



Université de Rennes 1 - ENSAI

Machine Learning

# Predicting diagnosis of malignant pleural mesothelioma with machine learning

*Authors:*
Romane Le Goff
Diane Maillot

*Supervisors:*
Dorothée Dolonay
Valérie Monbet

December 2021

# Abstract

Abstract will be written here.


# Contents


**1 Introduction** **2**

**2 Dataset overview** **2**

**3 Modelling / Methods** **5**

3.1 Cramer's V and Correlation Matrix . . . . . 5

3.2 Principal Component Analysis . . . . . 5

3.3 Choice of models . . . . . 5

3.4 Hyperparameter tuning . . . . . 6

**4 Results** **6**

**5 Conclusion** **6**

**6 References** **7**

# 1 Introduction

- Problematic: mesothelioma diagnosis using ML algorithms
- Use of artificial intelligence techniques for diagnosis of malignant pleural mesothelioma, Orhan Er, A. Çetin Tanrikulu, Abdurrahman Abakay, Bozok university (Turkey), 2015
- Computational prediction of diagnosis and feature selection on mesothelioma patient health records, Davida Chicco, Cristina Rovelli, 2019

MPM is a highly aggressive tumor of the serous membranes, which in humans is caused by exposure to asbestos and asbestiform fibers. It is a fatal cancer and a malignancy that is resistant to the common tumor directed therapies. Around half of people diagnosed with mesothelioma will live at least a year after the diagnosis, and around 10% of people with mesothelioma will live at least 5 years after diagnosis.

The symptoms of mesothelioma develop usually gradually over time and don't appear until several decades (typically 20 years) after exposure to asbestos For mesothelioma in the lining of the lungs, they include : chest pain, shortness of breath, fatigue, fever and sweating, couch, loss of appetite and unexplained weight loss, swollen fingertips. As for mesothelioma in the lining of the tummy, they include : swelling or tummy pain, feeling sick, loss of appetite and unexplained weight loss, diarrhoea or constipation.

Diagnostics of mesothelioma:

- X-ray of chest or tummy
- CT scan
- fluid drainage
- fluid thoracoscopy or laparoscopy

# 2 Dataset overview

The dataset used is composed of 324 real electronic health records from patients having mesothelioma symptoms in Turkey. Each record has 34 features, and there is no missing value. Diagnostic tests of every patient were recorded by an attending physician. The diagnosis of the mesothelioma disease is our target variable (*class of diagnosis*), which states if the patient is **healthy** or **unhealthy** (has mesothelioma or not). 33 other variables are considered for modelling. Out of the 324 patients, 228 were diagnosed with mesothelioma and 96 were not. In other words, 70.37% of the patients in the dataset have been diagnosed healthy by physicians despite having mesothelioma symptoms, while 29.54% are considered unhealthy and therefore ill. The challenge here is to make the difference between those two kinds of patients, which are similar in symptoms but not in illness.

The dataset owners published a first study in 2011. They put the dataset public on the UCI Machine Learning Repository afterwards, in 2016. No metadata was provided with the dataset. We hence managed to describe the variables in more details in Table 1, next page, thanks to another paper that reused the same dataset in 2019 (D. Chicco, C. Rovelli).

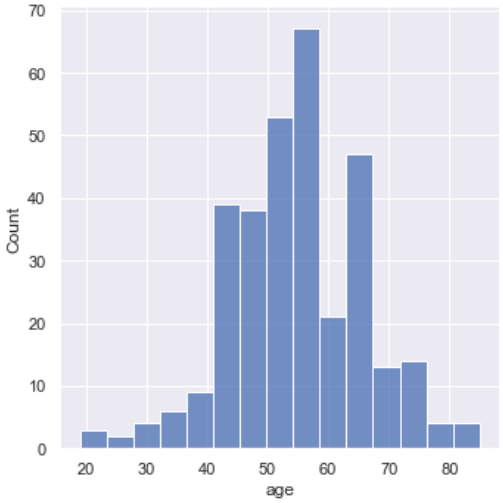


Figure 1: Age distribution of the dataset

Figure 1 highlights a distribution of the patients age centered between 40 and 60 years old.

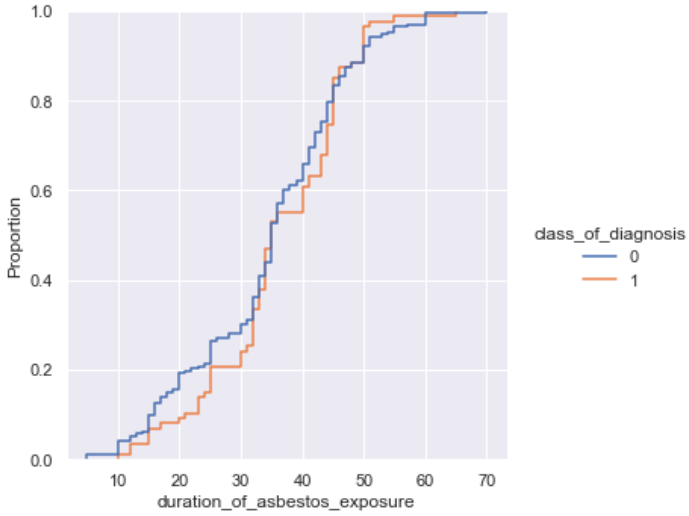


Figure 2: Duration of asbestos exposure (per class of diagnosis)

Figure 2 stresses that there is almost no difference in the duration of asbestos exposure between the healthy and unhealthy patients.

