Body signals of smoking

Team 01: Gaia Corbetta, Florin Andrei Rusu

Abstract

The purpose of this project is to use personal data to predict whether a person is a smoker or not through a classification model. To do this, it is necessary to perform a careful analysis of the variables while also considering domain knowledge and perform some cleaning and feature engineering.
Machine Learning models will be selected based on their predictive goodness.
A specially provided training dataset was used for this project while the original dataset is available on Kaggle where it can be downloaded.

Table of contents

Introduction

Smoking is known to be a major factor in the development of many potentially life-threatening diseases. Therefore, being able to understand whether a person is a smoker or not could be crucial in making diagnoses and improving treatment.
The most complicated aspect is to understand which variables are relevant to obtain an accurate predictor.
The dataset provided is a sample referring to the Korean population that contains 27939 observations for 27 different variables.
The variables and their variable type in Knime are listed as follows:

- **ID**: index
- **gender**: Male or Female - string (M/F)
- **age**: 5-years gap - Number
- **height**(cm) - Number
- **weight**(kg) - Number
- **waist**(cm) : circumference - Number
- **eyesight**(left) - Number
- **eyesight**(right) - Number
- **hearing**(left) - Number
- **hearing**(right) - Number
- **systolic**: Blood pressure - Number
- **relaxation**: Blood pressure - Number
- **fasting blood sugar** - Number
- **cholesterol**: total - Number
- triglyceride
- **HDL**: cholesterol type - Number
- **LDL**: cholesterol type - Number
- **hemoglobin** - Number
- **urine protein** - string (1/0)
- **serum creatinine** - string (1/0)
- **AST**: glutamic oxaloacetic transaminase type - Number
- **ALT**: glutamic oxaloacetic transaminase type - Number

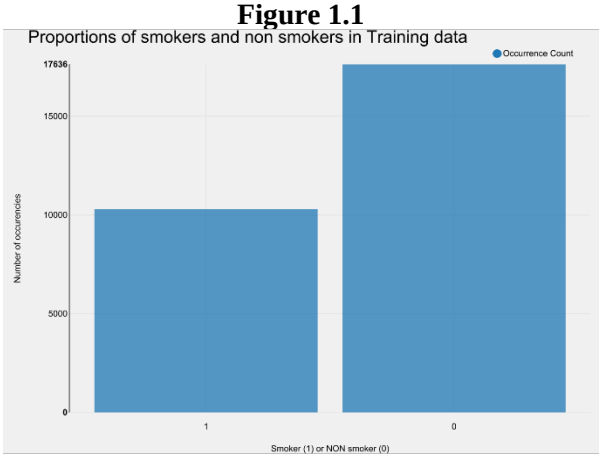
- **Gtp** : γ-GTP - Number
- **oral** : Examination status - string (1/0)
- **dental caries** - string (1/0)
- **tartar** : status - string (1/0)
- **smoking**: target variable - string (1/0)

In the following sections all the processes necessary to use Machine Learning tools and produce models capable of classifying smokers will be illustrated. The evaluation process will be crucial to select the model that has the best prediction performance.

1. Data Exploration

The first issue was to ensure that every variable was the correct data type, we also checked for possible missing values, none were found in the dataset.
Secondly, the binary response variable smoking was analyzed and an imbalance in the two classes emerged: within the training dataset there are 10303 smokers and 17636 nonsmokers. This is a problem that needs to be addressed and will be further discussed.

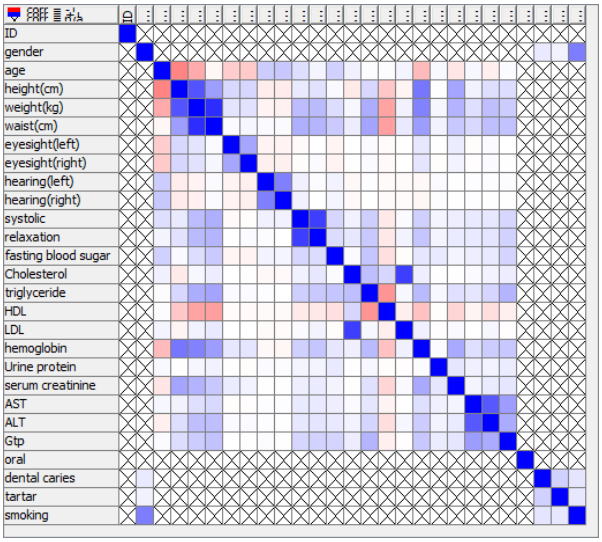
Figure 1.1



Finally, the relevance and redundancy of the explanatory variables was addressed.
The variance-covariance matrix was computed for the numerical variables; a few were found to be strongly correlated with each other such as height and age or cholesterol, LDL, HDL and triglycerides. The decision to not modify some variables was taken as correlation was preferred to the possible loss of information.
This decision was motivated by a better performance of models, including those robust in regard to collinearity problems, such as RandomForest.

