Health indicators of smoking

Kateřina Bártová, Vojtěch Balek, Jan Hes

Introduction

Our machine learning model is designed to distinguish smokers from non-smokers. This could be of great value for insurance companies as surgeries and treatments of diseases that are proved to be caused by smoking are among the most costly treatments available and take about 20% of annual health expenditure[1]. In South Korea, which is the country of origin of our dataset, health expenditure on treatments of diseases caused by smoking attributes for approx. 0.59 % - 0.78 % of GDP or around $2269.42 million - $2956.75 million[2]. Insurance companies would naturally seek to set their pricing policy accordingly for smokers and non-smokers. This model could help those companies with identification of said clients and allowing for adequate pricing measures.

Our dataset contains medical records of people in South Korea with information whether they smoke or not.

Link to the dataset: https://www.kaggle.com/datasets/kukuroo3/body-signal-of-smoking

All original columns are explained at the end of the report in section "Column explanation".

Our target attribute is "smoking" column as we want to train our model to predict smokers.

As our instance of interest, we chose a person with attributes:
gender = 1, age = 20, height(cm) = 170, weight(kg) = 70, waist(cm) = 78.0, eyesight(left) = 1.0, eyesight(right) = 1.2, hearing(left) = 1.0, hearing(right) = 1.0, systolic = 119.0, relaxation = 73.0, fasting blood sugar = 91.0, cholesterol = 179.0, triglyceride = 71.0, hdl = 67.0, ldl = 98.0, hemoglobin = 13.7, urine protein = 1.0, serum creatinine = 1.1, ast = 24.0, alt = 16.0, gtp = 20.0, dental caries = 1, tartar = 1, oral = Y, smoking = 0

We selected "hemoglobin" as our attribute of interest.

Our subset of rows is defined as males with dental carries who are listed as 20 years old, this yields 371 rows.

We chose cost-benefit matrix as follows:
TP: 100, FP: -90, TN: 20, FN: -100

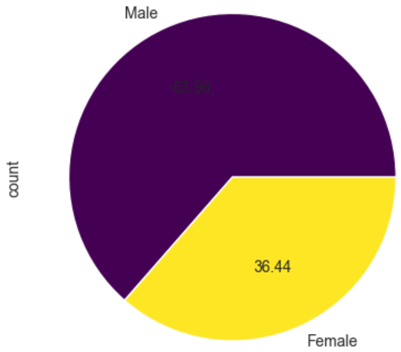
Where positive values mean benefit for the company, whereas negative mean loss. Our reasoning for this matrix is mainly given by insurance companies' desire to identify smokers (true positive), while trying to avoid identifying non-smokers as smokers (false-positives) as that would tarnish their reputation and make them lose clients. In addition, failure to identify a smoker (false-negatives) can prove costly in the long term.

[1] https://ct24.ceskatelevize.cz/ekonomika/2136702-restaurace-bez-koure-za-lecbu-se-mohly-usetrit-miliardy

[2] https://en.wikipedia.org/wiki/Smoking_in_South_Korea

Exploration

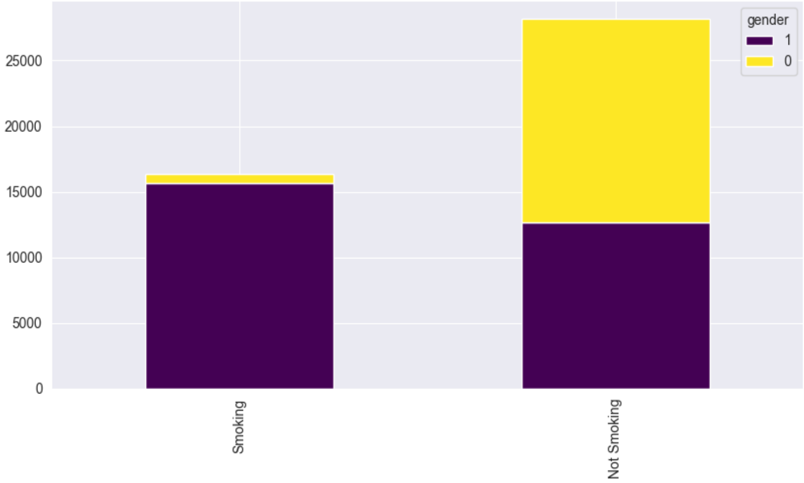
We performed data exploration after test-train split and initial pre-processing as it makes working with the data much easier. We used the train dataset with 44553 rows.



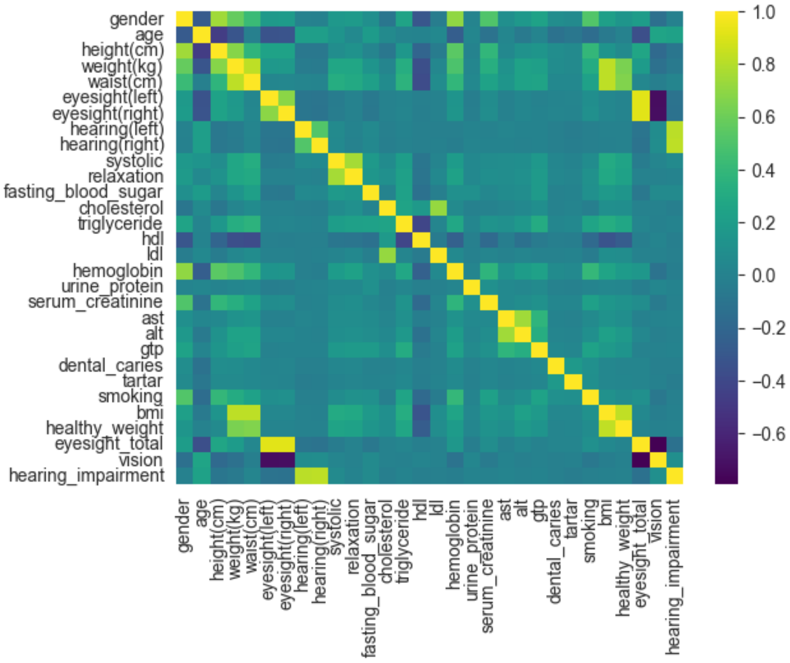
In our train dataset, there is a higher ratio of men than women, who make up approx. 63.50 % of medical logs. Exact numbers are 28316 males and 16237 females.

