the error.

We analyzed the four accuracy metrics for each of the three models trained on the feature set: Logistic Regression, Decision Tree, and XGBoost. The overall metrics for each model are shown in Figure 3. We generated the same metrics for each age group using the Fairlearn Toolkit[BDE^{+}20], which are visualized in Figures 5, 4, and 6.

While the overall accuracy for each of these predictors is not very high, we see in Figure 3 that XGBoost performs the best in terms of accuracy metrics. Similarly, Figure 6 shows that the XGBoost model also performs the best at a group level, with consistently lower error rates for all three groups compared to the other two models.

We see in all cases that the accuracy is lower and the error rate is higher for the young age group. We hypothesize that this may be due to the fact that fewer young people are represented in the dataset. We further hypothesize that a pre-existing bias may be causing this decreased representation as fewer young people tend to go to the doctor and therefore there is likely less medical data about this age group available.

4.3 Performance Across Fairness Metrics

We evaluate the fairness of these models with a continued focus on each model's ability to generate accurate predictions. We assess equalized odds difference and equalized odds ratio, again using the

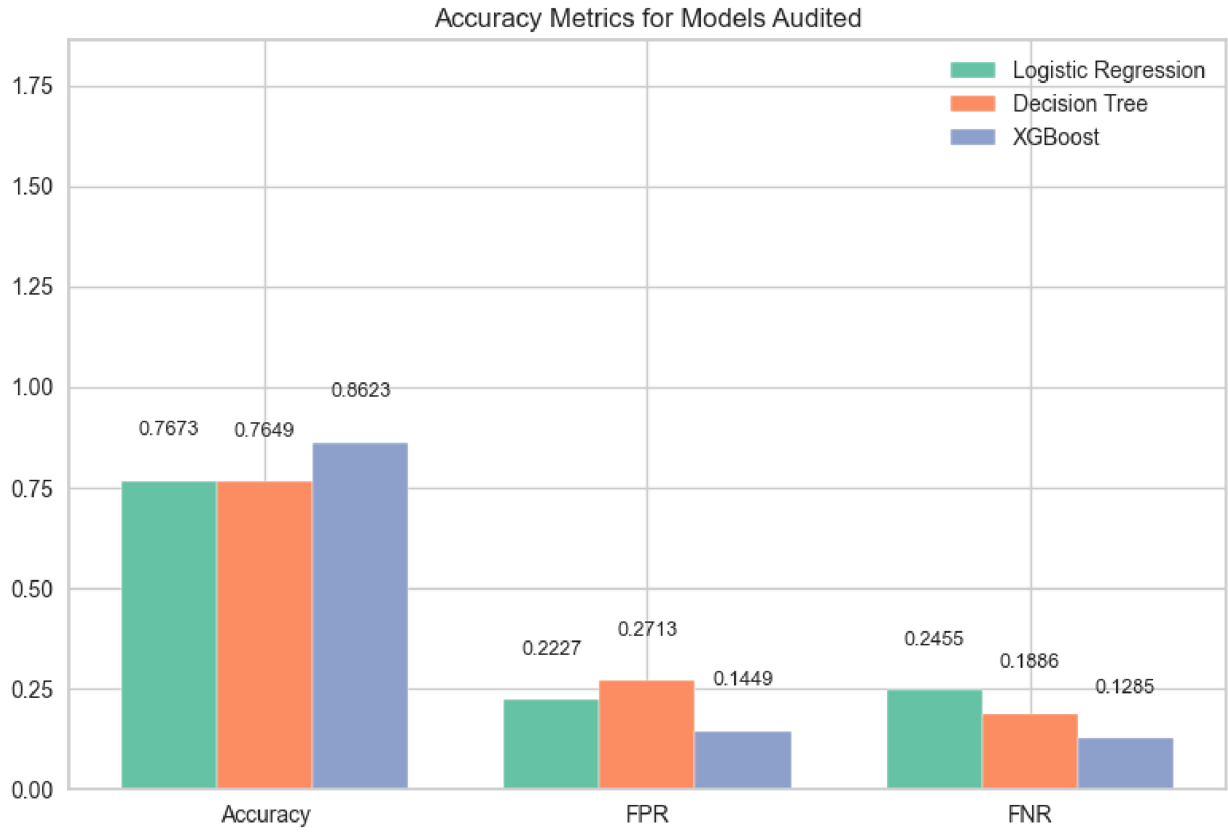

Figure 3: Accuracy Metrics for Each Audited Model

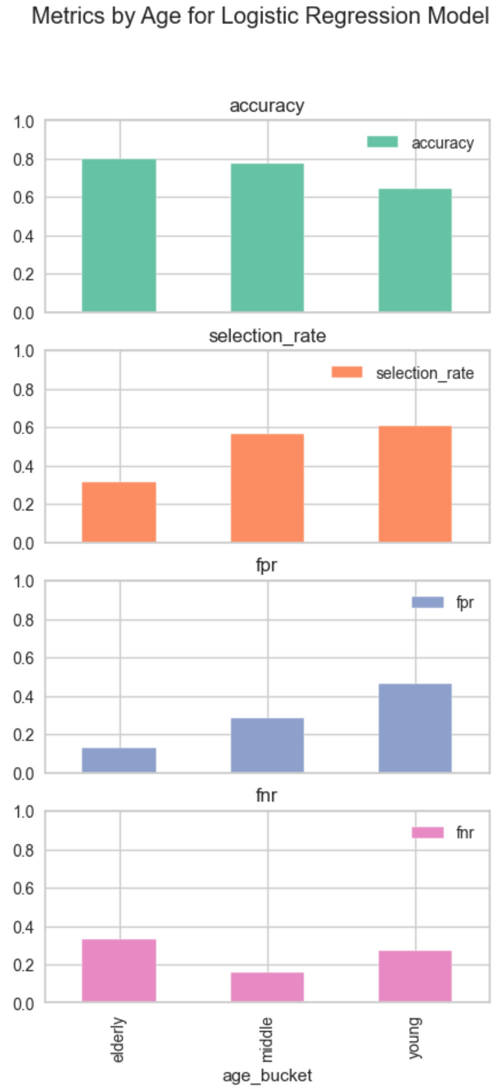

