



Université de Rennes 1 - ENSAI

Machine Learning

# Predicting diagnosis of malignant pleural mesothelioma with patient health records

*Authors:*
Romane Le Goff
Diane Maillot

*Supervisors:*
Dorothée Delaunay
Valérie Monbet

December 2021

# Abstract

Abstract will be written here.


# Contents

# 1 Introduction

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 can be done with the following technologies and tests:

- X-ray of chest or tummy;
- CT scan;
- Fluid drainage;
- Fluid thoracoscopy or laparoscopy.

It is usually difficult to scientifically differentiate healthy patients with symptoms from patients with mesothelioma. Machine learning algorithms are renowned in scientific research for tumour predictive diagnosis. Hence, to follow the steps used by the original dataset authors [1] in 2016 and by another paper published in 2019 [2], we started this study by building a multi-layer neural network. We then compared this algorithm with other machine learning models such as logistic regression, decision trees, random forests and SVM.

First of all, we will make a dataset overview and outline the particularities of the data. Then, we will detail the steps used for modelling and expose the results. Finally, we will conclude and make suggestions of improvements.

# 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 **??**, next page, thanks to another paper that reused the same dataset in 2019 (D. Chicco, C. Rovelli).

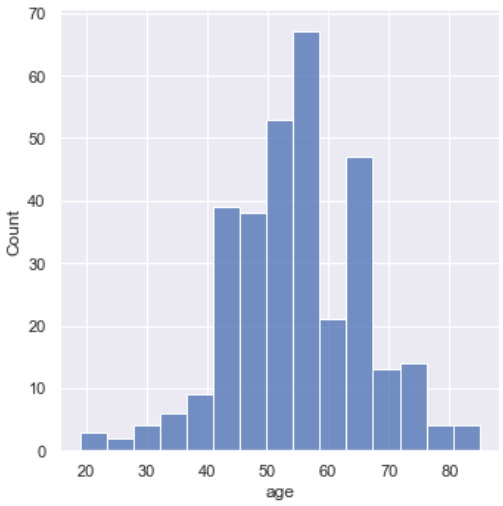



Figure 1: Age distribution of the dataset

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

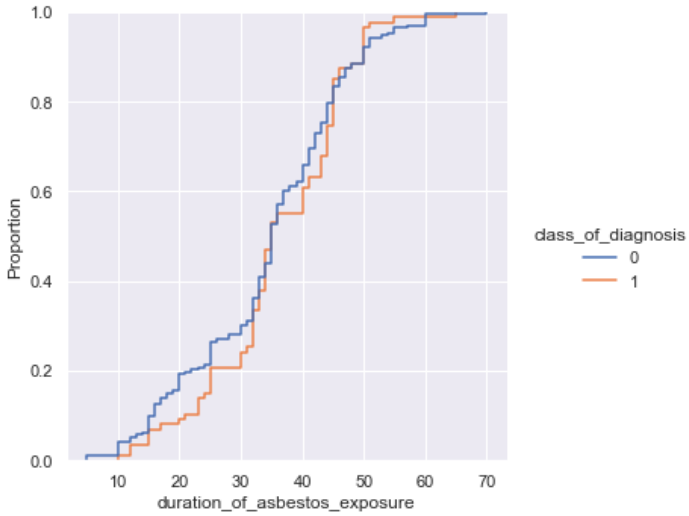



Figure 2: Duration of asbestos exposure (per class of diagnosis)

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

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.

*type of MM*'s feature was also removed since that most of the values are 0s and only 15 are 2s or 1s (out of 96 ill patients!). 32 features are remaining. We will see later that some of them should be removed for modelling.

# 3 Modelling / Methods

In the first part of this study, as a benchmark for feature selection, we computed association coefficients between the features as well as a Principal Component Analysis. Then, in the second part of this study, we built models using machine learning to make supervised binary predictions, before removing the variables that we judged not relevant. To this end, we used random forest feature selection.

## 3.1 Cramer's V and Correlation Matrix

We computed Cramer's V as well as a correlation matrix between the features of the mesothelioma dataset. Cramer's V is calculated between multi-modal (categorical) variables and results in a low association between the features. Our target variable *class of diagnosis* is not correlated with any variable whereas some dependent variables are moderately to strongly correlated. Among the most correlated : *asbestos exposure* is strongly correlated with *duration of asbestos exposure* ; *pleural protein* is 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%.

## 3.3 Choice of models

We chose to test four types of model - artificial neural network, parametric and non-parametric algorithms, and ensemble methods. The following models are implemented: Multilayer Perceptron, SVM, Logistic Regression, Decision Tree, Random Forest and XGBoost.

