DETECCIÓN DE LA CALIDAD DEL AGUA EN SISTEMAS LÓTICOS MEDIANTE TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

Edwin Ferney Castillo Quintero

Director:
MSc. Ing. David Camilo Corrales Muñoz

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Departamento de Telemática
Popayán, 2016

DETECCIÓN DE LA CALIDAD DEL AGUA EN SISTEMAS LÓTICOS MEDIANTE TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

Edwin Ferney Castillo Quintero

Anexos

Director:
MSc. Ing. David Camilo Corrales Muñoz

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Departamento de Telemática
Popayán, 2016

Tabla de contenido

Listado de Figuras

Listado de Tablas

Anexo A

1.1 Articulo generado

Como resultado de la fase investigativa del presente proyecto, se ha generado un artículo denominado "***Water quality warnings based on cluster analysis in Colombian river basins***" que contribuye a la construcción del conjunto de datos de entrenamiento para detectar la calidad del agua.

Water quality warnings based on cluster analysis in Colombian river basins

Alertas de calidad del agua basadas en análisis de agrupamiento en las cuencas de los ríos Colombianos

Edwin Ferney Castillo / efcastillo@unicauca.edu.co
Wilmer Fernando Gonzales / wfgtulcan@unicauca.edu.co
Iván Darío López / navis@ unicauca.edu.co
Apolinar Figueroa, Ph.D. / apolinar@unicauca.edu.co
David Camilo Corrales / dcorrales@unicauca.edu.co
Miller Guzmán Hoyos / mguzman@unicauca.edu.co
Juan Carlos Corrales, Ph.D. / jcorral@unicauca.edu.co
Universidad Del Cauca, Popayan-Colombia

ABSTRACT Fresh water is considered one of the most important renewable natural resources in the world. Among all the countries, Colombia is one of the places with the highest water supply, and has five watersheds: the Caribbean, Orinoco, Amazon, Pacific and Catatumbo. It is therefore vital to study and evaluate the water quality of the rivers and/ or lotic systems. In recent studies, some scientists made use of biological indices to calculate water quality, while others detected water quality through machine learning techniques. However, these studies do not allow users to easily interpret the results. These investigations motivated us to propose a dataset for generating water quality alerts in Piedras river basin based on the analysis of the K-Means clustering algorithm and C.4.5 classification technique.

KEYWORDS Clustering; water quality data; aquatic macro-invertebrates; taxon; C.4.5 decision tree.

Anexo B

En las siguientes secciones son presentados los registros (Log's) de los clasificadores evaluados

1.2 Dataset Rio Las Piedras original

C.4.5

```
Correctly Classified Instances         537               83.2558 %
Incorrectly Classified Instances       108               16.7442 %
Kappa statistic                          0.7486
Mean absolute error                      0.1314
Root mean squared error                  0.3183
Relative absolute error                 29.619  %
Root relative squared error             67.5796 %
Total Number of Instances              645

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.846    0.093      0.819     0.846     0.832      0.889    1
                 0.821    0.086      0.813     0.821     0.817      0.883    2
                 0.83     0.072      0.864     0.83      0.847      0.897    3
Weighted Avg.    0.833    0.083      0.833     0.833     0.833      0.89
```

RB

```
Correctly Classified Instances         417               64.6512 %
Incorrectly Classified Instances       228               35.3488 %
Kappa statistic                          0.469
Mean absolute error                      0.2532
Root mean squared error                  0.4241
Relative absolute error                 57.0539 %
Root relative squared error             90.0313 %
Total Number of Instances              645

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.682    0.23       0.596     0.682     0.636      0.782    1
                 0.537    0.171      0.587     0.537     0.561      0.775    2
                 0.709    0.128      0.755     0.709     0.731      0.866    3
Weighted Avg.    0.647    0.175      0.65      0.647     0.647      0.81
```

MVS

```
Correctly Classified Instances         425               65.8915 %
Incorrectly Classified Instances       220               34.1085 %
Kappa statistic                          0.4884
Mean absolute error                      0.317
Root mean squared error                  0.41
Relative absolute error                 71.4252 %
Root relative squared error             87.0475 %
Total Number of Instances              645

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
```

	0.528	0.148	0.638	0.528	0.578	0.744	1
	0.682	0.225	0.578	0.682	0.626	0.762	2
	0.761	0.135	0.758	0.761	0.759	0.852	3
Weighted Avg.	0.659	0.168	0.662	0.659	0.657	0.788	

RNA

Correctly Classified Instances	492	76.2791 %
Incorrectly Classified Instances	153	23.7209 %
Kappa statistic	0.6437	
Mean absolute error	0.1732	
Root mean squared error	0.3599	
Relative absolute error	39.0315 %	
Root relative squared error	76.4016 %	
Total Number of Instances	645	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.715	0.135	0.725	0.715	0.72	0.86	1
	0.701	0.14	0.695	0.701	0.698	0.855	2
	0.861	0.08	0.857	0.861	0.859	0.942	3
Weighted Avg.	0.763	0.117	0.763	0.763	0.763	0.888	

K-NN

Correctly Classified Instances	450	69.7674 %
Incorrectly Classified Instances	195	30.2326 %
Kappa statistic	0.5462	
Mean absolute error	0.2028	
Root mean squared error	0.4478	
Relative absolute error	45.6986 %	
Root relative squared error	95.0629 %	
Total Number of Instances	645	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.64	0.181	0.637	0.64	0.639	0.715	1
	0.612	0.194	0.589	0.612	0.6	0.697	2
	0.826	0.075	0.86	0.826	0.843	0.87	3
Weighted Avg.	0.698	0.147	0.701	0.698	0.699	0.765	

1.3 Dataset Rio Las Piedras procesado

RB

Correctly Classified Instances	445	81.3387 %
Incorrectly Classified Instances	102	18.6613 %
Kappa statistic	0.6013	
Mean absolute error	0.2257	
Root mean squared error	0.3325	
Relative absolute error	63.9503 %	
Root relative squared error	79.2118 %	
Total Number of Instances	547	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.951	0.403	0.796	0.951	0.866	0.765	1
	0.826	0.047	0.865	0.826	0.845	0.898	3
	0	0	0	0	0	0.601	2
Weighted Avg.	0.813	0.264	0.727	0.813	0.766	0.783	

MVS

Correctly Classified Instances 449 82.1501 %
Incorrectly Classified Instances 98 17.8499 %
Kappa statistic 0.6138
Mean absolute error 0.2736
Root mean squared error 0.3542
Relative absolute error 77.5112 %
Root relative squared error 84.3822 %
Total Number of Instances 547

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.971	0.414	0.795	0.971	0.874	0.778	1
	0.811	0.03	0.907	0.811	0.856	0.897	3
	0	0	0	0	0	0.636	2
Weighted Avg.	0.822	0.266	0.738	0.822	0.773	0.794	

