Article

Differentiation of Multi-Parametric Groups of Groundwater Bodies through Discriminant Analysis and Machine Learning

Ismail Mohsine [1], Ilias Kacimi [1], Vincent Valles [2], Marc Leblanc [1,2], Badr El Mahrad [1,3,4], Fabrice Dassonville [5], Nadia Kassou [1], Tarik Bouramtane [1], Shiny Abraham [6], Abdessamad Touiouine [1,7], Meryem Jabrane [7], Meryem Touzani [8], Abdoul Azize Barry [9], Suzanne Yameogo [9] and Laurent Barbiero [10,*]

1   Geosciences, Water and Environment Laboratory, Faculty of Sciences Rabat, Mohammed V University, Rabat 10000, Morocco; ismail.mohsine@um5r.ac.ma (I.M.); i.kacimi@um5r.ac.ma (I.K.); marc.leblanc@univ-avignon.fr (M.L.); badr.elmahrad@gmail.com (B.E.M.); n.kassou@um5r.ac.ma (N.K.); tarik_bouramtane@um5.ac.ma (T.B.); touiouineabdessamad@yahoo.fr (A.T.)
2   Mixed Research Unit EMMAH (Environnement Méditerranéen et Modélisation des Agro-Hydrosystèmes), Hydrogeology Laboratory, Avignon University, 84916 Avignon, France; vincent.valles@univ-avignon.fr
3   Murray Foundation, Brabners LLP, Horton House, Exchange Street, Liverpool L2 3YL, UK
4   CIMA, FCT-Gambelas Campus, University of Algarve, 8005-139 Faro, Portugal
5   ARS (Provence-Alpes-Côte d'Azur Regional Health Agency), 132, Boulevard de Paris, CEDEX 03, 13331 Marseille, France; fabrice.dassonville@ars.sante.fr
6   Electrical and Computer Engineering Department, Seattle University, Seattle, WA 98122, USA; abrahash@seattleu.edu
7   Laboratoire de Géosciences, Faculté des Sciences, Université Ibn Tofaïl, BP 133, Kénitra 14000, Morocco; meryem1jabrane@gmail.com
8   National Institute of Agronomic Research, Rabat, Morocco; meryem.touzani@inra.ma
9   Geoscience and Environment Laboratory, (LaGE), Department of Earth Sciences, Joseph KI-ZERBO University, Ouagadougou 7021, Burkina Faso; abdoul-azize.barry@alumni.univ-avignon.fr (A.A.B.); suzanneyameogo@yahoo.fr (S.Y.)
10 Institut de Recherche pour le Développement, Géoscience Environnement Toulouse, CNRS, University of Toulouse, Observatoire Midi-Pyrénées, UMR 5563, 14 Avenue Edouard Belin, 31400 Toulouse, France
*   Correspondence: laurent.barbiero@get.omp.eu

Abstract: In order to facilitate the monitoring of groundwater quality in France, the groundwater bodies (GWB) in the Provence-Alpes-Côte d'Azur region have been grouped into 11 homogeneous clusters on the basis of their physico-chemical and bacteriological characteristics. This study aims to test the legitimacy of this grouping by predicting whether water samples belong to a given sampling point, GWB or group of GWBs. To this end, 8673 observations and 18 parameters were extracted from the Size-Eaux database, and this dataset was processed using discriminant analysis and various machine learning algorithms. The results indicate an accuracy of 67% using linear discriminant analysis and 69 to 83% using ML algorithms, while quadratic discriminant analysis underperforms in comparison, yielding a less accurate prediction of 59%. The importance of each parameter in the prediction was assessed using an approach combining recursive feature elimination (RFE) techniques and random forest feature importance (RFFI). Major ions show high spatial range and play the main role in discrimination, while trace elements and bacteriological parameters of high local and/or temporal variability only play a minor role. The disparity of the results according to the characteristics of the GWB groups (geography, altitude, lithology, etc.) is discussed. Validating the grouping of GWBs will enable monitoring and surveillance strategies to be redirected on the basis of fewer, homogeneous hydrogeological units, in order to optimize sustainable management of the resource by the health agencies.

Keywords: groundwater bodies; machine learning; discriminant analysis; chemical composition; bacteriological composition; PACA region; France

1. Introduction

Our societies are putting increasing pressure on water resources for a variety of agricultural and industrial uses, but because they are less susceptible than surface water to various forms of pollution, groundwater resources are an essential source for supplying people with drinking water. The chemical and bacteriological composition and quality of groundwater depend on factors such as geology [1], climate [2,3], land use [4] and human activities [5,6]. A better understanding of the processes involved in the spatial and temporal variability of water quality is a prerequisite for sustainable management of the resource. In 2000, under the Water Framework Directive (WFD) [7–9], the European Union encouraged the mapping of groundwater bodies (GWBs) on the basis of major European river basins. A GWB is defined as a distinct volume of groundwater within one or more aquifers, a specific portion of groundwater within a larger hydrogeological system, typically delineated based on major hydrogeological characteristics, including depth, rock nature, flow type, karstic nature, riverine or coastal proximity, free, confined, etc. [10]. The mapping of these GWBs has stimulated considerable research efforts in the EU Member States [11–15]. Groundwater bodies have thus been considered as management units for the national implementation of the WFD. These units may consist of one, several or only part of an aquifer, and may be superimposed. However, because of the diversity of criteria used to designate these units (lithology, vertical and horizontal scale, catchment potential depending on yield), they are described in a heterogeneous manner [16].

On the basis of physico-chemical and bacteriological criteria in large databases, recent studies have proposed methods for homogeneously grouping these GWBs, using dimensionality reduction techniques such as principal component analysis (PCA), a statistical technique that brings together redundant information from different parameters and classifies sources of variability within the dataset [17]. Clustering methods, such as agglomerative hierarchical clustering (AHC) seeks to group similar data points into clusters based on their characteristics. These methods have been successfully applied in various administrative regions of France [18,19] with the aim of facilitating quality monitoring for human consumption for regional health agencies. Grouping into homogeneous groundwater bodies enables the processes responsible for physico-chemical and bacteriological variability to be better characterized [20]. The loss of information during the aggregation process from the catchment scale to that of the GWB, and then to that of a group of GWBs, has been quantified, and remains low compared with the total information initially contained in the datasets [21]. Nevertheless, despite the prevailing scarcity of applications on larger spatial scales, particularly within the scope of extensive European watersheds as suggested by the WFD (Danube, Rhine, Rhône-Méditerrannée, Po, Loire, Seine, Adour-Garonne, etc.) that encompass numerous groundwater bodies, this kind of applications should emerge as an alternative strategy for most state water agencies, which primarily adopt a sampling point as their main monitoring scale strategy. We should note that a recent work carried out in the Occitanie region, straddling the Rhône-Méditerranée and Adour-Garonne basins, has shown the relevance of separating these two basins when processing data [22]. Further work focused on the possibility of discriminating temporal and spatial variability within each group of GWBs, highlighting that each (spatial and temporal variability) varied according to the parameter considered. However, the proportion of each groundwater quality parameter within the GWB grouping has not been assessed, which impacts our overall understanding of the characteristics of these groundwater bodies. This research should be seen as a further step in this field.

The introduction of methods such as discriminant analysis, and more recently machine learning, has made a considerable contribution to the assessment and management (monitoring and surveillance) of groundwater resources [23]. Progress includes innovative applications of geographic information systems and statistical methods, improving contaminant management in particular [24,25]. The adoption of multivariate analysis and machine learning techniques, in particular ensemble learning [26,27], has improved the accuracy and efficiency of groundwater quality assessments, setting new benchmarks for robust and

accurate classification in diverse regions. Furthermore, examining the impacts of climate change on groundwater levels using advanced techniques such as geodesic observations and neural networks has revealed complex relationships between environmental changes and groundwater dynamics [28]. In addition, the integration of real-time data, stable water isotopes and microbial community analysis has significantly advanced monitoring methodologies beyond traditional practices [29]. These collective contributions mark an evolution in hydrogeology towards more nuanced, data-driven approaches to understanding and managing groundwater resources. These methods could be suitable for predicting whether a given water sample belongs to a GWB or a group of GWBs on the basis of its chemical and microbiological composition, and thus validating the procedure developed for grouping homogeneous GWBs, a procedure designed to optimize, simplify and reduce the cost of quality monitoring and surveillance by the regional health agencies. Dealing with multiple parameters, such as water quality in this case, requires the use of advanced statistical methods, such as linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) [30], which allow a thorough examination of linearity and non-linearity within the dataset. This analysis is necessary to guide the selection of machine learning methods suited to the complex and varied nature of the information contained in the dataset. A multi-faceted approach was then employed, combining ensemble methods [31,32] capable of handling both linear and non-linear relationships, here in a variety of hydrogeological contexts. Decision-based algorithms [33] were used for their accuracy in interpreting linear relationships, in order to disentangle simple environmental correlations. Neural networks [34] were deployed for their ability to model complex, non-linear interactions. Gradient enhancement methods, because of their versatility and high accuracy, were used to navigate efficiently through the various linearities of the regional data. This analytical arsenal has been reinforced by proximity-based models and probabilistic approaches [35], adapting to the linear and non-linear dynamic characteristics of the environmental data. Given the spatio-temporal variations in water quality at this regional scale, the accuracy of these various methods needed to be rigorously examined to ensure a complete and reliable analysis.

In this context, this study was carried out in the PACA region (south-east France) using a previously exploited database [21]. The aims of the study were twofold. The first was to identify the critical parameters for water quality classification at three distinct scales, namely that of observation points (sampling point), that of groundwater bodies (GWBs) and that of groups of GWBs. The second aim was to validate the grouping of GWBs by predicting the classification of water samples using discriminant analysis and Machine Learning methods.