Upon closer inspection, we can see that vast majority of people recorded in the dataset as smokers are men – 95.61 %. Only 4.39 % of all smokers are women.

Among non-smokers, the numbers are much more equal – 55.18 % for women and 44.82 % for men.

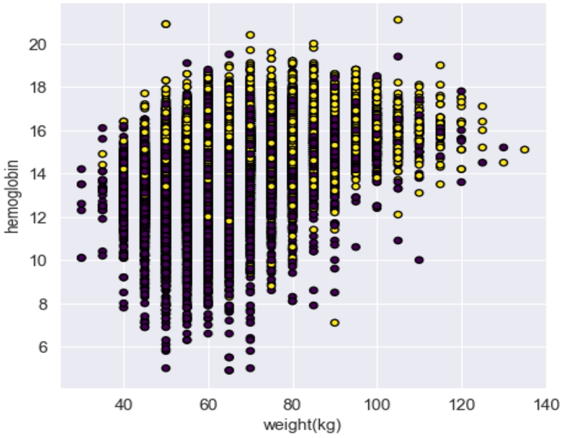


As there are more men than women in the data, we expected the ratio of men among smokers to be higher, however the ratio isn't even remotely close to the male/female ratio. This can be caused by various cultural and/or social factors specific to Korean society, which would need much more deeper analysis.



We then looked deeper into correlations. There were no strong (> 0.70) correlations with our target attribute "smoking". Strongest positive correlation could be seen with "gender" where r = 0.51, then with "hemoglobin" with r = 0.40, "height(cm)" r = 0.40 and "weight(kg)" r = 0.30. Correlation with height and weight could be caused by the fact that overwhelming majority of smokers are men, who, on average, tend to weigh more and be taller than women. Other positive correlations were weak (< 0.30).

Strongest negative correlation could be seen with "hdl" where correlation coefficient = -0.18, which is considered a weak correlation.



Lastly, we looked closer upon correlation between hemoglobin, weight and smoking. From the scatterplot on the left where yellow points show smokers and purple non-smokers, we can see that there indeed is a trend where people who smoke are more likely to have higher levels of hemoglobin and weigh more.

Data Preprocessing

Our dataset was already heavily preprocessed from our data source. Still, some changes needed to be made in order for it to be viable training and testing dataset. Source code is available in attached jupyter notebooks.

Preprocessing for supervised learning

After loading the data, we first split it randomly into two parts – testing and training. We chose 20 % of the dataset for testing and 80 % for training. We selected seed 1337 in order to get reproducible results.

We then check for NA/null values, which there are none, so there was no need to impute the missing values.

While checking for non-numerical variables, it was revealed that columns “tartar”, “gender” and “oral” are in fact not numerical. As all of these variables are categorical - “tartar” and “oral” have Y/N values - we used 1/0 to represent those boolean values in numerical form. The column “gender” has M/F values to distinguish males and females, we use 1 for males and 0 for females.

After further check, we see that every instance in both train and test dataset have “oral” value set to 1, which means that this column bears no information and we can drop it.

We then decided to add new column “bmi”, which allows us to more precisely track respondents’ health. While Body Mass Index is not a perfect representation of healthy weight/height ratio, we believed that it could give us more insight into body signals of smoking than weight and height alone. Based on BMI, we also added column “healty_weight” which had values -1, 0, 1 for malnourished, healthy and obese people.

Columns “eyesight(left)” and “eyesight(right)” bear information about respondents’ eyesight in left and right eye respectively. The unit for the measurement is visual acuity, which ranges from 2.0 (best vision) to 0.1 (nearly blind), and 9.9 represents total blindness. First, we changed 9.9 values to 0.0, as it more closely represents blindness and fits the scale. After that we added up these columns, to calculate total vision acuity.

From total vision acuity, we used WHO source data to create a new column “vision”. Vision is a categorical variable with values as follows:
0 - abnormally good vision - total visual acuity > 2.5
1 - unimpaired vision - total VA between (1.1;2.5)

2 - mildly impaired vision - total VA between (0.66;1.0)
3 - severely impaired vision - total VA between (0.2;0.66)
4 - blindness - total VA < 0.2

Analogical to eyesight, we decided to look into hearing. Columns “hearing(left)” and “hearing(right)” represent ability to hear in left and right ear respectively with “1” meaning perfect hearing and “2” meaning imperfect hearing. We decided to create a new column “hearing_imparment” which has value “0” if a person has perfect hearing in both ears and “1” otherwise.

The final step of pre-processing for supervised learning was to change column names so they are all in lowercase and follow snake_case standard.

After that, both test and train data were dumped into their respective .csv files for further use.

Pre-processing for unsupervised learning

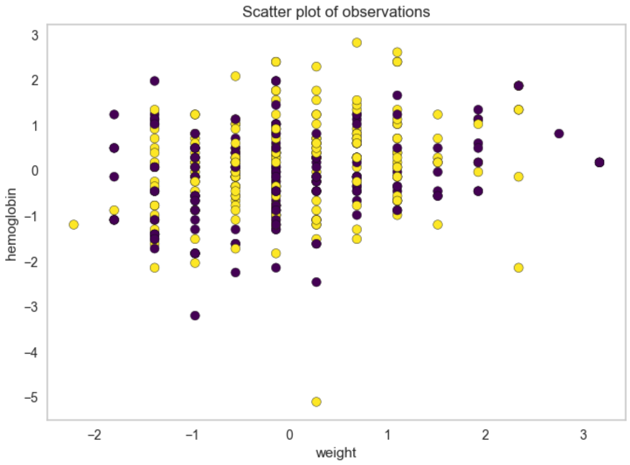
For unsupervised learning, we are starting with the pre-processed dataset for decision trees. In many visualisations connected with unsupervised learning, we are using “weight” and “hemoglobin” as axes as those are the two columns with the highest correlation with the target.