RNA

Correctly Classified Instances 447 81.7444 %
Incorrectly Classified Instances 100 18.2556 %
Kappa statistic 0.6047
Mean absolute error 0.1777
Root mean squared error 0.3167
Relative absolute error 50.341 %
Root relative squared error 75.4392 %
Total Number of Instances 547

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.967	0.43	0.788	0.967	0.868	0.779	1
	0.803	0.019	0.938	0.803	0.865	0.907	3
	0	0.007	0	0	0	0.619	2
Weighted Avg.	0.817	0.274	0.742	0.817	0.772	0.796	

K-NN

Correctly Classified Instances 396 72.4138 %
Incorrectly Classified Instances 151 27.5862 %
Kappa statistic 0.4808
Mean absolute error 0.1857
Root mean squared error 0.4274
Relative absolute error 52.5942 %
Root relative squared error 101.8244 %
Total Number of Instances 547

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.795	0.323	0.803	0.795	0.799	0.758	1
	0.818	0.069	0.812	0.818	0.815	0.884	3
	0.093	0.116	0.089	0.093	0.091	0.515	2
Weighted Avg.	0.724	0.232	0.727	0.724	0.726	0.765	

C.4.5

Correctly Classified Instances 457 83.57 %
Incorrectly Classified Instances 90 16.43 %
Kappa statistic 0.6398
Mean absolute error 0.1813
Root mean squared error 0.3034
Relative absolute error 51.3613 %
Root relative squared error 72.2712 %
Total Number of Instances 547

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.993	0.425	0.794	0.993	0.883	0.756	1
	0.811	0.003	0.991	0.811	0.892	0.868	3
	0	0.002	0	0	0	0.608	2
Weighted Avg.	0.836	0.265	0.76	0.836	0.788	0.77	

1.4 Dataset Estuario California Original

RB

Correctly Classified Instances	2496	99.6407 %
Incorrectly Classified Instances	9	0.3593 %
Kappa statistic	0.995	
Mean absolute error	0.0021	
Root mean squared error	0.0412	
Relative absolute error	0.5795 %	
Root relative squared error	9.7206 %	
Total Number of Instances	2505	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	0	1	1	1	0	
	0.982	0	1	0.982	0.991	0.999	1
	1	0.004	0.982	1	0.991	0.999	2
	1	0	1	1	1	1	3
Weighted Avg.	0.996	0.001	0.996	0.996	0.996	1	

C.4.5

Correctly Classified Instances	2504	99.9601 %
Incorrectly Classified Instances	1	0.0399 %
Kappa statistic	0.9994	
Mean absolute error	0.0004	
Root mean squared error	0.0141	
Relative absolute error	0.1108 %	
Root relative squared error	3.3337 %	
Total Number of Instances	2505	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	0	1	1	1	1	0
	1	0	0.998	1	0.999	1	1
	0.998	0	1	0.998	0.999	0.999	2
	1	0	1	1	1	1	3
Weighted Avg.	1	0	1	1	1	1	

RNA

Correctly Classified Instances	2504	99.9601 %
Incorrectly Classified Instances	1	0.0399 %
Kappa statistic	0.9994	
Mean absolute error	0.0022	
Root mean squared error	0.0124	
Relative absolute error	0.6012 %	
Root relative squared error	2.9156 %	
Total Number of Instances	2505	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	0	1	1	1	1	0
	1	0	0.998	1	0.999	1	1
	0.998	0	1	0.998	0.999	1	2
	1	0	1	1	1	1	3
Weighted Avg.	1	0	1	1	1	1	

MVS

Correctly Classified Instances	2504	99.9601 %
Incorrectly Classified Instances	1	0.0399 %
Kappa statistic	0.9994	
Mean absolute error	0.25	
Root mean squared error	0.3119	
Relative absolute error	69.4485 %	
Root relative squared error	73.5056 %	
Total Number of Instances	2505	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	0	1	1	1	1	0
	1	0	0.998	1	0.999	1	1
	0.998	0	1	0.998	0.999	1	2
	1	0	1	1	1	1	3
Weighted Avg.	1	0	1	1	1	1	

K-NN

Correctly Classified Instances	2504	99.9601 %
Incorrectly Classified Instances	1	0.0399 %
Kappa statistic	0.9994	
Mean absolute error	0.0009	
Root mean squared error	0.0141	
Relative absolute error	0.2398 %	
Root relative squared error	3.332 %	
Total Number of Instances	2505	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	0	1	1	1	1	0
	1	0	0.998	1	0.999	1	1
	0.998	0	1	0.998	0.999	0.999	2
	1	0	1	1	1	1	3
Weighted Avg.	1	0	1	1	1	1	

1.5 Dataset Estuario California procesado

RB

Correctly Classified Instances	1448	84.7239 %
Incorrectly Classified Instances	268	15.2761 %
Kappa statistic	0.7855	
Mean absolute error	0.0786	
Root mean squared error	0.2596	
Relative absolute error	22.1294 %	
Root relative squared error	61.5975 %	
Total Number of Instances	1757	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class

	0.97	0.167	0.72	0.97	0.827	0.981	0.0
	0.58	0.013	0.951	0.58	0.72	0.947	1.0
	0.964	0.024	0.947	0.964	0.956	0.986	2.0
	0.953	0.012	0.871	0.953	0.91	0.997	3.0
Weighted Avg.	0.847	0.064	0.872	0.847	0.84	0.974	

C.4.5

Correctly Classified Instances 1749 99.5706 %
Incorrectly Classified Instances 8 0.4294 %
Kappa statistic 0.994
Mean absolute error 0.0028
Root mean squared error 0.0461
Relative absolute error 0.775 %
Root relative squared error 10.9456 %
Total Number of Instances 1757

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0.004	0.992	1	0.996	0.998	0.0
0.996	0.002	0.996	0.996	0.996	0.999	1.0
0.992	0.001	0.998	0.992	0.995	0.997	2.0
0.992	0	1	0.992	0.996	0.996	3.0
Weighted Avg. 0.996	0.002	0.996	0.996	0.996	0.998	

RNA

Correctly Classified Instances 1698 96.6258 %
Incorrectly Classified Instances 59 3.3742 %
Kappa statistic 0.9526
Mean absolute error 0.0219
Root mean squared error 0.1135
Relative absolute error 6.1747 %
Root relative squared error 26.9225 %
Total Number of Instances 1757

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.982	0.011	0.976	0.982	0.979	1	0.0
0.934	0.008	0.981	0.934	0.957	0.963	1.0
0.982	0.022	0.952	0.982	0.967	0.989	2.0
0.969	0.006	0.932	0.969	0.95	1	3.0
Weighted Avg. 0.966	0.013	0.967	0.966	0.966	0.985	