2. Materials and Methods

2.1. Study Area

The study area is the administrative region of Provence-Alpes-Côte d'Azur (PACA), covering an area of 31,400 km^2 in the extreme south-east of France [20]. It has a diverse geological landscape, comprising coastal and high-altitude crystalline massifs, alluvial plains, a vast Jurassic and Cretaceous sedimentary formation composed of limestone marls with karst development and other sedimentary formations (Figure 1). In terms of topography, altitudes range from sea level to 3143 metres in the Mercantour crystalline massif. This wide range of altitudes is reflected in a wide diversity of natural environments and agricultural activities. For more details on the study site, readers can refer to our previous works on this region [17].
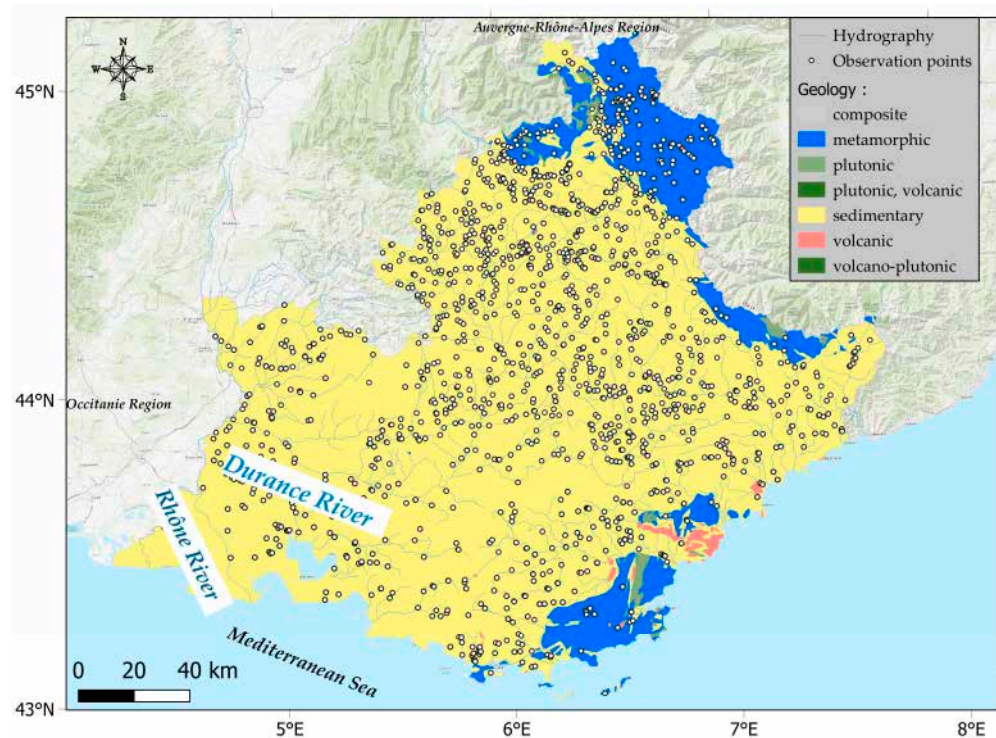
Figure 1. Study area (Provence—Alpes—Côte d'Azur region) in south-eastern France and location of the groundwater observation points.

2.2. SISE EAUX Database and Preliminary Processing

The SISE EAUX database [36,37] supplied by the Regional Health Agencies as part of health monitoring, contains analyses of water intended for human consumption. It includes both raw, untreated water analysis data and data on water after treatment. For the purposes of this study, only data relating to raw water was taken into account. These geo-referenced and dated data include a wide range of quality parameters. Database extraction over 15 years (2006–2020) resulted in 9121 analyses and 24 parameters. Sorting of the data produced a full matrix comprising 8673 analyses on 18 parameters, namely:

- Classical physico-chemical parameters (Electrical conductivity at 25 °C, pH, Total Dissolved Solids);
- Major ions (Ca^{2+}, Mg^{2+}, Na^+, K^+, HCO_3^-, SO_4^{2-}, Cl^-, NO_3^-) resulting from water-rock interactions and urban/agricultural pollution;
- Bacteriological parameters (*Enterococcus*, *Escherichia coli*), major indicators of faecal contamination.
- Trace elements such as metallic contaminants (Fe, Mn) sensitive to redox conditions, metalloids (As, B), and fluorine F.

This complete matrix comprises 1765 sampling points, equivalent to an average of 4.9 water analyses per observation point. In this way, the potential temporal variability of water quality at each point is taken into account. The observation points in the SISE-Eaux database were assigned to a groundwater body by cross-referencing with the French reference system for groundwater bodies (https://www.sandre.eaufrance.fr/jeu-de-donnees/param%C3%A8tres-sise-eau, accessed on 17 March 2022), on the basis of their geographical coordinates and the depth of the catchment. These 1765 sampling points are distributed in 63 GWB throughout the PACA region. The map showing the distribution of sampling points is presented in Figure 1.

This dataset was the subject of a previous study in order to obtain a grouping of homogeneous groundwater bodies, characterized by similar characteristics and similar processes responsible for the variability in chemical and bacteriological composition [21]. The procedure was as follows: Kolmogorov–Smirnov normality tests were carried out to

determine whether the data had a normal distribution; following this test, a logarithmic conditioning of the data was performed according to the formula: $y = \log(x + DL)$, x being the value of the physico-chemical or bacteriological parameter X, and DL the detection limit of this same parameter; the log-transformation was applied to all the data (except pH which is already on a logarithmic scale), with two objectives, i.e., to approximate a normal distribution and to reduce the weight of extreme values, which, during analysis, could mask or blur certain processes responsible for the variability of the water, thereby making the analysis more delicate [18].

A principal component analysis (PCA) was performed on all the data and the centroid coordinates of each GWB on the factorial axes (CPs) were calculated. The factorial axes (CPs) accounting for 85% of the total variance were retained, considering that the remaining 15% corresponded mainly to statistical background noise [38]. A hierarchical clustering was performed to group the GWBs into homogeneous sets [39]. It resulted in 11 GWB groups presented in Figure 2.
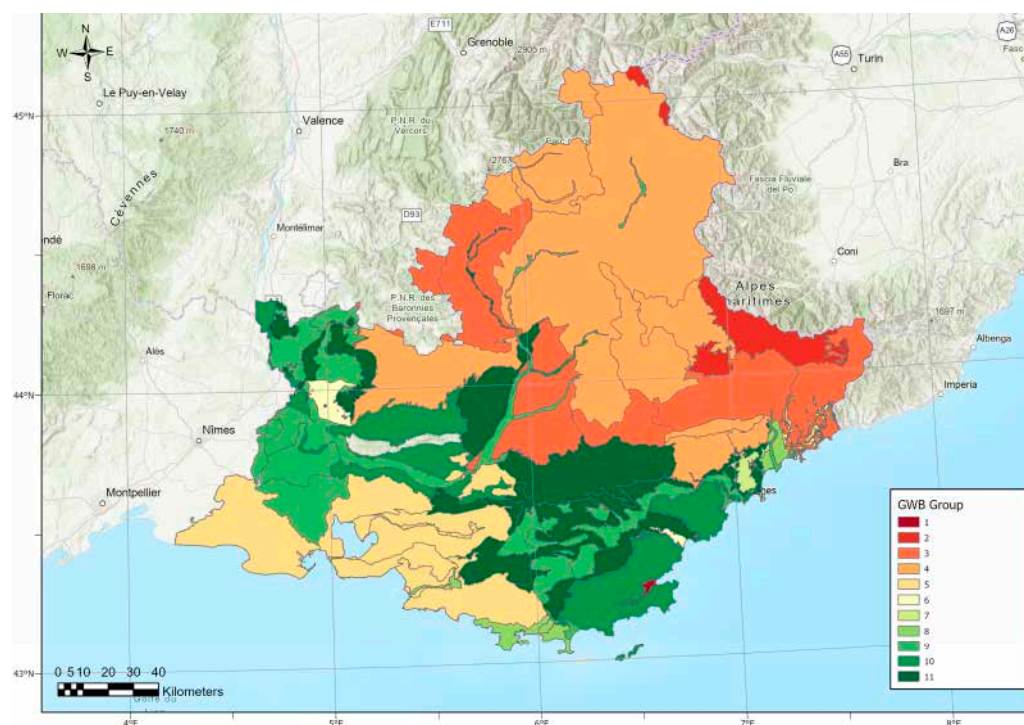
Figure 2. Distribution map of the 11 groups of groundwater bodies (GWB) in the PACA region, France (modified from [21]).

2.3. Statistical and Machine Learning Methods

Discriminant analysis and various machine learning algorithms were used to predict the membership of a given water sample to an observation point (sampling point), a groundwater body or a group of groundwater bodies on the basis of chemical and bacteriological composition. At each scale, the classification of each observation was considered as the dependent variable and various groundwater quality parameters as explanatory variables.

2.3.1. Discriminant Analysis Method

Discriminant analysis (DA) was performed to determine which variables best discriminate between groups or classes based on a set of parameters [40]. DA creates a combination of parameters that maximizes the separation between groups. It then uses this combination to make predictions of belonging of new observations to a given class. DA has been widely used for the water quality assessment [41–44] because it can handle both continuous and categorical variables. Here, discriminant function analysis were used to identify the most

significant parameters affecting groundwater characteristics and their contribution to GWB classification [45,46].

