Project Report: Asbestos Classification

Aaron Zumdick, Tilman Bockhacker

Summer 2025

1 Introduction

The deconstruction and removal of asbestos-containing materials, such as special cement mixes, pose a serious health risk to construction workers if not carried out with the proper equipment. Construction companies are therefore required to follow strict safety protocols, including laboratory testing of materials prior to deconstruction, to ensure worker safety and the appropriate handling of toxic waste.

Although this procedure minimizes the risk of workers unknowingly handling asbestos without protection, it also results in long waiting times for laboratory results. A faster method for testing suspected asbestos-containing material would therefore help to optimize workflows at construction sites. Even the reliable detection of true positives alone would substantially reduce the number of required laboratory tests. Nonetheless, our aim was to develop a model that could also rule out samples and classify them as true negatives with very high confidence.

The data were collected using the *Bruker S1 Titan 800*, a handheld X-Ray fluorescence analyzer with high accuracy, capable of quantifying the concentrations of elements from magnesium to uranium in under one minute of measurement time. The cement samples originated primarily from old bridges, particularly from spacers. The asbestos-free data points were obtained from a broad range of similar cement samples, with the aim of covering the variety of materials likely to be encountered in practical applications.

All data used in the final analysis was measured by Steffen Hinerasky from the University of Münster, and was not used in a similar analysis yet.

2 Data Preparation

In this section, we briefly describe the process of preparing and cleaning the data for further analysis. The dataset was provided in an Excel table, with the concentrations of the measured elements stored in the columns. The presence or absence of asbestos was encoded as 1 and 0, respectively. At first, we implemented a helper function to correctly import the data, since the German and English decimal separators differ. Without this adjustment, some concentrations initially appeared to lie above 100%, which is, of course, impossible.

In the first preparation step, we split the data into training and test sets before performing any cleaning or augmentation. This ensured that the test set remained completely unseen during model development and could therefore be interpreted as truly new data during evaluation. We applied an 80/20 train–test split with stratification, resulting in 27 asbestos-containing and 101 asbestos-free samples in the training set.

In the second step, we removed elements for which more than 50% of the measured values in the training set were below the analyzer's detection limit, indicated as *NA* in the data table. This resulted in 22 of the original 43 elements being marked for removal. It is worth noting that most of the removed elements had *NA* percentages above 90%, with only Antimony being close to the threshold with 53.9%. For each removed element, we also examined the class distribution of the samples with non-zero values to assess whether the presence or absence of that element could serve as a predictor of asbestos content. The majority of the distributions was similar to the overall class-distribution, so we decided to proceed with the removal. The only exception here was chlorine, but since the research on the ground-truth chemistry of asbestos indicated no established scientific correlation between chlorine concentration and asbestos content in cement, we decided to remove it as well.

In the final preparation step, we examined the low-variance attributes. Our assumption was that attributes with very low variance are unlikely to provide predictive power for this problem, since an element with an almost constant concentration across all samples is not a likely candidate to distinguish asbestos-containing from asbestos-free data points. Nevertheless, we wanted to set the threshold very low, so that we do not filter out elements with a low variance that may be meaningful. We calculated the median m of the remaining elements' variances, and chose the threshold to be $0.01 * m$. For the experiment given in the notebook, this resulted in the removal of Arsenic and Yttrium, both of which had a variance less than 0.000003%.

Although the second and the third preparation steps of heavily depend on the train-test split, we reran our pipeline with different seeds and did not observe a change in the removed elements.

3 Principal Component Regression

3.1 Principal Component Analysis

After preparing the data for processing and analysis, we conducted a principal component analysis (PCA) under the following assumption: Since all data points originate from cement samples, the concentrations of the fundamental compounds unrelated to asbestos content are expected to exhibit relatively low variability. In contrast, the concentrations of elements that differ between the two classes should display greater variation.

Prior to performing the PCA, we standardized the data so that each variable had a mean of zero and a variance of one. This pre-processing step ensured that elements with larger absolute concentrations did not dominate the first principal components, allowing the analysis to focus on relative variation. We then examined the proportion of variance explained by each component, as well as pairwise scatter plots for the first ten principal components. In particular, the first and second principal components provided strong support for our assumption, as their joint plot revealed a clear separation between the two classes.