MVS

Correctly Classified Instances 1534 87.3006 %
Incorrectly Classified Instances 223 12.6994 %
Kappa statistic 0.8166
Mean absolute error 0.2617
Root mean squared error 0.33
Relative absolute error 73.6413 %
Root relative squared error 78.2913 %
Total Number of Instances 1757

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.882	0	1	0.882	0.937	0.975	0.0
0.958	0.052	0.89	0.958	0.923	0.945	1.0
1	0.131	0.772	1	0.872	0.934	2.0
0	0	0	0	0	0.826	3.0
Weighted Avg. 0.873	0.056	0.818	0.873	0.84	0.942	

K-NN

Correctly Classified Instances 1707 97.1779 %

Incorrectly Classified Instances	50	2.8221 %
Kappa statistic	0.9603	
Mean absolute error	0.0151	
Root mean squared error	0.1186	
Relative absolute error	4.2475 %	
Root relative squared error	28.1477 %	
Total Number of Instances	1757	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.974	0.012	0.972	0.974	0.973	0.981	0.0
	0.964	0.014	0.968	0.964	0.966	0.975	1.0
	0.982	0.009	0.98	0.982	0.981	0.987	2.0
	0.953	0.004	0.953	0.953	0.953	0.974	3.0
Weighted Avg.	0.972	0.011	0.972	0.972	0.972	0.98	

Anexo C

En esta sección se presentan los artefactos generados en el proceso de ingeniería que se llevo a cabo para el desarrollo del prototipo.

1.6 FASE DE INICIACIÓN

En esta fase se determina la viabilidad y alcance del presente proyecto, mediante las bases conceptuales, investigación de presuntas alternativas y demás actividades que permiten estructurar el presente trabajo. Esta fase se compone de los siguientes artefactos:

3.1.1 Captura y análisis de requisitos funcionales

Los requisitos funcionales identificados para el prototipo son los siguientes:

- **RF1**. El usuario puede cargar un archivo con extensión .arff desde el computador donde se encuentra instalado el programa.
- **RF2**. El usuario tiene la opción de deshacer el archivo cargado.
- **RF3**. El prototipo es capaz de verificar si se ha cargado un archivo .arff, y notificárselo al usuario.
- **RF4**. Una vez cargado el archivo con extensión .arff el sistema permite al usuario seleccionar el procesamiento de datos que desee (Reducción de atributos e instancias y balanceo de clases).
- **RF5**. El prototipo es capaz de procesar un archivo .arff y detectar la calidad del agua, a partir del archivo .arff cargado.
- **RF6**. El usuario tiene la opción de elegir visualizar los resultados de la detección de la calidad del agua obtenidos por el prototipo.
- **RF7**. El aplicativo es capaz de construir los dataset procesados, es decir, a partir del archivo cargado, el prototipo podrá crear un nuevo archivo .arff con los datos procesados.

3.1.2 Identificación de casos de uso

En la Figura 1 se presenta el diagrama de casos de uso para el prototipo desarrollado. Los casos de uso se identificaron a partir de las funcionalidades que ofrece el sistema, a continuación se realiza una breve descripción de cada uno de ellos:

- **CU-001 Reducir atributos**: este caso de uso permite reducir la cantidad de atributos dentro del dataset.
- **CU-002 Reducir instancias**: este caso de uso permite reducir la cantidad de instancias del dataset.
- **CU-003 Balancear clases**: este caso de uso permite al usuario emparejar o balancear la cantidad de instancias de las clases del dataset.

- **CU-004 Todos los procesos**: este caso de uso permite al usuario reducir el número de atributos e instancias del dataset, además permite realizar el balanceo de las clases (combinación de los casos de uso **CU-001-CU-003**.
- **CU-005 Visualizar detección de la calidad del agua**: este caso de uso determina y visualiza la detección de la calidad del agua de los casos de uso individual **CU-001-CU-003** y en conjunto **CU-004.**

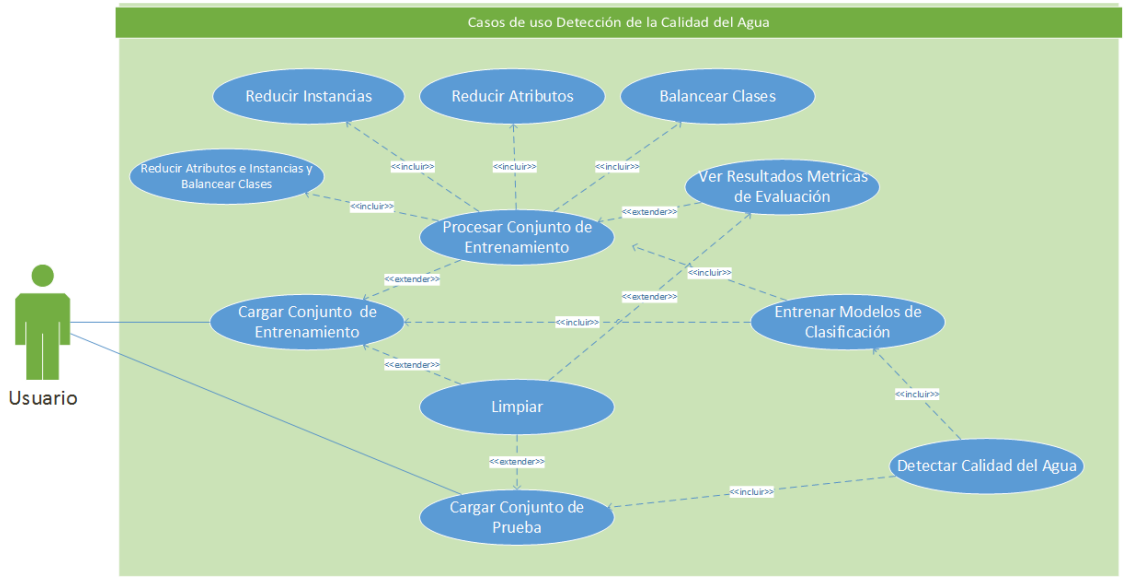

Figura 1. Diagrama de casos de uso del sistema

3.1.3 Cronograma de actividades

Para el desarrollo del proyecto fue utilizada la metodología PUA, la cual consta de 4 fases (iniciación, elaboración, construcción y transición). A su vez estas fases cuentan con n iteraciones definiendo un modelo en casada sobre las actividades: análisis, diseño, implementación y pruebas. Para el actual proyecto se definió una iteración por cada fase a excepción de la fase de construcción donde se fijaron dos iteraciones. La Tabla 1, corresponde al cronograma del proyecto.