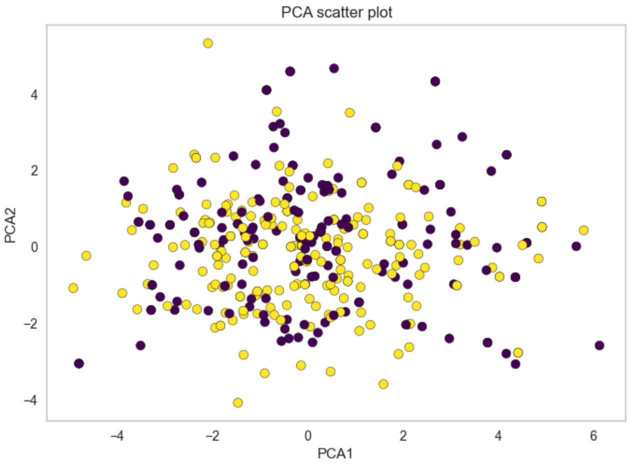
We decided to use subset of males who are in their early 20s and have dental caries (371 observations, 200 smokers + 171 non-smokers). After separating said rows, columns “gender”, “dental_caries”, and “age” lost their value so we dropped them.

Because unsupervised learning is based on finding distances, we needed to rescale features so that they have similar weights and one variable cannot influence the whole dataset. We used the StandardScaler() function from package sklearn on all numerical features.

Since our dataset has many columns, we decided to use principal component analysis to deal with multidimensionality and compare models with and without PCA.

Right before dumping the data into csv files, we tried to plot the data so we can later compare the actual data with predicted. Then we removed the target column “smoking”. Just from looking at the scatter plots, we presume that clustering is not going to be very successful.





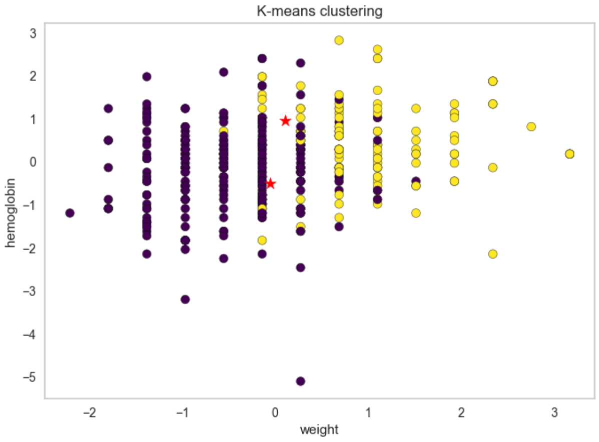
Modelling

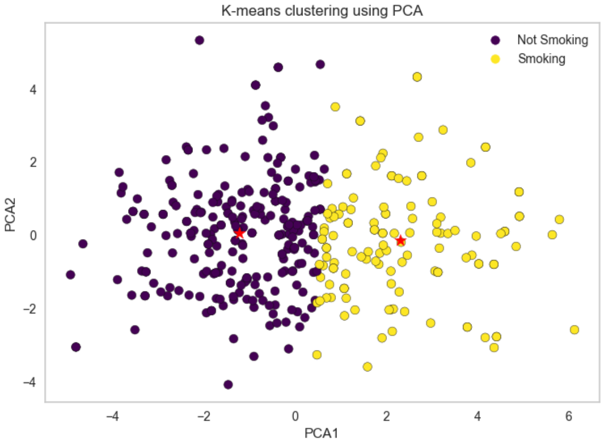
Supervised model

The two machine learning algorithms we used are decision trees and random forests. For each of them, we created a “bigger” and a “smaller” one by using different metaparameters. For the random forests, we changed the number of trees. The bigger one used 100 trees and the smaller forest only 10. For the trees, we altered their depth. We limited the depth of the smaller one to 3, so that we are able to visualize it properly. The depth of the bigger tree was not limited, which means that the nodes are expanded until all leaves are pure or until all leaves contain less than two samples.

Clustering model

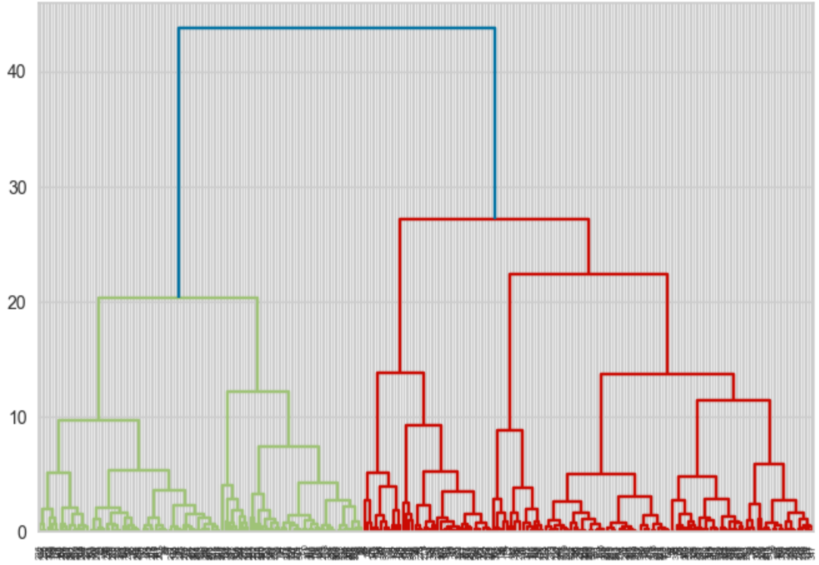
First, we tried to fit the k-means algorithm to both our sets of data. We used two as k (the number of clusters) because we are mostly interested in splitting out subjects into two groups (smokers/non-smokers). Then we plotted both approaches (see below). On the first glance, it seems that the model without PCA is a bit closer to the actual data. Red stars in plots represent cluster centroids.