Figure 1.2

Linear correlation table



The variance of numerical variables and the number of observations in the different classes of nominal variables were evaluated to decide which ones could be useful predictors. In fact, a lack of variance implies that a variable cannot be useful in explaining the variance in the data.
The presence of outliers was detected, this was handled by deleting all values bigger than Q3 + 5(IQR) or smaller than Q1 - 5(IQR), where Q3 is the third quartile of the considered variable, Q1 is the first quartile and IQR is the interquartile range.
The choice of an extremely conservative multiplier for the interquartile range was motivated by the intent to exclude only impossible values.

2. Pre-Processing

Feature Selection

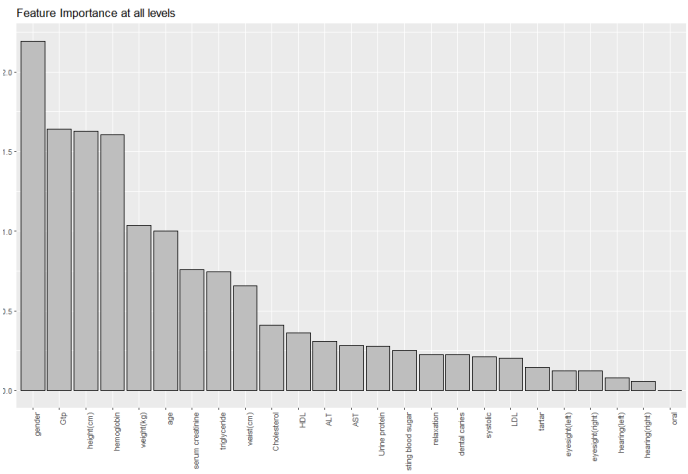
After evaluating the correlations between variables and the variance, the decision to eliminate all variables that can be considered zero or near zero variance predictors was taken.
The following low variance variables were identified:

- eyesight both left and right (removed at 0.3% variance threshold),
- hearing both left and right (removed at 0.2% variance threshold),
- urine protein (removed at 0.3% variance threshold),
- serum creatinine (removed at 0.1% variance threshold),
- oral (nominal variable with all observations from the same class 'Y')

All the other variables displayed variance bigger than 5%.
The ID column was removed as it is not relevant for the prediction problem, while 'weight' and 'height' were summarized in a different variable that will be further discussed and so they were consequently removed.
All the variables were also standardized.
In order to evaluate the importance of variables the KNIME Tree Ensemble was employed, the node offers a second outport which gives details about the variable importance. The node shows how often a variable was used for building a decision tree at the first, second or third level. (KNIME Forum, 2015)

Figure 2.1

Feature importance (all levels)



As can be seen the most important variables at all levels seem to be gender, gtp, height and hemoglobin. The least important are oral, hearing, eyesight, and tartar.
To summarize, the goal of variable selection is to reduce as much noise as possible so as to improve the performance of the classifier. To do so two different approaches were followed: the evaluation of the variance of explanatory variables and feature importance analysis.
This was to ensure a correct selection of the explanatory variables and increase the robustness of predictive models.

Feature Engineering

A synthetic index was created to incorporate height and weight into a single variable, given the high correlation between the two. In this regard, the BMI - Body Mass Index variable was created.

$$BMI = Weight(kg) / Height(cm)^2$$

BMI is commonly used as an indicator of body fatness and a screening tool for weight categories that may lead to health problems. Knowing the BMI of individuals can provide insights into their overall health status, as extreme values (both low and high) are associated with various health risks. (Piirtola et al., 2018)

