Smoke Signals: Unveiling the Bio-Signal Indicators of Smoking Behavior in a Machine Learning Framework

Julius Maliwat[a]

[a] *University of Milano-Bicocca, Data Science*

Abstract

This study focuses on analyzing the relationship between smoking and bio-signal indicators using machine learning, specifically employing the KNIME Analytics Platform. Smoking is a major cause of chronic diseases and presents a significant public health challenge. Our goal was to classify individuals as smokers or non-smokers and identify key health indicators affected by smoking. A Random Forest model was used to process a dataset containing demographic and biomedical data. The model achieved high performance in classifying smokers, evidenced by strong AUC, F1 score, and accuracy metrics. The study identified critical health indicators linked to smoking, such as hemoglobin levels, GTP, serum creatinine, triglycerides, and HDL cholesterol.

Contents

1. Introduction

The impact of smoking on public health is a subject of extensive research and concern. Smoking is known to be a leading cause of various chronic diseases, including heart disease, stroke, lung disease, and various types of cancer. The World Health Organization has identified tobacco use as one of the biggest public health threats the world has ever faced, killing more than 6 million people a year around the world (1). Understanding and mitigating the health impacts of smoking requires innovative approaches to identify smokers and understand the physiological and biological changes associated with tobacco use.

In this context, our study aims to explore the relationship between smoking and a range of bio-signal indicators using machine learning techniques, implemented through the KNIME Analytics Platform. We will use a dataset that includes a wide array of health metrics — from basic demographic data to detailed biomedical markers. This data offers a comprehensive snapshot of an individual's health profile, encompassing factors like cardiovascular health, metabolic function, and general physical condition.

The goal of our project, is not just to classify individuals as smokers or non-smokers based on their health data, but more importantly, to identify which health indicators are most affected by smoking. This understanding is crucial for several reasons. First, it can help in the early detection of individuals at risk of smoking-related health issues, enabling timely intervention. Second, it can provide insights into how smoking influences various biological processes and health conditions, contributing to broader research in public health and preventive medicine.

By employing machine learning models to analyze this data, we aim to uncover patterns and correlations that might not be immediately apparent through traditional statistical analysis.

2. Dataset Description

This rich dataset provides a multi-dimensional view of individual health statuses, crucial for our objective of predicting smoking behavior through machine learning techniques.

- ID: This column serves as an index, uniquely identifying each record in the dataset.

- Gender: This categorical variable indicates the gender of the individual.

- Age: grouped in 5year intervals, provides a broad overview of the age distribution of the participants.

- Height (cm): This metric records the height of the individual in centimeters

- Weight (kg): The weight of the individual is recorded in kilograms.

- Waist (cm): Waist circumference in centimeters, an important indicator of abdominal obesity.

- Eyesight (left) and Eyesight (right): These columns record the eyesight capabilities of the left and right eyes, respectively.

- Hearing (left) and Hearing (right): These variables measure the hearing abilities of the left and right ears.

- Systolic: This column indicates systolic blood pressure, a critical cardiovascular health indicator.

- Relaxation: This metric records relaxation blood pressure readings.

- Fasting Blood Sugar: A key indicator of metabolic health, particularly relevant for diabetes risk.

- Cholesterol (Total): Total cholesterol level in the blood, a vital marker for cardiovascular health.

- Triglyceride: Level of triglycerides, which are a type of fat found in the blood.

- HDL (High-Density Lipoprotein) Cholesterol: Often referred to as ’good’ cholesterol.

- LDL (Low-Density Lipoprotein) Cholesterol: Known as ’bad’ cholesterol, high levels are associated with health risks.

- Hemoglobin: A measure of the hemoglobin in the blood, essential for carrying oxygen.

- Urine Protein: Presence of protein in urine, which can indicate kidney health.

- Serum Creatinine: A marker used to assess kidney function.

- AST (Aspartate Aminotransferase) and ALT (Alanine Aminotransferase): Enzymes that help gauge liver function.

- Gtp (Gamma-Glutamyl Transferase): Another enzyme indicative of liver health.

- Oral Examination Status: A categorical variable indicating the results of an oral health examination.

- Dental Caries: This column records the presence of dental caries.

- Tartar: Indicates the status of tartar buildup, a dental health metric.

- Smoking: The target variable, indicating the smoking status of the individual.