Figure 4: Group Accuracy Metrics with Logistic Regression Model

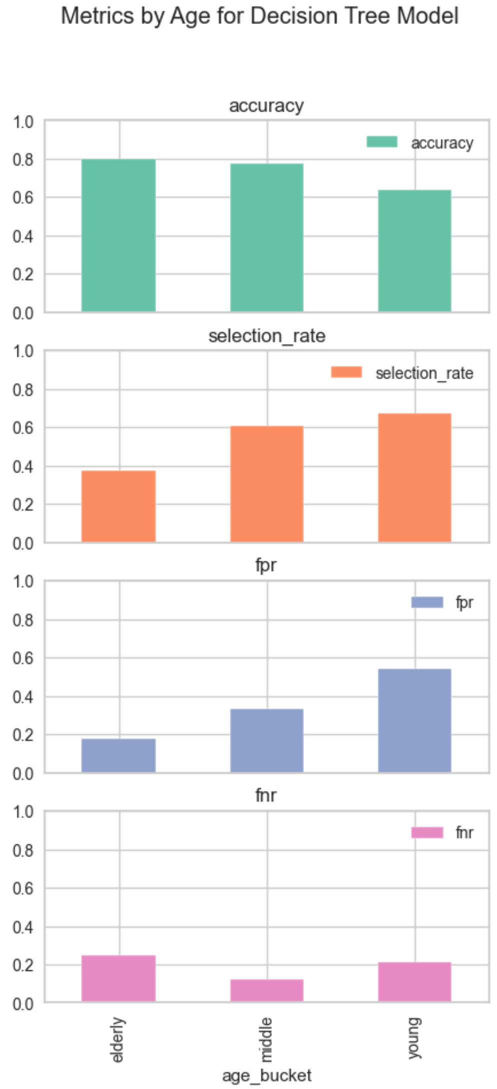

Figure 5: Group Accuracy Metrics with Decision Tree Model

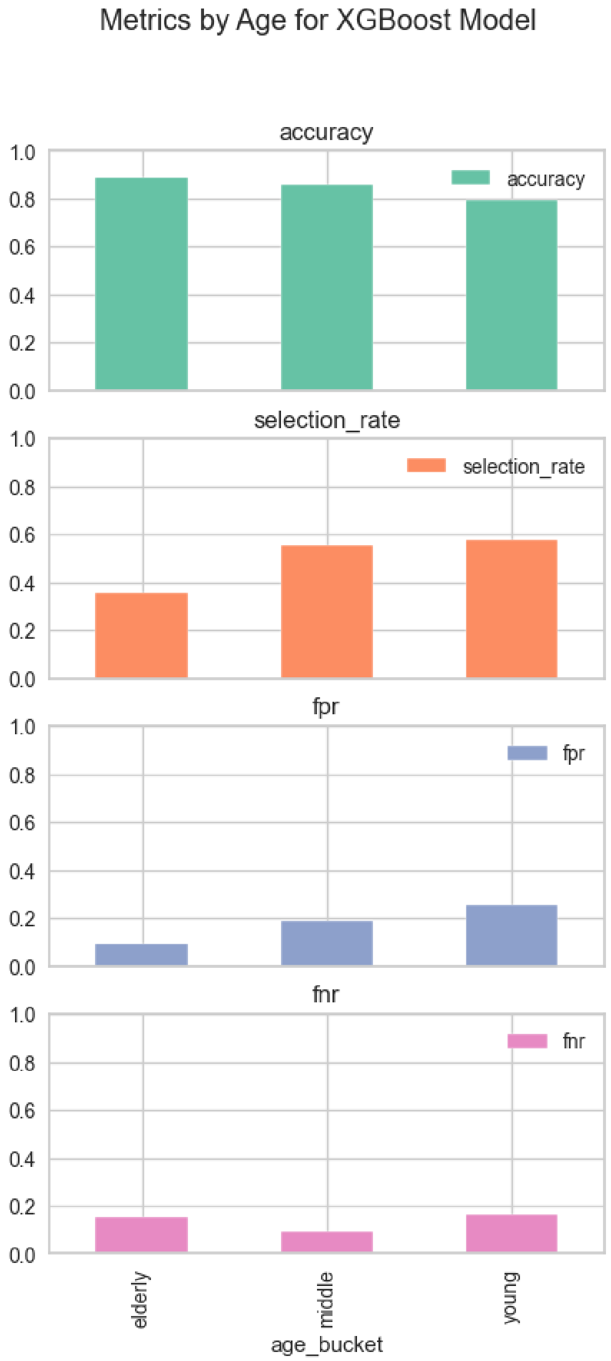

Figure 6: Group Accuracy Metrics with XGBoost Model

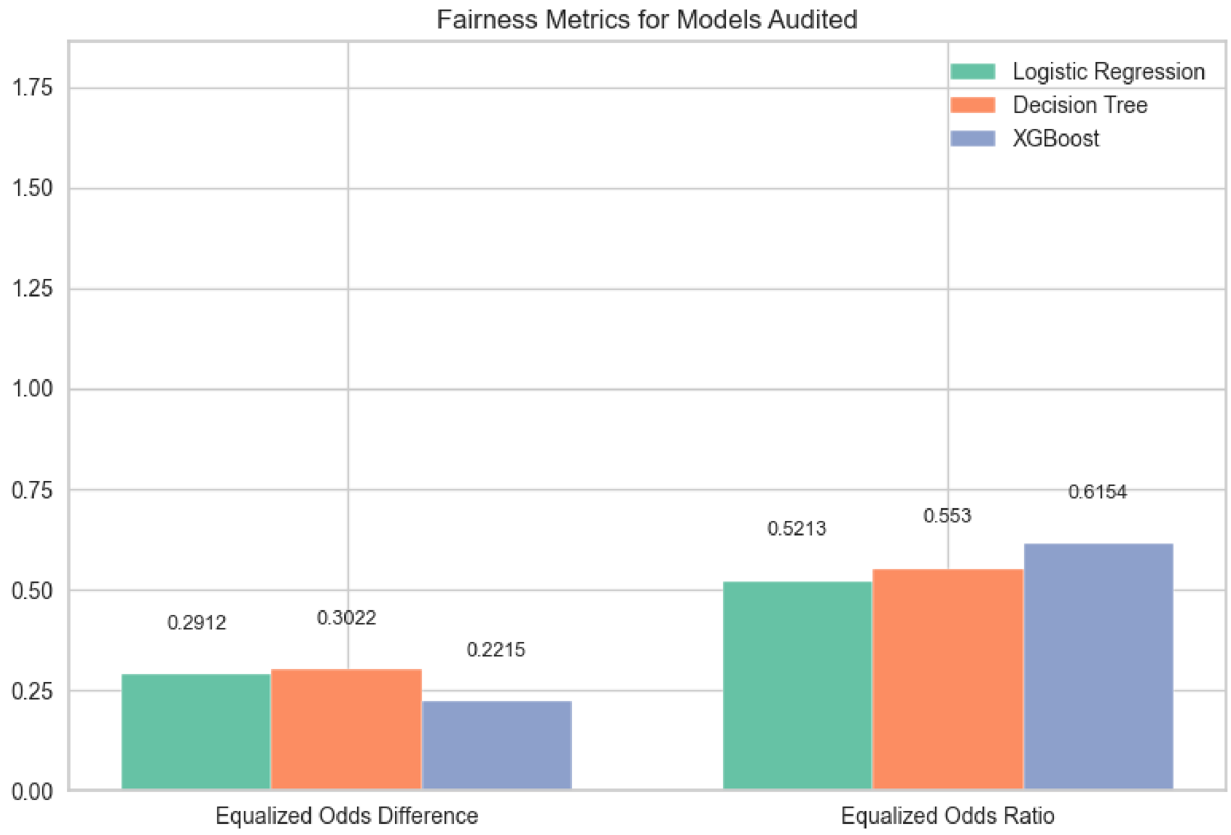

Figure 7: Fairness Evaluation of Three Models Over the Sensitive Attribute Age

Fairlearn Toolkit [BDE+20], because these fairness metrics allow us to compare errors made by the model across groups. We consider these as opposed to demographic parity ratio and demographic parity difference as those focus on selection rate.

Note that because there are three groups for the sensitive attribute, age, these Fairlearn metrics are calculated between the largest and smallest error rates of all of the three age groups. In other words, these metrics represent the worst case scenario.

As with the accuracy metrics, we see in Figure 7 that the XGBoost model performs better than the other two on both fairness metrics.

4.4 Stability and Robustness

To assess model stability, we performed ten different train/test splits each with a different random seed. We then predict with each model on the new train/test set and record accuracy metrics. Figures 8, 9, and 10 show the results of these ten trials for each of the three models we tested. In general we see that XGBoost continues to perform the best with the Decision Tree showing the most instability, especially in its error rates.

