Université de Rennes 1 - ENSAI

Machine Learning

Predicting diagnosis of malignant pleural mesothelioma with patient health records

Authors:
Romane Le Goff
Diane Maillot Tchofo

Supervisors:
Dorothée Delaunay
Valérie Monbet

January 2022

Abstract

The subject of this article is predicting malignant pleural mesothelioma from electronic health records of symptomatic patients in Turkey. The aim of this study is to determine the most significant features allowing diagnosis of mesothelioma and to find the best classification model for prediction. The results demonstrated the importance of eleven features, including four dummy categories, in the creation of a random forest. Among them, the side of the lungs which is experiencing pleural plaques, as well as the duration of asbestos exposure, the duration of symptoms and the quantity of C-reactive protein in the body have a significant role in mesothelioma diagnosis. Four types of models (Regression, Decision Tree, Support Vector Machine, Ensemble Methods) were applied to the data. Many performance measures, especially the Matthews Correlation Coefficient, helped to decide which model is to prefer. The Random Forest and the Multi-Layer Perceptron, among the models selected for prediction, appeared to be the best choices for modelling mesothelioma. However, the small amount of data is a constraint to ensure reliability, as the results obtained in the metrics performance are very variable and there are too many false negatives. Finally, the results obtained in this article extend the work of H. Osman and E. Celik [1], the original datasets authors, that used deep learning in their study.

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 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], this study started building a simple parametric algorithm Logistic Regression. Then, other machine learning algorithms are compared, which are decision trees, random forests, XGBoost, Support Vector Machine and Multilayer Perceptron.

First of all, in Section 2, a dataset overview is presented and the particularities of the data are higlighted. Then, in Section 3, the modelling steps are detailled. In section 4, results of the modelling are exposed. Finally, in Section 5, concluding remarks and suggestions of improvements are proposed.

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.

The dataset owners published a first study in 2011. They subsequently put the dataset public on the UCI Machine Learning Repository, in 2016.

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.

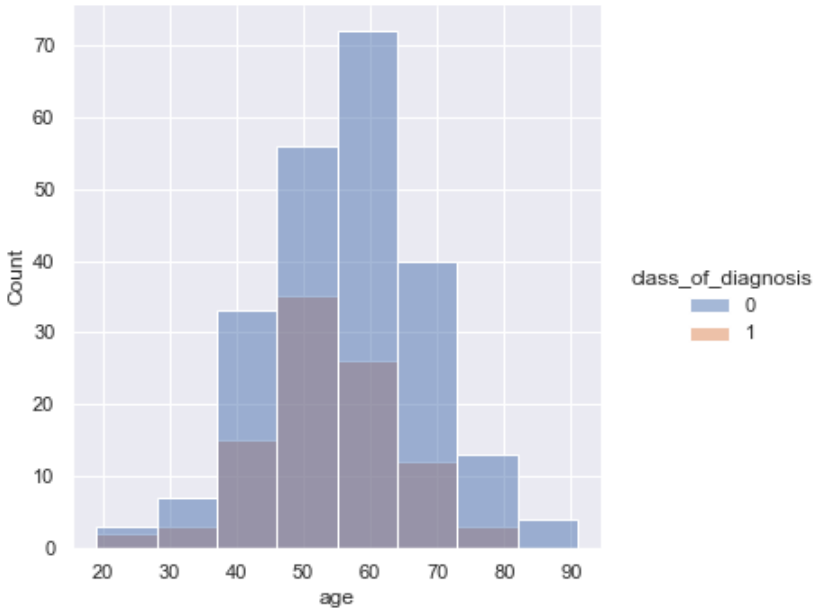


Figure 1: Age distribution of the dataset

Figure 1 highlights a distribution of the patients age centered between 40 and 70 years old. This confirms the rarity of the disease below 40. Hence, the risk of mesothelioma increases with age and diminishes from approximately 57 years old.

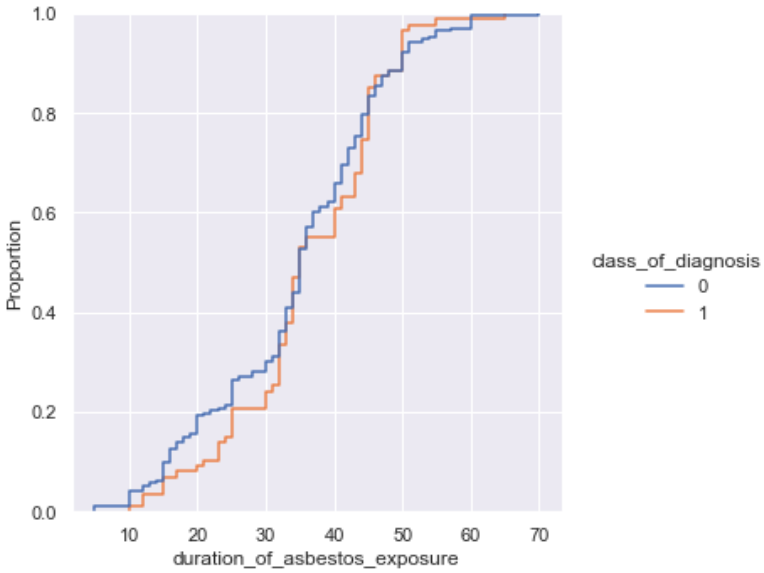


Figure 2: Duration of asbestos exposure (per class of diagnosis)