To load this comprehensive dataset into the analysis workflow, the ’File Reader’ node in the KNIME Analytics Platform was utilized.

3. Exploratory Data Analysis

Exploratory Data Analysis (EDA) involves summarizing the main characteristics of a dataset, often with visual methods, to gain an understanding of the data’s structure, anomalies, patterns, and relationships. In the context of our smoking behavior study, EDA helps us to comprehend the dataset before applying machine learning models.

The dataset, encompassing a total of 55692 rows, comprises 27 columns. The target variable, ’smoking’, is binary, signifying that our task is a classification problem. There are 26 explanatory variables, consisting of 22 numerical and 4 categorical variables.

Target Distribution

A barplot analysis of the target variable reveals an imbalance: 63% of the entries belong to the ’negative’ (non-smokers) class, and 37% are smokers, as can be seen in Figure 1, using the node ’Bar Chart’. This distribution highlights the need for careful consideration of evaluation metrics in our classification models, as imbalanced classes can bias the model performance.

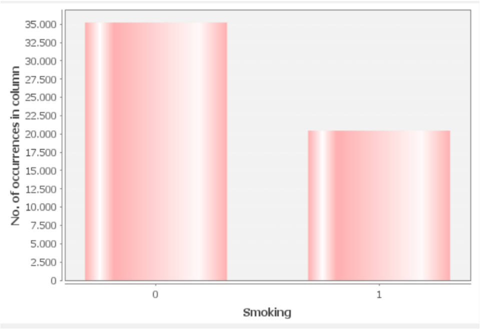

Figure 1: *Smoking Distribution*

Missing Values and Outliers

The dataset has no missing values. However, a critical observation from the boxplot analysis of the numerical variables is the presence of outliers in all 22 of them. These outliers are not anomalies or errors but represent natural variations in the data.

In health data, such variations are common and represent individual health differences. For example, variables like 'GTP', 'AST', and 'Fasting blood sugar' show significant right skewness, indicating higher values in some individuals, as can be seen in Figure 2 using the 'Box Plot' node. These outliers, reflecting unique health profiles, are crucial for a comprehensive analysis. Handling these outliers in the preprocessing phase is important. Proper management ensures that the analysis is balanced, capturing the dataset's nuances without being skewed by extreme values. Additionally, the wide ranges of numerical variables necessitate normalization for uniformity.

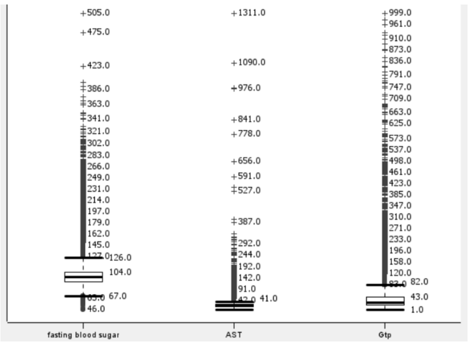

Figure 2: *Distribution of 'Fasting Blood Sugar', 'AST', 'GTP'*

Correlation Findings

We conducted a correlation analysis using linear correlation coefficients, a statistical measure that quantifies the degree of linear relationship between two variables. One of the most significant findings was the high correlation between waist circumference and weight. This is an intuitive result as both variables are measures of an individual's physical build and body composition. A larger waist circumference is often associated with higher body weight. The presence of multicollinearity, as observed between waist and weight, necessitates careful consideration in the preprocessing phase. When variables are highly correlated, they carry redundant information which can distort the true relationship between independent variables and the target variable. In machine learning models, particularly those based on linear assumptions, such redundancy can degrade model performance. In the preprocessing phase, we will address this issue by potentially removing one of the correlated variables

A deeper investigation into how numerical variables relate to smoking status uncovers nuanced differences, indicating possible physiological impacts of smoking. This analysis utilized conditional boxplots for a detailed comparison:

- Weight and Smoking: The conditional boxplots for weight show a clear distinction between smokers and non-smokers. Smokers, on average, have higher weight values compared to non-smokers. This could be indicative of lifestyle factors associated with smoking or potential metabolic changes due to tobacco use.
- Triglyceride Levels and Smoking: Similarly, triglyceride levels are higher in smokers. Elevated triglyceride levels in smokers may suggest a correlation with smoking-induced metabolic changes.
- Hemoglobin and Smoking: The analysis of hemoglobin levels in relation to smoking status, using the 'Conditional Box Plot' node, reveals significant differences between smokers and non-smokers (see Figure 3).