### 3.4 Implementation and hyperparameter tuning

We split the dataset between a training (80%) and a testing (20%) dataset before implementing the models and tuning the hyperparameters.

The Random Forest model was the preferred algorithm in Chicco and Rovelli's paper. We consequently decided to use it as a tool for feature selection. Tree algorithms usually are marginally affected by feature correlations, and therefore they are robust when applied to patient health records datasets, as in our case. We used SHAP [3] to determine the feature importance of the Random Forest and accordingly select the most relevant features. Indeed, RF is an algorithm sensitive to the choice of variables so we need to assure that we only keep the informative ones. We hence ended up with 19 variables. Among them, platelet count, age, duration of symptoms and C-reactive protein are the most informative ones, before duration of asbestos exposure.

The parameters of the best estimators - MLP, RF and SVM - used to apply these methods are optimised by cross-validated grid-search over a parameter grid. We used a stratified 10-fold cross-validation splitting strategy. However, the hyperparameter tuning is not really conclusive since we do not have enough data. The models are consequently not too sensitive to the set of parameters.

## 4 Results

The two classes of patients do not have the same probability to occur. Hence, we need to be careful about the value of the accuracy computed by each model, and look at other metrics such as the confusion matrices, the f1-scores, the Matthews Correlation Coefficients, the ROC curves and the AUC. The Matthews correlation coefficient (MCC) is an interesting statistical rate for binary classification since it produces a high score only if the prediction achieved good results in all of the four confusion matrix categories, in proportion both to the size of the positives (1s) and the size of the negatives (0s) in the dataset. It ranges in the real unit interval [0,1], and is calculated as below, where TP: True Positives, TN: True Negatives, FP: False Positives, FN: False Negatives.

$$MCC = \frac{TP.TN - FP.FN}{\sqrt{(TP + FP).(TP + FN).(TN + FP).(TN + FN)}}$$

Figure 4, next page, shows the comparison of accuracy scores between models using stratified 10-fold cross-validation. Looking as well at the MCC (Figure 5 next page) it is clear that SVM, RF and MLP does a better job that the other classifiers to predict the diagnosis of mesothelioma. However, AUC scores vary a lot (Figure 3).

| | MLP | SVM | CART | LR | RF | XGB |
|---|---|---|---|---|---|---|
| AUC | 0.66 (±0.31) | 0.58 (±0.26) | 0.63 (±0.18) | 0.54 (±0.30) | 0.57 (±0.25) | 0.37 (±0.26) |

Figure 3: Area Under the ROC Curve (AUC)

This can be explained by the very low prediction rate of unhealthy patients predicted by the models, which attains sometimes 0. There are furthermore more false negatives than false positives in every model. All the models have a tough time predicting mesothelioma patients, due to the unbalanced dataset.

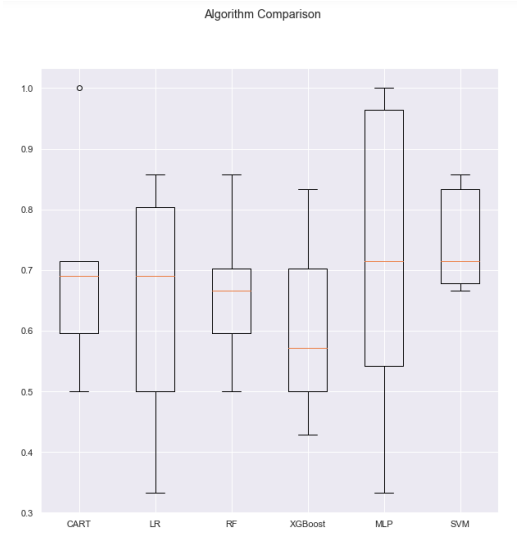



Figure 4: Algorithm comparison using cross-validation

| | MLP | SVM | CART | LR | RF | XGB |
|---|---|---|---|---|---|---|
| MCC | 0.38 | 0.36 | 0.18 | 0.08 | 0.41 | 0.34 |

Figure 5: Matthews Correlation Coefficients

Finally, performance in accuracy of MLP model is very variable (cf Figure 4). We will consequently prefer SVM and RF models - their ROC curves are visible in appendix 6. SVM seems to be the more accurate choice given all the metrics listed above. However, RF is a more interpretable model as a tree-like graph model. Combining those two methods can be a good trade-off for predicting mesothelioma.

