# PREDICTING A HEART ATTACK: A MODERN STATISTICAL APPROACH

Matilde Dolfato and Pablo Manuel Ruiz

March 2024

## 1 Introduction

Cardiovascular diseases (CVD) are the main cause of death worldwide. According to the World Health Organization, 17.9 million people die each year because of these diseases. Among them, ischemic CVD like heart attacks are the first cause of death, accounting for 16% of the world's total deaths, followed by strokes that account for another 11%. [1] Given the relevance of this matter, this study aims at developing a valid model considering both simple regressions and machine learning (ML) methods, to predict whether an individual will suffer a heart attack throughout their life, in order to aid researchers and doctors at least in a primary skimming of patients.

The data is on a sample of 1319 individuals, of whom it provides information about their age, gender, and some physiological parameters related to heart diseases. First, some descriptive analysis on the data is performed, to provide some technical knowledge and reference values on the variables and try to anticipate which ones will be significant for the prediction. Then, we fit logistic regressions and find that *age*, *gender* and *kcm* are the significant regressors, and compare a simple linear model with ones with interaction and nonlinear terms. Based on the estimations of the models, some conclusions are drawn on the role of the variables and the low suitability of the logit models for our purpose, finding a high miss-classification error in all of them. Successively, we fit a random forest model, and, similarly, estimate it and assess its accuracy. In the end, it is found that the random forest with *age*, *gender* and *kcm* as predictors is our best specification, with only a 2% miss-classification error and a good degree of interpretability to be explained to doctors and other non-computer scientists.

## 2 Data

To perform this analysis, the dataset is downloaded from the Kaggle library[2] containing demographic information and observations of medical parameters on a sample of 1319 individuals. The dataset we use comprises 8 variables, of which 7 are the potential inputs and 1 is the output variable, and in particular:

- *class*, which is a dummy variable, the output variable, with value equal to 1 if the individual has had an heart attack, and 0 otherwise;

[1] https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[2] https://www.kaggle.com/datasets/bharath011/heart-disease-classification-dataset/data

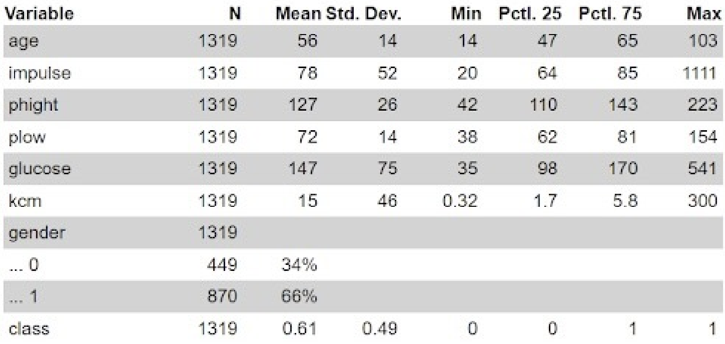
- *age*;
- *gender* (equal to 1 for males and 0 for females);
- *impulse*, which is the heart rate and is measured in beats per minute;
- *phight*, which is the systolic blood pressure, i.e. the pressure when the heart is contracting, measured in mmHg;
- *plow*, which is the diastolic blood pressure, when the heart rests between beats;
- *glucose*, the level of glucose in the blood, measured in mg/dL;
- *kcm*, which is the output of the CK-MB test measuring the level of the Creatine-Kinase enzyme (CK) , associated with damages to the heart muscle.

The original dataset contained a 9th variable, the level of troponin, which is a protein whose concentration in the human body increases a lot *while* a heart attack is happening. We decided to exclude this variable from our analysis, as it turned out to be highly correlated with the outcome variable, distorting our results, as it violates the requirement of chronological sequence of the independent variable before the dependent one[3]. In fact, it is not useful to predict the probability of having a heart disease, but it is instead used by doctors to diagnose it afterwards. [4]

# 3 Descriptive analysis

To get started, we analyse of the dataset and attempt to anticipate the degree of association between variables, and their relevance.