Two types of discriminant analysis methods have been used: linear discriminant analysis (LDA), which produces a linear boundary between classes and assumes that the data follow a normal distribution, and quadratic discriminant analysis (QDA), which produces a quadratic decision boundary and assumes that each class has its own covariance matrix [30]. In the realm of discriminant analysis, both linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) find their roots in the foundational principles of Bayes' theorem. The theorem, expressed as:

$$\forall k \in [1, K] : \; P(Y = K | X = x) = \frac{P(X = x | Y = k) \cdot P(Y = k)}{P(X = x)}. \tag{1}$$

This equation underpins the probabilistic approach to discerning class memberships, with: $P(X = x | Y=K)$ is the likelihood, modelled as a multivariate Gaussian distribution; $P(Y = K)$ is the prior probability of class, it is equal to π_k which is the estimated prior probability of class k; and $P(X = x)$ is the marginal probability of observing features x across all classes. The crux of discriminant analysis lies in the likelihood function, $P(X = x | Y = K)$, which quantifies the probability of observing features x given class K. Within a Gaussian context, this likelihood is mathematically framed as:

$$\forall k \in [1, K] : \; P(X = x | Y = K) \propto e^{-\frac{1}{2}(x - \mu_k)^t \Sigma_k^{-1} (x - \mu_k)}, \tag{2}$$

which incorporates the mean vector $\mu_k \in \mathcal{R}^d$ and covariance matrix Σ_k. The discriminant functions $\delta k(x)$ emerge as pivotal tools for classifying observations; for LDA, assuming a common covariance matrix Σ, the discriminant function takes the form:

$$\forall k \in [1, K] : \; \delta_k(x_i) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + log(\pi_k). \tag{3}$$

For QDA, conversely, accommodating distinct covariance matrices adopts the discriminant function:

$$\delta_k(x_i) = -\frac{1}{2} log(|\Sigma_k|) - \frac{1}{2} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) + log(\pi_k), \tag{4}$$

where x_i is an unknown measurement vector for a sample i.

To facilitate optimal classification, the log-ratio of posterior to prior probabilities is deployed in scikit-learn's implementation of discriminant analysis. It is implicitly used through the decision function or predict log-proba methods to facilitate decision-making in the classification process. For each class $k \in [1,K]$, this is expressed as follows:

$$\delta_k(x_i) = log\left(\frac{P(Y = k | X = x_i)}{P(Y = j | X = x_i)} \right), \tag{5}$$

Which encapsulates the logarithmic odds ratio of the posterior probability of class k to the prior probability of each of $j \in [1,k]$ classes.

While the discriminant function is used for classification, the log-likelihood function is part of the probabilistic framework. In the scikit-learn implementation, in addition to Equations (3) and (4) which are the original forms of the LDA and QDA, the log-likelihood plays a pivotal role in the decision-making process for both LDA and QDA. The log-likelihood for class k, denoted as $log(P(Y = k | X = x))$ is expressed differently for LDA and QDA:

For LDA:

$$log(P(Y = k | X = x)) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + log(\pi_k) + Cst, \tag{6}$$

Here, Σ is a shared covariance matrix among all classes. For QDA:

$$log(P(Y = k|X = x)) = -\tfrac{1}{2}log(|\Sigma_k|) - \tfrac{1}{2}(x - \mu_k)^T\Sigma_k^{-1}(x - \mu_k) + log(\pi_k) + Cst \tag{7}$$

Here, each class k has its own covariance matrix Σ_k, and $log(|\Sigma_k|)$ is the natural logarithm of determinant of variance-covariance matrix Σ_k. These mathematical formulations extend across k distinct classes, accommodating the intricacies of various groundwater bodies in our specific case. LDA and QDA are limited in their ability to handle high-dimensional datasets, with performance degrading as the number of parameters increases [47].

2.3.2. Machine Learning Methods

A selection of ML algorithms was used to perform multi-class classification, choosing the algorithm based on the characteristics of the problem and the available data. These algorithms belong to different categories: tree-based, linear, non-parametric, kernel-based, probabilistic, deep learning and ensemble methods. Tree-based models such as decision trees [33] partition the feature space to predict the values of target variables, and despite their potential for over-fitting, this can be mitigated through pruning or ensemble methods. Linear models such as logistic regression [48] assume a linear correlation between the input features and the target variable and, although computationally efficient, can be hampered by complex non-linear relationships. Non-parametric models, such as K-nearest neighbours [49], estimate the relationship between the input features and the target variable without any a priori assumptions about the functional form of the relationship, allowing them to capture complex patterns in the data. Kernel models, notably the support vector machine [50] and Kernel SVM [51], are able to identify complex non-linear correlations through transformation of input features into a high-dimensional space. Probabilistic models, such as Gaussian Naive Bayes and Bernoulli Naive Bayes [35], compute the probability of the target variable with respect to the input features, providing insight into the uncertainty of the prediction. Deep learning models, including neural networks (multilayer perceptron) [34], use complex functions for predictions, allowing intricate data patterns to be captured. Finally, ensemble methods combine multiple prediction models to enhance accuracy and robustness by reducing the variance and bias of individual models. Examples of these include Random Forest, XGBoost [32], LightGBM [52] and Subspace KNN [53].

The weight of the various parameters in determining the prediction of whether observations belong to their group of GWB, GWB and sampling point, with maximum accuracy, was evaluated using a hybrid approach [54] combining, on the one hand, recursive feature elimination (RFE) techniques, i.e., is a systematic and iterative feature selection technique in machine learning that progressively removes the least important features from a dataset, until an optimal accuracy is obtained [55] and on the other hand, the random forest feature importance (RFFI) estimated from the average decrease in the Gini impurity [56] for each node of the decision tree. The Gini impurity of a node can be defined as follows:

$$Gini\ impurity\ (i) = 1 - \sum_{j=1}^{C}[p(j|i)]^2 \ , \tag{8}$$

where $p(j|i)$ denotes the probability of an instance in node i being classified as class j, and C refers to the number of classes in the dataset. The average decrease in the Gini impurity can be defined as:

$$\Delta Gini(k) = \sum_i Gini(i) - Gini(f), \tag{9}$$

where *Gini(i)* is the Gini impurity of a node before the split, and *Gini(f)* is the weighted sum of the Gini impurities of the two daughter nodes created by the split.

3. Results

3.1. Discriminant Analysis

The first three discriminant functions (LD1, LD2 and LD3) accounted for 56%, 24% and 10% of the inertia (the part of the overall differences in groundwater characteristics that is explained), respectively (Figure 3), totalling 90% of the discrimination of the 11 groups of groundwater bodies.
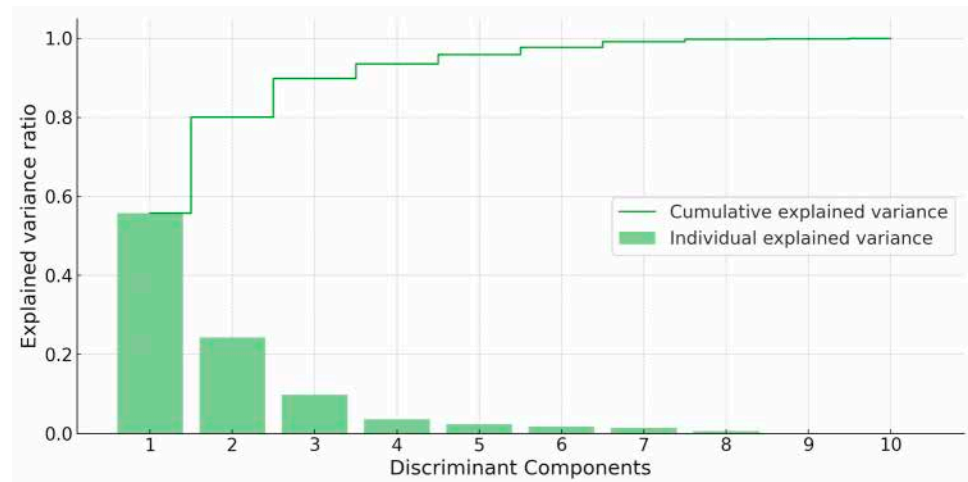
Figure 3. Scree-Plot of the discriminant analysis performed on 8673 observations and 18 parameters.

Figure 4 illustrates the relative contribution of various water quality parameters to discrimination between groups. The orientation and magnitude of the parameter vectors indicate their influence on the discriminant functions. In the first discriminant plane (LD1-LD2), total dissolved solids (TDS), Ca, Cl and pH play a major role in the discrimination. More specifically, it is the differential between TDS, on the one hand, and Ca, Cl and pH on the other, which accounts for almost (80%) of the discrimination. Trace elements, such as As, and bacteriological indicators, although important for assessing water quality, have less discriminating power due to their greater local and temporal variability, as reflected by their shorter vectors.