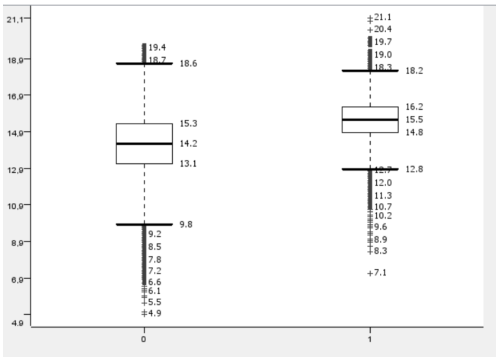

Figure 3: *Comparison of Hemoglobin Levels in Non-Smokers vs. Smokers*

These observations, particularly regarding weight, triglyceride levels and hemoglobin levels, are crucial for the subsequent stages of model development. They suggest that these variables are likely to be more predictive of smoking status and should be considered carefully during feature selection and modeling.

Categorical Variable Analysis

The analysis of categorical variables has led to insightful observations. The 'Oral' variable was found to have no variation, with all records showing the same value. This lack of diversity in the dataset renders it ineffective for analysis or prediction and thus was removed from further consideration. For 'Tartar', the distribution suggests a balanced representation of individuals with and without tartar. Similarly, 'Gender' shows a near-equal split between male and female participants. This balance is beneficial as it allows for a more nuanced analysis of how these variables interact with smoking status. A different scenario is observed with the 'Dental Caries' variable. There is a notable imbalance, with a greater proportion of the dataset indicating the absence of dental caries. This skewness could influence the model's ability to discern patterns related to dental caries and smoking.

A key part of our analysis involved examining the distribution of smokers across these categorical variables. A particularly striking finding is the low prevalence of smoking among female participants in the sample. The analysis indicates that almost no females in the dataset are smokers. Given this significant disparity and the aim of our study to focus on biomedical factors influencing smoking behavior, we have decided to exclude the 'Gender' variable from our predictive model. Including a variable with such an imbalanced distribution could introduce bias into the model, potentially skewing the results and interpretations. The analysis suggests a higher proportion of smokers among individuals with dental caries. This observation indicates a potential correlation between smoking and dental health, specifically the presence of dental caries, which could be a significant predictor of smoking behavior. These insights from the categorical variables analysis are integral to the feature selection process for our predictive model.

4. Preprocessing

Preprocessing in data science refers to the process of preparing raw data for analysis and modeling. It involves cleaning, transforming, and organizing data to enhance its quality and efficiency for extracting insights.

Before initiating the preprocessing steps, a fundamental task is to divide the dataset into training and test sets. We chose a 70:30 split, allocating 70% of the data for training the model and 30% for testing its performance. Given the slight imbalance in our target variable 'smoking', we have employed stratified sampling in our data split. Stratified sampling is crucial in maintaining the proportion of the target variable's classes in both training and test sets. This method ensures that both sets are representative of the overall dataset, preserving the underlying structure and distribution of the target variable. This division, done prior to engaging in any form of data cleaning or transformation, plays a pivotal role in preventing data leakage. Data leakage occurs when information from the test dataset, which is supposed to be unseen and unknown, inadvertently influences the training process. This can happen in subtle ways, such as when summary statistics computed from the entire dataset are used during the preprocessing of the training data, or when the data splitting process itself inadvertently allows some information to seep from the test set into the training set. The consequences of data leakage are significant. It can lead to overly optimistic performance estimates during model training and validation stages. So all preprocessing operations will be applied first to the training set, and subsequently, the same transformations will be applied to the test set using parameters derived from the training data.

4.1. Feature Selection

In our study, feature selection is guided by the objective of focusing on biomedical factors potentially influencing smoking behavior. We began by identifying and excluding non-biomedical variables. The ID variable, being an identifier, holds no analytical value for our modeling purposes and was thus removed. Similarly, we excluded gender and age, despite their potential significance, to maintain our focus on biomedical factors.

Further refining the dataset, we addressed variables that are not strictly biomedical in nature. We chose to remove weight, height, and waist circumference. These variables, while relevant in general health studies, do not directly contribute to our specific analytical goal of identifying biomedical factors influencing smoking behavior.