## 3.1 Summary statistics



| Variable | N | Mean | Std. Dev. | Min | Pctl. 25 | Pctl. 75 | Max |
|---|---|---|---|---|---|---|---|
| age | 1319 | 56 | 14 | 14 | 47 | 65 | 103 |
| impulse | 1319 | 78 | 52 | 20 | 64 | 85 | 1111 |
| phight | 1319 | 127 | 26 | 42 | 110 | 143 | 223 |
| plow | 1319 | 72 | 14 | 38 | 62 | 81 | 154 |
| glucose | 1319 | 147 | 75 | 35 | 98 | 170 | 541 |
| kcm | 1319 | 15 | 46 | 0.32 | 1.7 | 5.8 | 300 |
| gender | 1319 | | | | | | |
| ... 0 | 449 | 34% | | | | | |
| ... 1 | 870 | 66% | | | | | |
| class | 1319 | 0.61 | 0.49 | 0 | 0 | 1 | 1 |

Figure 1: Summary statistics

[3] Galit Shmueli, To Explain or to Predict?, Statistical Science, 25(3) 289-310, August 2010, https://doi.org/10.1214/10-STS330

[4] https://www.bhf.org.uk/informationsupport/heart-matters-magazine/medical/ask-the-experts/troponin

First, some summary statistics of our variables is displayed in figure 1. The data is on individuals from 14 to 103 years old, and roughly 2/3 of the sample are men. Moreover, about 60% of our individuals have had a heart attack in their lifetime.

As for the medical parameters, please notice that in general they present many extreme values, so looking at the central 50% of the data (between 0.25 and 0.75 percentiles) is useful to get an idea of these values relative to normal ones. The impulse variable central portion ranges between 64 and 85 bpm, very close to an individual's normal range (between 60 and 100 bpm), and there are very extreme values also shown by the outliers in the boxplot in figure 3. The systolic blood pressure (*plow*) is also close to normal values, which would be lower than 80 mmHg, whereas the diastolic one (*phigh*) presents high values, as it should be lower than 120 mmHg. The sampled individuals' levels of glucose are noticeably high, as normal values range between 70 and 100 mg/dL, and also the CK levels vary more than usual, as they are normally between 3 and 5%.

## 3.2 Relationship between dependent and independent variables

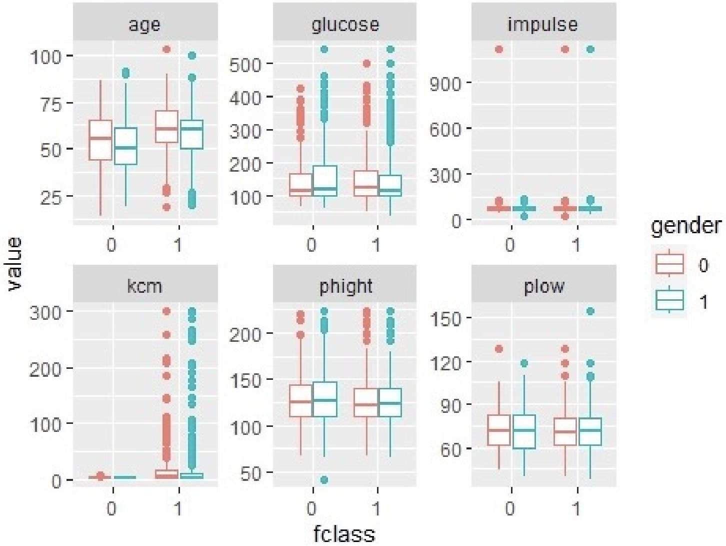


Figure 2: Boxplots of independent variables vs. outcome

The analysis goes on by visualizing the distribution of the data. More specifically, we plot each of our variables of interest against our dependent variable *class*, in order to get an idea of which predictors will be useful for our analysis. As it can be seen from the boxplots above, variables *kcm* and *impulse* present extreme values. This makes visualizing our data difficult so, to solve this, we apply a logarithmic transformation to those variables. The result is in the figure below (figure 3).