(a)

(b)
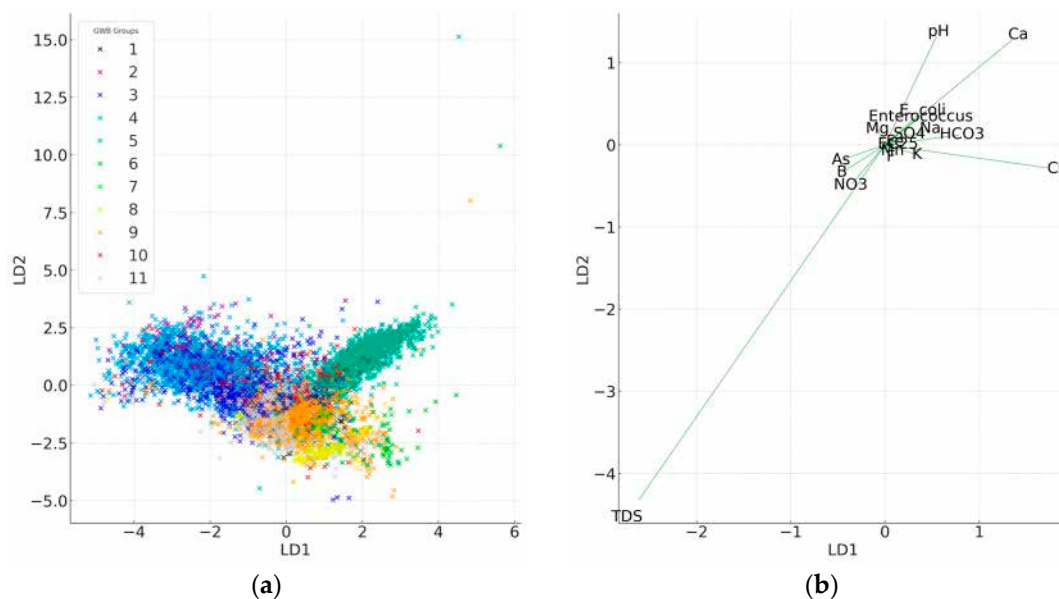
Figure 4. (**a**) Sample coordinates in LD1 and LD2; (**b**) LDA feature importance of each parameter in LD1 and LD2.

Graphs of the discriminant functions are shown in Figure 5. For the sake of readability, four parameters were selected, namely *Escherichia coli* characterising bacterial contamination, total dissolved solids (TDS), i.e., approximately the sum of the major ions, iron (Fe), a common metallic contaminant sensitive to redox processes, and nitrates (NO_3), an indicator of anthropogenic pollution. The bivariate plots involving TDS showed fairly uniform zones occupied by the classes. No class occupied more than 45% of the bivariate region, i.e., the space between the minimum and maximum value of each parameter. Evaluation of these pairs of parameters highlights varying degrees of differentiation. For the pair *E. coli* vs. TDS, the decision limits were well defined, suggesting a fairly good degree of differentiation between GWB groups. In contrast, for the pairs *E. coli* vs. Fe and TDS vs. Fe, the decision limits appeared to be very close to each other, supporting a limited potential for discrimination. Concerning the pair *E. coli* vs. NO_3, a more obvious separation between the groups was observed, whereas overlaps persisted for the pair TDS vs. NO_3. Finally, the decision boundaries for the pair Fe vs. NO_3 were very close to each other, suggesting that it is very difficult to achieve a clear differentiation of the groups along this axis.
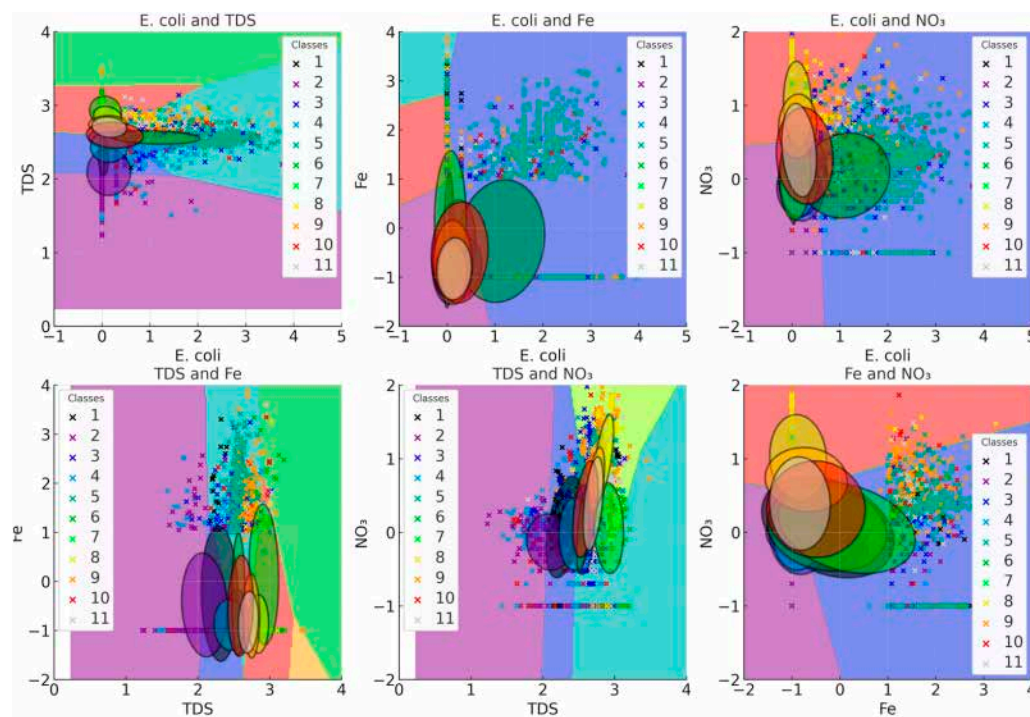
Figure 5. Graphical representation of the discrimination functions of each GWB group by linear discriminant analysis for 4 pairs of selected parameters (*E. coli*, TDS, Fe, NO_3). On the background of each biplot, the colour palette delimits the boundaries of the discriminant functions, i.e., the decision limits separating the different classes. Ellipses represent covariance ellipsoids for each group (log units).

The confusion matrix for the LDA, based on 1765 water samples that correspond to 20% of the dataset, is presented in Table 1. About 67% of the samples were correctly classified but disparities were observed between groups. Groups 5, 6 and 7 had a high rate of good classification (89%, 83% and 100%, respectively), while GWB groups 8, 10 and 11 were the least well discriminated, partially overlapping mainly with GWB group 9. GWB groups 4 and 5, which corresponded to medium mountain areas and accounted for more than half of the water analyses, achieved correct classification rates of about 70% and 89%, respectively. Results of the LDA performed not on groups but on GWB showed a lower rate of well classified, around 38%.

Table 1. Confusion matrix for the LDA test sample. The last column is the percentage of well classified samples.

	GWB Groups											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**%**
1	6	0	0	0	0	0	0	0	2	0	0	75
2	2	12	1	6	2	0	0	0	0	0	0	52.2
3	0	4	70	53	9	0	0	0	13	1	15	42.4
4	0	32	45	307	18	1	0	0	24	4	8	69.9
5	1	0	0	0	446	2	0	1	33	4	15	88.8
6	0	0	0	0	0	10	0	0	2	0	0	83.3
7	0	0	0	0	0	0	2	0	0	0	0	100
8	0	0	0	2	3	1	0	12	15	0	5	31.6
9	1	1	3	2	10	10	2	5	215	1	36	75.2
10	4	2	2	5	19	1	0	0	26	12	11	14.6
11	0	0	11	27	9	3	0	5	54	4	65	36.5

3.2. QDA Results

Quadratic discriminant analysis (QDA) is likely to take into account more complex relationships between parameters for group differentiation (Figure 6). For the *Escherichia coli* vs. Total dissolved solids pair, complex decision boundaries emerged, but the ability to distinguish GWB groups based on these parameters remained moderate. Similarly, for the *Escherichia coli* vs. iron, *Escherichia coli* vs. nitrate, and Iron vs. nitrate pairs, the complex decision boundaries suggested that the quadratic nature of QDA may offer a better fit and potential differentiation between groups compared to LDA. Furthermore, for the TDS vs. iron pair, the LDA showed high degrees of overlap between groups, and the QDA decision boundaries showed greater flexibility. The confusion matrix is shown in Table 2.

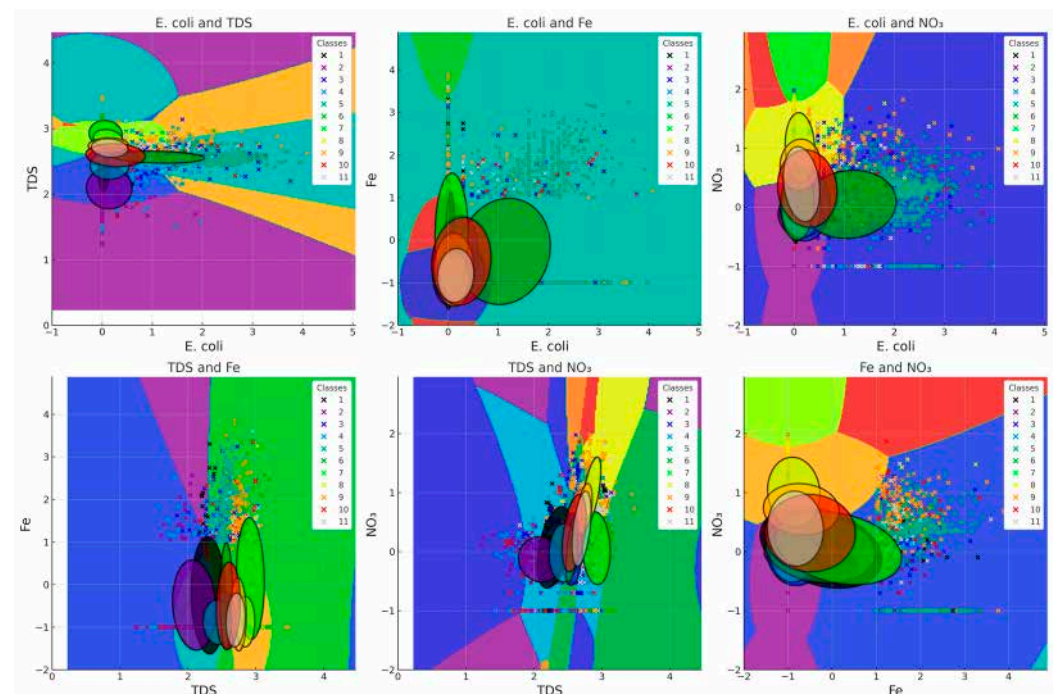**Figure 6.** Graphs of QDA discrimination functions between GWB groups for 4 pairs of selected parameters (*E. coli.*, TDS, Fe, NO$_3$) (see Figure 5 for details of representation).

