جامعة بيرزيت

BIRZEIT UNIVERSITY

Electrical and Computer Engineering Department
ENCS5341 - Machine Learning and Data Science

Smoking Status Detection: A Binary Classification Approach to Body Signal Analysis

Prepared by:
Hala Ziq 1191637
Zakiya AbuMurra 1191636

Instructor:
Yazan AbuFarha

Date:
January 26, 2024

Contents

List of Figures

List of Tables

1 Introduction

This project aims to identify whether individuals are smokers or non-smokers using their health data. We've used various machine learning methods for this task, including K-nearest neighbors (KNN) as a Basline model, eXtreme Gradient Boosting (XG-BOOST),and Random Forest(RF). Our focus was on binary classification, which means we only had two groups to identify: smokers or non-smokers.

To measure how well our models work, we've used a few key metrics. Accuracy tells us the percentage of correct predictions. We've also looked at precision and recall, which respectively tell us how many of our positive predictions are true and how many actual positives we can capture. The F1 score combines precision and recall to give a single measure of quality.

In the following sections, we'll go into the data, how we've prepared it, how each model performed, and what we've learned from the results.

2 Dataset

2.1 Data Collection

The dataset used in this project was sourced from Kaggle, a platform that hosts datasets provided by users and organizations. It consists of health-related features that are potentially indicative of an individual's smoking status.The dataset can be accessed through the link

2.2 Data Preprocessing

In the data preprocessing phase of our study, we ensured the dataset was free of duplicates and missing values to maintain data integrity. We selected important features with assistance from a medical expert and created new ones, such as the HDL/LDL ratio and Body Mass Index (BMI), derived from height and weight data. Additionally, we transformed categorical variables like 'gender', 'oral', and 'tartar' using one-hot encoding to make them suitable for machine learning analysis. The dataset was imbalanced as shown inFigure 2.1 in smoking plot , so we applied the SMOTE algorithm to achieve balance. Finally, we strategically split the dataset into training, validation, and testing segments in a 70:15:15 ratio, facilitating a comprehensive model training and evaluation process

As shown in Table 1 the important numbers for each health feature we looked at. On average, people in the study were about 44 years old, and there was a good mix of men and women. We can see how much people weigh on average and how big around their waist is, which tells us about their body size. The table also gives us an idea of everyone's general health, with numbers on blood sugar, cholesterol, and blood pressure. We've looked at specific ratios like the one between the good and bad cholesterol and other tests that tell us about liver health. Lastly, the table shows how many people in the study smoke. All these numbers help us understand the health of the group before we dive deeper into the data.

Feature	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**
Age	44.18	12.07	20.00	40.00	40.00	55.00	85.00
Gender (F)	0.36	0.48	0.00	0.00	0.00	1.00	1.00
BMI	24.17	3.48	14.27	21.60	23.88	26.12	42.45
Waist (cm)	82.05	9.27	51.00	76.00	82.00	88.00	129.00
Blood Pressure	1.68	0.51	1.00	1.00	2.00	2.00	3.00
fasting blood sugar	99.31	20.80	46.00	89.00	96.00	104.00	505.00
Cholesterol	196.90	36.30	55.00	172.00	195.00	220.00	445.00
triglyceride	126.67	71.64	8.00	74.00	108.00	160.00	999.00
LDL/HDL ratio	2.14	0.98	0.02	1.53	2.04	2.63	51.71
hemoglobin	14.62	1.56	4.90	13.60	14.80	15.80	21.10
serum creatinine	0.89	0.22	0.10	0.80	0.90	1.00	11.60
AST/ALT ratio	1.15	0.45	0.01	0.85	1.10	1.38	20.00
Gtp	39.95	50.29	1.00	17.00	25.00	43.00	999.00
smoking	0.37	0.48	0.00	0.00	0.00	1.00	1.00

Table 1: Descriptive statistics of the dataset features.