From these plots, variables *age* and *kcm* seem to be the most useful predictors of having a heart attack. Indeed, these seem to be the two variables whose distribution varies the most when we compare the group of people who suffered a heart attack with the people who didn't. The rest of the boxplots, instead, seem to be very similar across groups. While not evident at

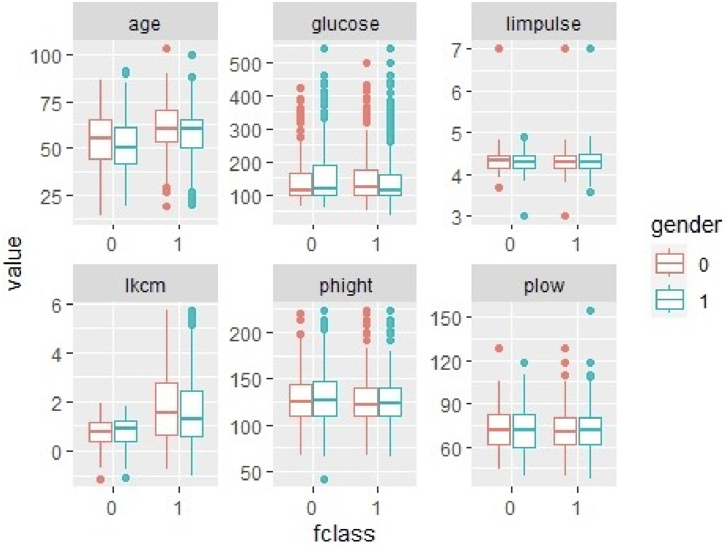


Figure 3: Boxplots of independent variables vs. outcome (with log transformations)

first sight, individuals' gender also seems to be relevant. Moreover, *gender* and *kcm* appear to be correlated, as the mean of *kcm* varies between genders, and this difference is bigger for the observations of those who had a heart attack. The rest of the variables don't seem to capture much information, possibly due to the quality of the data (as discussed more in depth in section 6.1).

## 3.3 Relationship between all the independent variables

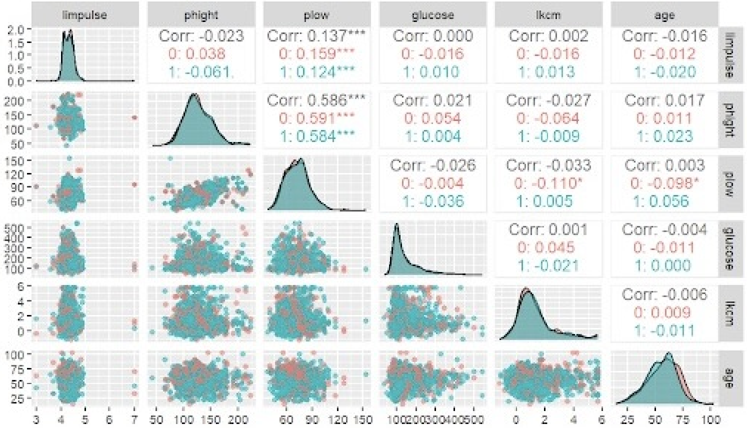


Figure 4: Pairwise relationship between the independent variables

Figure 4 below shows the relationship between all the independent variables graphically and numerically. The story this plot tells is very simple: the correlation between all the variables is very low, the only exception being the one between *plow* and *phigh*. This makes sense as, if a person has a higher blood pressure on average than another, it is intuitive that that its maximum and minimum are also higher than the other person's ones.

# 4 Methodology

The methods used in this study include boxplots, scatterplots and density plots for the descriptive analysis and model assumptions checking. Our core analysis looks at both simple generalized linear models and deep learning ones as possible models of choice. First, we consider a logistic regression, also adding interaction and non-linear terms, and then a random forest model. The subset of variables to be included in the logit models is selected with the *bestBIC* procedure, computing the Bayesian Information Criterion (BIC) for all the possible combinations of variables and choosing the model with the lowest BIC. Then, we also compare the goodness of the obtained models with BIC and Akaike information criterion (AIC). Moreover, we use Chi-squared tests, included in the *anova* analysis, to assess the significance of the additional interaction terms in our model.

We implement the k-fold cross validation procedure, using the measure of accuracy of % of miss-classification, to assess the out-of-sample accuracy of our models and choose the best one. This, in the case of the random forest, will be computed for each number of trees used by the model, as it uses bagging between many subsets of trees, with different subsets of the features.

# 5 Results

We estimate the logit and random forest models and assess their suitability, in order to get the best model to predict heart attacks.

## 5.1 Logit models

First, we consider three different logit models: a simple one with just the covariates, a second one with interactions, and a third one including non-linearities.