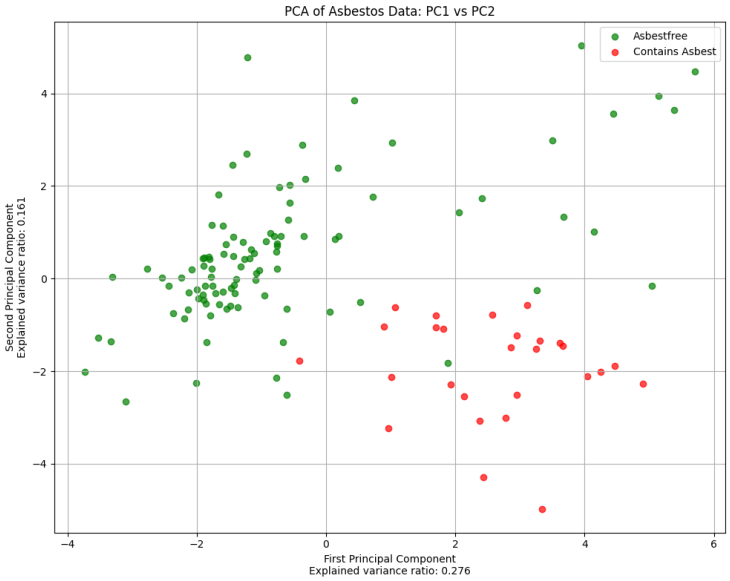

Figure 1: Plot of the first against the second principal component of the training data. The green points form a dense cluster, and the region where the first PC is greater than zero and the second PC is smaller than zero contains almost all asbestos points.

The portion of variance explained further supports the assumption: The first two principal components explain around 40% of the variance of the data, while the fourth to tenth components all have similar percentage values around 5%. In combination with the clear separation of the classes by the first two components the assumption we made in the beginning is well supported.

3.2 Weighted Logistic Regression

In the next step we fitted two different logistic regression models on the first three principal components of the data: A logistic regression with automatically balanced weights, where the weights are chosen so that both classes have the same cumulated weights, and a custom weighted logistic regression where we emphasized the asbestos class even more, by assigning five times the weight to the training points from the asbestos class. This resulted in the negative log-likelihood

$$\mathcal{L}(\beta) = -\sum_{i=1}^{n} w_i \, \ell_i(\beta), \tag{1}$$

where $\ell_i(\beta)$ is the standard per-sample log-likelihood term for logistic regression:

$$\ell_i(\beta) = y_i \log \sigma(\mathbf{x}_i^\top \beta) + (1 - y_i) \log \left(1 - \sigma(\mathbf{x}_i^\top \beta)\right) \tag{2}$$

and $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function. For the custom weighted logistic regression we chose the weights

$$w_{custom_i} = \begin{cases} 5, & \text{if } y_i = \text{asbestos}, \\ 1, & \text{if } y_i = \text{non-asbestos}. \end{cases}$$

while the automatically balanced regression resulted in the weights

$$w_{balanced_i} = \begin{cases} 101/27 = 3.\overline{704}, & \text{if } y_i = \text{asbestos}, \\ 1, & \text{if } y_i = \text{non-asbestos}. \end{cases}$$

We evaluated the models' performance using both cross-validation and a held-out test set. Across different random seeds for the train–test split, both classifiers consistently achieved high accuracy. In the worst observed case, the balanced classifier misclassified a single asbestos-containing sample as asbestos-free, whereas no configuration of the custom weighted regression produced a false negative in the test set. The number of false positives never exceeded two for either classifier, and for many random seeds, both models achieved perfect classification.

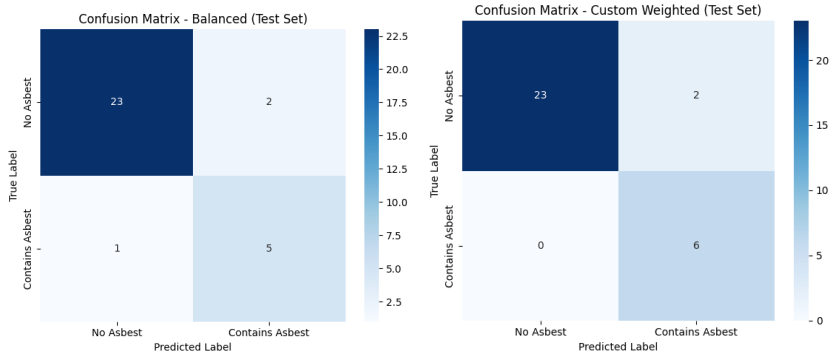

Figure 2: Confusion Matrices of the Classifiers for a suboptimal train-test split: Note that the false negative is not occurring in the custom balanced classifier, while the number of false positive remains unchanged.