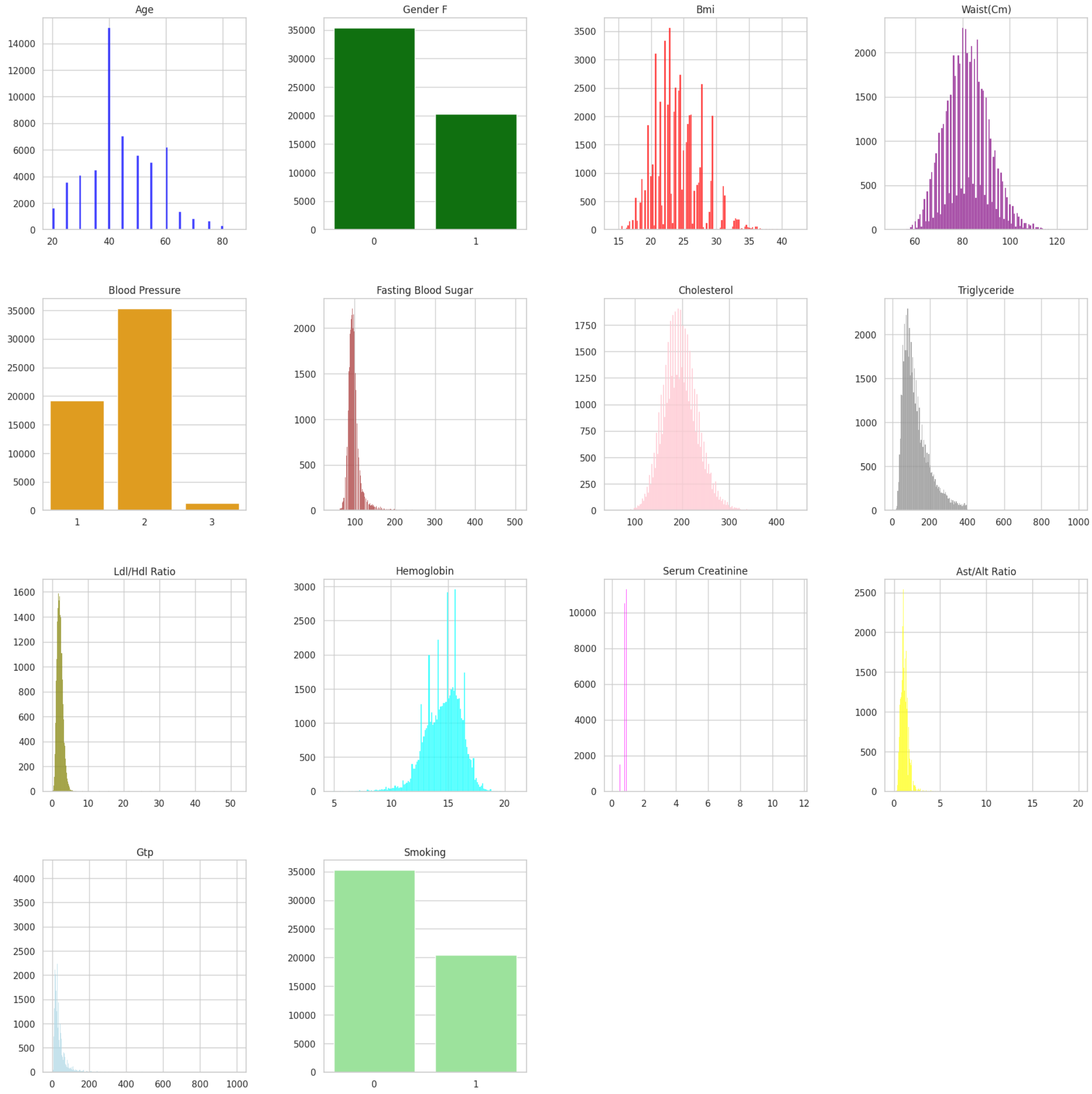

Figure 2.1: Comprehensive Health Metrics Overview.