Using the *bestBIC* algorithm, we end up with the three models displayed in table 1:

- **Model 1**: For the model without interactions and non-linearities the most relevant variables are *age*, *lkcm* and *gender*.
- **Model 2**: For the model with interactions, the best specification seems to use *age*, *lkcm*, *gender* and the interaction between *gender* and *lkcm*.
- **Model 3**: The model with non-linearities uses the same three variables and finds nonlinear effects in the form of third degree polynomials for variables *age* and *lkcm*

Something interesting is that the values of AIC and BIC of the three models aren't actually that different, and the estimations between the first and the second model are actually very similar. Are these additional variables really significant? To answer this question we perform two maximum likelihood ratio tests between **Model 1** and the other two models. The result of the Chi-squared test indicates that the interaction term between *age* and *lkcm* (the only

| | **Model 1** | **Model 2** | **Model 3** |
|---|---|---|---|
| (Intercept) | -3.737 (0.341) | -4.963 (0.517) | 7.670 (1.151) |
| age | 0.050 (0.005) | 0.071 (0.008) | |
| lkcm | 0.916 (0.077) | 1.965 (0.329) | |
| I(gender)1 | 0.601 (0.135) | 0.604 (0.136) | 0.679 (0.144) |
| age × lkcm | | -0.019 (0.006) | |
| poly(age, 3)1 | | | 26.977 (2.842) |
| poly(age, 3)2 | | | -6.048 (2.981) |
| poly(age, 3)3 | | | -3.967 (3.328) |
| poly(lkcm, 3)1 | | | 753.074 (109.202) |
| poly(lkcm, 3)2 | | | 535.195 (80.385) |
| poly(lkcm, 3)3 | | | 167.381 (27.824) |
| AIC | 1452.9 | 1443.6 | 1314.4 |
| BIC | 1473.687 | 1469.492 | 1355.856 |

Table 1: Logit models

added variable in **Model 2**) is significant at the 99.9% confidence level[5], and the same could be said about the non-linearities. Thus, they can be considered a better fit than **Model 1**. As expected, all the regressors are significant above the 99% confidence level. In brief, the estimated coefficients in the three models all indicate that being a man, being older and having a higher level of *kcm* increase the probability of having a heart attack. Moreover, from the *age* x *lkcm* coefficient it arises that the effect the level of CK has on the probability of suffering a heart attack decreases with age. This could reflect the fact that it is more common to have high levels of CK when older, especially after 40 years old. [6]

As the actual goal of this project is to create a good enough model for prediction, table 2 below shows the level of out-of-sample accuracy of the three models.

| | **Model 1** | **Model 2** | **Model 3** |
|---|---|---|---|
| % of miss-class | 0.3108415 | 0.3040182 | 0.2835481 |

Table 2: Logit models % of miss-classification

The percentage of miss-classification is around 30%, which is really big. This implies that our additional specifications don't improve the prediction that much.

To see why this is the case, we check some of the assumptions of the logit model on **Model 1**. In the logit model, comparing the prediction with the error gives us plots difficult to analyse, thus we compute the logit errors and compare them with each of the three main variables we are studying. The results are displayed in figure 5: it tells us that there's dependence between the errors and the variables we are using. The *lkcm* graph is particularly interesting : after a certain value of *lkcm* (2), the error goes down towards 0. If the value is under (2), we tend to over or under estimate the probability of a heart attack.

[5] This can be checked by dividing the estimated coefficients by the standard deviation indicated between brackets, and seeing that it is higher than the critical value of 2.576

[6] See for reference https://pubmed.ncbi.nlm.nih.gov/7237844/

Looking at the plot of *age*, moreover, it's evident that there's a relationship between the variance of the error and the variables we are looking at. This means that there must be some variable not included in our model that have predictive power. These could be other physiological parameters that were not available in the original dataset, like the level of cholesterol, which is generally associated with heart diseases, or other individual characteristics, like weight and height, that would allow to detect obesity, also linked to CVD. [7]

