Executive Summary: Machine Learning Project Report

[**Introduction: 1**](#_kha2yu5p63o7)

[**Approach: 1**](#_eyix4xcwwj27)

[Data exploration 1](#_mt0q6vga0uvc)

[Feature engineering 2](#_p22rqtg6ebpn)

[Our machine learning algorithms 2](#_7ajbz8uwjwx2)

[**Results: 2**](#_oj7evqyiwhfj)

[**Challenges: 2**](#_dfcypkol6b7k)

[**Conclusions: 3**](#_ihpebotr3dg1)

# Introduction:

Our team undertook a machine learning project with the goal of addressing a patient’s smoking status given information about other various health indicators. The dataset we worked with was provided by Kaggle and included numerical and categorical data types that required minimal data cleaning and some data preprocessing. The machine-learning task involved classifying the smoker status using Logistic Regression, Random Forest, and XGBoost to achieve the evaluation metric of area under the ROC curve. We will be presenting our methodology regarding data preparation, data preprocessing, model training and evaluation, and evaluation metric interpretation.

# Approach:

To tackle this challenge, our approach involved a thoughtful exploration of the data, various experiments in feature engineering, and the utilization of multiple machine learning algorithms, including Logistic Regression, Random Forest and Xgboost.

## Data exploration

**Univariate:**

We started with a comprehensive understanding of the dataset, identifying key features and potential challenges. After initial exploration of the dataset, we realized that it was not missing any data and there were no duplicates so we were able to jump directly into data preparation. To start with, we made a univariate analysis, plotting a bar chart to see the distribution of each variable, discovering that there were some extremely high values in some columns, like a 9.9 in eyesight and 9.9 in serum creatinine.

To gain insights into the significance of the visual acuity values, we referred to eye examination standards which indicated that visual acuity scores ranged from 0.1 to 2.0, with values of 9.9 denoting blindness and those below 0.1 being excluded from evaluation. To verify this, we focused on right eyesight as a representative example and constructed heatmaps and correlation matrices. Our analysis revealed no discernible correlations between eyesight and other variables. Consequently, we retained values of 9.9 for further analysis.

For the numerical variables encompassing systolic and diastolic blood pressure, fasting blood sugar, cholesterol levels, triglyceride levels, HDL and LDL cholesterol levels, hemoglobin levels, serum creatinine levels, AST, ALT, and GTP, we scrutinized their distributions using histograms. Remarkably, they displayed semblances of normal distributions. Recognizing the potential impact of outliers on model parameters, we opted to compute the z-scores for these variables and eliminated values exceeding a z-score of 3.

Notably, in the case of serum creatinine, where only one entry exhibited a value of 5.9 while all others were below 3, we opted to exclude this particular row from further analysis for consistency.

Subsequent to the data cleansing process, we revisited the distribution plots for each variable. Encouragingly, we observed significantly improved distributions, rendering the dataset primed for subsequent analyses and modeling endeavors.

**Bivariate:**

In our analysis, we generated box plots to visualize the distribution of various variables, with smoking status as the independent variable along the x-axis and the respective variable values plotted along the y-axis. To ascertain the presence of a significant effect of smoking status on these variables, we employed a criterion wherein if the 25th percentile line of the box plot exceeded the median value of the non-smoking group, it indicated a discernible effect in prediction.

Consequently, we identified a notable impact of smoking on the following columns: Height, age, weight, waist circumference, triglyceride levels, HDL cholesterol levels, hemoglobin levels, serum creatinine levels, and GTP levels. This observation underscores the potential influence of smoking on these physiological parameters, warranting further investigation into its implications on health outcomes.

Given the categorical nature of our target variable, we opted to categorize our dataset based on the types of data present. Numerical variables were subjected to analysis using ANOVA tests to explore potential correlations with smoking status, while categorical variables underwent chi-square tests to quantitatively assess their interrelation with smoking.