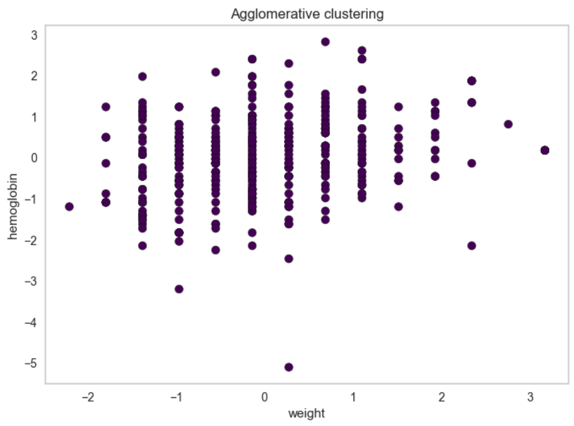


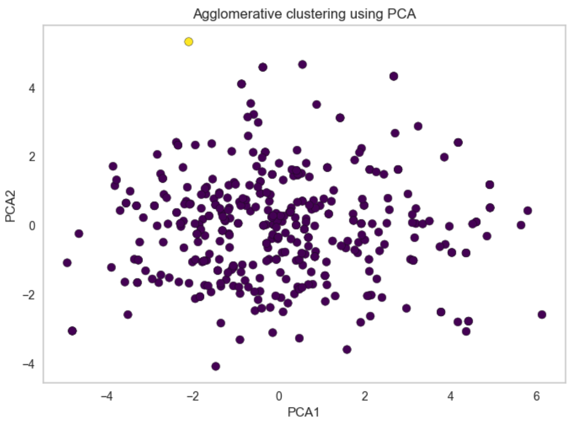
To try to fit the model better, we used the elbow visualizer to determine an optimal number of clusters. We re-ran both fits with improved number of clusters.

For hierarchical approach, we used agglomerative clustering. Unlike k-means method, they start from the bottom – single leaves – and group them together based on their proximity. Picture below shows a dendrogram of how our final PCA agglomerative model was linked.



Because we are interested in two groups, we set the function to create two clusters. We tried to change the settings of 'linkage' method in AgglomerativeClustering function, however many options did not seem to make sense (most of the observations were in one large cluster and very few observations in the other) so in the end we decided to use 'ward' option. Next two pictures show results with 'single' as value of said parameter.

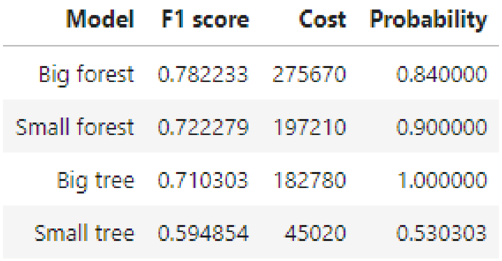

Agglomerative clustering


Agglomerative clustering using PCA

Model Evaluation

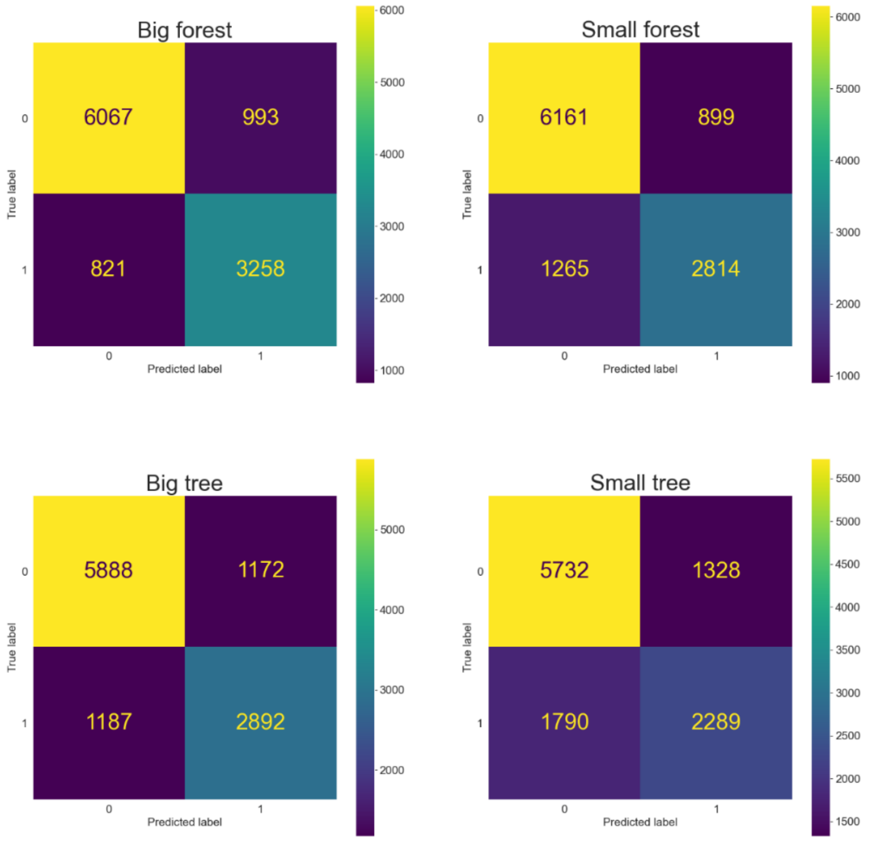
Supervised model

As both false negatives and false positives are expensive for us according to the cost matrix, we chose F1 score as the metric we want to measure. In the table below, we can see that the bigger random forest is by far the best performing model. The smaller forest performs worse, closely followed by the big decision tree. The small tree performs poorly, because we limited it to only three nodes.