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

The "diagnosis method" feature is strongly correlated with "class of diagnosis". We hence removed it for classification and feature selection purposes; *type of MM*'s feature is also removed since that most of the values are 0s and only 15 are 2s or 1s (out of 96 ill patients). Out of the 32 remaining features, 10 features are boolean, 14 are real values, 3 are time values, and 6 are categorical.

3 Modelling

The dataset is coded with dummy variables out of the categorical variables, in order to differentiate the effect of each modality. The feature *performance status*, two modalities of *city* and two modalities of *habit of cigarette* are removed for proportion issues, as some modality did not have enough corresponding patients (which could add noise to any model).

Firstly, the dataset is split into a training (80%) and a test dataset (20%) before implementing the models inference. Secondly, in order to compare the variation in accuracy and to set the hyper-parameters, cross-validation is used.

Four types of classification model were tested - a regression algorithm, a tree algorithm, two ensemble methods and an artificial neural network. The following models are implemented: Logistic Regression (LR), Classification And Regression Tree (CART), Random Forest (RF), XGBoost, Support Vector Machine (SVM) and Multilayer Perceptron (MLP). LR is known to be good in performing for small datasets while CART has the advantage to be non-parametric as well as easy to explain. RF and XGBoost are two ensemble methods that have the advantages to integrate multiple weaker models

(trees). SVM is excellent in performing on non-linear separable problems, as in our problematic, where differences in diagnostics are unclear. MLP is tested in reference to the reference papers: [1] and [2].

In addition, because the first attempt to include all variables was not successful, a variable pre-selection method was used to reduce, for some models, the number of variables to be taken into account.

Feature importance and selection

The Random Forest model was the preferred algorithm in [2]. Indeed, 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.

SHAP (SHapley Additive exPlanations) [3] is used to determine the importance of each feature of the Random Forest. Shapley values computation is a method that tells us the importance of each feature for a given prediction. It is used for non linear models (e.g. Ensemble Methods Based on Decision Trees), as they are more complex. The estimate of the effects of each feature on the predictions is hence not straightforward. In particular, feature importance when class of diagnosis = 1 is calculated for each variable - *platelet count*, *C-reactive protein*, *blood* and *pleural lactic dehydrogenise*, *sedimentation*, *glucose*, *duration of symptoms*, *age*, *alkaline phosphatise*, *pleural glucose*, *duration of asbestos exposure*, *albumin*, *cell count*, *total protein*, *pleural protein*, *pleural albumin*, *white blood*, *lung side* are the most important ones for classifying the disease using RF.

Furthermore, permutation feature importance [4] is computed. The permutation feature importance measures the increase in model prediction error after swapping the values of the feature, breaking the relationship between the feature and the true outcome. It highlights which features contribute the most to the generalization power of the inspected model. Hence it tells how important the features are after permutation. A feature is considered important if shuffling its values increases the model error (i.e. the model relied on the feature for the prediction), and is unimportant if the model error remains unchanged. In our case, some variables do not contribute at all to the generalization power of the random forest (permutation importance equals zero). As a result, using this criteria, 11 variables (including 4 dummies) were kept for modelling (cf. Figure 4a in the Appendix). Among them, *lung side* (modality 2), *age*, *gender* (both modalities), and *C-reactive protein* are again identified as the most informative ones, before *duration of asbestos exposure*. *duration of symptoms*, *sedimentation*, *pleural lactic dehydrogenise*, *total protein* and *city* (modality 1) are also contributing significantly to the random forest's creation and are hence kept. The permutation feature importance graph after removing the unimportant variables can be seen in Figure 4b in the Appendix and stresses the importance of both the side of the lungs (modality 2) and duration of asbestos exposure for prediction.

Hyper-parameter tuning

Hyper-parameter tuning was also tried on the relevant models, i.e. MLP, RF and SVM. A cross-validated grid-search over a parameter grid was run with the help of a stratified 3-fold cross-validation as a splitting strategy. However, this was not really conclusive since the data sample was not large enough. The models are consequently not too sensitive to the set of parameters, and do not improve.