Upon scrutinizing the data, a notable observation emerged: individuals who smoke exhibited tendencies towards greater height and weight, prompting speculation regarding potential gender-related influences. To investigate this hypothesis, we constructed scatter plots depicting the relationships between height, weight, and smoking status. Intriguingly, the density of non-smokers skewed towards the lower left quadrant, hinting at a potential correlation between smoking and gender. Subsequently, we computed Body Mass Index (BMI) values and found no significant correlation with smoking status, suggesting that gender may indeed exert a substantial influence. Additionally, we explored correlations between variables such as triglyceride levels and hemoglobin levels, recognizing the challenge of selecting suitable medical variables for comparison without specialized expertise.

## Feature engineering

For feature engineering, we chose to focus on dimension reduction and experimented with three methods. The first method was observing the p-value for certain features. The second method was applying scikit-learn functions to a baseline Logistic Regression model to extract the values of each feature. The third method was to utilize decision trees to evaluate feature importance to the model. After experimenting with these features, we decided to keep the top 10 important variables shown in random forest.

Furthermore, we computed the Body Mass Index (BMI) for each entry to explore potential correlations with smoking status. However, upon closer examination, we found that incorporating BMI into our analysis introduced multicollinearity issues, ultimately diminishing the performance of our predictive models. Consequently, we deemed BMI as not contributing significantly to our analytical objectives and excluded it from further consideration. It is not included in our final notebook.

Upon inspecting the dataset, we identified numerical and categorical columns and were able to separate the data according to each category. It was important to separate categorical data from numerical data in order to apply One Hot Encoding. One-hot encoding is crucial in machine learning as it transforms categorical data into a binary format, enabling algorithms to effectively process and interpret categorical variables. This method eliminates the ordinal relationship assumption in categorical data, ensuring that the model can accurately capture distinctions between different categories without introducing unintended numerical relationships.

## Feature selection

Our methodology centered around the RandomForest algorithm, specifically its feature importance rankings, to guide our selection process. We first found the optimal max\_depth setting for the RandomForest model and then proceeded with a targeted selection of features. This process involved sequentially adding features based on their ranked importance, closely monitoring the impact on the model's ROC AUC score to determine the optimal feature set.

Analysis indicated a plateau in performance improvement beyond the inclusion of the top 10 features, establishing a threshold for feature selection. The top 10 features were: height, hemoglobin, Gtp, triglyceride, weight, serum creatine, age, HDL, LDL and ALT.

These features were not only statistically validated through ANOVA tests, exhibiting the highest F-scores, but also demonstrated significant predictive power in univariate analysis. This approach helped in avoiding the pitfalls of overfitting and underfitting by ensuring a balanced model that leverages the most informative features without unnecessary complexity.

Further validation was conducted through correlation and variance analyses. The correlation matrix was instrumental in assessing the inter-feature relationships, ensuring minimal redundancy among the selected features. Variance analysis complemented this by confirming that the chosen features exhibited sufficient variability to contribute meaningfully to the model's predictive capabilities.

In the end our feature selection process was driven by a combination of RandomForest feature importance analysis and rigorous statistical validation. This ensured the development of an efficient, interpretable, and robust model, demonstrating a methodical and data-driven approach to achieving optimal model performance.

## Our machine learning algorithms

We were able to create code for all of our machine learning models using ChatGPT. This code, in addition to code snippets from our lab practicals, gave us the framework for our project.

We utilized a Logistic Regression model as the baseline model. There is research to suggest that this could have been our only model, as it works well for classification problems. However, we wanted to explore decision trees further since these are common in Kaggle competitions. We were able to derive feature importance from this model.

For the decision tree models, we captured feature importance from the gini index calculations. We pruned the tree to determine optimal split to improve generalization and regulate the complexity of the model. Although fine-tuning the hyperparameters of an individual decision tree may yield some enhancements, a more potent approach involves combining the outcomes of multiple decision trees trained with slightly diverse parameters. To achieve this, we tried the Random Forest model after evaluating our single Decision Tree model. We defined various functions to visualize the optimal hyperparameters of the model in an attempt to avoid overfitting.