Model	F1 score	Cost	Probability
Big forest	0.782233	275670	0.840000
Small forest	0.722279	197210	0.900000
Big tree	0.710303	182780	1.000000
Small tree	0.594854	45020	0.530303

Confusion matrices:



Big forest

True label \ Predicted label	0	1
0	6067	993
1	821	3258

Small forest

True label \ Predicted label	0	1
0	6161	899
1	1265	2814

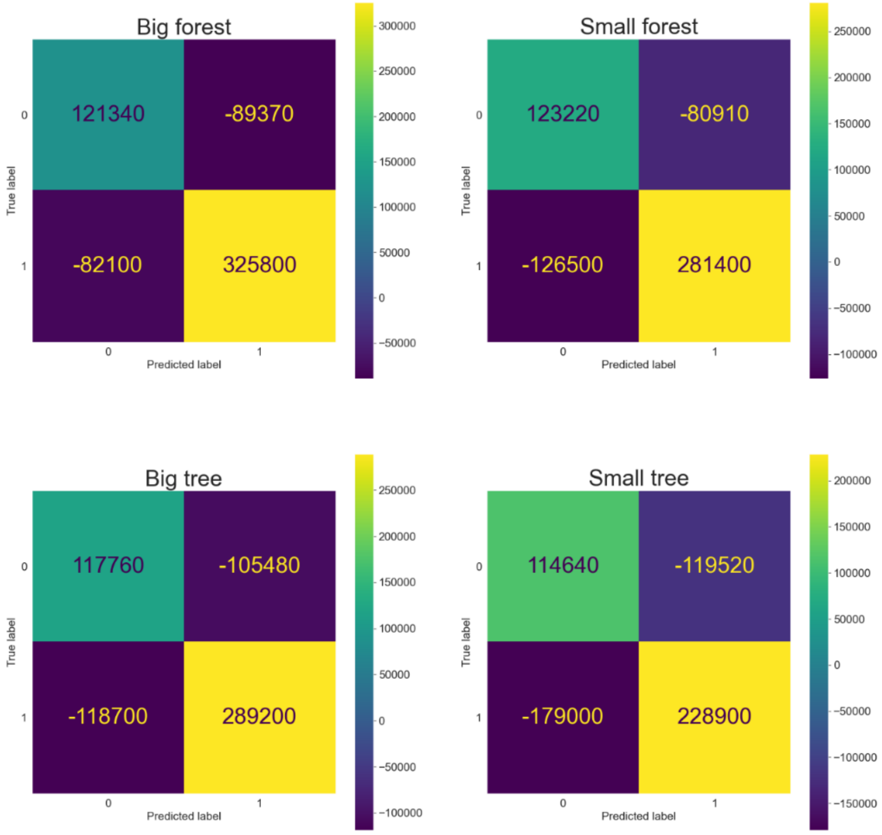
Big tree

True label \ Predicted label	0	1
0	5888	1172
1	1187	2892

Small tree

True label \ Predicted label	0	1
0	5732	1328
1	1790	2289

Combined confusion and cost matrices (positive values signify profit and negative values signify loss):



Big forest

True label \ Predicted label	0	1
0	121340	-89370
1	-82100	325800

Small forest

True label \ Predicted label	0	1
0	123220	-80910
1	-126500	281400

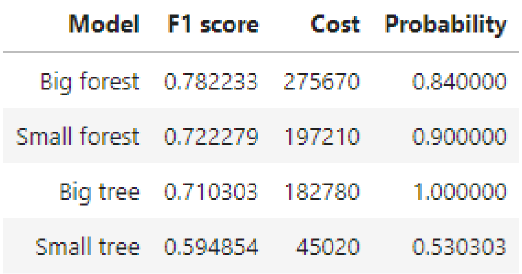
Big tree

True label \ Predicted label	0	1
0	117760	-105480
1	-118700	289200

Small tree

True label \ Predicted label	0	1
0	114640	-119520
1	-179000	228900

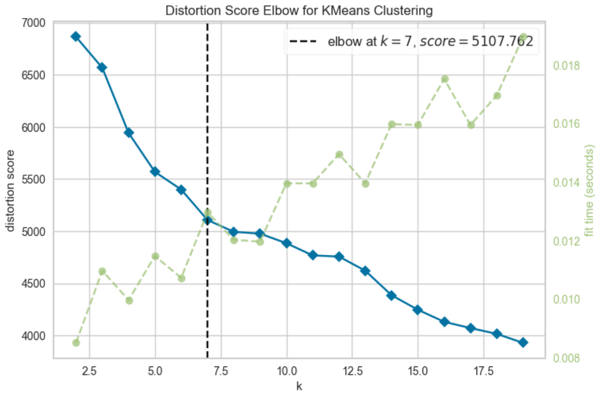
As we can see from the total sum of the combined matrix (the “Cost” column in the table below), the big forest model has by far the highest value, therefore it is the model that brings the insurance company the most profit.

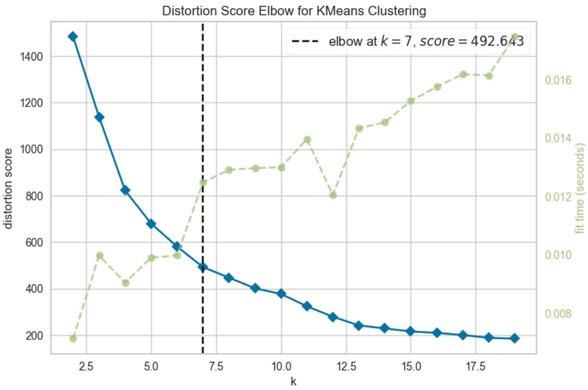


Model	F1 score	Cost	Probability
Big forest	0.782233	275670	0.840000
Small forest	0.722279	197210	0.900000
Big tree	0.710303	182780	1.000000
Small tree	0.594854	45020	0.530303

Unsupervised model

In both our approaches, using plain rescaled data and principal components, the elbow graph based on inertia suggested seven as the optimal number of clusters.