Imbalanced Class

When one class heavily outnumbers the other, prediction problems can arise as this can lead to predictions biased towards the majority class.
In terms of performance evaluation and model comparison the problem needs to be addressed as accuracy becomes a not reliable evaluator: the model can achieve high accuracy by simply predicting the majority class most or all of the time.
In light of this, the problem of class imbalance must be addressed to improve forecasting performance.
The **S.M.O.T.E.** (Synthetic Minority Oversampling Technique) filter was implemented through the Knime 'SMOTE' node to balance the number of observations in the two classes of the response variable. (*Smote*. KNIME Community Hub)
It creates synthetic samples for the minority class starting from the existing instances. This is possible by selecting a minority class

instance and its k nearest neighbors. Synthetic instances are then generated along the line segments connecting the selected instance to its neighbors. The result is two balanced classes.
The filter was only applied to training data to assure the best performance of the models, while also maintaining an unbiased estimation on the test set.
The equal size sampler, a Knime node that undersamples the majority class, was also evaluated but the performance of the models was worse.
When sampling partitions of the dataset, both for hold-put and cross-validation, stratified sampling was employed to maintain the proportions between the two classes of the output variable.

3. Models

Different models were tested to understand which one had the best performance.
The following models were employed:

- Naive Bayes Tree
- Spegasos
- SMO poly
- Decision Tree
- J48
- Random Forest
- Logistic
- XGBoost

These models exhibit diverse characteristics in terms of complexity, interpretability and handling of different types of data, allowing the choice of the most appropriate one for the classification problem at hand.

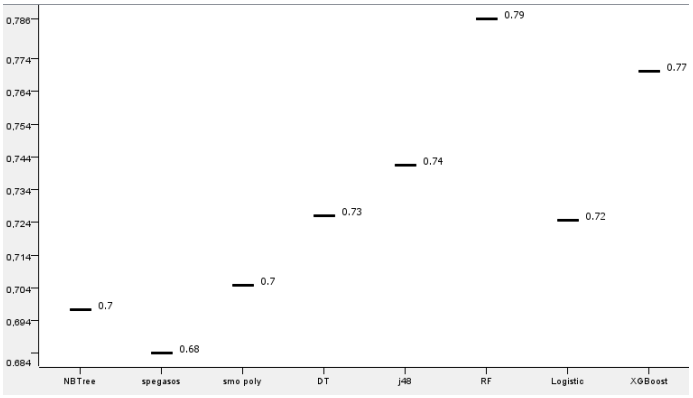
4. Evaluation

Overall Accuracy

As a first step to evaluate the performance of the models, the overall accuracy of the models was compared. The dataset was partitioned in training (80% of the observations) and test set (20% of the observations).

Figure 4.1

Models overall Accuracy



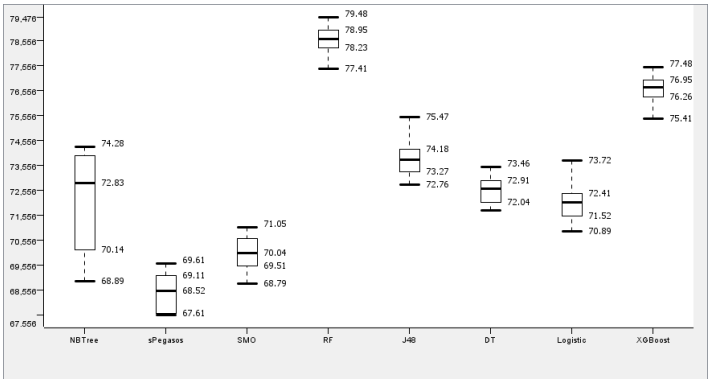
As previously stated, accuracy is not the most reliable metric for performance evaluation when there is an imbalance in the response variable distribution, nevertheless it is useful when compared to other metrics.
The best model according to this evaluation is the ***Random Forest*** model which achieves an overall accuracy of 0.79 followed by the XGBoost with 0.77.

Cross Validation

This method provides a robust assessment of a model's performance by testing it on multiple subsets of the dataset. This helps understanding how well the model generalizes to new, unseen data.
Cross-validation is a crucial technique for obtaining a more accurate and stable evaluation of a machine learning model, contributing to better model selection, hyperparameter tuning, and overall performance assessment. (*Cross-validation,* 2023)
K-fold cross validation was employed, with k = 10, meaning that the training set (80% of the original observations) was split in 10 subsets. This choice was made to have a good comparison to the performance of the models obtained through the hold-out method trained on the same subset.