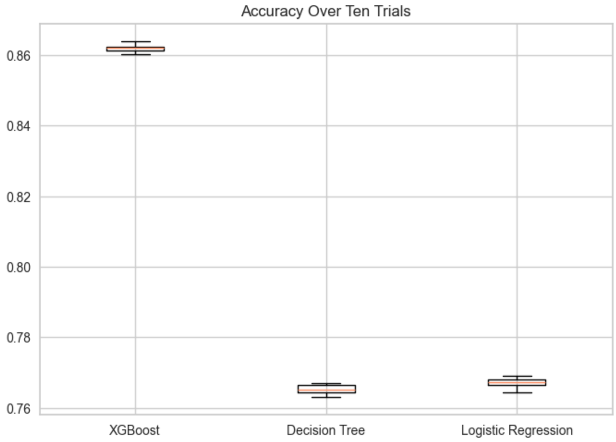

Figure 8: Stability of Accuracy Over Ten Trials

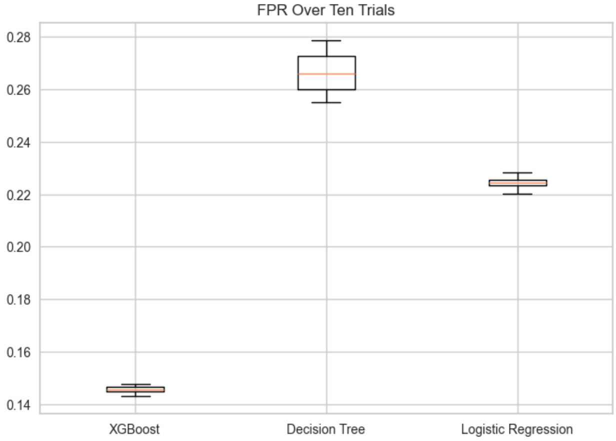

Figure 9: Stability of FPR Over Ten Trials

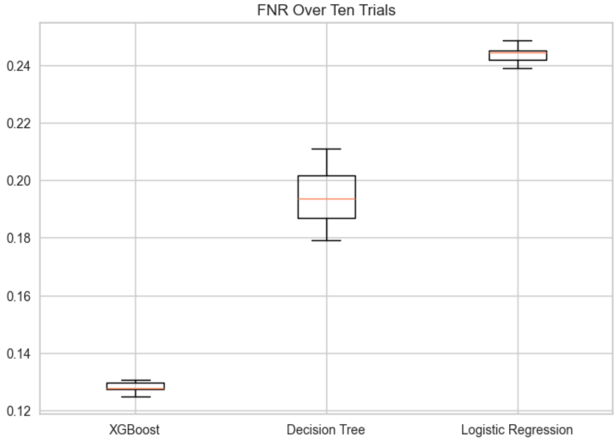

Figure 10: Stability of FNR Over Ten Trials

5 Reflection

5.1 Summary

It is always an inherently risky decision to implement an ADS in a medical context, but we conclude based on our analyses that the implementation of this system is overall appropriate for the decision task.

In particular, we show that the XGBoost model performs well in terms of accuracy (in Figure 3), has consistently low error rates across each group (in Figure 6), and performs the best in the fairness-specific evaluation (in Figure 7).

5.2 Stakeholders

We focused the majority of our analysis on error rates because an error in either direction (false positive or false negative) would burden the individual subject to that error.

A false positive would be a case where an individual is not a smoker but is identified as a smoker. This may result in additional unnecessary medical oversight which would cost the individual in time, money, and other resources. A false negative would be a case where an individual is actually a smoker but is not identified as such. This could result in the individual not being tested or monitored for diseases or ailments to which smokers are subject and could ultimately result in the individual's death.

Both false positives and false negatives would additionally burden the medial professionals treating the individuals subject to the error. The medical professionals would either be wasting resources on a patient who does not need them (in the case of a false positive) and thereby taking them away from patients who require more care. In the case of a false negative, they might not account for serious conditions to which smokers are susceptible, and therefore might be caught off guard or understaffed when an individual unexpectedly comes down with a smoking-related disease.

5.3 Future Improvements

In Section 4.1 we mentioned that young people are underrepresented in this dataset and throughout our analysis we noted that young people consistently see the highest error rates. Our recommendation before considering deploying this ADS in the public sector is that data collection for this dataset be better distributed across all three age groups.

References

[BDE+20] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical report MSR-TR-2020-32, Microsoft, May 2020 (cited on pages 7, 11).

[Dut25] Gaurav Dutt. Smoker status prediction using biosignals, 2025. Accessed: 2025-04-09 (cited on page 2).

[LZW+20] Bo Li, Li Zeng, Wei Wang, Yujie Wang, Yawei Zhang, Hui Jiang, et al. Predictive modeling of smoker status based on biosignals. *Frontiers in Public Health*, 8:432, 2020 (cited on page 1).

[RC23] Walter Reade and Ashley Chow. Binary prediction of smoker status using bio-signals. https://kaggle.com/competitions/playground-series-s3e24, 2023. Kaggle (cited on page 1).