We then explored the training process of our (GBM) model. The term "gradient" in GBM signifies that each decision tree aims to minimize the loss from the preceding iteration, akin to the concept of gradient descent. Meanwhile, the term "boosting" encapsulates the overarching technique of training new models to enhance the performance of an existing model. In contrast to Random Forest, which attains diversity by independently training each tree with random subsets of features, Gradient Boosting Machine (GBM) takes a sequential approach. For consistency across our machine learning workflow, we trained our XGBoost model on top 10 features established from the feature importance of our Random Forest model. We hoped that this would make it easier to manage and compare the models. Again, we defined various functions to visualize the optimal hyperparameters of the model in an attempt to avoid overfitting.

# Results:

# The results of our experiments revealed distinct performances among the algorithms. While we won't delve into technicalities, it's worth noting that : height, hemoglobin, Gtp, triglyceride, weight, serum creatine, age, HDL, LDL and ALT. We achieved a commendable Receiver Operating Characteristic Area Under the Curve (ROC\_AUC) rate of approximately 86.5% and key metrics such as precision, recall, and accuracy were considered, providing a comprehensive evaluation of the models.Challenges:

Throughout the project, our team encountered various challenges that tested our analytical prowess.The complexity of the data, coupled with the inherent difficulty of binary classification in health-related contexts, underscored the need for continuous improvement and exploration of new methodologies. Apart from it, the presence of aberrant data points, and striking the delicate balance between model complexity and performance count as major challenges as well. Addressing these challenges demanded a methodical and iterative refinement of our approach, wherein we continuously fine-tuned our methodologies to ensure the robustness and reliability of our models.

# Conclusions:

In conclusion, our work underscores the significance of 10 variables in predicting smoking status. Our findings demonstrate the efficacy of a meticulous approach focused on identifying optimal model parameters. By leveraging techniques such as Gridsearch to fine-tune parameters, we maximized our modeling efforts while minimizing computational overhead. Insights gleaned from this endeavor highlight the importance of variables such as height, hemoglobin, GTP, triglyceride, and weight in predicting smoking behavior.

To conclude our study on figuring out if someone smokes or not by using body signals, we've seen that going from basic data to helpful findings isn't easy but really eye-opening. Our work showed how powerful a tool like machine learning can be, especially with something called XGBoost, in dealing with tricky data to guess outcomes that matter a lot for health studies.

Looking ahead, there's a lot more we can do. We could try new ways to pick the most telling features or test out other machine learning models for even sharper predictions. Also, adding more varied data or different body signals could make our model work better across different groups of people. In addition, we advocate for the inclusion of gender as a variable for further investigation. Our analysis revealed a notable disparity in smoking prevalence, particularly among individuals with lower height and weight, suggesting a potential gender-related influence. This discovery presents promising avenues for future research and underscores the importance of considering gender dynamics in understanding smoking behavior. By addressing these recommendations, we can enhance the efficacy and comprehensiveness of predictive models in this domain. Furthermore, while XGBoost proved to be effective, there remains the possibility that alternative methods or combinations thereof could yield even better results. Therefore, maintaining a mindset of innovation and openness to novel approaches is essential for advancing our understanding and predictive capabilities in this domain.

Lastly, since this is our first formal attempt at more advanced machine learning algorithms, we see that our Kaggle score does not reflect the train/validation scores. This could indicate that we have overfitted our model, despite our best efforts to tune the hyperparameters appropriately. We conclude that in future iterations of machine learning models, such as Random Forest and XGBoost, we could research alternative hyperparameters to achieve a higher ROC AUC score.

In the end, mixing machine learning with studying body signals has a lot of promise for improving how we understand health behaviors. What we learned not only shows that our methods were on the right track but also opens doors for more research that could dig deeper into the complex info hidden in our biological data. This executive summary aims to provide a high-level overview of our machine learning project, emphasizing key outcomes and insights without delving into technical intricacies. If further details or clarifications are desired, we are readily available for a more in-depth discussion.