Tabla 1. Cronograma de actividades

	Meses											
Actividades	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1. Revisión bibliográfica												
1.1 Fase Preparatoria												

Revisión del estado del arte del proceso de reducción de la dimensionalidad.	X	X										
Revisión del estado del arte de del proceso de balanceo de clases		X	X									
Revisión del estado del arte de clasificadores para la detección de la calidad del agua.			X	X								
1.2 Fase Descriptiva												
Explorar mecanismos que puedan ser usados como referente para reducir el número de atributos	X	X										
Explorar mecanismos que puedan ser usados como base para reducir el número de instancias		X	X									
Explorar mecanismos que puedan ser usados como base para balancear las clases			X	X								
Explorar modelos de clasificación para la detección de la calidad del agua				X	X							
2. Construcción del prototipo (2 Iteraciones para la fase de construcción)												
Iniciación												
Lista de requisitos funcionales y no funcionales					X							
Diagrama de casos de uso					X							
Lista de riesgos					X							
Diseño de pruebas					X	X						
Elaboración												
Diagrama de clases						X						
Diagrama de paquetes						X						
Construcción						X	X	X	X			
I Iteración												
Caso de uso Reducir Atributos				X								
Caso de Reducir Instancias				X								
Caso de uso Visualizar Detección Calidad del Agua				X								
Caso de uso Balancear Clases					X							
Caso de uso Combinar Procesos					X							
Ejecución de pruebas					X							
II Iteración												
Caso de uso Evaluar procesos por operación						X						
Caso de uso Evaluar procesos por entradas						X						
Caso de uso Evaluar procesos por salidas						X						
Caso de uso clasificar procesos recuperados y evaluados							X					
Caso de uso evaluación total de procesos							X					
Ejecución de pruebas							X	X				
Transición												
Presentación y divulgación										X	X	
3. Documentación y divulgación												
3.1 Fase de construcción teórica y global												

Elaboración del documento final y anexos			X	X	X	X	X	X	X	X	X	
3.2 Fase de extensión y publicación												
Análisis de los resultados								X	X	X	X	

3.1.4 Identificación de riesgos

Tabla 2. Riesgos

No. Riesgo	Descripción	Disparador	Acciones preventivas	Impacto	Probabilidad	Mitigación
R1	Integrantes del equipo no están concentrados en el proyecto.	Los integrantes del proyecto incumplen con su tarea o entregan lo que no se les pidió.	Reuniones frecuentes.	Medio	Baja	Realizar una charla de motivación para los integrantes.
R2	Mala organización de los integrantes de la tesis.	Se repite el trabajo o no se realiza el trabajo por los integrantes de la tesis.	Verificar los roles y responsabilidades.	Medio	Baja	Reunión general para definir nuevos roles y responsabilidades.
R3	Mal diseño.	En desarrollo no se puede realizar algún tipo de funcionalidad que debería estar.	Verificar si el diseño refleja los requisitos.	Alto	Media	Modificar el diseño según los requisitos planteados.
R4	Mala implementación.	Errores en tiempo de ejecución o resultados incoherentes.	Pruebas bien diseñadas y ejecutadas.	Alto	Baja	Realizar depuración del módulo donde se encontró el error.
R5	Pruebas del producto mal diseñadas y ejecutadas.	La ejecución de las pruebas es correcta pero se presentan errores al usuario final del sistema.	Realizar pruebas alfa para evitar los defectos en la aplicación.	Alto	Media	Rediseñar las pruebas de la aplicación.
R6	Perder personal.	Un integrante de la tesis se retira del proyecto.	Cada uno de los tesistas debe mantenerse enterado de todo el proyecto.	Alto	Baja	Solicitar un tiempo de prórroga para terminar el proyecto.
R7	Construir producto que no se ha pedido ni especificado en los requerimientos	Funciona todo pero no es lo que se necesita.	Utilizar prototipos de papel para verificar los requisitos.	Alta	Baja	Verificar los requerimientos e informarle de los avances del producto al director de la tesis.
R8	Interfaz mal diseñada.	Inconformidad en la interfaz de usuario.	Desarrollar la aplicación basándose en los prototipos de interfaces, de tal manera que el usuario no tenga problemas de interacción con la herramienta.	Medio	Baja	Rediseñar la interfaz.
R9	Mal manejo de la metodología de desarrollo.	Roles inexistentes, fases y/o artefactos no correspondientes.	Estudio inicial de la metodología, en cuanto a roles, fases y artefactos.	Alto	Media	Replantear la metodología con roles, fases y artefactos específicos.
R10	Director de tesis no satisfecho	Los tesistas no se reúnen	Programar reuniones	Medio	Baja	Acordar nuevas reuniones y llevar

	por ausencia los tesistas.	frecuentemente con el director.	periódicas con el director de tesis.			una agenda con el director.
R11	Baja disponibilidad del director de tesis.	El director de la tesis no se reúne con los tesistas.	El director de tesis debe comprometerse a realizar reuniones con los tesistas.	Alto	Alta	Acordar nuevas reuniones con el director de tesis y en caso que sea necesario, programar nuevas fechas de entrega de artefactos.

Tabla 3. Matriz probabilidad/impacto

		Probabilidad		
		Alto	**Medio**	**Bajo**
Impacto	**Alto**	R10, R11	R3, R5, R9	R6, R4, R7
	Medio			R1, R2, R8
	Bajo			

La priorización se realizó de la siguiente manera:

- El primer aspecto a tener en cuenta es el impacto del riesgo, es decir, los riesgos de mayor impacto tendrán mayor prioridad.
- El segundo aspecto es la probabilidad de ocurrencia del riesgo, es decir, en caso de que dos riesgos tengan igual impacto, tendrá mayor prioridad aquel que tenga mayor probabilidad de presentarse.

La lista priorizada ordenada de mayor a menor prioridad es la siguiente:

Tabla 4. Lista priorizada de riesgos

Prioridad	**Riesgos**
1	R10,R11
2	R3,R5,R9
3	R6,R4,R7
4	R1,R2,R8

1.7 FASE DE ELABORACION

3.2.1 Arquitectura

A partir de las características de la solución a desarrollar se procedieron a realizar el análisis y diseño de las funcionalidades necesarias para dar cumplimiento a las necesidades identificadas anteriormente. A continuación se expone la manera en que se relacionan entre si los componentes del prototipo.

A continuación se presenta una breve descripción de cada uno de los subsistemas que componen el prototipo.

- **Usuario**: Corresponde a los usuarios que interactúan con el prototipo. Los usuarios cargan un archivo con extensión .arff, con el fin de detectar la calidad del agua que

representa el mismo. El prototipo entrega al usuario los resultados de la detección de la calidad del agua de manera visual tanto del dataset procesado como del dataset sin procesar (original).