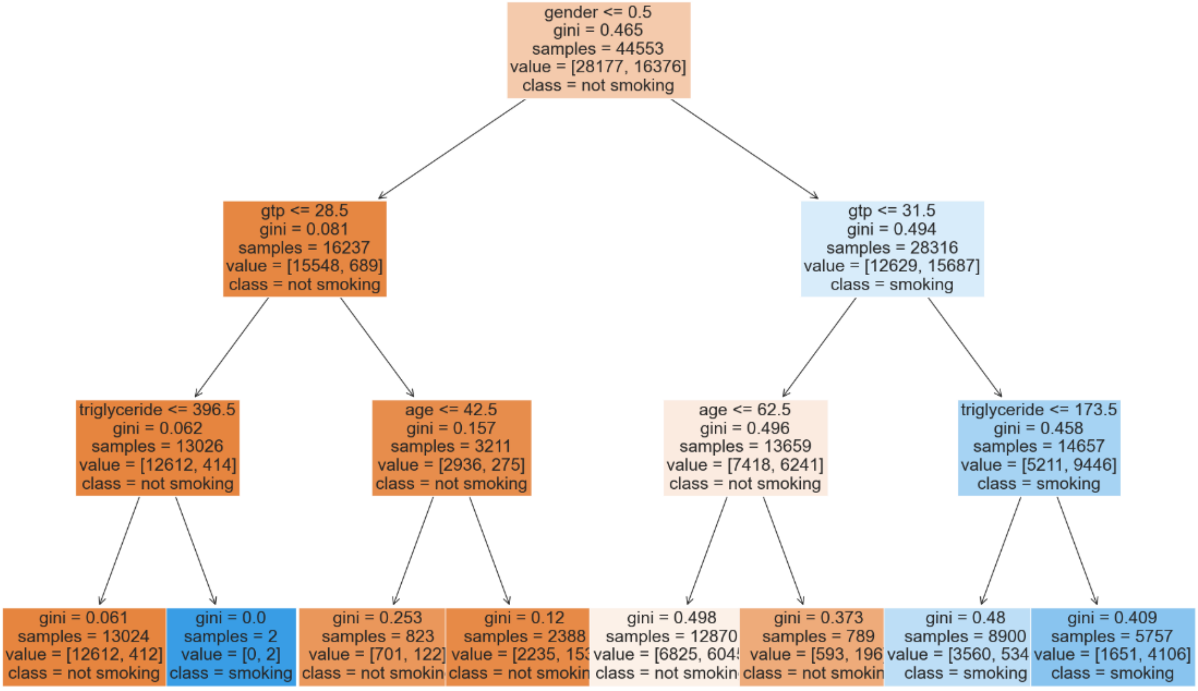
The silhouette scores of models with 7 clusters was 0.08 from data and 0.35 from PCA, which means that in the first model, there are relatively large distances between observations within one cluster and small distances between clusters. The second model seems to have more distinct groups.

In case of agglomerative clustering, the final two groups have silhouette scores 0.17 and 0.33 which is an improvement only in case of the model built from plain data. Adjusted rand index for data and PCA was approx. 0.00 for both, which means the clustering did poorly no matter what data it was fed and would be of little use in a task such as this.

Explanation

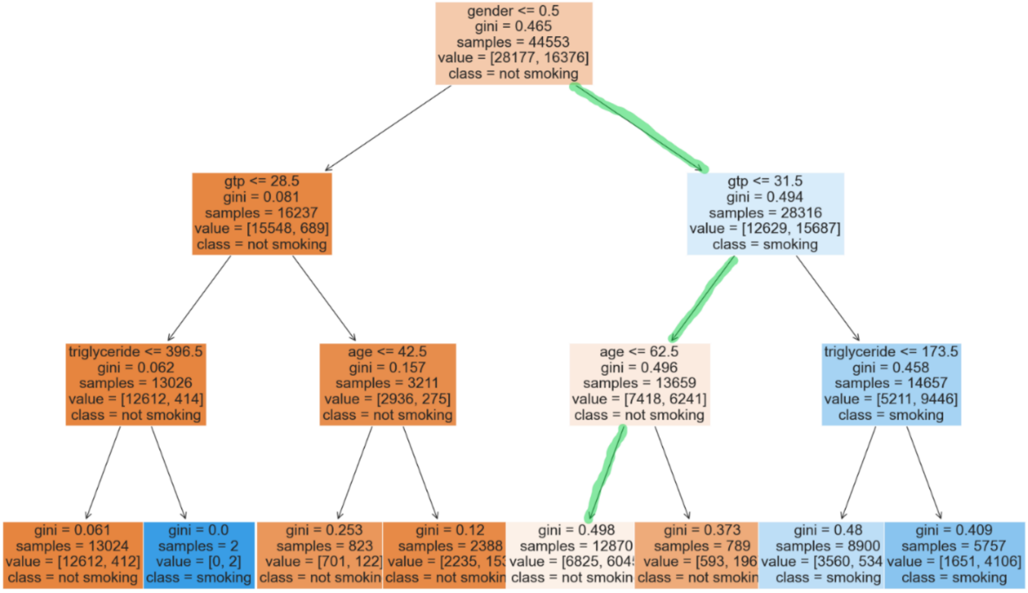
Supervised models

Visualisation of the smaller tree:



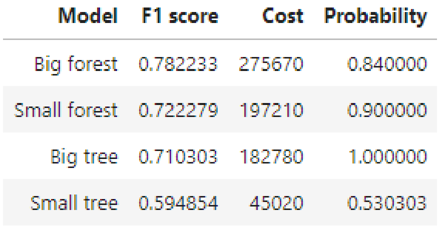
The most important variable in our model is gender, as 95 % of the smokers are men (gender = 1). Other important variables are gtp, age and triglyceride.

Our chosen instance went through following nodes:



Probability of correct prediction

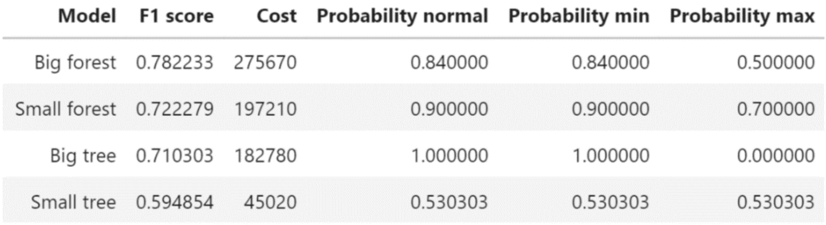
The forest models assigned the correct class with high probability, the big tree model as well, but the capabilities of the small tree are too limited because of how short it is.