Conversely, we decided to retain variables like tartar and dental caries. Including these variables provides a broader perspective on possible health indicators related to smoking. Their inclusion is based on the premise that oral health conditions like tartar and dental caries, though often overlooked, could offer additional insights into smoking behavior.

We also scrutinized binary variables for their variability and relevance to smoking. Both Hearing(left) and Hearing(right) were removed, as they predominantly contained a single value with minimal entries in the alternate category. Their limited variability, coupled with the absence of a significant relationship with the smoking variable, led to their exclusion.

Through this methodical process of feature selection, we ensured that the dataset was streamlined, focusing only on those variables that are most likely to yield meaningful insights into the biomedical factors influencing smoking behavior. This approach not only improves the computational efficiency of our models but also enhances the interpretability of our results.

4.2. Feature Transformation

Feature transformation involves applying mathematical transformations to variables to make them more suitable for analysis and modeling.

Log Transformation

In our dataset, identified during the EDA phase, several variables demonstrated right-skewed distributions, indicating the presence of outliers. These outliers, while representing natural variations, can disproportionately influence the model's performance if not addressed. To mitigate this, we applied log transformations to several variables. The log transformation is a powerful technique that helps in normalizing data distributions, particularly useful for right-skewed data. By applying a logarithmic scale, extreme values are pulled closer to the mean, reducing the impact of outliers. This transformation makes the data more symmetrical, improving the model's ability to learn from it. The variables that underwent log transformation include Systolic, Relaxation, Fasting Blood Sugar, Cholesterol, Triglyceride, AST, HDL, ALT, and GTP. These variables were chosen because their skewed nature could potentially bias the predictive model. In Figure 4, you can see Boxplot of 'Fasting Blood sugar,' 'AST', and 'GTP' after the log transformation.

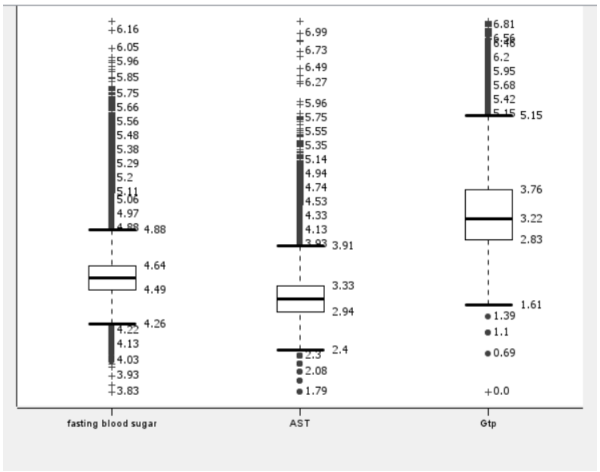

Figure 4: *Boxplot of 'Fasting Blood Sugar', 'AST', 'GTP' after log transformation*

Serum Creatinine presented a unique case where it exhibited a right skew but also included values between 0 and 1. For this variable, a log(1+x) transformation was applied. This variant of log transformation is particularly effective for datasets that include zero and small positive values. The addition of 1 ensures that zero values do not become undefined when logged, while still normalizing the distribution.

Binning

To further refine the dataset and reduce the influence of outliers, we employed binning on the variables that continued to show outliers even after log transformation. Binning is the process of converting continuous data into categorical intervals, which helps in smoothing out minor fluctuations in the data. By dividing these variables into six bins, we struck a balance between maintaining the integrity of the original data and minimizing the impact of outliers. This process was applied to the same variables that underwent the log transformation.

Binary Encoding

For the nominal variables in our dataset, binary encoding was implemented. For Tartar, categorized as 'Y' and 'N', we converted it into a binary format with 1 representing 'Y' and 0 for 'N'. This encoding simplifies the data, making it more suitable for use in most statistical models that expect numerical input.

Standardization

The final step in our feature transformation phase is standardization, an essential process for preparing the dataset for effective model training. Standardization is a technique that involves rescaling the features of our dataset so that they have a mean of 0 and a standard deviation of 1. This is achieved using the formula 1.

$$\text{Standardized Value} = \frac{\text{Value - Mean}}{\text{Standard Deviation}} \tag{1}$$

Standardization is particularly important in our context for several reasons. It helps to equalize the influence of each feature in the model, especially critical when dealing with variables that span different scales. Without this step, variables with larger scales could disproportionately affect the model, overshadowing the contribution of those with smaller scales. By standardizing, we ensure a level playing field for all features, making each one's contribution to the model's predictions equally important.