- **Dataset:** este componente contiene los datos de calidad del agua a detectar, y tiene como extensión el formato .arff.
- **Dataset procesado:** este componente contiene la información del dataset generado por el módulo de procesamiento de datos (Reducción de atributos e instancias y balanceo de clases).
- **Procesamiento de datos**: este módulo permite procesar los datos del dataset ya sea: reducción de atributos, reducción de instancias y balanceo de clases de manera individual y conjunta.
- **Validación cruzada**: este módulo evalúa los resultados del análisis estadístico y garantiza que la partición de datos de entrenamiento y prueba sean independientes.
- **Librería Weka**: este componente es la librería que se utilizó para modelar los mecanismos de procesamiento de datos y los algoritmos de clasificación.
- **Modelo de clasificación**: recibe como parámetros el conjunto de datos de entrenamiento y el conjunto de datos de prueba, resultado del módulo de validación cruzada. Además, este componente es el encargado de realizar la detección de la calidad del agua, a partir de dataset que recibió como parámetro.
- **Detección de la calidad del agua**: este bloque recibe los datos resultantes del modelo de clasificación y los muestra al usuario. Muestra tanto los resultados de la detección de la calidad del agua del dataset original (sin procesar) como del procesado.

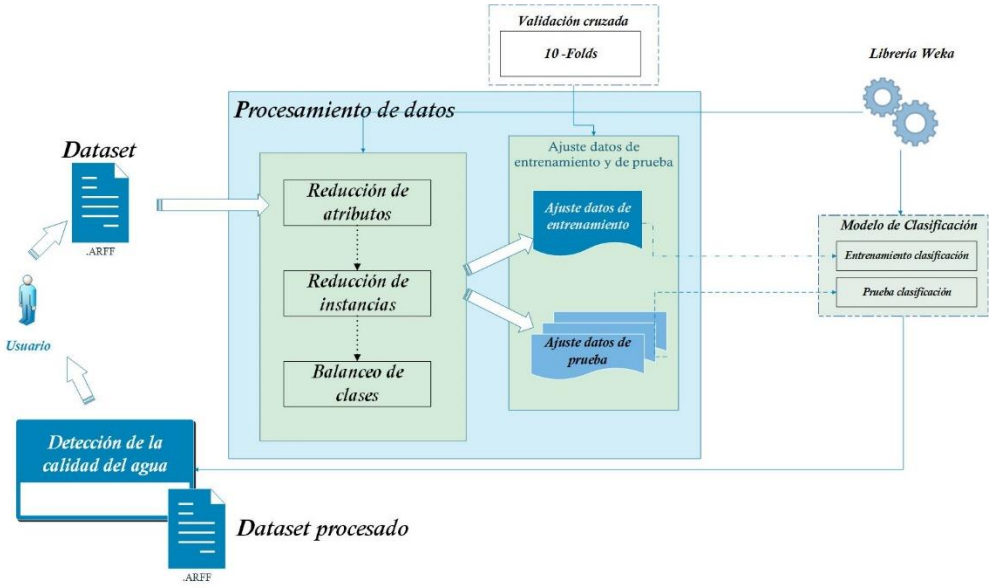

Figura 2. Arquitectura del prototipo

1.8 FASE DE CONSTRUCCIÓN

3.3.1 Diagrama de clases del sistema

Esta sección se enfoca en el desarrollo del diagrama de clases del prototipo implementado. La Figura 3 presenta el diagrama de clases de la aplicación, clases que son descritas a continuación.

- **GUI**: implementa la lógica necesaria para la presentación de la interfaz gráfica del prototipo.
- **PrincipalComponents**: esta clase realiza el proceso de Análisis de Componentes Principales, calculando la varianza y transformando el dataset en componentes ortogonales.
- **ACP**: esta clase contiene la lógica suficiente y necesaria para realizar la reducción de atributos del dataset.
- **ArffSaver**: esta clase es la encargada de gestionar la creación de un archivo con extensión .arff.
- **FilesArff**: la función principal de esta clase consiste crear un directorio con extensión arff y de crear el listado de componentes más adecuado.
- **EvaluationClassifier**: esta clase calcula el promedio de las precisiones de los clasificadores.
- **ConverterUtils**: esta clase permite la manipulación de las características y atributos de los archivos de tipo .arff.
- **Instances**: la función principal de esta clase es la de aplicar los diferentes tipos de filtros a las instancias del dataset.
- **LoadBoostingIS**: esta clase contiene toda la configuración inicial de las variables del prototipo.
- **ManagerFileRips, Drop3Algo, Ib3Algo, MsAlgo, RnnAlgo y RandomAlgo**: estas clases implementan su respectivo algoritmo de reducción de instancias.
- **BoostingIS**: esta clase contiene la lógica suficiente y necesaria para realizar la reducción de instancias del dataset.
- **SMOTE**: esta clase implementa la definición del algoritmo con el mismo nombre.
- **SelectClases**: calcula el grado de desbalanceo que tiene las clases del dataset y las organiza de manera descendente.
- **LogicSmote**: la función principal de esta clase es encontrar la cantidad más adecuada de instancias sintéticas, para sobre-muestrear con ellas la clase minoritaria del dataset.

Figura 3. Diagrama de clases del prototipo

3.3.2 Diagrama de paquetes

Según [28], los diagramas presentados en esta sección, exponen la vista lógica de las aplicaciones software que componen el sistema. Dichos diagramas están organizados en paquetes, subsistemas y capas (Presentación y lógica del negocio, acceso a datos), mostrando la interacción existente entre capas así como también los paquetes más relevantes que las componen. La Figura 4 presenta el diagrama de paquetes del prototipo.