Table 2. Confusion matrix for the QDA test sample. The last column is the percentage of well classified samples.

					GWB Groups							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**%**
1	2	0	0	0	0	0	0	1	1	4	0	25
2	0	16	2	3	0	0	0	0	0	2	0	70
3	0	3	122	14	7	0	0	2	11	4	2	74
4	0	23	173	185	3	0	0	2	36	7	10	42
5	1	0	1	3	455	0	0	10	24	6	2	91
6	0	0	0	0	0	9	0	0	3	0	0	75
7	0	0	0	0	2	0	0	0	0	0	0	0
8	0	0	2	0	3	0	0	27	5	1	0	71
9	0	1	10	2	30	3	0	55	172	6	7	60
10	0	2	5	2	28	1	0	1	18	21	4	26
11	0	0	32	11	7	3	0	28	76	8	13	7

QDA's ability to potentially capture more complex patterns within the data could enable better differentiation of GWB groups. However, group differentiation remained difficult to estimate from a 2-dimensional projection of an 18-dimensional space. Compared with the LDA, there was a clear improvement for groups 2, 3, 5 and 8. The two well-ranked samples in group 7, on the other hand, were confused with group 5, reducing the success rate for this group from 100% to 0%. As with the LDA, there was a lot of confusion from observations of group 10 and to groups 9 and 5. The overall comparison between LDA and QDA is summarized in Table 3. In general, the results show that linear discriminant analysis performs better at all scales. The prediction of a water sample belonging to a GWB group fell from 67% to 59% between LDA and QDA, respectively. In addition, while LDA was able to predict with an accuracy of 41% whether an observation belonged to a sampling point, the QDA results were practically nil.

Table 3. Accuracy scores of LDA and QDA classifications.

Classifier	Observation ID	GWB	GWB Groups
LDA	0.41	0.40	0.67
QDA	0.001	0.17	0.59

3.3. Machine Learning Based Methods

At the scale of the collection point (Observation ID, Table 4), the best accuracy rates were obtained with the Random Forest algorithm (0.44), followed by XGboost (0.34) and Neural Network (0.33). The least efficient algorithms were Bernouilli Naive Bayes (0.057) and Kernal SVM (0.059). Random Forest remained the best performing algorithm (0.67) on the groundwater body scale, followed by XGboost (0.57) and Sub-space KNN (0.49). Finally, looking at the GWB groups scale, we observe that the highest accuracy score was achieved using the Random Forest model (0.83), followed by the LightGBM (0.80) and XGBoost (0.76) models.

Table 4. Accuracy results for the different ML methods at collection points, GWBs and GWB groups scales.

Classifier	Scale		
	Observation ID	**GWB**	**GWB Groups**
Random Forest	0.43	0.67	0.83
Decision Tree	0.26	0.49	0.74
Neural Network	0.33	0.46	0.71
XGBoost	0.34	0.57	0.75
LightGBM	0.09	0.14	0.80
K-Nearest Neighbours	0.20	0.48	0.73
Ensemble (Subspace KNN)	0.20	0.49	0.72
Support Vector Machine	0.22	0.42	0.68
Kernal SVM	0.05	0.33	0.66
Logistic Regression	0.14	0.39	0.65
Gaussian Naive Bayes	0.19	0.19	0.62
Bernoulli Naive Bayes	0.05	0.28	0.56

Feature Contribution

Since Random Forest had the best accuracy score at all scales, this algorithm was used to identify the key parameters for predicting whether a water sample belonged to an observation point, GWB or group of GWBs. The random forest recursive feature elimination (RF-RFE) method was applied in 18 iterations, adding up the parameters one by one. The results are summarized in Tables 5–7 for observation points, GWBs and GWB groups, respectively. Note that when considering GWB groups (Table 7), shifts appeared in the matrix between 5 and 9 selected parameters in the recursive feature elimination (RFE) parameter selection process. This is due to a change in the sequence due to the equality in the order of importance of the selected parameters, as illustrated by the GINI index.

Table 5. Random forest feature importance per the number of RFE selected features (Observation ID scale).

		TDS	Mg	SO$_4$	HCO$_3$	Na	Ca	Cl	NO$_3$	EC25	pH	K	F	*E. coli*	Ent.	B	Fe	Mn	As
	1	1																	
	2	0.57	0.43																
	3	0.37	0.32	0.31															
	4	0.26	0.26	0.24	0.24														
	5	0.21	0.21	0.2	0.19	0.19													
	6	0.17	0.18	0.17	0.16	0.17	0.15												
	7	0.14	0.16	0.15	0.14	0.14	0.14	0.13											
	8	0.13	0.14	0.13	0.13	0.12	0.12	0.12	0.11										
	9	0.11	0.13	0.12	0.12	0.11	0.11	0.1	0.1	0.1									
	10	0.1	0.12	0.11	0.1	0.1	0.1	0.1	0.1	0.09	0.08								
	11	0.1	0.11	0.1	0.1	0.09	0.09	0.09	0.09	0.08	0.08	0.07							
	12	0.09	0.11	0.1	0.09	0.09	0.09	0.08	0.08	0.08	0.07	0.07	0.05						
	13	0.09	0.1	0.09	0.09	0.08	0.08	0.08	0.08	0.08	0.07	0.06	0.06	0.04					
	14	0.08	0.1	0.09	0.09	0.08	0.08	0.08	0.08	0.08	0.07	0.06	0.05	0.03	0.03				
	15	0.08	0.1	0.09	0.09	0.08	0.08	0.08	0.07	0.07	0.07	0.06	0.05	0.03	0.03	0.02			
	16	0.08	0.1	0.09	0.09	0.08	0.08	0.08	0.08	0.07	0.06	0.05	0.04	0.03	0.03	0.03	0.01		
	17	0.08	0.1	0.09	0.09	0.08	0.08	0.08	0.08	0.06	0.06	0.05	0.04	0.03	0.03	0.02	0.02	0.01	
	18	0.08	0.09	0.09	0.08	0.08	0.08	0.07	0.07	0.07	0.06	0.06	0.04	0.03	0.03	0.03	0.02	0.01	0.004

Column heading (vertical): Total Number of Selected Parameters in RFE per iteration

Table 6. Random forest feature importance per the number of RFE selected features (GWB scale).

		Selected Features

		TDS	Cl	Mg	SO$_4$	HCO$_3$	Na	Ca	NO$_3$	EC25	pH	K	F	E. coli	Ent.	B	Fe	Mn	As
Total Number of Selected Parameters in RFE per iteration	1	1																	
	2	0.6	0.4																
	3	0.36	0.33	0.32															
	4	0.27	0.26	0.24	0.24														
	5	0.19	0.23	0.21	0.19	0.18													
	6	0.17	0.18	0.18	0.17	0.16	0.15												
	7	0.14	0.16	0.16	0.14	0.14	0.14	0.13											
	8	0.12	0.14	0.14	0.13	0.12	0.13	0.12	0.10										
	9	0.10	0.13	0.14	0.12	0.11	0.12	0.10	0.09	0.09									
	10	0.09	0.13	0.13	0.11	0.10	0.11	0.09	0.09	0.08	0.08								
	11	0.08	0.12	0.12	0.10	0.09	0.10	0.09	0.08	0.08	0.08	0.07							
	12	0.08	0.11	0.11	0.10	0.09	0.09	0.08	0.07	0.07	0.07	0.07	0.05						
	13	0.08	0.11	0.10	0.09	0.09	0.09	0.08	0.07	0.07	0.07	0.07	0.05	0.04					
	14	0.08	0.10	0.10	0.09	0.08	0.09	0.08	0.07	0.07	0.07	0.07	0.05	0.03	0.03				
	15	0.08	0.10	0.10	0.09	0.08	0.08	0.07	0.07	0.07	0.06	0.06	0.05	0.03	0.03	0.03			
	16	0.07	0.10	0.10	0.09	0.08	0.09	0.07	0.07	0.07	0.06	0.06	0.05	0.03	0.03	0.02	0.01		
	17	0.07	0.10	0.10	0.09	0.08	0.08	0.07	0.07	0.07	0.06	0.06	0.05	0.03	0.03	0.02	0.01	0.01	
	18	0.07	0.10	0.10	0.09	0.08	0.08	0.07	0.07	0.07	0.07	0.06	0.05	0.03	0.03	0.02	0.01	0.01	0.006

Table 7. Random forest feature importance per the number of RFE selected features (GWB group scale). Note that the RFE in this scale modifies the parameter elimination sequences from one iteration to the next, which explains the shifts in the matrix.