Another key benefit of standardization lies in its impact on the performance of certain algorithms. Models like logistic regression and Support Vector Machines (SVMs) are sensitive to the scale of the data. These models perform better when the features are on a similar scale, as standardization helps to avoid biases in the model's learning process and improves its overall predictive performance.

5. Evaluation Metrics

In machine learning, particularly in classification tasks, evaluation metrics are crucial for assessing the performance of models. They provide quantifiable measures to gauge how well a model classifies data into the correct categories. For our project on smoking behavior, we focused on three metrics, ordered by importance: Area Under the Curve (AUC), F1 Score, and Accuracy. We utilized the 'Scorer' node to calculate the F1 Score and Accuracy, and the 'ROC Curve (legacy)' node to evaluate the Area Under the Curve (AUC).

5.1. Area Under Curve

The Area Under the Curve (AUC) in the context of a Receiver Operating Characteristic (ROC) curve is a critical evaluation metric in classification problems, particularly in studies like ours where the primary goal is to distinguish between two classes - smokers and non-smokers. The ROC curve itself is a plot that illustrates the performance of a binary classification system as its discrimination threshold is varied. It plots two parameters: True Positive Rate (TPR) or Recall on the Y-axis and False Positive Rate (FPR) on the X-axis. The AUC measures the entire two-dimensional area underneath the entire ROC curve. It provides an aggregate measure of the model's performance across all possible classification thresholds. Essentially, it tells us how well the model can distinguish between two classes.

We chose AUC as our primary metric for several compelling reasons. In our dataset, there is a slight imbalance with a skew towards non-smokers. Traditional metrics like accuracy can often be misleading in such scenarios. However, AUC remains reliable as it is unaffected by the class imbalance and offers insight into how well the model can distinguish between smokers and non-smokers, irrespective of the classification threshold.

5.2. F1 Score

The F1 Score harmonizes Precision and Recall (see formula 2) , providing a single measure that balances both the accuracy of positive predictions and the completeness of capturing actual positives.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2}$$

5.3. Accuracy

Accuracy is a ratio of correctly predicted observations to the total observations. It is useful when the classes are symmetrically balanced and the costs of false positives and false negatives are roughly the same. The formula is the following:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FN}} \tag{3}$$

where TP (True Positives) and TN (True Negatives) are the positive and negative records correctly classified while FP (False Positives) and FN (False Negatives) are the positive and negative records incorrectly classified.

In our case, given the slight imbalance in the dataset, Accuracy is not as critical as AUC or F1 Score. While still a useful measure, it doesn't provide the nuanced view required for our analysis, hence its position as the third metric in our order of importance.

6. Models

In this chapter, we will examine four distinct models used for the data classification: Random Forest, Gradient Boosted Trees, Multi-Layer Perceptron, and Logistic Regression. Each of these models has its own characteristics and applications.

6.1. Random Forest

The Random Forest is an ensemble model that constructs several decision trees, each trained on random subsets of the training data and features. The 'Random Forest Learner' node is responsible for this training process. Once the trees are developed, their predictions are synthesized by the 'Random Forest Predictor' node, through majority voting, leading to the final model output. This model is known for its robustness and ability to handle data with many features and potential outliers.

The main hyperparameters that will be used for tuning are:

- n_estimators: This parameter defines the number of trees in the forest. Increasing n_estimators generally improves the model's performance but also increases computational cost.

- max_depth: This defines the maximum depth of the trees. Deeper trees can model more complex patterns but might lead to overfitting.

- min_samples_split: This parameter determines the minimum number of samples required to split an internal node. Higher values prevent the model from learning overly specific patterns, thus avoiding overfitting.

6.2. Gradient Boosting Tree

Gradient Boosted Trees is another ensemble model that builds a sequence of trees. In this case, each subsequent tree is trained to correct the errors of the previous models. It is known for being highly accurate and suitable for complex problems. In our analysis, the 'Gradient Boosted Tree Learner' node was employed to build the Gradient Boosted Trees model, and the 'Gradient Boosted Tree Predictor' node was used to generate predictions.

The main hyperparameters used are:

- n_estimators: Similar to Random Forest, this parameter defines the number of sequential trees to be modeled.