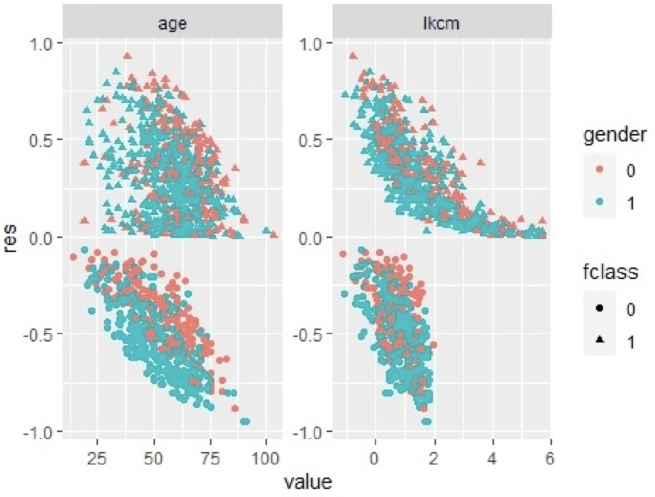


Figure 5: Residuals vs. age and lkcm plots

## 5.2 Random forest model

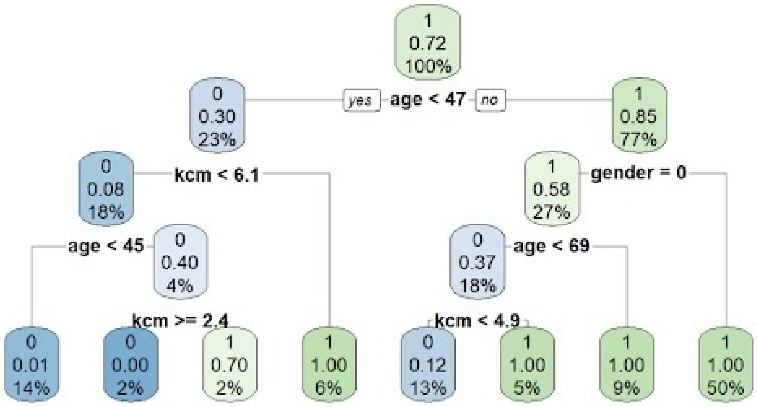


Figure 6: Tree for random forest prediction

As the miss-classification error is so high, and because we don't have the option of getting another sample, we are going to consider a different approach and use a random forest model

[7] See for reference https://www.cdc.gov/heartdisease/

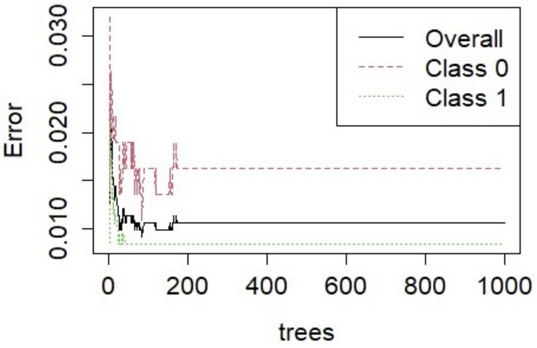


Figure 7: Random forest miss-classification plot

for the prediction. As this method doesn't assume a specific distribution of the data, we go back to consider the original variable of *kcm* rather than the log-transformed one.

When including all the variables in the dataset, it is found that using other variables apart from *kcm*, *gender* and *age* actually worsens the predictive power of our model, so we go on considering only these three and fit a random forest model. The resulting tree is shown above (figure 6). The tree, taking as output the predictions of the random forest, can be inspected to understand the decision algorithm implemented by our model and the thresholds of the features that are useful to predict the presence of a heart attack.

Interestingly, we get that this model predicts that male individuals older than 46 will have a heart attack independently of their levels of CK: this is a quite strong statement, and far from reality, but it is an inevitable one due to our model having only 3 features. For female individuals, the model predicts a heart attack for women older than 70, and instead it looks at their levels of CK if younger, considering the threshold of 4.9 (when the average *kcm* for individuals that did not have a heart attack was 2.1). Moreover, a critical level of *kcm* is 6.1, as for all individuals younger than 47, the model forecasts the occurrence of a heart attack with *kcm* higher than 6.1. A non-trivial result is that, for individuals between 45 and 47 years old, a heart attack is predicted if *kcm* is lower than 2.4, and the absence of the disease if it is higher. This is somehow in contrast with the previous result that a higher value of *kcm* generally indicates the presence of a heart attack, and it could be due to having too few observations of individuals in the narrow age range of 45-47.

Because of the randomness of the method, that involves bagging, the prediction error is given with the plot in figure 7. The prediction error is around 1%, much lower than that of the logit model. Moreover, this model seems to predict better when the individuals have a heart attack rather than when they don't have it, probably as there are more observations of individuals that had a heart attack in our sample (see figure 1). A more sophisticated method seems thus to solve for the shortcomings of our previous specification.

