TRABAJO SOLUCIÓN INTELIGENCIA DE NEGOCIO

Descripción
El trabajo realizado (documentación, presentación, código y datos) deberá subirse a la actividad correspondiente de la plataforma virtual. La fecha tope de entrega será el día del examen.

A continuación se pasa a detallar las características de las entregas del documento, el código, los datos y la presentación.

Documentación

Organizacón del documento

El documento debe constar de las siguientes secciones:

1. Introducción (extensión máxima 1 carilla). Debe explicar el problema que se pretende solucionar y qué beneficios se obtendría de desarrollar una solución de IN..
2. Plan. Debe cubrir los aspectos fundamentales sobre la etapa plan mencionados en el apartado anterior. Sólo deben desarrollarse los puntos que se adecuen al trabajo que se va a desarrollar.
3. Análisis. Debe cubrir los aspectos fundamentales sobre la etapa análisis mencionados en el apartado anterior. Sólo deben desarrollarse los puntos que se adecuen al trabajo que se va a desarrollar.
4. Diseño. Debe incluir los requerimientos identificados en el análisis para crear las especificaciones detalladas del diseño. Dichas especificaciones son:
    1. Capacidad de memoria de la organización. Estas especificaciones conciernen a la identificación de los datos de origen tanto de la organización como externa a ella. Además incluirá los procesos ETL necesarios para mover los datos de origen a la solución de IN.
    2. Capacidad de integración de información. Estas especificaciones conciernen a las posibles incidencias que puedan encontrarse en las fuentes de datos así como las formas de afrontar dichas inconsistencias para poder realizar una integración efectiva de la información.
    3. Capacidad de crear conocimiento. Estas especificaciones se centran en los distintos análisis de los datos. Es importante identificar los procesos que serán usados para describir el pasado, resumir la situación actual, e identificar predicciones futuras. Esto deberá realizarse para las distintas áreas (inventario, ventas, etc.), departamentos (recursos humanos, IT, etc.), roles (supervisor, vendedor, presidente de un departamento, etc.) y tareas (predicción de ventas, visualización de de cambios de productos, etc.) de la solución de IN.
    4. Capacidad de presentación. Estas especificaciones se centran en las formas de las presentaciones (panel de control, reportes, etc.), tiempos (intervalos regulares, bajo demanda, etc.), y medios (aplicación de escritorio y/o web, internet, intranet, etc.).
5. Implementación. Esta etapa incluye las tareas necesarias para desarrollar una solución de IN. Esta etapa incluirá información de las características técnicas necesarias:
    1. Pasos para la instalación y configuración de los datos.

2. Software necesario para el análisis de los datos.
3. Software necesario para crear los reportes.
4. Etc.

6. Despliegue. Esta etapa incluye casos de uso de la solución de IN desarrollada.
7. Conclusiones. Debe indicarse de forma clara y breve (extensión máxima media carilla) qué problema se resuelve y qué ventajas ofrece la IN.

Se recomienda que en la medida de lo posible, el documento incluya figuras y diagramas que faciliten su lectura.

Extensión

El documento que se entregue debe tener una extensión mínima de 20 páginas y máxima de 35 por ambas caras. Dicha extensión se refiere a los puntos destacados en el punto Organización del documento.

Código y datos

La entrega deberá incluir la solución de IN desarrollada por el alumno así como los datos necesarios para poder ejecutarla. En caso de utilizar datos sensibles de una empresa, contactar con el profesor a través del correo mgarciat@upo.es.

Se deberá incluir un manual breve con indicaciones paso a paso para la instalación de la solución de IN desarrollada.

Solución de Inteligencia de Negocio

BOMBAS DE AGUA DE TANZANIA

Autores

Manuel Herrera Pulido

José Manuel Fernández Labrador

Daniel Barciela Rueda

INDICE

1. Introducción

El problema que hemos elegido para nuestro proyecto de Inteligencia de Negocio es sobre las bombas de agua sobre el territorio de Tanzania, país situado en la costa este de África central, donde el fallo de alguna bomba de agua puede provocar numerosos estragos sobre la población que se abastece de dicha bomba, provocando en el peor de los casos la pérdida de vidas humanas.

La necesidad de realizar un proyecto de inteligencia de negocio sobre estos datos recae sobre que, si realizamos una compresión inteligente de los puntos de agua, podremos obtener la información de la distribución de las bombas de agua repartidas por todo el territorio, con lo que mejoraremos la toma de futuras decisiones.

Hemos abordado el problema, repartiendo el trabajo en diferentes etapas (estudio, análisis, diseño, implementación), abordando cada etapa con diferentes tecnologías como son Pentaho, Weka y R Studio, con las que hemos conseguido el objetivo de mostrar la información necesaria del estado de cada bomba de agua.