- max_depth: Controls the depth of each tree, impacting the model's ability to capture complex patterns.

- learning_rate: This is a crucial parameter that scales the contribution of each tree. Lower values generally lead to more robust models but require more trees to be effective.

6.3. Multi-Layer Perceptron

Multi-Layer Perceptron is a feedforward artificial neural network consisting of one or more hidden layers between the input and output layers. It is capable of capturing complex relationships in the data but requires careful hyperparameter tuning. The Multi-Layer Perceptron model was trained using the 'Rprop MLP Learner' node and predictions were made with the 'Multilayer Perceptron Predictor' node,

The main hyperparameters used are:

- hidden_neurons: This parameter defines the number of neurons in each hidden layer, which impacts the model's ability to capture complex relationships in the data.

- hidden_layer_sizes: This refers to the number of hidden layers in the network. More layers can model more complex functions but increase the risk of overfitting and computational cost.

6.4. Logistic Regression

Logistic Regression is a linear classification model that uses the logistic function to estimate the probability of a sample belonging to one of the two classes. It is a simple yet effective model, especially when the relationships between features and the target variable are linear. The Logistic Regression model was developed with the 'Logistic Regression Learner' node and applied using the 'Logistic Regression Predictor' node.

The main hyperparameters used are:

- C: This is the inverse of regularization strength. Smaller values specify stronger regularization, which can help prevent overfitting but might lead to underfitting if too strong.

- epsilon: This parameter influences when the optimization algorithm decides to stop iterating, potentially affecting the model's convergence.

7. Classification and Results

In the classification phase, our focus shifts to a methodical evaluation and comparison of the four machine learning models: Random Forest, Gradient Boosted Trees, Multi Layer Perceptron, and Logistic Regression. This phase is structured into several key steps, each critical for ensuring the robustness and accuracy of our findings.

The first step involves hyperparameter tuning for each model. This process is essential for optimizing the models' performance, where we adjust various parameters to find the most effective settings for our specific dataset. The tuning is guided by a structured approach, ensuring that each model is evaluated on a level playing field.

Once the models are tuned, we proceed to compare their performances in the validation stage. This comparison is based on their ability to accurately classify data they have not been trained on, a crucial test of their generalizability and effectiveness. The validation results provide a clear picture of how each model fares in real-world conditions.

The final step is selecting the best-performing model based on validation results. This model, having demonstrated superior performance during validation, is then used to evaluate its effectiveness on the test set.

7.1. Hyperparameter Tuning

The process of hyperparameter tuning, particularly within the scope of this project, had to be carefully navigated, considering the significant computational constraints we faced. Our approach, therefore, was tailored not only for optimization but also for efficiency. Within the KNIME Analytics Platform, we initiated this process with the 'Parameter Optimization Loop Start' node, setting the stage for a series of iterative training and validation steps that would be managed within the loop. Given these considerations, we selected Bayesian search as our primary method for hyperparameter tuning. A critical factor influencing our choice of the Bayesian search method was the intensive computational demand of our models. A single iteration of training could take up to 2 minutes. Given this, the entire process for 20 iterations would stretch to approximately 40 minutes per model. This duration was a key consideration, as it directly impacted the feasibility of extensive hyperparameter search strategies.

In light of these constraints, Bayesian search offered an efficient alternative to more computationally demanding methods like grid search or randomized search. Its ability to intelligently navigate the hyperparameter space, learning from each iteration and thereby reducing the need for exhaustive search, was particularly advantageous in our context. The configuration of the Bayesian search was set like this:

- Max Number of Iterations (20): This limit was set not only to guide the optimization process but also to keep the computational time within manageable limits.

- Number of Warm-Up Rounds (8): These initial rounds were crucial for gaining insights into the hyperparameter space without committing extensive computational resources to optimization.

- Gamma Value (0.25): This setting was crucial in balancing exploration and exploitation, especially important in our context to avoid unnecessary iterations that would increase computational load.

- Number of Candidates per Round (5): Evaluating multiple candidates in each round allowed us to explore a diverse range of hyperparameters while keeping the computational demands in check.