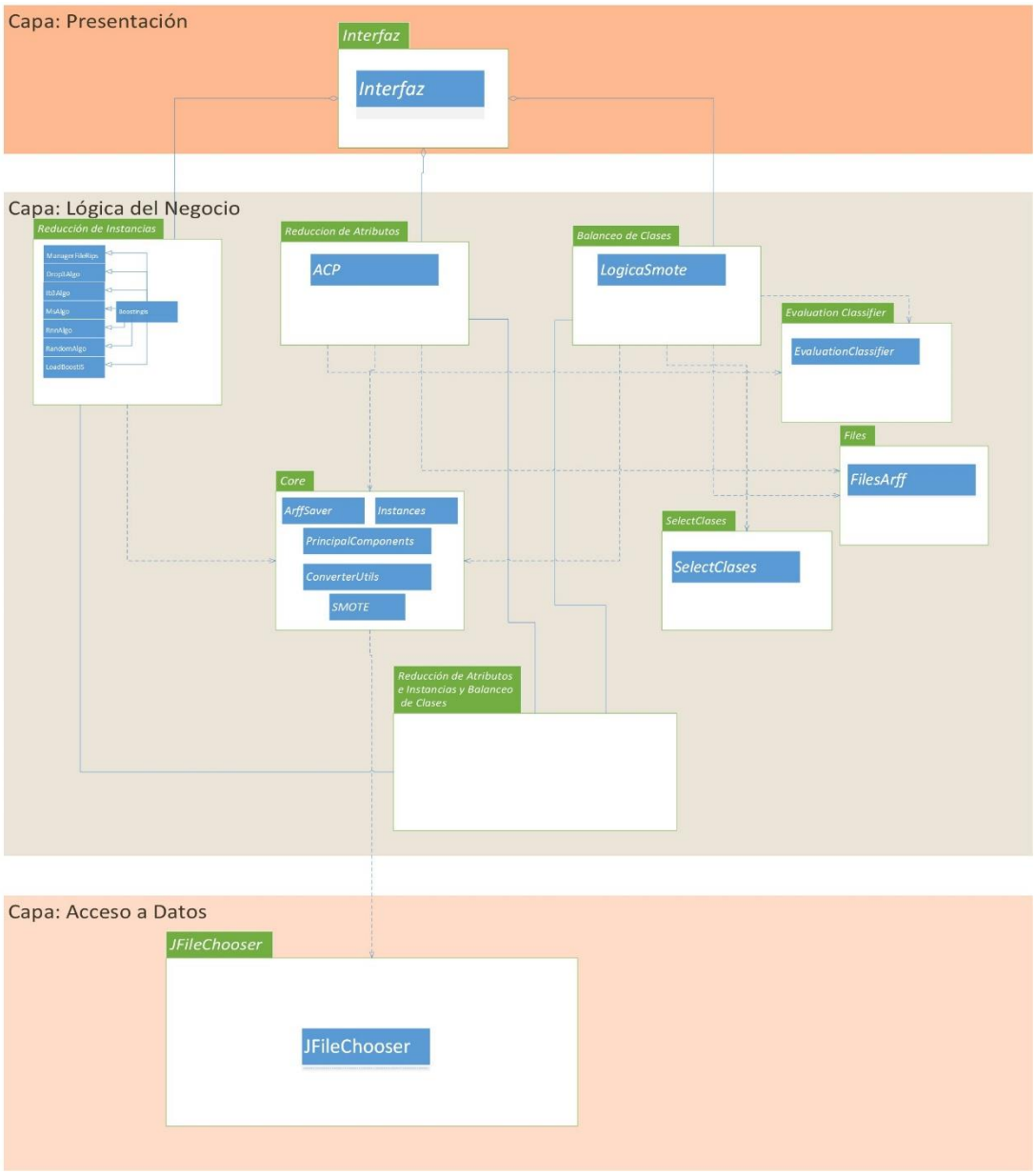

Figura 4. Diagrama de paquetes de la aplicación

A continuación se describen cada una de las capas y su respectiva interacción.

Capa de Presentación

Esta capa es implementada por medio del paquete Interfaz, el cual se describe a continuación.

- **Interfaz**: Ofrece una interfaz gráfica al usuario, que permite cargar un archivo .arff (dataset) con el fin de detectar la calidad del agua.

Capa Lógica del Negocio

Los paquetes que implementan esta capa, representan las funcionalidades más importantes del sistema propuesto. En esta sección se realiza una breve descripción de cada uno de estos paquetes.

- **Core**: Paquete con las clases e interfaces que conforman la infraestructura de WEKA [29]. Define las estructuras de datos que contienen los datos a manejar por los algoritmos de aprendizaje. Este paquete encapsula un dataset. Los atributos e instancias junto con los métodos para manejarlos (creación y copia, división en sub-datasets, aleatorización, gestión de pesos, etc.).
- **SelectClases**: contiene la clase encargadas de seleccionar la clase minoritaria y la clase mayoritaria del dataset, así como también calcular el grado de desbalanceo de estas clases.
- **Files**: este paquete representa la clase necesaria para seleccionar un conjunto de atributos que se desean suprimir del dataset. Además, permite crear el datasets procesado con extensión .arff.
- **Evaluation Classifier**: Representa a un paquete contenedor de la clase que permite evaluar el comportamiento de un conjunto de modelos de clasificación de manera simultáneas, también calcula y promedia la precisión de los algoritmos de aprendizaje supervisado utilizados.
- **Reducción de instancias**: este paquete contiene toda la funcionalidad para reducir las instancias irrelevantes y redundantes del dataset.
- **Reducción de atributos**: representa un paquete contenedor de la clase que permite reducir el número de atributos del dataset, eliminando las características más irrelevantes, redundantes y que representan poca información.
- **Balanceo de clases**: contiene la clase que permite determinar la cantidad de instancias sintéticas adecuada para balancear las clases del dataset.

Capa de acceso a Datos

Esta capa tiene por objetivo servir como puente entre la capa lógica de negocio y el dataset cargado en memoria. Los paquetes que hacen parte de esta capa son descritos a continuación:

- **JFileChooser**: Representa a un paquete contenedor de la clases que proporciona un mecanismo sencillo para que el usuario seleccione un archivo y lo cargue en memoria.

1.9 Fase de pruebas

Esta fase es la encargada de evaluar el prototipo construido, sin embargo es importante revisar el proceso, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso anterior, en el que se haya posiblemente cometido algún error. Las pruebas de calidad y rendimiento, se hacen sobre los módulos que realizan las funcionalidades más importantes del prototipo: Modulo de Reducción de la dimensionalidad, Modulo de Balanceo de Clases y Modulo del modelo de clasificación. En los capítulos 3, 4 y 5 de la monografía son explicadas las métricas utilizadas para medir la calidad y el rendimiento del sistema, además son presentados los resultados obtenidos.

1.10 Fase De Transición

En esta fase se presenta al usuario una breve descripción del manual de usuario del prototipo ejecutable desarrollado, que le permite detectar la calidad del agua mediante técnicas de aprendizaje automático, a partir de un dataset con información relacionada al agua.

3.5.1 Interfaz de usuario

En esta sección son presentadas las interfaces graficas de usuario (GUI), del prototipo desarrollado basado en la metodología PUA. En la Figura 5 se puede observar la interfaz principal de usuario del prototipo.

Interfaz principal

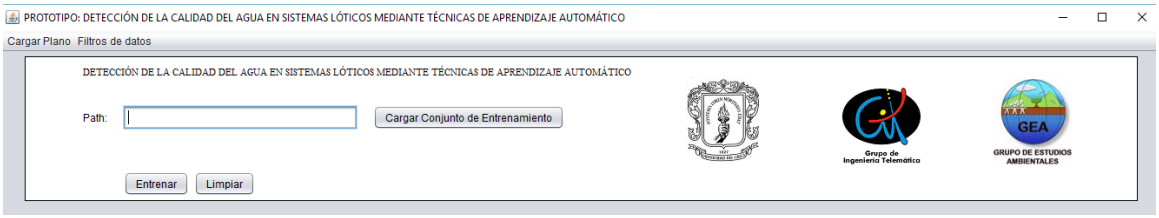

Figura 5. Interfaz principal de usuario

De manera seguida, se describe los principales controles de la interfaz gráfica principal del usuario:

- **Caja de texto Path**: esta caja de texto permite al usuario ingresar la dirección donde se encuentra el dataset.
- **Botón Cargar Conjunto de Entrenamiento**: este botón es utilizado para cargar el dataset de entrenamiento en memoria.
- **Botón Limpiar**: permite inicializar el estado del prototipo.
- **Botón Entrenar**: este botón da inicio al procesamiento de los datos y el respectivo proceso de entrenamiento de los modelos de clasificación.