3 Experiments and Results

3.1 K-nearest neighbors (KNN)

In this part, we used k-NN as the baseline models for k=1 and k=3, and it yielded the following results:

Table 2: Validation Set Evaluation of KNN Model (k=3)

Metric	**Class 0**	**Class 1**
Precision	0.75	0.59
Recall	0.76	0.57
F1-Score	0.76	0.58

Overall Accuracy: 69.43%

Table 3: Validation Set Evaluation of KNN Model (k=1)

Metric	**Class 0**	**Class 1**
Precision	0.80	0.65
Recall	0.79	0.65
F1-Score	0.80	0.65

Overall Accuracy: 74.16%

Table 4: Confusion Matrices for Validation Set for KNN

Confusion Matrix (k=3)		**Confusion Matrix (k=1)**	
Predicted Negative	**Predicted Positive**	**Predicted Negative**	**Predicted Positive**
4040 (TN)	1243 (FP)	4189 (TN)	1094 (FP)
1311 (FN)	1760 (TP)	1065 (FN)	2006 (TP)

Table 5: Test Set Evaluation of KNN Model(k=1)

Metric	**Class 0**	**Class 1**
Precision	0.78	0.63
Recall	0.78	0.63
F1-Score	0.78	0.63

Overall Accuracy: 72.68%

Table 6: Test Set Evaluation of KNN Model (k=3)

Metric	**Class 0**	**Class 1**
Precision	0.74	0.58
Recall	0.77	0.54
F1-Score	0.76	0.56

Overall Accuracy: 68.72%

The KNN model with $k = 1$ outperforms the $k = 3$ model in both the validation and test sets. In the validation set, the $k = 1$ model achieves an accuracy of 74.16%, while the $k = 3$ model achieves an accuracy of 69.43%. This suggests that the $k = 1$ model makes more accurate predictions on this dataset.

Looking specifically at Class 0 (non-smokers), the $k = 1$ model has a higher precision of 80%, indicating that it correctly identifies a larger proportion of non-smokers among its predictions, compared to the $k = 3$ model's precision of 75%. Additionally, the $k = 1$ model has a slightly higher recall of 79% for non-smokers, suggesting it misses fewer actual non-smokers compared to the $k = 3$ model's recall of 76%.

Table 7: Confusion Matrices for Test Set KNN

k=3		**k=1**	
Predicted Negative	**Predicted Positive**	**Predicted Negative**	**Predicted Positive**
4127 (TN)	1147 (FP)	4063 (TN)	1211 (FP)
1135 (FN)	1945 (TP)	1402 (FN)	1678 (TP)

For Class 1 (smokers), the $k = 1$ model maintains a precision of 65%, while the $k = 3$ model has a lower precision of 59%. This means that the $k = 1$ model is better at correctly identifying smokers among its predictions. In terms of recall, the $k = 1$ model has a recall of 65%, while the $k = 3$ model has a lower recall of 57%, indicating that the $k = 1$ model captures more actual smokers.there is a notable bias towards predicting negative instances more accurately than positive instances. This could be due to class imbalance.

The $k = 1$ KNN model performs better than the $k = 3$ model on this dataset, achieving higher accuracy and better precision and recall for both non-smokers and smokers. This suggests that a lower value of k results in improved performance for this classification task.

3.2 eXtreme Gradient Boosting (XGBOOST)

The selected **hyperparameters** for the XGBoost model, including a learning rate of 0.1, a maximum depth of 15, 250 estimators, and a scale pos weight of 2, have been fine-tuned to optimize its performance. These settings strike a balance between precision and recall, resulting in a highly accurate and effective classification model for the given dataset.

Table 8: Classification Reports for Validation and Test Sets XGBOOST model

Metric	**Validation Set**		**Test Set**	
	Class 0	**Class 1**	**Class 0**	**Class 1**
Precision	0.88	0.72	0.87	0.72
Recall	0.82	0.81	0.82	0.79
F1-Score	0.85	0.76	0.84	0.75
Accuracy	81.26%		80.89 %	

The test results are consistent with the validation results, indicating that the model generalizes well to unseen data. It maintains a good trade-off between precision and recall for both classes.

the XGBoost model with the specified hyperparameters achieved high accuracy and performed well in classifying both Class 0 and Class 1.and the results suggest that it can effectively identify instances of both classes while minimizing false predictions.