		Selected Features

		TDS	Cl	Ca	Na	EC25	pH	Mg	HCO$_3$	SO$_4$	K	NO$_3$	F	E. coli	Ent.	B	Fe	Mn	As
Total Number of Selected Parameters in RFE per iteration	1	1																	
	2	0.57	0.43																
	3	0.33	0.38	0.29															
	4	0.26	0.28	0.24	0.22														
	5	0.19	0.26		0.19	0.18	0.18												
	6	0.16	0.22	0.16	0.17		0.12	0.13	0.15										
	7	0.15	0.21	0.12	0.16		0.12		0.12	0.12									
	8	0.12	0.19	0.12	0.14	0.11	0.11	0.10	0.11										
	9	0.10	0.18	0.10	0.13	0.10	0.11	0.10	0.11		0.09								
	10	0.10	0.17	0.09	0.12	0.09	0.10	0.09	0.09	0.08	0.08								
	11	0.10	0.16	0.08	0.11	0.08	0.09	0.08	0.08	0.08	0.08	0.06							
	12	0.09	0.15	0.08	0.11	0.08	0.09	0.07	0.09	0.08	0.07	0.06	0.04						
	13	0.08	0.15	0.08	0.10	0.08	0.08	0.07	0.08	0.07	0.07	0.06	0.04	0.03					
	14	0.08	0.14	0.08	0.10	0.08	0.08	0.07	0.08	0.07	0.07	0.06	0.04	0.03	0.02				
	15	0.08	0.13	0.08	0.11	0.08	0.08	0.07	0.07	0.07	0.07	0.06	0.04	0.03	0.02	0.02			
	16	0.07	0.14	0.08	0.11	0.08	0.08	0.07	0.08	0.07	0.07	0.06	0.04	0.03	0.02	0.02	0.01		
	17	0.08	0.14	0.07	0.10	0.08	0.08	0.07	0.07	0.07	0.07	0.05	0.04	0.03	0.03	0.02	0.01	0.01	
	18	0.07	0.14	0.07	0.10	0.08	0.08	0.07	0.07	0.07	0.08	0.05	0.04	0.03	0.02	0.02	0.01	0.01	0.006

The results show that major ions, electrical conductivity, nitrates and pH were the most significant parameters for water samples classification at both the observation point and GWB scales (Figure 7). On the other hand, at the GWB group level, electrical conductivity and pH appeared as early as when the number of parameters selected by RFE was 5. The weight of nitrates was much lower, appearing only when more than 10 parameters are selected (Table 7 and Figure 6). Bacteriological parameters and trace element metals had little influence on the classification of water samples, whatever the scale of observation. The inclusion of these parameters only marginally improved the accuracy of the classification from 0.44 to 0.46 at the observation point scale, from 0.65 to 0.66 at the GWB scale and from 0.82 to 0.83 at the GWB group scale.
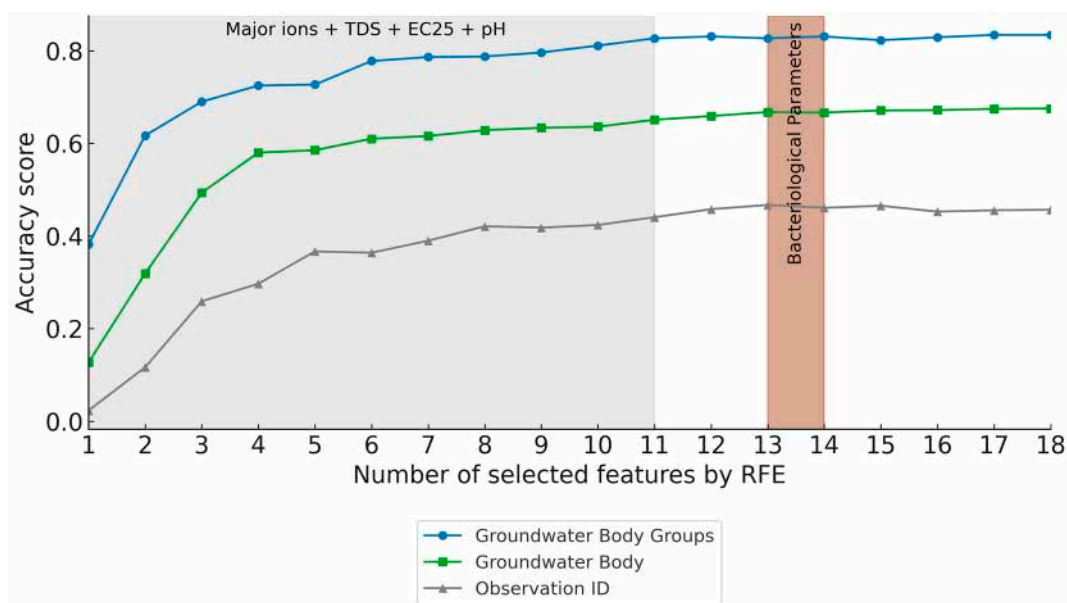
Figure 7. Accuracy Score per number of selected features in RFE in 3 different scales: Observation ID; GWB; GWB groups.

4. Discussion

4.1. Contrasting Water Quality

Linear discriminant analysis was used to assign a water sample to a GWB group with a success rate of 67%. At the same time, the Random Forest algorithm had a success rate of around 83%. Some machine learning methods give slightly better results than linear DA, but these algorithms are known to favour discrimination efficiency over interpretability [57,58]. The results obtained by linear DA are more easily interpreted into an understanding of the processes or parameters that differentiate spatial units. Overall, these results are very satisfactory for a region as vast as the PACA region, with a wide diversity of environments in terms of lithology, altitudinal gradient, land use, etc. In particular, these results are much better than those obtained previously from a fairly similar extraction of the same SISE-Eaux database [20]. This author mentioned a rate of well classified samples within GWB groups of the order of 30% using the Naive Bayesian algorithm, whereas here we obtain a rate of 62% using the same method. Several factors may explain this difference in results: (1) a 15-year data series in our study compared with 10 years in the previous study, resulting in a 30% larger matrix and therefore a more solid learning phase, (2) a number of 18 parameters compared with only 15 in the previous study, even though the added parameters (F, B) play only a very minor role in the discrimination of GWB groups (Figure 6), and (3) the pre-processing of the data by log transformation, which reduces the impact of extreme values [18,21] at different stages in the process of grouping the groundwater bodies (PCA, calculation of the mean coordinate of each GWB on factorial axis, hierarchical clustering). The good result, as well as the concentration of 90% of the variance explained by LDA on the first 3 discriminant functions, shows the strong structuring of the space in terms of water quality, which is reflected by heterogeneity within the dataset. The first discriminant axis clearly illustrates the contrast between calcic, mineralised carbonate waters and much more dilute waters showing signs of faecal contamination. This major contrast in groundwater quality has already been noted in previous studies of the PACA region [17,21], more specifically in the riverine aquifers, but also in the Occitanie region in southwest France [22]. In these diluted waters, the high levels of potassium and chloride ions suggest the influence of sewage plant effluents, which generally contribute to runoff. These results show that vulnerability to contamination varies greatly from one region to another. Karst aquifers are generally sensitive to pollutants due to their specific characteristics. The epikarst zone has thin layers of soil that generally do

not provide adequate filtration and allow rapid transit of contaminants [59]. Colloidal ferric iron is a carrier for bacteria, which can be rapidly transported through the vast network of fractures and conduits typical of these systems. The high degree of connectivity within these aquifers further exacerbates the risk, leading to episodes of widespread contamination that can affect vast areas and significantly compromise water quality. In contrast, deep aquifers with moderate porosity and thicker layers of soil and rock are much less vulnerable to contamination. These aquifers benefit from natural attenuation processes during a much longer transfer, through soils that filter pollutants more effectively before they reach the water table [19]. This distinction is essential when considering monitoring strategies and protection measures for groundwater resources. Sustainable groundwater management therefore necessarily involves approaches that are adapted to the increased risks of contamination, implementing rigorous protection protocols to safeguard these fragile systems.

The selection of various machine learning methods revealed the supremacy of ensemble methods in terms of classification accuracy. Their robustness in handling complex linear and non-linear relationships may be a consequence of the diversity of parameters in the dataset, including major ions, trace elements, and bacteriological parameters, each resulting from distinct processes, and manifested by distinct statistical distributions.

This good overall result, which validates the GWB grouping method, nevertheless conceals disparities. The success rate is high for sedimentary aquifers in coastal areas in the southwest of the PACA region (group 5). This good result can be explained by the specific nature of these aquifers, which are the most mineralised in the region and contain stagnant water in reducing conditions. On the other hand, the riverine aquifers, particularly in the downstream part of rivers, are largely confused with the water resources of the slopes that border them. This confusion reflects the degree of similarity in terms of hydrochemistry and vulnerability to bacterial contamination and is illustrated by the confusion between geographically and geologically neighbouring GWB groups 9 and 10. Group 10 corresponds to the aquifers that accompany the lower valleys of the Rhône and Durance rivers and some of their tributaries, while group 9 corresponds to the middle valleys of the Durance and some of its tributaries, the riverine aquifers of mid-mountain and hilly rivers, and the porous medium aquifers of the region's lower valleys. Groups 2 and 3 show a proportion of observations that are misclassified by most ML algorithms and by LDA analysis. These groups contain mountain and high mountain GWBs, with consequent high temporal variability in mineral load. The water is much diluted when the snow melts, particularly in group 2, which is located on crystalline rock, whereas it can be much more mineralized at the end of the summer. The alternating mineral content of the water may be the reason for the misclassification. To this must be added a high vulnerability to contamination, the temporal variability of which is difficult for the algorithms to take into account.