Interfaz de configuración de procesos

En la Figura 6 se expone la interfaz de configuración del prototipo, en donde el usuario podrá seleccionar el o los procesamientos de datos de manera individual y combinada.

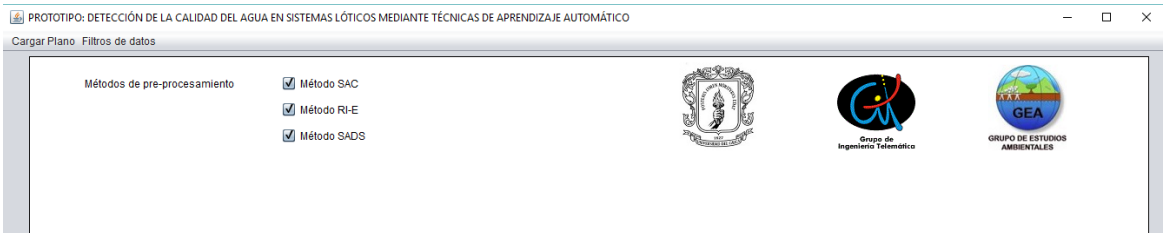

Figura 6. Interfaz de configuración de procesos

Una vez seleccionados los métodos de procesamiento de datos deseados, el usuario debe dirigirse a la interfaz principal (Figura 5) y presionar el botón *Entrenar* para dar inicio al entrenamiento de los clasificadores.

De igual manera, una vez finalizado el proceso de entrenamiento en la interfaz principal se mostrara los botones que se indican en la Figura 7.

- **Botón Cargar Conjunto de Prueba**: este botón es utilizado para cargar el dataset de prueba (dataset al que se le detectara la calidad del agua) en memoria.
- **Botón Ver Resultados de Entrenamiento**: permite visualizar los resultados del proceso de la detección de la calidad del agua.
- **Botón Volver a Entrenar**: este botón re-direcciona a la interfaz principal (Figura 5) y permite volver a entrenar los modelos de clasificación.
- **Botón Detectar la Calidad del Agua**: da inicio al proceso de detección de la calidad del agua.

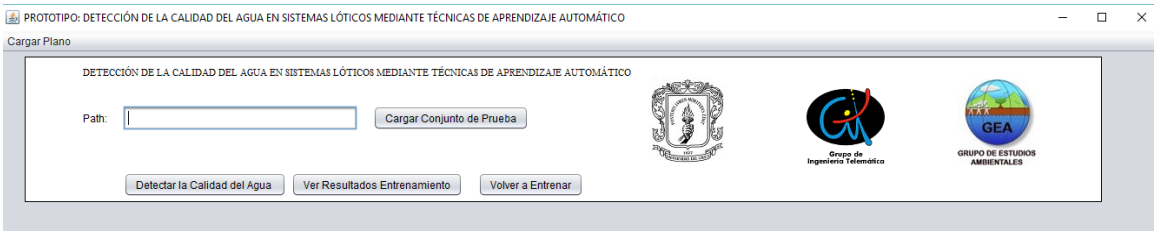

Figura 7. Interfaz principal – Resultados

Al presionar sobre el botón *Ver Resultados de Entrenamiento* se despliega el panel donde se exponen los resultados del proceso de entrenamiento obtenidos por el prototipo (Figura 8).

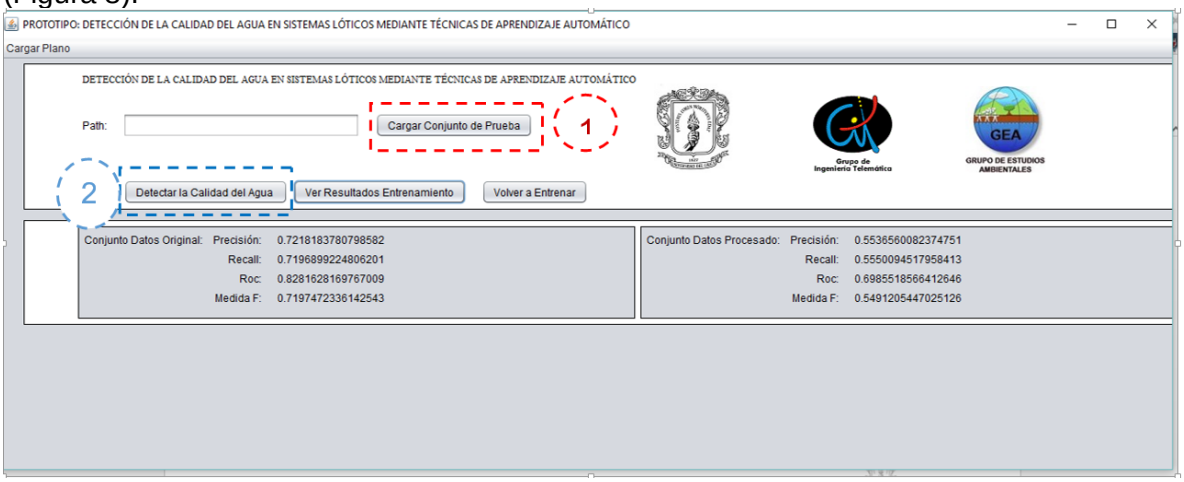

Figura 8. Interfaz principal – resultados de la detección de la calidad del agua

Como se indica en la figura anterior, el prototipo contrasta de manera visual los resultados obtenidos de entrenar los clasificadores con el dataset sin procesar (original) y el dataset procesado.

Por otra parte, el usuario puede realizar una limpieza de la interfaz principal presionando sobre el *Botón Limpiar*, y así volver al estado inicial del prototipo para poder empezar otro proceso.

Ahora, para detectar la calidad del agua el usuario debe cargar en el dataset de *prueba*, presionando sobre el botón *Cargar Conjunto de Prueba* y seleccionado el dataset respectivo. Una vez cargado dicho dataset, se procede a detectar la calidad del agua presionando sobre el botón *Detectar la Calidad del Agua (Figura 8)*

Figura 9. Resultado proceso de la detección de la calidad del agua

Como resultado final se despliega una ventana emergente que contiene la detección de la calidad del agua, tal y como se indica en la Figura 9.