Table 9: Confusion Matrices for Validation and Test Sets XGBOOST model

Matrix	**Validation Set**		**Test Set**	
	Class 0	**Class 1**	**Class 0**	**Class 1**
Predicted Negative	4313 (TN)	595 (FN)	4314 (TN)	636 (FN)
Predicted Positive	970 (FP)	2476 (TP)	960 (FP)	2444 (TP)

3.3 Random Forest (RF)

Table 10: Confusion Matrix for Validation and Test Sets for the RF

Set	**Validation Set**		**Test Set**	
	Class 0	Class 1	Class 0	Class 1
Predicted Negative	4222 (TN)	482 (FN)	4203 (TN)	510 (FN)
Predicted Positive	1061 (FP)	2589 (TP)	1071 (FP)	2570 (TP)

Table 11: Classification Report for Validation and Test Sets

Metric	**Validation Set**			**Test Set**		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Class 0	0.90	0.80	0.85	0.89	0.80	0.84
Class 1	0.71	0.84	0.77	0.71	0.83	0.76
Accuracy	81.53%			81.07%		

In a **comparison of RandomForest and XGBoost** for a classification problem, RandomForest slightly edges out in accuracy and recall for the minority class, and has higher precision for the negative class. Both models exhibit similar precision for Class 1 and F1-scores, indicating comparable balance between precision and recall. The consistent performance from validation to test sets in both models demonstrates good generalization capabilities. Overall, **RandomForest may be the preferable choice**, especially for its effectiveness in classifying the minority class and the negative class.

4 Analysis

The performance of the Random Forest model can be determined based on the Table 10andTable 11.The accuracy for both the validation and test sets it's indicating a good overall performance, where the model correctly predicts 81% of the outcomes, and the similarity in the both sets indicates there is no over-fitting.

According to the confusion matrix in theTable 10 shows TN value it's high where this indicating that the model seems better at predicting the negative class(non-smoker 0) correctly but has a relatively high FP, indicating it tend to mistakenly predict the positive class when it's not, it's also shows that The model has higher precision for the negative class (0) but better recall for the positive class (1), where the F1-scores are good for both classes but slightly favor the negative class.

For over all the model performs consistently across the validation and test sets, which is a good in generalization (unseen data), It is better at identifying the negative class (0) than the positive class (1), However, the high recall for the positive class (1) indicates that it successfully identifies most of the positive instances.

To make improvements on the model performance, extract the Misclassified samples by the model by collecting the relevant information about these instance, including the input features, predicted outputs, and true labels, then looks for patterns or similarities among them, by creating to highlight the differences between it, Figure 4.1 below shows a histograms for the distribution of various features in the misclassified instances from the Random Forest model, categorized by their true labels (0 and 1).

From the analysis of the misclassified cases in the data shows interesting patterns. the age of individuals is different across the two classes, suggesting that age plays a big role in deciding the class. there are fewer 'Female' entries among these cases, hinting at a possible bias in the data or the way the model works based on gender. Also, body metrics like BMI and waist size, along with health indicators like blood pressure, show different trends in the two classes, This might be affecting how the cases are classified, the other health measurements such as blood sugar, cholesterol, and various ratios and levels in the blood show distinct trends for each group, indicating they may be important in understanding why these cases are misclassified.

In summery, appears that several features, including age, BMI, waist circumference, and the other features, show different distribution patterns between the true labels, also there was an factors that limit this model and make it performance not as needed because the model is sightly biased on one of the classes because of imbalanced dataset, and the feature elimination may be not as well.