While the performance of the classifiers looks satisfying at first, we need to be careful: In practice, misclassifying an asbestos-containing sample as asbestos-free could lead to severe health hazards and potential exposure of workers to the toxic material. Therefore, to develop a model suitable for reliably predicting the absence of asbestos, further refinement is necessary.

4 Constrained Gaussian

4.1 Motivation

One problem of logistic regression is that it struggles to generalize to out-of-distribution data. For example, if new data exhibit principal component values much higher or lower than any observed in the training set, the model might classify it as asbestos-free with high probability, due to the linear dependence between the predictor variables and the outcome. In practice, classifying such a sample as asbestos-free may be problematic. If there are no similar samples in the training data, the measured material may differ too greatly to draw reliable conclusions based on its components.

With this problem in mind, we sought a way to incorporate the similarity of a test sample to the training data into the model. The basic idea was to introduce a sub-classification for samples labeled as asbestos-free. If a sample lies close to the distribution of the training set, it should be classified as *almost surely asbestos-free*. If, however, it falls into a sparser region not well represented by the training data, it should instead be classified as *probably asbestos-free*.

4.2 Implementation

An approach to implementing this emerged from examining the principal component plot in Figure 2: Recall that the asbestos-free points form a dense and homogeneous cluster, that could correspond to the region one may use to classify the *almost surely asbestos-free* points. It remained to find a way of mathematically deriving this cluster and its confidence. We developed an iterative method to find this *high confidence cluster* of asbestos-free points.

Algorithm 1: Iterative constrained Gaussian

Input:

A: asbestos-free points
B: asbestos-containing points
p: probability threshold for the asbestos points
q: probability threshold for the asbestos-free ellipse

Output: Mean vector μ, covariance matrix Σ of final Gaussian.

Initialize: $X_A \leftarrow A$, $X_B \leftarrow B$, $\tau \leftarrow$ distance for confidence level p, $\pi \leftarrow$ distance for confidence level q

repeat
1. Compute μ as mean of X_A
2. Estimate Σ by maximizing likelihood of Gaussian under constraint: $\forall$ y $\in X_B : d_y \geq \tau$
3. $\forall$ x $\in X_A$, compute Mahalanobis distance d_x to $\mathcal{N}(\mu, \Sigma)$
   3a. outliers $= \emptyset$
   3b. $d_x > \pi \rightarrow$ add x to outliers
   3c. $X_A = X_A -$ outliers

until *no outliers in step 3b*;

return μ, Σ

The algorithm iteratively estimates multidimensional Gaussian distributions, subject to a constraint on how likely the asbestos-containing training samples are permitted to be under the estimated distributions and the chosen p-value. For computational convenience, this probability threshold is transformed into a value τ, which serves as a cutoff for the Mahalanobis distance between the asbestos samples and the estimated Gaussian. Since jointly optimizing both the mean and the covariance matrix is not feasible, we adopted an iterative procedure in which outliers are removed at each step and the mean is recalculated from the remaining points. The algorithm terminates once no further outliers are detected with respect to the estimated distribution and the chosen q-value, and it returns the mean and covariance matrix of the multidimensional Gaussian.

The resulting distribution has two key properties: (i) the probability that any asbestos-containing training point (or a point located even further away) is sampled from the distribution remains below the specified p-value, and (ii) the q-confidence ellipse of the Gaussian is densely populated with asbestos-free samples.

4.3 Results

We implemented and evaluated the algorithm in the same manner as the logistic regression, using the first three principal components of the data and an 80/20 train–test split. Using this reduced representation instead of the full feature set helps to avoid issues associated with high-dimensional Gaussian distributions, particularly the sparsity of data points around the mean and the curse of dimensionality. We chose a q-value of 0.95 for the asbestos-free points, and a p-value of 0.001 for the asbestos-containing points, and again repeated the experiments for different random seeds for the train-test split.