Figure 4.2

Models overall Accuracy with Cross Validation

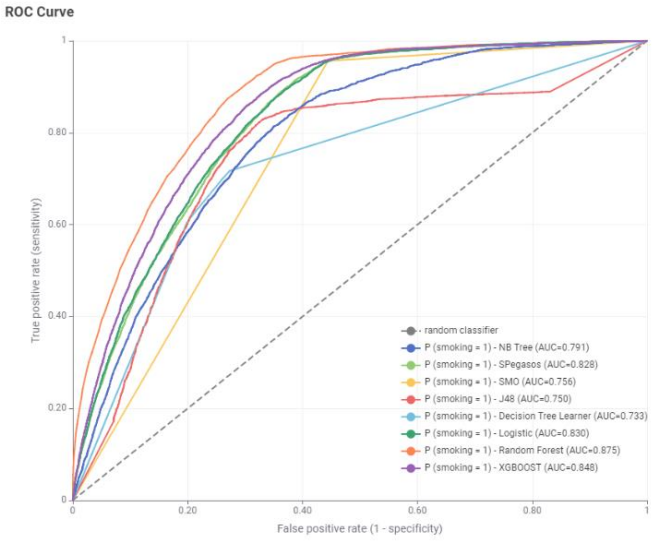


Random Forest continues to prove the model with the best accuracy reaching a median of 78.599% followed by XGBoost with 76.668%.

ROC curve and AUC

The ROC curve or 'Receiver operating characteristic curve' plots the true positive rate against the false positive rate according to each model, AUC can be defined as the 'Area under the ROC curve'.
The minority class is selected as the positive one.
ROC curve and AUC offer valuable insights into the overall performance of binary classification models, especially when considering varying decision thresholds and dealing with class imbalances. They are versatile tools for model evaluation and comparison.

Figure 4.3



RandomForest is the model that achieves better results also according to this metric.

Other metrics

Finally, the focus was placed on different metrics to test the models including: Recall, Precision, F-measure.
These metrics are much more reliable when an imbalance is present in the distribution of the response variable. Recall, Precision and F-measure check respectively the amount of observations correctly classified as positive out of all the positive observations and the amount of observations that are actually positive out of all the observations that are classified as positive. The last metric, F-measure, represents a synthesis of the first two as it is the harmonic mean between Recall and Precision.
In all cases the minority class is selected as the positive class.
The previously described metrics were computed for each model, as shown:

Figure 4.4

Models performance table

RowID	Accuracy Overall *String*	Recall *String*	Precision *String*	F-measure *String*	AUC *String*
NBTree	0.697	0.407	0.623	0.492	0.791
SPegasos	0.684	0.972	0.534	0.689	0.828
SVM-poly	0.705	0.966	0.552	0.702	0.756
DT	0.726	0.709	0.602	0.651	0.733
J48	0.741	0.749	0.617	0.676	0.750
RF	0.786	0.821	0.665	0.735	0.875
Logistic	0.724	0.911	0.575	0.705	0.830
XGBoost	0.770	0.743	0.661	0.700	0.848

The best model according to accuracy, precision, f-measure and AUC is RandomForest, while the best model according to recall is sPegasos. (Desai, 2024)

5. Conclusion

The process of developing a smoking prediction model involved: data preprocessing, feature engineering, and model selection. Notably, the integration of BMI, addressing

class imbalance with SMOTE, and employing Random Forest as the primary machine learning algorithm led to an overall accuracy of 78.9%.

The other evaluation metrics, particularly ROC curves and AUC, provided valuable insights for model selection.

In conclusion, the selected model achieves a good performance in classifying smokers and non-smoker. Nevertheless, future studies may include new variables and some more feature engineering in order to get a higher performance in the classification process.

References

Desai, U. (2024, January 4). *Demystifying classification evaluation metrics: Accuracy, precision, recall, and more*. Medium. https://utsavdesai26.medium.com/demystifying-classification-evaluation-metrics-accuracy-precision-recall-and-more-613dc7cc44b2

Piirtola, M., Jelenkovic, et al. (2018). Association of current and former smoking with body mass index: A study of smoking discordant twin pairs from 21 twin cohorts. PLOS ONE, 13(7), e0200140. https://doi.org/10.1371/journal.pone.0200140

Smote. KNIME Community Hub. https://hub.knime.com/knime/extensions/org.knime.features.base/latest/org.knime.base.node.mine.smote.SmoteNodeFactory

Wikimedia Foundation. (2023, December 19). *Cross-validation (statistics)*. Wikipedia. https://en.wikipedia.org/wiki/Cross-validation_(statistics)

KNIME Forum. (2015, September 17). *How to get the variable importance from the Random Forest Model?*. KNIME Community Forum. https://forum.knime.com/t/how-to-get-the-variable-importance-from-the-random-forest-model/8613