We have integrated 10-fold cross-validation into our model evaluation and hyperparameter tuning strategy, utilizing the 'X-Partitioner' and 'X-Aggregator' nodes within KNIME. This method involves dividing the training dataset into ten distinct parts. The 'X-Partitioner' node systematically manages the data segmentation, ensuring that the model is trained and validated on different data subsets, and is configured for stratified sampling to maintain proportionate class distribution in each fold. This approach allows each segment of the data to act as a validation set once, while the model is trained on the remaining portions. By employing the 'X-Aggregator' node, we then consolidate the results from each fold, providing a holistic view of model performance across all segments. This 10-fold cross-validation strategy, supported by these nodes, yields a more dependable measure of the model's predictive power.

The effectiveness of hyperparameter tuning is ultimately measured by the model's performance, specifically using the Area Under the Curve (AUC) metric.

7.2. Results in Validation and Model Selection

In the validation phase, our analysis focused on evaluating the performance of the four models: Random Forest, Gradient Boosted Trees, Multi Layer Perceptron, and Logistic Regression. In table 1, we can see the results of the models based on the metrics we choose.

Model	**AUC**	**F1 Score**	**Accuracy**
Random Forest	0.881	0.777	0.796
Gradient Boosting	0.815	0.72	0.741
MLP	0.812	0.719	0.739
Logistic Regression	0.794	0.7	0.725

Table 1: Performance metrics of classifiers in validation

The analysis revealed that the Random Forest model consistently achieved higher scores across the key metrics of AUC, F1 score, and accuracy, compared to the other models. This performance aligns with the characteristics of Random Forest, particularly its effectiveness in managing outliers and handling imbalanced data, which are prevalent in our dataset. The validation results underscored Random Forest's suitability for our project, leading to its selection for further evaluation on the test set.

The Logistic Regression model showed the weakest performance across all metrics, including AUC, F1 score, and accuracy. This can be attributed to its linear nature, which may not effectively capture the complex, non-linear relationships in our dataset, especially given the class imbalance between smokers and non-smokers. Logistic Regression's limitations in handling such complexities and imbalanced data are likely reasons for its lower performance in comparison to the Random Forest model.

7.3. Results on Test set

In the test phase of our project, we used the Random Forest model, which emerged as the best performer in the validation stage. This model was retrained on the entire training set using the optimal hyperparameters identified during the hyperparameter tuning process.We then evaluated its performance on the test set, and the results are represented in table 2.

Model	**AUC**	**F1 Score**	**Accuracy**
Random Forest	0.885	0.78	0.797

Table 2: Performance metrics of the Random Forest classifier in Test set

The test results demonstrate that the Random Forest model, with its optimized hyperparameters, performs effectively on unseen data. An AUC close to 1 indicates excellent model performance. In our case, an AUC of 0.885 suggests that the Random Forest model has a high capability in distinguishing between smokers and non-smokers.

A F1 score of 0.78 is quite strong, especially in a scenario with class imbalance. It indicates that the Random Forest model achieved a good balance in minimizing false positives and false negatives.

An accuracy of 79.7% is fairly high, indicating that the model correctly predicts the smoking status in about 80% of the cases in the test set

7.4. Feature Importance

After evaluating our Random Forest model on the test set, our analysis aimed to identify the key biomedical factors that impact the classification between smokers and non-smokers. Feature importance is a crucial metric in Random Forest models, ranking each feature by its ability to improve model accuracy through the reduction of impurity within the decision trees.

To quantify feature importance, we utilized the 'Math Formula' node, applying the expression detailed in formula 4, as sourced from the KNIME community forum (2).

$$\text{Importance} = \frac{s_0}{c_0} + \frac{s_1}{c_1} + \frac{s_2}{c_2} \tag{4}$$

In equation 4:

- s_i is the number of models where the attribute is used as a split at level i, with level 0 being the root.
- c_i s the number of times an attribute was in the candidate set for a split at level i, which varies from node to node in a random forest. Even if an attribute is selected for a split at level 0 and is also a candidate at level 1, it is still counted as a candidate for level 1 despite not being used for splitting at that level (3).