Model	F1 score	Cost	Probability
Big forest	0.782233	275670	0.840000
Small forest	0.722279	197210	0.900000
Big tree	0.710303	182780	1.000000
Small tree	0.594854	45020	0.530303

Changing the attribute of interest

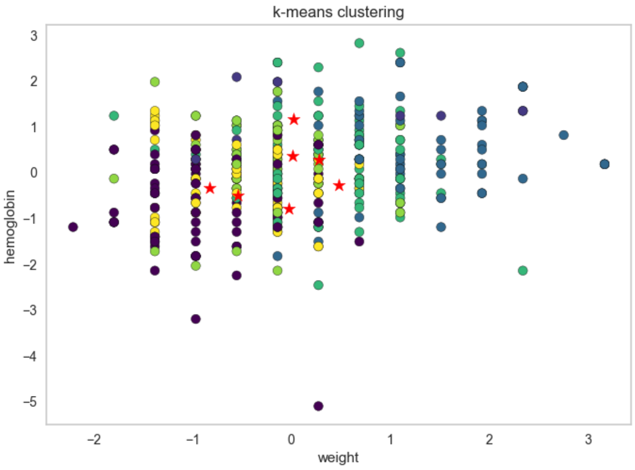
When trying to explore the effect of the attribute of interest (hemoglobin), we changed the instances' attribute to maximum and minimum values in the dataset. Originally, the value was 13.7, minimum is 6.2 and maximum is 20. Changing to the minimum did not have any impact, but bringing it to the highest value had severe impact on our models. We can see in the table below that the big forest model now has only 50% confidence and the big tree considers this person a smoker. The small tree remains unchanged as hemoglobin is not included in its very few nodes. This would suggest that people with high hemoglobin levels could be more likely to be smokers.

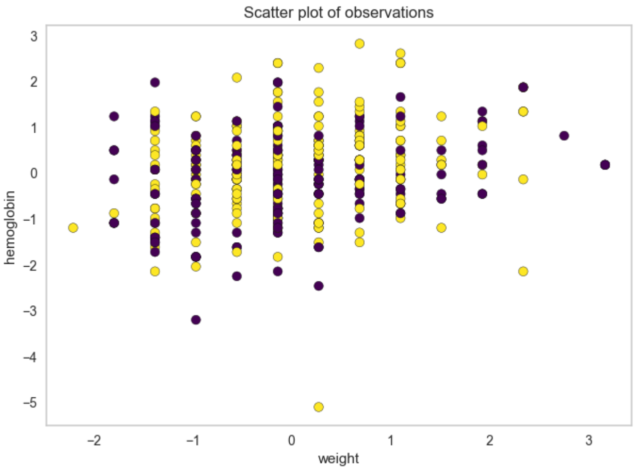


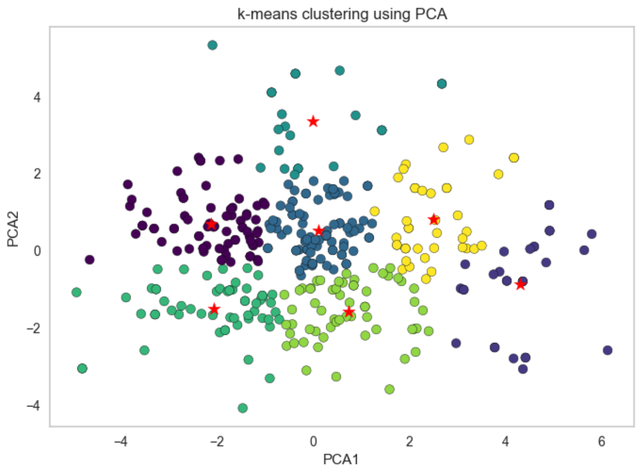
Model	F1 score	Cost	Probability normal	Probability min	Probability max
Big forest	0.782233	275670	0.840000	0.840000	0.500000
Small forest	0.722279	197210	0.900000	0.900000	0.700000
Big tree	0.710303	182780	1.000000	1.000000	0.000000
Small tree	0.594854	45020	0.530303	0.530303	0.530303

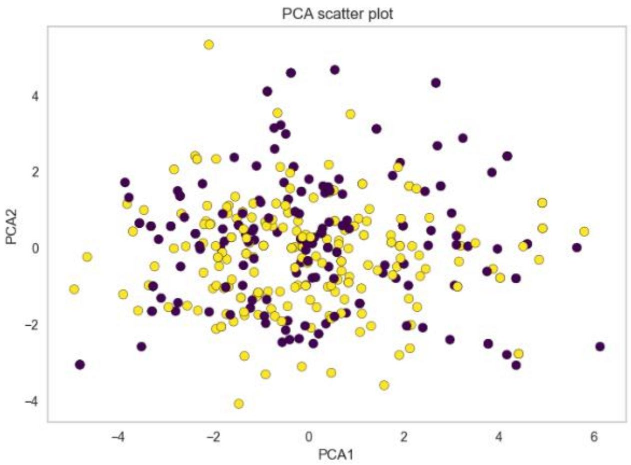
Unsupervised models

Since we are trying to find only two groups, using seven clusters is very difficult to interpret even when we know what the true labels of data are. In the following plots, we can see the comparison between the modelled data on the left with colors depending on predicted clusters, and the actual input data on the right with yellow colored smokers and purple non-smokers. Red stars again mark cluster centroids.