| Variable | Details |
|---|---|
| ache on chest | presence or absence of pain in the chest area |
| asbestos exposure | if a patient has been exposed to asbestos during life |
| cytology exam of pleural fluid | test to detect cancer cells in the area that surrounds the lung |
| dead or not | if the patient is still alive |
| diagnosis method | if the patient has had a mesothelioma diagnosed by a common diagnosis method |
| dyspnoea | shortness of breath |
| hemoglobin | test that measures how much hemoglobin is in blood |
| pleural effusion | presence of effusion, commun symptom that can inhibit the normal function of the organ |
| pleural level of acidity (pH) | if the pleural fluid pH is lower than the normal pleural fluid pH, that it's neutral |
| pleural thickness | any form of thickening involving either the partial or visceral pleura |
| weakness | lack of strength |
| city | pace of prevenance of the patients |
| gender | female or male |
| habit of cigarette | four categories for the habit of smoking |
| lung side | the side of the lungs which is experiencing pleural plaques or mesothelioma traces |
| performance status | patient's ability to perform normal tasks |
| type of malignant mesothelioma | mesothelioma stage to which the symptoms seem to belong |
| age | age of the patient |
| duration of asbestos exposure | how long has been the environmental exposure to asbestos |
| duration of symptoms | the time period, in year, in which the patients show symptoms |
| albumin | level of blood albumin |
| ALP | test used to help detect liver disease or bone disorders |
| CRP | acute phase reactant, significantly elevated in patients with MPM |
| glucose | test which measures the amount of glucose in a sample of blood |
| LDH | protein that helps produce energy in the body |
| PLT | test to measure how many platelets patients have in the blood |
| pleural albumin | level of albumin in the pleural fluid |
| pleural fluid WBC count | the count of leukocytes in the pleural fluid |
| pleural fluid glucose | low level can be linked to infection or malignancy |
| pleural lactic dehydrogenase | its level indicates if the fluid is exudate or transudate |
| pleural protein | fluid protein level classifies pleural effusions as transudates or exudates |
| sedimentation rate | test to measure how quickly erythrocytes settle in a test tube (in 1 hour) |
| total protein | biochemical test for measuring the total amount of protein in serum |
| white blood cells (WBC) | test measures the number and quality of WBC |

Table 1: Variables' description

The “keep side” feature was renamed "lung side" feature, as suggested by D. Chicco and C. Rovelli in their paper. The “diagnosis method” feature is strongly correlated with “class of diagnosis”. We hence removed it for classification and feature selection purposes. Of the 33 remaining features, 10 features are boolean, 14 are real values, 3 are time values, and 6 are categorical.

Other variables removed:

- *type of MM*: most of the values are 0s and only 15 are 2s or 1s (out of 96 ill patients!)
- *dead of not*: there are more healthy people that died than unhealthy people (strange)

So: 31 remaining features.

# 3 Modelling / Methods

## 3.1 Cramer’s V and Correlation Matrix

- Cramer’s V between multi-modal (categorical) variables are low
- *class of diagnosis* is not correlated with any variable
- some variables are moderately correlated
- *asbestos exposure* strongly correlated with *duration of asbestos exposure* (thinking about removing one of them)
- *pleural protein* strongly correlated with *pleural albumin*

## 3.2 Principal Component Analysis

A dataframe is built using one-hot encoding. As a first step, we scaled this dataframe and used it (without *class of diagnosis*) to do a PCA which was not conclusive. Indeed, it requires 19 components (still fewer than 31 features) to get a sum of the percentage of variance explained by each of the selected components above 70%. Perhaps trying a FAMD (for mixed data) would be more appropriate.

## 3.3 Choice of models

Our data are labelled and we have few individuals. Hence, we can apply supervised classification models to our data, that can be adapted to small samples. For now, the following models are implemented:

- Multilayer Perceptron (authors use the bayesian approach)
- SVM (not mentionned in the papers)
- Decision Tree
- Logistic Regression

- Random Forest
- AdaBoost

### 3.4 Hyperparameter tuning

Hyperparameter tuning for RF and Multilayer Perceptron are not really conclusive since we do not have enough data. The models are consequently not too sensitive to the set of parameters.

## 4 Results

- SVM does better than other classifiers in terms of accuracy (but is less interpretable than tree algorithms)
- Classification tree is not far from RF in terms of accuracy and prediction (weird since this is not what was highlighted by the second paper cited in the introduction)
- Logistic Regression computes similar results as Classification tree
- RF is good at predicting 0s
- More false negatives than false positives in every model (not good)
- Importance features (Random Forest) using SHAP

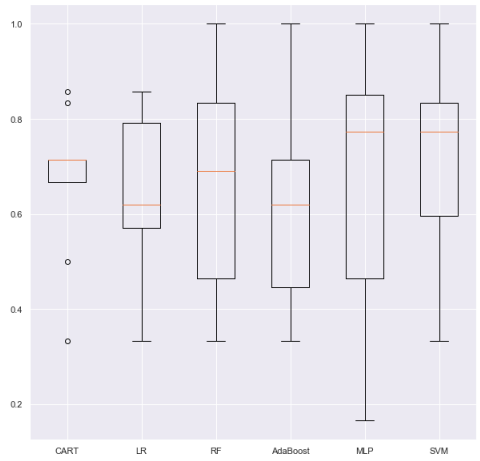


Figure 3: Algorithm comparison using cross-validation

## 5 Conclusion

# 6 References


- Mesothelioma, NHS
- SHAP: Explain Any Machine Learning Model in Python, Towards Data Science, K. Tran, 2021
- scikit-learn: Machine Learning in Python