4.2. Disparities in Discrimination Depending on Parameters

The good prediction result also highlights that the linearity of the discriminant analysis is not a major constraint to the recognition of aquifer types (GWB groups) based on water quality. Quadratic discriminant analysis, although better able to capture complex boundaries between GWB groups in the data hyperspace, does not improve prediction, suggesting that it is not necessary to consider a specific variance for each parameter. If QDA shows more distinct decision limits than LDA, this means that discrimination is better with a non-linear model. This should guide us towards a learning model suitable for discrimination according to this pair of variables. Here, while TDS is a key variable in discrimination (Figure 4) and follows a linear aspect (Figure 5), this is not the case for Fe and NO3, which have a much lower weight in discrimination between groups of GWBs. This explains the good results obtained with LDA. As the different parameters do not depend on the same processes responsible for variability in water composition, they are not affected in a similar way by the linearity or non-linearity of the discrimination method. The

results of both the LDA and ML algorithms suggest the major ions that account for most of the discrimination, since they determine more than 80% of the discrimination between GWB groups. These are essentially lithologically determined parameters with a high spatial range [21,22]. In contrast, metals and metalloids generally show very local variability, at the scale of the observation point. Arsenic, for example, was identified as contributing around 8% of the variance in the dataset [20] but plays virtually no role in group discrimination. This is due to a local variability that relates to only a few sampling points, but does not concern an entire GWB, let alone a group of GWBs. The same applies to bacteriological parameters. The variability is not only local, but also seasonal, with contamination often caused by late summer storms, a classic meteorological phenomenon occurring in the Mediterranean climate, inducing run-off of turbid water that can carry a large bacterial load. Thus, although these parameters play a significant role in the variability of the dataset on a regional scale, they contribute very little to the discrimination of GWB groups. In addition to this seasonal aspect, groundwater contamination is multifactorial. The presence of *E. coli* is linked to suspended matter, generally clay particles, but can also be associated with iron particles. Transport to the water table depends on the soil's filtration capacity, its mineralogical nature, which is particularly slowed by the presence of flocculent cations, and human activities (agriculture, livestock farming, quality of water treatment plants, etc.). Nitrates make a moderate contribution to discrimination, which is consistent with an intermediate spatial range between major ions on the one hand, and trace elements and bacteriological parameters on the other. When considering the Fe vs. NO_3 couple, these two parameters are frequently in opposition due to redox conditions. [19,20,22]. NO_3 is stable under oxidizing conditions and is reduced during denitrification under reducing conditions. Dissolved iron is soluble in reducing environments and insoluble in oxidizing environments, although it can be found in colloidal form in oxidizing environments (for example karst). The closeness of the decision limits for the Fe vs. NO_3 pair suggests that the range of redox conditions is not an axis allowing clear discrimination between the groups.

The disparity in the importance of the parameters for GWB group discrimination is an aspect that has already been addressed in earlier stages of the development of this procedure, notably by Tiouiouine et al. [20], but only implicitly. Here, we explicitly show this disparity, which stems from the range of each parameter, but also from the redundancy of the information carried by each of these parameters. Major ions, TDS and EC reflect lithology, a key feature not only in the acquisition of the chemical characteristics of water, but also sometimes in land use, which can have an impact on vulnerability to various forms of pollution (nitrates, faecal contamination, etc.). Thus, classifying observations within GWB groups is mainly based on major ions, which have low temporal variability and high spatial range. The findings thus underscore the importance of recognizing the unique statistical characteristics and local/broad or even regional behaviours of parameters when employing various machine learning methods for the classification of groundwater quality. It is based on all these interrelations, taken into account in the grouping into homogeneous GWBs, that the monitoring and surveillance of the quality of the resource must be considered.

5. Conclusions

In order to facilitate the monitoring of groundwater resources in France, a multi parametric classification of groundwater bodies has been carried out, based on physico-chemical and bacteriological characteristics. In the Provence-Alpes-Côte d'Azur (PACA) region, previous studies have shown that the classification into 11 groups of GWBs obtained made sense, grouping together GWBs where the processes responsible for the variability of characteristics were similar. Based on a 15-year sampling of 8673 observations (water samples) and 18 parameters, we establish the legitimacy of this grouping using various techniques for classifying observations within the 11 groups, with a success rate of between 67% and 83%. The multi parametric clustering method proposed by Touiouine et al. [20] and improved by Jabrane et al. [18] and Mohsine et al. [21] is robust and efficient. For the health

agencies responsible for monitoring the quality of groundwater resources, each group of GWBs with a well-defined specificity is a relevant spatial unit. Despite the relatively large number and disparity of parameters, the study shows that the number of discriminant functions required to establish GWB groups is relatively low. The major ions, including the two related parameters, namely total dissolved charge and electrical conductivity, exhibit low temporal variability and high spatial range. They are the key parameters in the discrimination of GWB groups. Although trace elements and bacteriological parameters play a significant role in the variability of water quality in the dataset and account for most of the non-conformities in water intended for human consumption, they have only a very marginal influence on the establishment of groups due to their local and high temporal variability. The GWBs delineated by the French Geological Survey in accordance with the guidelines of the European Union's Water Framework Directive are too numerous to be a practical unit for monitoring groundwater quality by health agencies. However, a better understanding of the factors influencing groundwater quality, based on GWB groups, which are fewer and more homogeneous units, opens the way to improved monitoring and protection strategies, to ensure the sustainability of the resource.

References

1. Korkka-Niemi, K. Cumulative Geological, Regional and Site-Specific Factors Affecting Groundwater Quality in Domestic Wells in Finland. *Boreal Environ. Res. Monogr.* **2001**, *20*, 1–20.
2. Earman, S.; Dettinger, M. Potential Impacts of Climate Change on Groundwater Resources—A Global Review. *J. Water Clim. Chang.* **2011**, *2*, 213–229. [CrossRef]
3. Barbieri, M.; Barberio, M.D.; Banzato, F.; Billi, A.; Boschetti, T.; Franchini, S.; Gori, F.; Petitta, M. Climate Change and Its Effect on Groundwater Quality. *Environ. Geochem. Health* **2023**, *45*, 1133–1144. [CrossRef] [PubMed]
4. Lerner, D.N.; Harris, B. The Relationship between Land Use and Groundwater Resources and Quality. *Land Use Policy* **2009**, *26*, S265–S273. [CrossRef]
5. Motlagh, A.M.; Yang, Z.; Saba, H. Groundwater Quality. *Water Environ. Res.* **2020**, *92*, 1649–1658. [CrossRef] [PubMed]
6. Burri, N.M.; Weatherl, R.; Moeck, C.; Schirmer, M. A Review of Threats to Groundwater Quality in the Anthropocene. *Sci. Total Environ.* **2019**, *684*, 136–154. [CrossRef]
7. European Commission Directive 2014/80/EU Amending Annex II to Directive 2006/118/EC of the European Parliament and of the Council on the Protection of Groundwater Against Pollution and Deterioration. *Off. J. Eur. Union* **2014**, 52–55.
8. European Commission Directive 2006/118/EC of the European Parliament and of the Council of 12 December 2006 on the Protection of Groundwater against Pollution and Deterioration. *Off. J. Eur. Union* **2006**, *372*, 19–31.
9. European Commission Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 Establishing a Framework for Community Action in the Field of Water Policy. *Off. J. Eur. Communities* **2000**, *22*, 2000.
10. Allan, I.J.; Vrana, B.; Greenwood, R.; Mills, G.A.; Knutsson, J.; Holmberg, A.; Guigues, N.; Fouillac, A.-M.; Laschi, S. Strategic Monitoring for the European Water Framework Directive. *TrAC Trends Anal. Chem.* **2006**, *25*, 704–715. [CrossRef]
11. Irish Working Group on Groundwater. Approach to Delineation of Groundwater Bodies, Guidance Document No.2. 2005. Available online: https://www.gsi.ie/documents/Groundwater%20Body%20Delineation.pdf (accessed on 28 November 2023).
12. European Commission. Guidance Document No. 26. Guidance on Risk Assessment and the Use of Conceptual Models for Groundwater. 2010. Available online: https://op.europa.eu/en/publication-detail/-/publication/ab5b2e26-dabc-43aa-96ea-ef554b78eb09/language-en (accessed on 28 November 2023).

13. European Commission. *Guidance Document No. 22. Guidance on Implementing the Geographical Information System (GIS) Elements of the EU Water Policy. Tools and Services for Reporting under RBMP within WISE. Guidance on Reporting of Spatial Data for the WFD (RBMP)*; European Commission: Brussels, Belgium, 2009.

14. European Commission. *Guidance Document No 2: Identification of Water Bodies*; European Commission: Brussels, Belgium, 2003.

15. Duscher, K. Compilation of a Groundwater Body GIS Reference Layer. In Proceedings of the WISE GIS Workshop, Copenhagen, Denmark, 16–17 November 2010.

16. Wendland, F.; Blum, A.; Coetsiers, M.; Gorova, R.; Griffioen, J.; Grima, J.; Hinsby, K.; Kunkel, R.; Marandi, A.; Melo, T.; et al. European Aquifer Typology: A Practical Framework for an Overview of Major Groundwater Composition at European Scale. *Environ. Geol.* **2008**, *55*, 77–85. [CrossRef]

17. Tiouiouine, A.; Yameogo, S.; Valles, V.; Barbiero, L.; Dassonville, F.; Moulin, M.; Bouramtane, T.; Bahaj, T.; Morarech, M.; Kacimi, I. Dimension Reduction and Analysis of a 10-Year Physicochemical and Biological Water Database Applied to Water Resources Intended for Human Consumption in the Provence-Alpes-Cote d'azur Region, France. *Water* **2020**, *12*, 525. [CrossRef]

18. Jabrane, M.; Touiouine, A.; Bouabdli, A.; Chakiri, S.; Mohsine, I.; Valles, V.; Barbiero, L. Data Conditioning Modes for the Study of Groundwater Resource Quality Using a Large Physico-Chemical and Bacteriological Database, Occitanie Region, France. *Water* **2022**, *15*, 84. [CrossRef]

19. Lazar, H.; Ayach, M.; Barry, A.A.; Mohsine, I.; Touiouine, A.; Huneau, F.; Mori, C.; Garel, E.; Kacimi, I.; Valles, V.; et al. Groundwater Bodies in Corsica: A Critical Approach to GWBs Subdivision Based on Multivariate Water Quality Criteria. *Hydrology* **2023**, *10*, 213. [CrossRef]