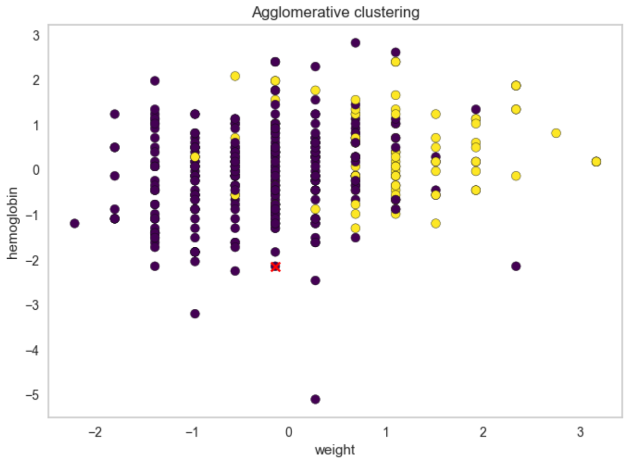


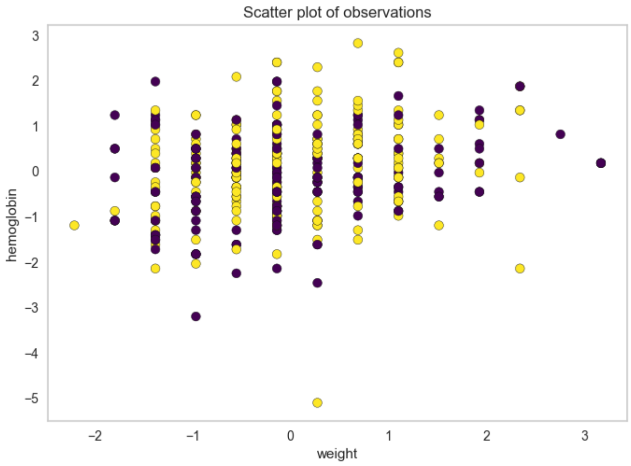


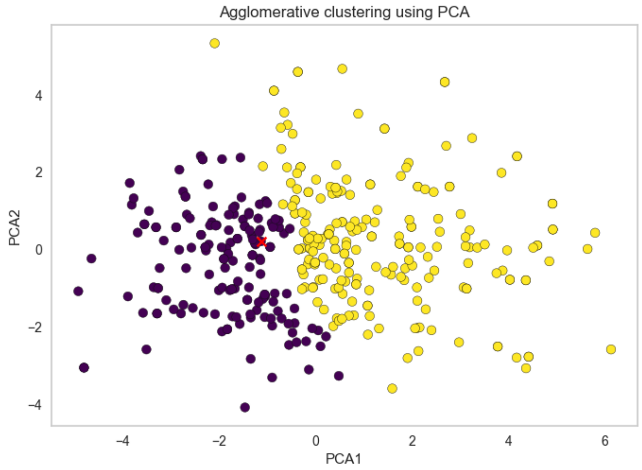

k-means clustering using PCA

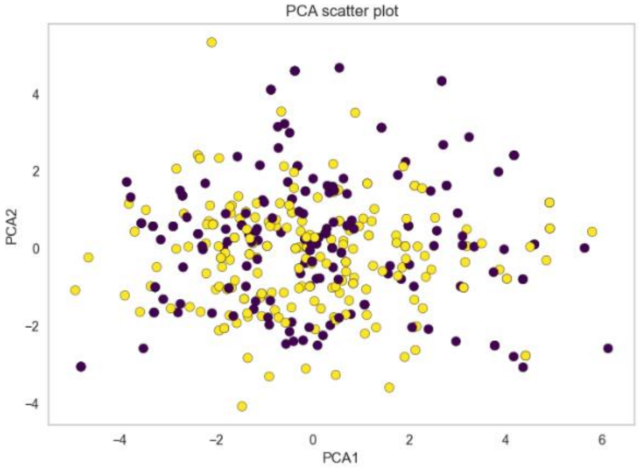

PCA scatter plot

However, if we have only two clusters, which was the case of hierarchical clustering, the clusters seem much less chaotic. Again, next pictures show models on the left and true values on the right.


Agglomerative clustering


Scatter plot of observations


Agglomerative clustering using PCA


PCA scatter plot

The true counts of each class are 200 smokers (in plots coloured yellow) and 171 non-smokers (purple). The ratio of observations in clustering based on plain data between classes is 91 to 280, while the ratio in the model using PCA is 155 to 216 and after comparing with true values, we interpret the yellow class in plots (labelled 1) in both models as smokers.

The red cross in agglomerative clustering plots symbolizes our instance of interest – in both cases classified as a smoker, which corresponds to the true value.

Conclusion

As the dataset is already heavily processed, we needed to complete only few basic pre-processing steps. We did not need to impute any data, but some columns were dropped, as they contained no information relevant to our research. Other pre-processing steps included creation of our own columns and renaming all columns to oblige with snake case notation.

Out of all the supervised models, the bigger random forest (100 trees) achieved better results than the other models, even though it only reached F1 score of 0.78. We can see that lowering number of trees in a forest and number of nodes in a tree severely impacts the predictions. However, after applying the cost matrix on the confusion matrices, it is apparent that even the weakest model still predicts well enough to make a profit.

When looking at the attributes, we can see that gender is the most important, as most of the smokers in our dataset are men. Other important attributes except for age are rather niche medical parameters such as gtp or triglyceride levels. The columns that we added, such as bmi, vision and hearing did not have a big impact on the models.

Regarding unsupervised learning, after the data exploration we had the assumption that our model is not very well suited for unsupervised models. It may be due to high dimensionality of the dataset and a lot of variability between smokers that the clusters are chaotic and do not work well. We used different methods, all of which failed to point out different types of people who smoke. After all, we decided that clustering is indeed not suitable for this task.

Column explanation

ID – identification number of observation

gender – categorical variable (M – male, F – female)

age – group age of subject – groups in 5 year gaps

height(cm) - measured height in centimetres

weight(kg) - measured weight in kilograms

waist(cm) - measured waistline in centimetres

eyesight(left)/(right) - visual acuity of said eye (values from 0.1 to 2.0, 9.9 means blindness)

hearing(left)/(right) - quality of hearing in said ear (normal hearing = 1, abnormal hearing = 2)

systolic – systolic blood pressure in torrs

relaxation –diastolic blood pressure in torrs

fasting blood sugar – measure of blood sugar after overnight fast, measured in milligrams per decilitre

Cholesterol – total cholesterol measured in milligrams per decilitre

triglyceride – triglyceride levels measured in milligrams per decilitre

HDL – HDL cholesterol levels measured in milligrams per decilitre

LDL – LDL cholesterol levels measured in milligrams per decilitre

hemoglobin – amount of haemoglobin, measured in decagrams per litre

Urine protein – categorical variable of protein in urine (values from 1 to 6)

serum creatinine – amount of serum creatinine in blood in milligrams per decilitre

AST – AST level measured in units per litre

ALT – ALT level in units per litre

Gtp – GGT levels in units per litre

oral – oral examination status (Y – yes, N – no)

dental caries – boolean variable (1 – has cavities, 0 – does not)

tartar – boolean variable (1 – has tartar, 0 - does not)

smoking – target variable, boolean (1 – smoker, 0 – non smoker)