# 6 Conclusions

Given these results, our model of choice to predict the occurrence of a heart attack is a random forest with *age*, *gender* and *kcm* as predictors.

In fact, the random forest method provides a very good model to predict heart diseases, only with a 1% miss-classification error, as it is more complex and uses a more advanced computational algorithm compared to the logit regression, which had around a 30% error.

This indicates that given a new sample of 1000 patients, our model of choice would make a mistake for around 10 of them, and the most of these would be false positives (patients not having a heart attack, but predicted to have one), which could be considered a 'less dangerous' mistake than false negatives.

At the same time, the higher degree of complexity could make deep learning models less preferable, as these are more costly and difficult to interpret. A random forest, however, is more interpretable than other deep learning tools, like neural networks, as we have seen that the tree displayed in figure 6 allowed us to explain the decisional algorithm implemented by our model. Thus, it can be a good choice, especially in the medical field. Indeed, being able to explain to doctors the how the ML model predicted a certain output is crucial when dealing with individuals' lives and disease predictions.

Another difference with the logit model is that with a random forest we are directly providing doctors with the prediction of the presence a heart attack or not, rather than the probability of this happening. This can be, again, more practical, as in the end the aim is to either consider a patient at risk or not, but it could also be seen as losing some information. However, the accuracy of the model is more relevant, and makes this a better tool than the logit specification in any case.

## 6.1 Limitations and future perspectives

The main limitation of our model is the quality of data: we lack information on the sample selection criterias, as the individuals may be randomly chosen from a given geographical area, or they could be patients of a hospital, which would create a correlation with the probability of having heart diseases, biasing the inference of our results.

Moreover, we do not know precisely the timing of the collection of data, especially relative to the moment in which (part of) the individuals had a heart attack. Having this information would allow us to specify to a greater degree the purpose of our model and thus its usage for medical purposes, like predicting future heart attacks, or assessing the presence of heart attacks in the past.

Finally, the final model of choice uses a low amount of variables, of which only one is a medical parameter. This is good, as it provides us with a less complex method to be understood and less data to be collected, and so it allows to have an idea of the risk of a patient getting ill with a more rapid and practical method. At the same time, it could seem too simplify things too much, and finding other medical parameters that are relevant predictors could make the diagnosis performed by the model even more precise and reliable.

Consequently, our future perspectives to improve this study are to include new variables, like the level of cholesterol or indicators of obesity, that could improve the diagnosis and provide doctors with more accurate indications. Simultaneously, we want to replicate the data collection procedure in a more precise and clear way, ensuring to randomly select our sample and taking care of the timing of the measurements, and also to use more precise data, including also other features. By doing so, we aim at specifying to a greater degree the aim of our model, developing a specific one for the prediction of *future* heart attacks.

## References

[1] Mohammad Ziaul Islam Chowdhury and Tanvir C Turin (2020), Variable selection strategies and its importance in clinical prediction modelling, *Fam Med Community Health.* https://doi.org/10.1136/fmch-2019-000262

[2] Galit Shmueli (2010), To explain or to predict?, *Statistical Science*, 25(3) 289-310. https://doi.org/10.1214/10-STS330

[3] Guy Cafri, Luo Li, Elizabeth W. Paxton and Juanjuan Fan (2018), Predicting risk for adverse health events using random forest, *Journal of Applied Statistics*, 45:12, 2279-2294. https://doi.org/10.1080/02664763.2017.1414166

[4] Ramesh TR, Umesh Kumar Lilhore, Poongodi M, Sarita Simaiya, Amandeep Kaur and Mounir Hamdi (2022), Predictive analysis of heart diseases with machine learning approaches, *Malaysian Journal of Computer Science*, 132–148. https://doi.org/10.22452/mjcs.sp2022no1.10

[5] https://www.bhf.org.uk/informationsupport/heart-matters-magazine/medical/ask-the-experts/troponin

[6] https://www.cdc.gov/heartdisease/

[7] https://pubmed.ncbi.nlm.nih.gov/7237844/

[8] https://quantifyinghealth.com/report-random-forest/

[9] https://www.who.int/news-room/fact-sheets/detail/the-top-10-caus