This approach allowed us to quantify the significance of attributes like hemoglobin, GTP (Gamma-Glutamyl Transferase), serum creatinine, triglycerides, and HDL in the classification process. These features were identified as the most influential based on the importance scores calculated, highlighting their roles in the predictive power of the model. The detailed scores of these features are presented in Table 3. Following this discovery, we aligned each of these features with existing scientific research to validate our insights:

Feature	Importance (all levels)
hemoglobin	2.776
Gtp	2.455
serum creatinine	1.825
triglyceride	1.353
HDL	0.814

Table 3: Top 5 Features Based on Importance Scores in Random Forest Model

- Hemoglobin: Hemoglobin levels can be affected by smoking, as the carbon monoxide in tobacco smoke increases the amount of carboxyhemoglobin in the blood, which can lead to altered hemoglobin levels. This makes it a significant indicator in differentiating smokers from non-smokers (4).
- GTP (Gamma-Glutamyl Transferase): GTP levels are often elevated in smokers. This enzyme is linked to liver function and can be influenced by factors like alcohol consumption and smoking, both of which are common in smokers (5).
- Serum Creatinine: Smoking has been shown to influence serum creatinine levels. Creatinine is a waste product that's filtered by the kidneys, and changes in its levels can be reflective of the impact smoking has on kidney function (6).
- Tryglyceride: Smoking can impact lipid metabolism, leading to altered triglyceride levels. Higher levels of triglycerides are often found in smokers, making it a relevant factor in classification (7).
- HDL (High-Density Lipoprotein): Smoking is known to reduce HDL cholesterol levels. HDL is often referred to as 'good cholesterol', and its lower levels in smokers contribute to cardiovascular risks (8).

The insights from our feature importance analysis, supported by these scientific studies, provide a robust understanding of the physiological and biological changes associated with smoking. This multi-faceted approach, combining data-driven modeling with established research, offers a comprehensive view of smoking's impact on health.

8. Conclusions

As we reach the conclusion of our study, it's evident that we have successfully achieved the objectives set out in the introduction. Our investigation into the impact of smoking on public health, leveraging machine learning techniques, has yielded significant insights and results.

In aligning with the introduction's emphasis on the extensive research needed to understand smoking's impact, our project employed a Random Forest model to analyze a comprehensive dataset encompassing a wide array of health metrics. This model demonstrated strong performance in classifying individuals as smokers or non-smokers, evidenced by robust metrics such as AUC, F1 score, and accuracy.

Furthermore, our study successfully identified key health indicators that are most influenced by smoking. Factors like hemoglobin levels, GTP, serum creatinine, triglycerides, and HDL cholesterol were found to be significant markers in determining smoking status. These findings, supported by external research, provide valuable insights into the physiological and biological changes associated with tobacco use.

References

[1] W. H. Organization *et al.*, *WHO global report on trends in prevalence of tobacco smoking 2015*. World Health Organization, 2015.

[2] K. C. Forum, "How to get the variable importance from the random forest model." `https://forum.knime.com/t/how-to-get-the-variable-importance-from-the-random-forest-model/8613/3`, 2021. Accessed: 2024-01-21.

[3] KNIME, "Random forest classification learner." `https://hub.knime.com/knime/extensions/org.knime.features.ensembles/latest/org.knime.base.node.mine.treeensemble2.node.randomforest.learner.classification.RandomForestClassificationLearnerNodeFactory2`. Accessed: 2024-01-21.

[4] D. Nordenberg, R. Yip, and N. J. Binkin, "The effect of cigarette smoking on hemoglobin levels and anemia screening," *Jama*, vol. 264, no. 12, pp. 1556–1559, 1990.

[5] D. Lee, H. Kang, and Y. Kim, "Association between cigarette smoking and serum gam-ma-glutamyl transferase level," *Respir Pulmonary Med*, vol. 6, p. 125, 2019.

[6] J.-M. Halimi, B. Giraudeau, E. Cacès, H. Nivet, Y. Lebranchu, and J. Tichet, "Effects of current smoking and smoking discontinuation on renal function and proteinuria in the general population," *Kidney international*, vol. 58, no. 3, pp. 1285–1292, 2000.

[7] W. Willett, C. H. Hennekens, W. Castelli, B. Rosner, D. Evans, J. Taylor, and E. H. Kass, "Effects of cigarette smoking on fasting triglyceride, total cholesterol, and hdl-cholesterol in women," *American heart journal*, vol. 105, no. 3, pp. 417–421, 1983.

[8] R. Garrison, W. Kannel, M. Feinleib, W. Castelli, P. McNamara, and S. Padgett, "Cigarette smoking and hdl cholesterol the framingham offspring study," *Atherosclerosis*, vol. 30, no. 1, pp. 17–25, 1978.