REFERENCIAS ANEXOS

[1] V. Y. D. T. Ministerio De Ambiente, "Política Nacional para la Gestión Integral del Recurso Hídrico," *Ministerio de Ambiente, Vivienda y Desarrollo Territorial, Bogotá, D.C.: Colombia.,* p. 124, 2010 2010.

[2] L. F. Á. Arango María Cecilia, Gloria Alexandra Arango, Orlando Elí Torres & Asmed de Jesús Monsalve, "Calidad Del Agua De Las Quebradas La Cristalina Y La Risaralda, San Luis, Antioquia," *EIA,* pp. 121-141, Julio 2008 2008.

[3] A.-T. Javier, "Macroinvertebrados Acuaticos Y Calidad De Las Aguas De Los Rios," *IV Simposio del Agua en Andalucia,* vol. 2, pp. 203-213, 1996.

[4] D. M. G. Wilber Pino Chalá, Martha Lucia Mosquera, Kelly Patricia Caicedo, Jhon Arley Palacios, Anilio Alberto Castro & Jair Enrique Guerrero, "Diversidad De Macroinvertebrados Y Evaluación De La Calidaddel Agua De La Quebrada La Bendición, Municipio De Quibdó (Chocó, Colombia)," *Acta Biológica Colombiana,* vol. 8 No.2, p. 8, Octubre 2003 2003.

[5] M. P. N. F. Claudia Rico, "Modelación De La Estructura Jerárquica De Macroinvertebrados Bentónicos A Través De Redes Neuronales Artificiales Acta Biológica Colombiana," *Open Journal Systems,* vol. 3, pp. 71-96, 2009.

[6] T.-S. C. Young-Seuk Parka, Inn-Sil Kwak & Sovan Lek, "Hierarchical Community Classification And Assessment Of Aquatic Ecosystems Using Artificial Neural Networks," *Science of the Total Environment,* pp. 105-122, 2004.

[7] G. R. Pérez, *Bioindicación de la Calidad del Agua en Colombia: Propuesta Para el Uso del Método BMWP Col*, Primera ed. vol. 1: Universidad de Antioquia, 2003.

[8] K. P. S. S. Gupta, "Artificial Intelligence Based Modeling for Predicting the Disinfection by-Products in Water," *Chemometrics and Intelligent Laboratory Systems,* vol. 114, pp. 122–131, 15 May 2012.

[9] M.-J. B. Y.-S. Park, "Biological Early Warning System Based on the Responses of Aquatic Organisms to Disturbances: A Review," *Science of The Total Environment,* vol. 466-467, pp. 635–649, 1 January 2014 2014.

[10] H. T. Shuangyin Liu, Qisheng Ding, Daoliang Li, Longqin Xu & Yaoguang Wei, "A Hybrid Approach of Support Vector Regression With Genetic Algorithm Optimization for Aquaculture Water Quality Prediction," *Mathematical and Computer Modelling,* vol. 58, pp. 458-465, August 2013 2012.

[11] I. O. K. Bucak, Bekir, "Detection of Drinking Water Quality Using CMAC Based Artificial Neural Networks," *Ekoloji Dergisi,* vol. 20, pp. 75-81, 2011.

[12] G. R. Pérez, "Bioindicación De La Calidad Del Agua En Colombia: Uso Del Metodo Bmwp," vol. Primera edicion, p. 165, 2003.

[13] I. G. Olatz Arbelaitz, Javier Muguerza, Jesús María Pérez & Iñigo Perona, "An Extensive Comparative Study Of Cluster Validity Indices," *Pattern Recognition,* vol. 46, pp. 243-256 January, 2013 2013.

[14] J. M. Ibai Gurrutxaga , , Olatz Arbelaitz , Jesús M. Pérez & José I. Martín, "Towards A Standard Methodology To Evaluate Internal Cluster Validity Indices," *Pattern Recognition Letters,* vol. 32, pp. 505-515 February, 2011 2011.

[15] C.-R. L. M.-S. Chen, "Combining Partitional and Hierarchical Algorithms for Robust and Efficient Data Clustering with Cohesion Self-Merging," *IEEE Transactions on Knowledge and Data Engineering,* vol. 17, pp. 145-159, February 2005 2005.

[16] C. M. J. W. Guojun Gan, *Data Clustering: Theory, Algorithms, and Applications*. Philadelphia, Pennsylvania: Society for Industrial and Applied Mathematics, 2007.

[17] T. V. T. Santhanam, "Computational Complexity between K-Means and K-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points," *Journal of Computer Science,* vol. 6, pp. 363-368, 2010.

[18] S. Theodoridis, Koutroumbas & Konstantinos, *Pattern Recognition, Second Edition*. Crashing Rocks Books (Punta Gorda, FL, U.S.A.): Academic Press, 2003.

[19] M. S. V. K. Pang-Ning Tan, *Introduction to Data Mining*. Inc. Boston, MA, USA, 2005.

[20] A. H. Moreno, *La Clasificación Numérica y su Aplicación en la Ecología*: Santo Domingo, R.D. : Instituto Tecnológico de Santo Domingo, 2000, 2000.

[21] K. S. P.Baby, "Agglomerative Hierarchical Clustering Algorithm- A Review," *International Journal of Scientific and Research Publications,* vol. 3, March 2013 2013.

[22] A. H. S. S. Philipp Cimiano, "Comparing Conceptual, Divisive and Agglomerative Clustering for Learning Taxonomies from Text," *Proceedings of the 16th European Conference on Artificial Intelligence,* vol. 110, p. 435, 2004.

[23] T. S. Madhulatha, "An Overview on Clustering Methods," *IOSR Journal of Engineering,* vol. 2, pp. 719-725, Apr. 2012 2012.

[24] D. P. González, "Algoritmos de Agrupamiento basados en densidad y Validación de clusters " Doctoral, Departament de Llenguatges I Sistemas Informátics, Universitat Jaume I 2010.

[25] N. D. s. D. Sinan Saraçli, "Comparison of Hierarchical Cluster Analysis Methods by Cophenetic Correlation," *Journal of Inequalities and Applications,* vol. 2013, 23 Apr 2013 2013.

[26] J. M. Badia Contelles, Pla Bañón, Filiberto, Quirós Bauset, Ricardo Javier, Badia Contelles & José Manuel, "Métodos informáticos avanzados," *e-Treballs d'informàtica i tecnologia,* vol. 1, 2007.

[27] D. C. Corrales, "Toward Detecting Crop Diseases And Pest By Supervised Learning," *Ingeniería y Universidad,* vol. 19, 2015.

[28] S. W. Ambler, "**UML Package Diagrams** " 2005.

[29] E. F. Mark Hall, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations,* vol. 11, 2009.