20. Tiouiouine, A.; Jabrane, M.; Kacimi, I.; Morarech, M.; Bouramtane, T.; Bahaj, T.; Yameogo, S.; Rezende-Filho, A.T.; Dassonville, F.; Moulin, M.; et al. Determining the Relevant Scale to Analyze the Quality of Regional Groundwater Resources While Combining Groundwater Bodies, Physicochemical and Biological Databases in Southeastern France. *Water* **2020**, *12*, 3476. [CrossRef]

21. Mohsine, I.; Kacimi, I.; Abraham, S.; Valles, V.; Barbiero, L.; Dassonville, F.; Bahaj, T.; Kassou, N.; Touiouine, A.; Jabrane, M.; et al. Exploring Multiscale Variability in Groundwater Quality: A Comparative Analysis of Spatial and Temporal Patterns via Clustering. *Water* **2023**, *15*, 1603. [CrossRef]

22. Jabrane, M.; Touiouine, A.; Valles, V.; Bouabdli, A.; Chakiri, S.; Mohsine, I.; El Jarjini, Y.; Morarech, M.; Duran, Y.; Barbiero, L. Search for a Relevant Scale to Optimize the Quality Monitoring of Groundwater Bodies in the Occitanie Region (France). *Hydrology* **2023**, *10*, 89. [CrossRef]

23. Zhu, M.; Wang, J.; Yang, X.; Zhang, Y.; Zhang, L.; Ren, H.; Wu, B.; Ye, L. A Review of the Application of Machine Learning in Water Quality Evaluation. *Eco-Environ. Health* **2022**, *1*, 107–116. [CrossRef]

24. He, S.; Wu, J.; Wang, D.; He, X. Predictive Modeling of Groundwater Nitrate Pollution and Evaluating Its Main Impact Factors Using Random Forest. *Chemosphere* **2022**, *290*, 133388. [CrossRef]

25. Judeh, T.; Almasri, M.N.; Shadeed, S.M.; Bian, H.; Shahrour, I. Use of GIS, Statistics and Machine Learning for Groundwater Quality Management: Application to Nitrate Contamination. *Water Resour.* **2022**, *49*, 503–514. [CrossRef]

26. Salem, S.B.H.; Gaagai, A.; Ben Slimene, I.; Ben Moussa, A.; Zouari, K.; Yadav, K.K.; Eid, M.H.; Abukhadra, M.R.; El-Sherbeeny, A.M.; Gad, M.; et al. Applying Multivariate Analysis and Machine Learning Approaches to Evaluating Groundwater Quality on the Kairouan Plain, Tunisia. *Water* **2023**, *15*, 3495. [CrossRef]

27. Zounemat-Kermani, M.; Batelaan, O.; Fadaee, M.; Hinkelmann, R. Ensemble Machine Learning Paradigms in Hydrology: A Review. *J. Hydrol.* **2021**, *598*, 126266. [CrossRef]

28. Haji-Aghajany, S.; Amerian, Y.; Amiri-Simkooei, A. Impact of Climate Change Parameters on Groundwater Level: Implications for Two Subsidence Regions in Iran Using Geodetic Observations and Artificial Neural Networks (ANN). *Remote Sens.* **2023**, *15*, 1555. [CrossRef]

29. Lyons, K.J.; Ikonen, J.; Hokajärvi, A.-M.; Räsänen, T.; Pitkänen, T.; Kauppinen, A.; Kujala, K.; Rossi, P.M.; Miettinen, I.T. Monitoring Groundwater Quality with Real-Time Data, Stable Water Isotopes, and Microbial Community Analysis: A Comparison with Conventional Methods. *Sci. Total Environ.* **2023**, *864*, 161199. [CrossRef]

30. Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2009, Volume 2. Available online: https://link.springer.com/book/10.1007/978-0-387-84858-7 (accessed on 28 November 2023).

31. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

32. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794.

33. Quinlan, J.R. Induction of Decision Trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]

34. Schmidhuber, J. Deep Learning in Neural Networks: An Overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef]

35. Rish, I. An Empirical Study of the Naive Bayes Classifier. In Proceedings of the IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, Seattle, WA, USA, 4–6 August 2001; pp. 41–46.

36. Chery, L.; Laurent, A.; Vincent, B.; Tracol, R. Echanges SISE-Eaux/ADES: Identification Des Protocoles Compatibles Avec Les Scénarios d'échange SANDRE; Vincennes/Orléans, France. 2011. Available online: https://infoterre.brgm.fr/rapports/RP-59211-FR.pdf (accessed on 28 November 2023).

37. Gran-Aymeric, L. Un Portail National Sur La Qualite Des Eaux Destinees a La Consommation Humaine. *Tech. Sci. Méthodes* **2010**, *12*, 45–48. [CrossRef]

38. Pearson, K. LIII. On Lines and Planes of Closest Fit to Systems of Points in Space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1901**, *2*, 559–572. [CrossRef]

39. Day, W.H.E.; Edelsbrunner, H. Efficient Algorithms for Agglomerative Hierarchical Clustering Methods. *J. Classif.* **1984**, *1*, 7–24. [CrossRef]

40. Huberty, C.J. Discriminant Analysis. *Rev. Educ. Res.* **1975**, *45*, 543–598. [CrossRef]

41. Ha, D.H.; Nguyen, P.T.; Costache, R.; Al-Ansari, N.; Van Phong, T.; Nguyen, H.D.; Amiri, M.; Sharma, R.; Prakash, I.; Van Le, H.; et al. Quadratic Discriminant Analysis Based Ensemble Machine Learning Models for Groundwater Potential Modeling and Mapping. *Water Resour. Manag.* **2021**, *35*, 4415–4433. [CrossRef]

42. Singh, S.B.; Gupta, M.K.; Shukla, N.; Chaurasia, G.L.; Singh, S.; Tandon, P.K. Water purification: A brief review on tools and techniques used in analysis, monitoring and assessment of water quality. *Green Chem. Technol. Lett.* **2016**, *2*, 95–102. [CrossRef]

43. Amiri, V.; Nakagawa, K. Using a Linear Discriminant Analysis (LDA)-Based Nomenclature System and Self-Organizing Maps (SOM) for Spatiotemporal Assessment of Groundwater Quality in a Coastal Aquifer. *J. Hydrol.* **2021**, *603*, 127082. [CrossRef]

44. Wilson, S.R.; Close, M.E.; Abraham, P. Applying Linear Discriminant Analysis to Predict Groundwater Redox Conditions Conducive to Denitrification. *J. Hydrol.* **2018**, *556*, 611–624. [CrossRef]

45. Ilić, I.; Puharić, M.; Ilić, D. Groundwater Quality Assessment and Prediction of Spatial Variations in the Area of the Danube River Basin (Serbia). *Water Air Soil Pollut.* **2021**, *232*, 117. [CrossRef]

46. Ielpo, P.; Cassano, D.; Felice Uricchio, V.; Lopez, A.; Pappagallo, G.; Trizio, L.; de Gennaro, G. Identification of Pollution Sources and Classification of Apulia Region Groundwaters by Multivariate Statistical Methods and Neural Networks. *Trans. ASABE* **2013**, *56*, 1377–1386. [CrossRef]

47. Sifaou, H.; Kammoun, A.; Alouini, M.-S. High-Dimensional Quadratic Discriminant Analysis Under Spiked Covariance Model. *IEEE Access* **2020**, *8*, 117313–117323. [CrossRef]

48. DW Hosmer, D.J.; Lemeshow, S.; Sturdivant, R. *Applied Logistic Regression*; John Wiley & Sons: Hoboken, NJ, USA, 2013.

49. Cover, T.; Hart, P. Nearest Neighbor Pattern Classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. [CrossRef]

50. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]

51. Schölkopf, B.; Smola, A. Learning with Kernels Support Vector Machines, Regularization, Optimization, and Beyond. 2002. Available online: https://direct.mit.edu/books/book/1821/Learning-with-KernelsSupport-Vector-Machines (accessed on 28 November 2023).

52. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Guyon, I., Luxburg, U., Von Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.

53. Ho, T.K. Nearest Neighbors in Random Subspaces. In *Advances in Pattern Recognition*; Amin, A., Dori, D., Pudil, P., Freeman, H., Eds.; Springer: Berlin/Heidelberg, Germany, 1998; pp. 640–648.

54. Guyon, I.; Elisseeff, A. An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.

55. Li, F.; Yang, Y. Analysis of Recursive Feature Elimination Methods. In Proceedings of the 28th annual international ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, 15–19 August 2005; ACM: New York, NY, USA, 2005; pp. 633–634.

56. Strobl, C.; Boulesteix, A.-L.; Zeileis, A.; Hothorn, T. Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinform.* **2007**, *8*, 25. [CrossRef] [PubMed]

57. Baryannis, G.; Dani, S.; Antoniou, G. Predicting Supply Chain Risks Using Machine Learning: The Trade-off between Performance and Interpretability. *Future Gener. Comput. Syst.* **2019**, *101*, 993–1004. [CrossRef]

58. Freitas, A.A. Automated Machine Learning for Studying the Trade-Off Between Predictive Accuracy and Interpretability. In *Machine Learning and Knowledge Extraction*; Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 48–66.

59. Dussart-Baptista, L. Transport Des Particules En Suspension et Des Bactéries Associées Dans l'aquifère Crayeux Karstique Haut-Normand. 2003. Available online: https://books.google.com.au/books/about/Transport_des_particules_en_suspension_e.html?id=paUEzgEACAAJ&hl=en&output=html_text&redir_esc=y (accessed on 28 November 2023).