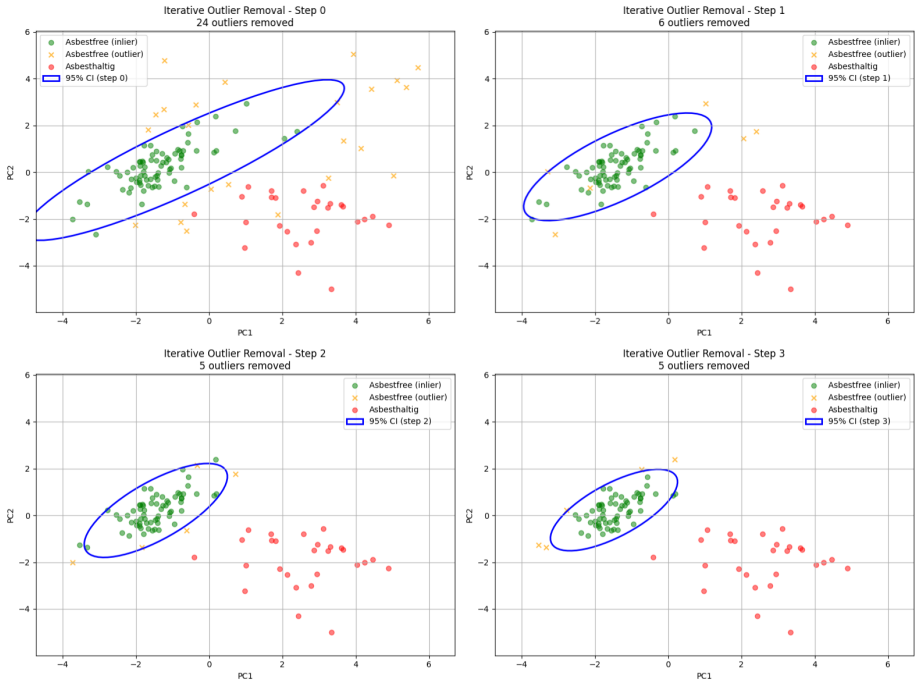

Figure 3: Visualization of the first three steps of the algorithm: The 95% confidence ellipse shrinks, while the outliers get removed.

Similar to the evaluation of the logistic classifier, we assessed this model using the held-out test data. The evaluation, however, is less straightforward than for logistic regression: the objective here is to distinguish test points that fall within regions densely populated by training samples from those located in sparse regions. Thus, we computed the principal components of the test points and plotted them together with the 95 % confidence ellipse of the final Gaussian estimated by the algorithm.

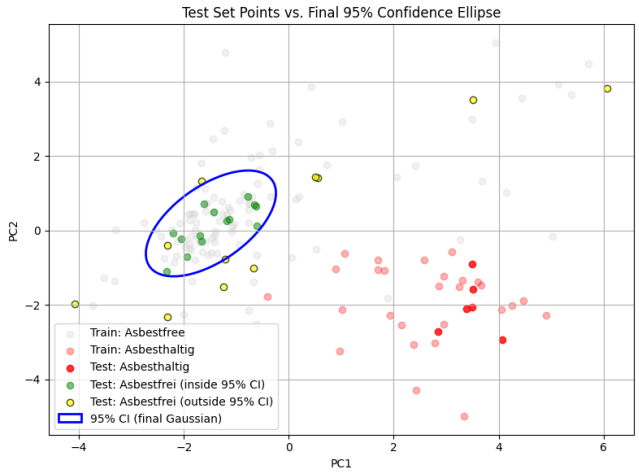

Figure 4: The final confidence ellipse and the test-points: The light green points are marked as *almost surely asbestos-free*, while the yellow points are too far away from the gaussian distribution and get marked as *probably asbestos-free*. Note that one yellow point lies inside blue ellipse, but its value in the third principal component is not inside the three-dimensional confidence ellipsoid.

When analyzing the visualization of the final results, the method seems to deliver the desired properties. The final ellipsoid is a dense cluster of asbestos-free training points, and matches the 'intuitive' asbestos-free region we located when we first inspected the PCA plot. Its well separated from the asbestos samples, and the sub-classification works as intended. The test samples that fall in a sparse region get marked as *probably asbestos-free*, and could be send to a laboratory for confirmation. A clear example is provided by the three yellow points in the third quadrant: They are too close to the asbestos-containing training point in the third quadrant to classify them as *almost surely asbestos-free*, based on the training data available.