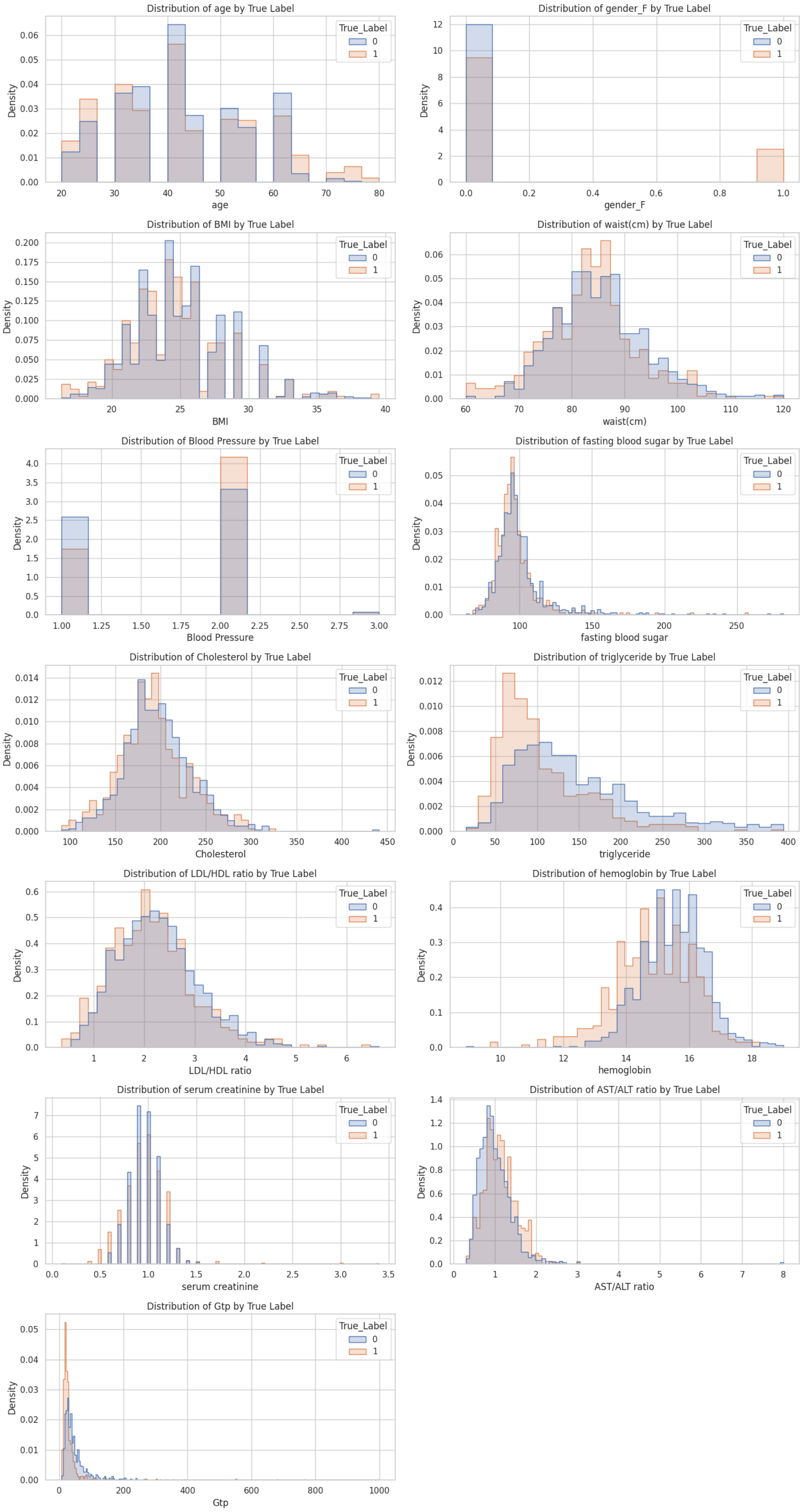

Figure 4.1: pattern between miss-classification examples for Random Forest.

5 Conclusions and Discussion

In our study, we found that the Random Forest method was a better than XGBoost and kNN model in accurately identifying if someone smokes , especially in less common cases. However, our study had some limits because we only looked at a specific group of people, which might not represent everyone. Also, it's hard to understand how these methods work inside, which is important for healthcare. In the future, we could try using a more varied group of people, test advanced methods that might spot complex trends better, and use more detailed ways to measure success that are more meaningful for health. Our study shows that using machine learning to predict health habits like smoking is promising and could lead to healthcare that's more tailored to each person's needs.