# 5 Conclusion

As a result, this study was not as conclusive as the papers on which we relied. Nevertheless, we could determine the most relevant features to predict mesothelioma and we could make two models - SVM and RF - with high MCC scores as well as respectively good and satisfactory average accuracy scores (0.76 for SVM and 0.66 for RF). Platelet count, age, C-reactive protein rate and the duration of symptoms have a high impact on the prediction of the disease, surprisingly more than the duration of asbestos exposure. To go further, we could think of downsampling and upsampling techniques that may overcome the unbalanced dataset issue we encountered.

# References

[1] Hamza Osman Ilhan and Enes Celik. The mesothelioma disease diagnosis with artificial intelligence methods. pages 1–5, 2016.

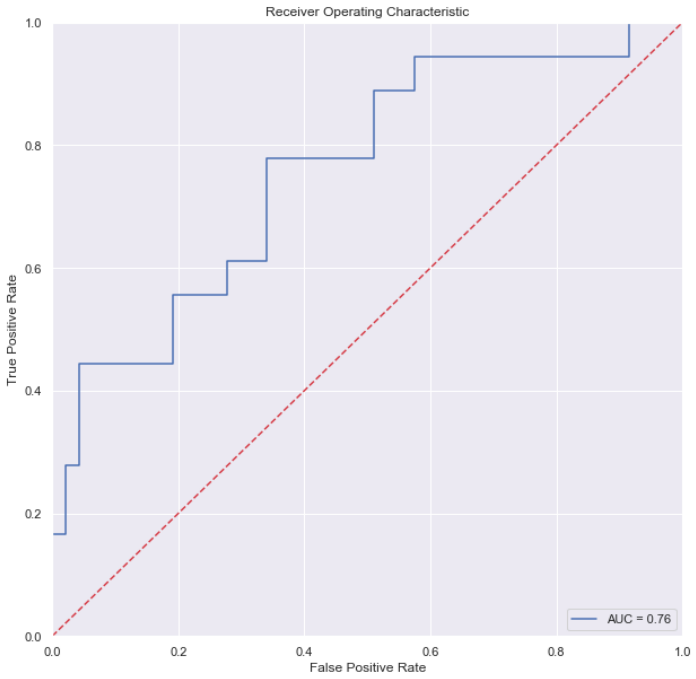
[2] Davide Chicco and Cristina Rovelli. Computational prediction of diagnosis and feature selection on mesothelioma patient health records. *PLOS ONE*, 14(1):1–28, 01 2019.

[3] Khuyen Tran. Shap: Explain any machine learning model in python. *Medium*, 24, 09 2021.

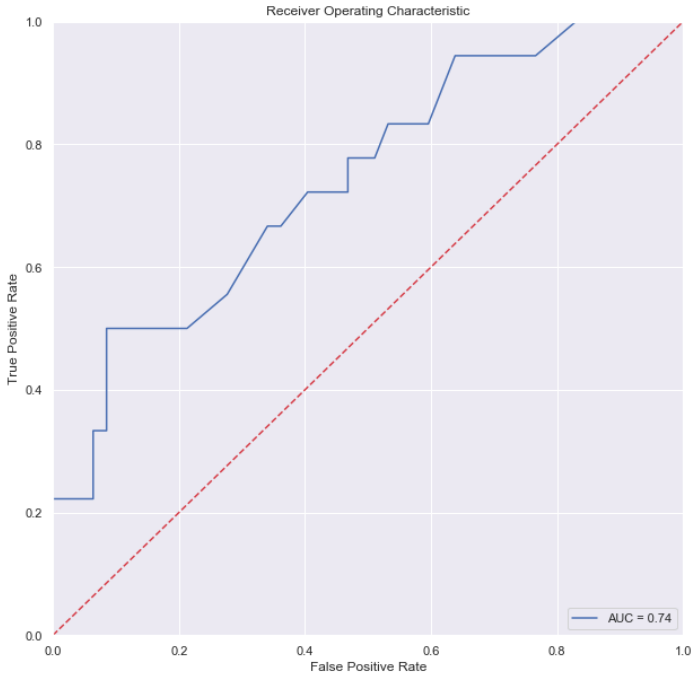
# A Description of the variables

| 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 method |
| dyspnoea | shortness of breath |
| hemoglobin | test that measures how much hemoglobin is in blood |
| pleural effusion | presence of effusion, common 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 (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 |

# B ROC curves




(a) ROC Curve of SVM model




(b) ROC Curve of RF model

Figure 6: ROC Curves of the two best models

## C F1-scores for label 1 (harmonic mean of the precision and recall of 1s predictions)

| | MLP | SVM | CART | LR | RF | XGB |
|---|---|---|---|---|---|---|
| f1-score | 0.52 | 0.29 | 0.5 | 0.36 | 0.29 | 0.42 |