A Gaussian Mixture Models

Since the constrained Gaussian model we constructed is closely related to a Gaussian Mixture Model (GMM), we additionally implemented two GMMs for comparison. As in the other experiments, the first three principal components of the training data were used as input. Specifically, we trained one GMM with a single component on the samples of each class, and another GMM with two components on the entire dataset.

The distributions estimated in the first approach exhibit several problems. The 95% confidence ellipsoid of the distribution estimated from the asbestos-free samples contains many data points clustered around its mean but is very thinly populated in the boundary regions. This is likely due to the influence of outliers on the estimation, which are mitigated in the constrained Gaussian approach. Another limitation is the overlap between the confidence ellipsoids: if these distributions were to be used for classification and uncertainty prediction on the test data, the probability thresholds would require careful selection and fine-tuning. This issue is also avoided in our approach.

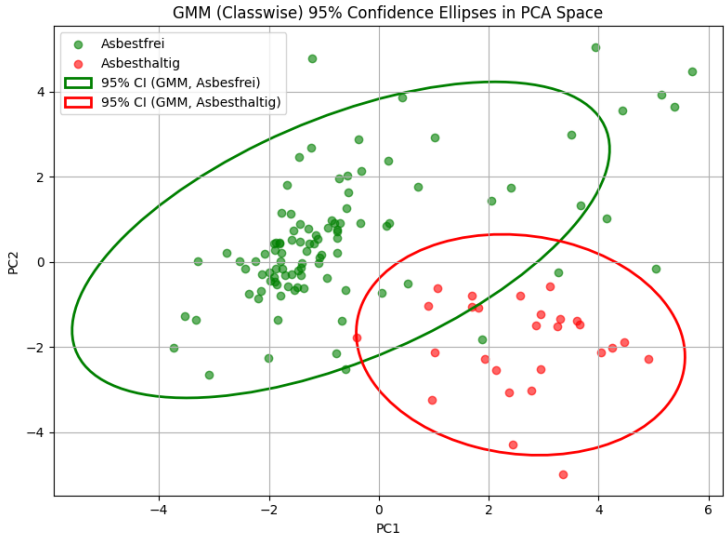

Figure 5: Confidence ellipsoids of the distributions estimated on the two classes. The asbestos-free datas' ellipsoid and the asbestos datas' ellipsoid are overlapping.

The second approach would also require refinement to be used in practice. While the first component of the GMM is just a Gaussian enclosing almost all of the data points regardless of their class affiliation, the second component of the model is covering the high-density cluster of asbestos-free points. The key difference from our model is that this distribution results from an unsupervised algorithm and therefore lacks the mathematical grounding provided by the p- and q-value constraints.

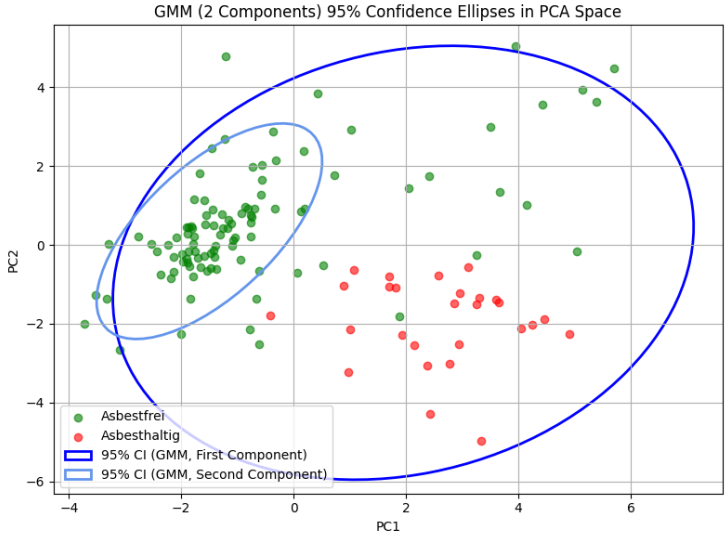

Figure 6: The 95% confidence ellipses of the estimated gaussian distributions of the GMM with two components: While the first component is covering almost all of the data points, the second component is centered around the high-density cluster of asbestos-free points.