4 Results

The two classes of patients do not have the same probability to occur. Hence, it is necessary to be careful about the value of the accuracy computed by each model, and look at other metrics such as the confusion matrices, the recall, the F_1-scores, the Matthews Correlation Coefficients, the AUC. The Matthews correlation coefficient (MCC) [5] 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 refer to True Positive, TN to True Negative, FP to False Positive and FN to False Negative.

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

In order the see the variation in the metrics listed above, the dataset composed of 11 variables is used for a stratified 20-fold cross-validation (except for MLP which uses all the features). Figure 3 shows the comparison of accuracy scores with this method. Looking as well at the MCC (Table 1) it is clear that SVM and RF do a better job than the other classifiers to predict the diagnosis of mesothelioma. Nevertheless, the harmonic mean of the precision and recall, or F_1-score, and the recall both visible in Table 1 highlight the good performance of both CART and MLP models at classifying unhealthy patients, which is essential in our clinical study, as we do not want to diagnose a patient healthy when he is ill.

On the other hand, AUC scores vary a lot. This can be explained by the very low prediction rate of unhealthy patients predicted by the models, which attains sometimes 0. Due to the highly unbalanced dataset, every model has a tough time predicting mesothelioma-affected patients, indeed, there are more false negatives than false positives in every model. However, LR has clearly a higher and satisfactory AUC (0.67) on average than the other models, while XGBoost is undoubtedly a bad predictor (AUC = 0.5) and seems to compute random guesses.

	LR	CART	RF	XGB	SVM	MLP
MCC	0.25 (±0.12)	0.13 (±0.13)	**0.29 (±0.11)**	0.21 (±0.11)	0.29 (±0.12)	0.22 (±0.11)
AUC	**0.67 (±0.19)**	0.54 (±0.14)	0.58 (±0.22)	0.50 (±0.22)	0.59 (±0.20)	0.62 (±0.17)
F_1-score	0.20 (±0.31)	0.30 (±0.23)	0.29 (±0.30)	0.23 (±0.28)	0.15 (±0.33)	**0.32 (±0.20)**
Recall	0.20 (±0.32)	0.33 (±0.30)	0.24 (±0.30)	0.21 (±0.29)	0.14 (±0.32)	**0.32 (±0.23)**

Table 1: Scoring metrics: average MCC, AUC, recall and F_1-score (±std.)

In the end, looking at all performance measures, RF algorithm seems to be the best compromise to predict mesothelioma. Indeed, it has a mean F_1-score of 0.29 (very close to the highest one of 0.32 attributed to the MLP) and it has the highest MCC and accuracy scores, on average. The multilayer

perceptron could also be a good alternative to the random forest, given the high values of its recall and its F_1-score (both equal to 0.32 on average).

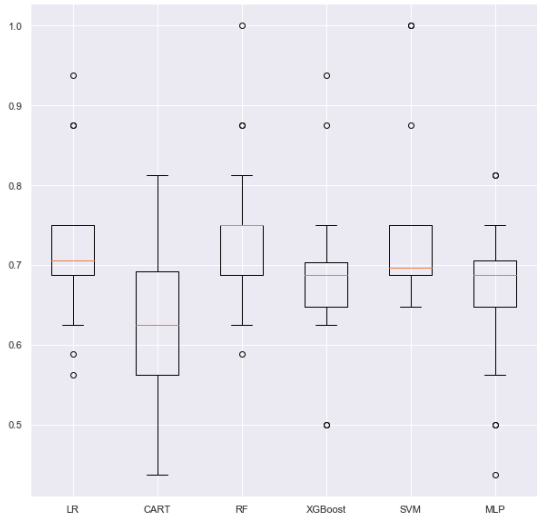


Figure 3: Algorithm accuracy comparison using cross-validation

5 Conclusion

In conclusion, we tempted to predict the diagnosis of malignant pleural mesothelioma with symptomatic Turkish patient health records using machine learning techniques. As a first step, we used feature importance and permutation feature importance to determine the most relevant features, and decided to keep 11 variables for modelling, including *lung side* (modality 2), *gender*, *age* and *C-reactive protein*. Obviously, the duration of asbestos exposure has a high impact in the prediction. Many decision criteria and performance measures allowed us to decide to keep the random forest as our preferred model. Indeed, on average it outperformed the other models in terms of MCC (0.41) and accuracy (0.74). However, this model is not very reliable as there is still a lot of false negatives. The amount of data was not sufficient to compute a decent AUC score (0.58 ± 0.22) and F_1-score (0.29 ± 0.30). MLP is a thoughtful alternative for predicting mesothelioma-affected patients, given its average recall and F_1 scores (both equal to 0.32) which are higher than every other models. To go further, we could think of downsampling and upsampling [6] techniques that may overcome the unbalanced dataset issue we encountered. However, our dataset is may be too small for that kind of solution. Another perspective could be to fit a Logistic Regression with interactions and a Lasso penalty, as it can be used as an alternative to perform variable selection in order to reduce the complexity of the model.

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] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[4] Christophe Molnar. Permutation feature importance. In *Interpretable Machine Learning, A Guide for Making Black Box Models Explainable.*, chapter 8, 5. 2021.

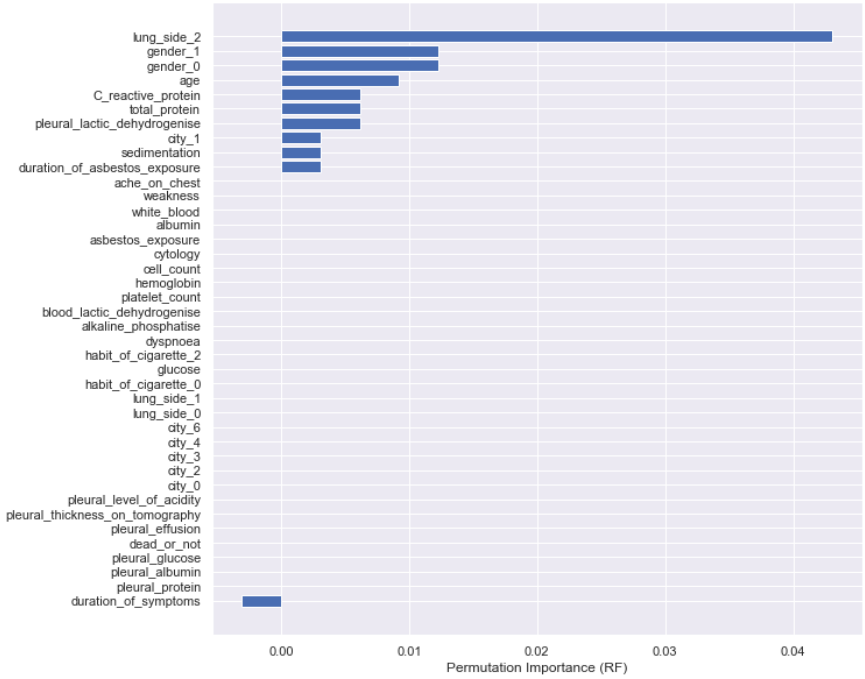
[5] Giuseppe Jurman Davide Chicco. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. 2019.

[6] Bruce H. Cottman. Balancing and augmenting structured data. In *The "paso" Project*. 2019.

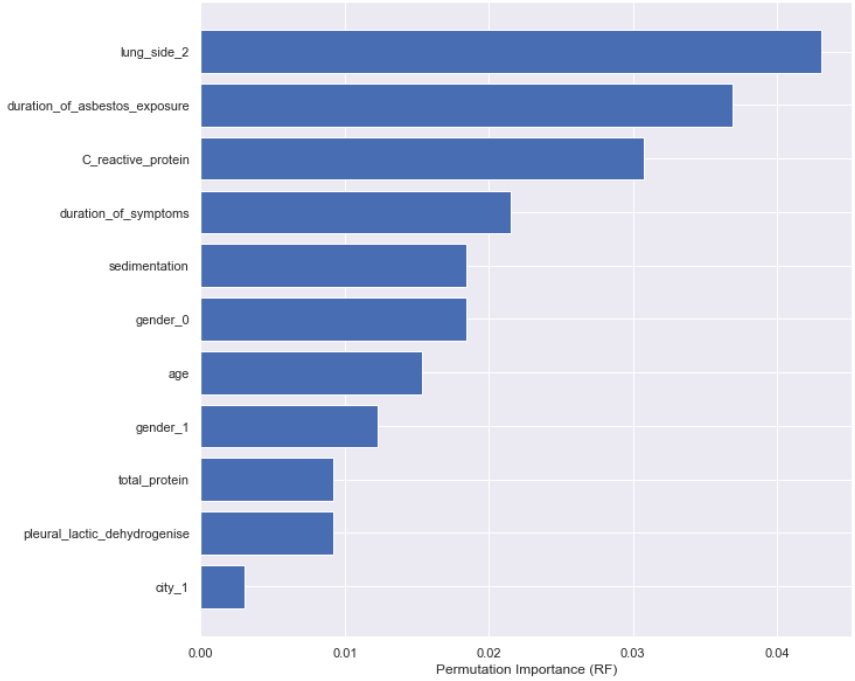
A Description of the variables

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
alkaline phosphatise	test used to help detect liver disease or bone disorders
C-Reactive Protein	acute phase reactant, significantly elevated in patients with MPM
glucose	test which measures the amount of glucose in a sample of blood
blood lactic dehydrogenise	protein that helps produce energy in the body
platelet count	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 Permutation feature importance



(a) Permutation feature importance in the RF before removing variables



(b) Permutation feature importance in the RF after removing variables