2. Plan

2.1 Objetivos

El objetivo general de nuestra propuesta de inteligencia de negocio consistirá en proporcionar una información útil, estructurada y con un conjunto de indicadores que muestren la situación actual de cada bomba, con lo que ayude al país de Tanzania a la toma de correctas decisiones en el futuro.

Objetivos específicos

- Investigar el entorno de la Inteligencia de Negocio y los procesos de extracción de agua, para una toma de conocimiento actual y para un uso futuro.
- Diseñar una solución de Inteligencia de negocio, relacionando las fuentes de datos, para la extracción de la información.
- Testear la solución obtenida.

2.2 Plan de trabajo

2.2.1 Solución de los objetivos

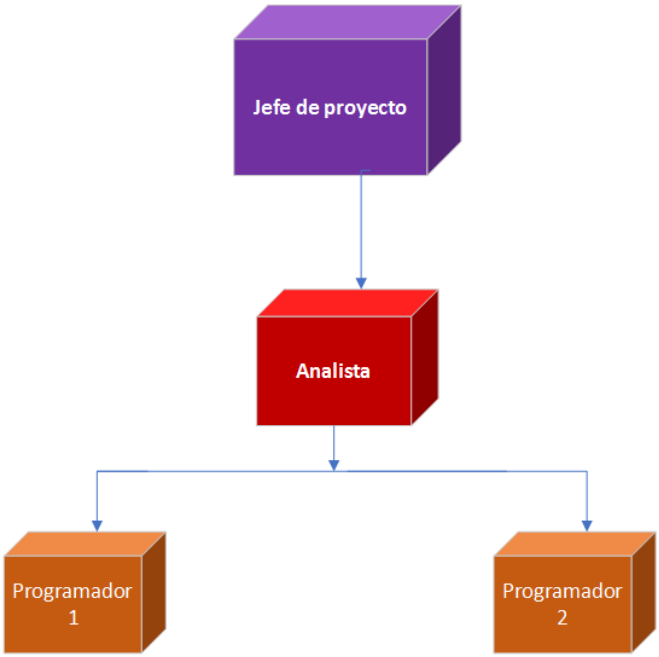
Para poder solucionar los diferentes problemas anteriormente planteados, se desea buscar una herramienta de inteligencia de negocios capaz de abordar todas esas dificultades.

Para realizar esa solución, se abordarán las siguientes soluciones:

- Optimizar el tiempo.
- Automatizar la extracción de datos.
- Automatización de procesos.
- Ayuda al análisis de la información.
- Visualizar la información.

2.2.2 Organización del proyecto

El equipo del proyecto estará formado por:



Las principales responsabilidades y funciones de los miembros del equipo de proyecto serán las siguientes:

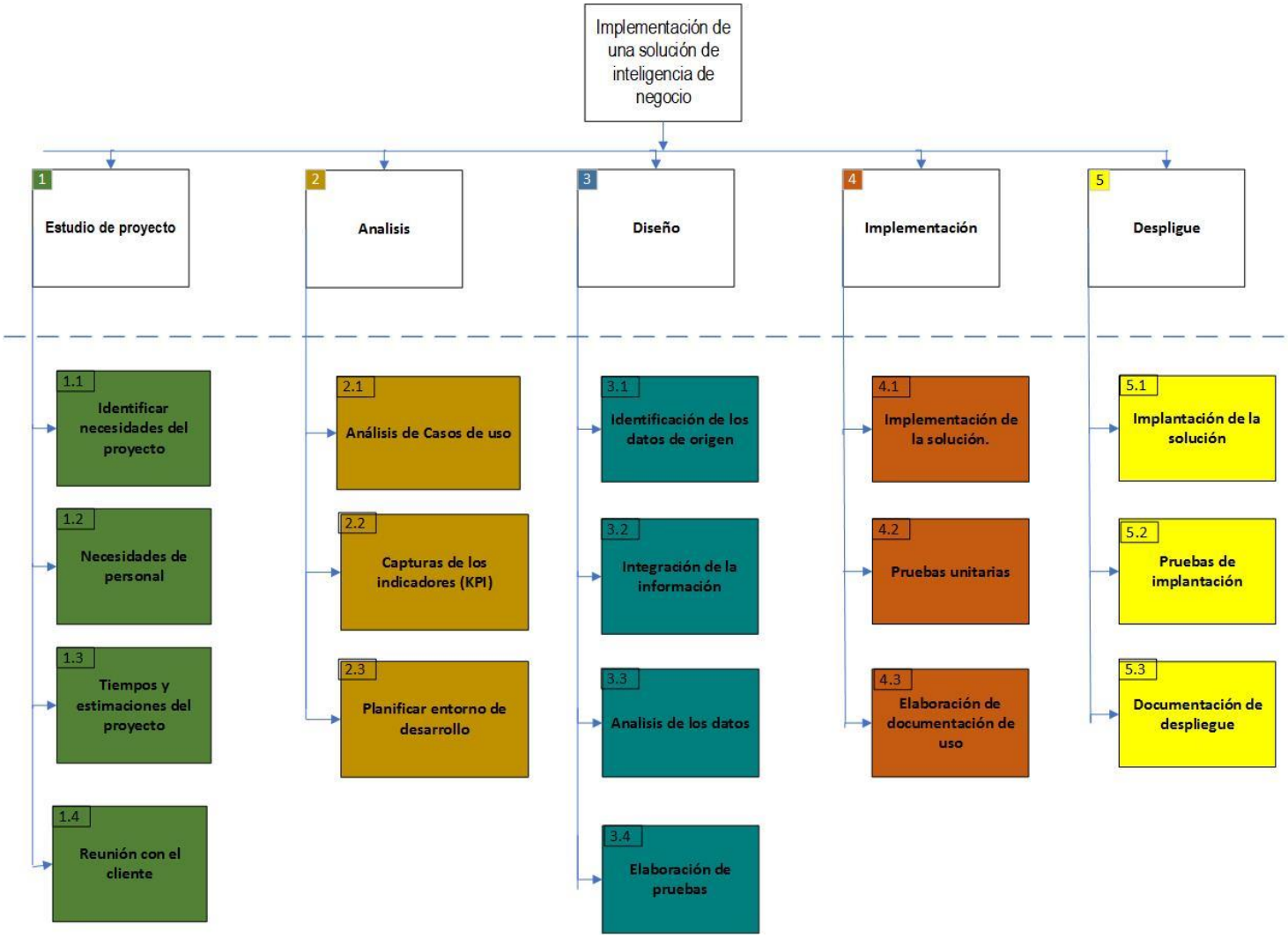
- Jefe de Proyecto:
  - Identificar necesidades del proyecto.
  - Identificar necesidades de personal necesario.
  - Seguimiento del proyecto.
  - Reuniones con el cliente.
  - Estimaciones de tiempos y costes.
  - Puesta en marcha del proyecto.

- Analista:
  - Análisis de los requerimientos del proyecto.
  - Seguimiento del proyecto (parte de los programadores).
  - Análisis de los datos.
  - Documentación del proyecto.
  - Aprobar los procesos de construcción ETL.
  - Desarrollo de casos de prueba.
  - Despliegue de la aplicación.

- Programadores:
  - Implementación de las funcionalidades del sistema.
  - Limpiado de los datos.
  - Construir los procesos ETL.
  - Desarrollo de pruebas unitarias.

2.2.3 Plan de tareas

En cuanto al desglose estructurado de tareas, se han identificado las siguientes fases y tareas en el proyecto, que se representan en el EDT que se muestra en la siguiente figura:



En lo que respecta a cada tarea, se detallan a continuación, incorporando la responsabilidad de llevarla a cabo y la estimación de cada una de ellas, teniendo en cuenta la distribución de las tareas por fases:

- **Estudio de proyecto**

Código	Nombre	Responsable	Estimación (horas)
1.1	Identificar necesidades del proyecto	Jefe de Proyecto, Analista	16
Descripción			
Tras conocer en detalle la propuesta del cliente a desarrollar, se realiza un estudio del proyecto en general, identificando las necesidades hardware y software, un planteamiento general de las funcionalidades que necesita la aplicación para llevar a cabo con éxito los distintos objetivos propuestos, así como la viabilidad de este.			

Código	Nombre	Responsable	Estimación (horas)
1.2	Necesidades de personal	Jefe de Proyecto, Analista	8

Descripción

Análisis detallado sobre el personal requerido para realizar el proyecto. En este caso se plantea que el proyecto necesita un jefe para supervisión de todas las fases de este. Un analista, para las fases de análisis y diseño, junto a la documentación. Dos programadores, puesto que en la fase de implementación y pruebas será necesario varios trabajadores al mismo tiempo para avanzar más rápido, aparte que la baja inesperada de alguno permitiría cumplir con los tiempos estimados.

Código	Nombre	Responsable	Estimación (horas)
1.3	Tiempos y estimaciones del proyecto	Jefe de Proyecto, Analista	8

Descripción

Estimación del tiempo dedicado en cada fase, y el coste del proyecto.

Se tiene en cuenta que puede haber factores inesperados que alteren los tiempos, por lo que se estiman más tiempo del necesario en cada tarea, para que, en caso de algún contratiempo, se puedan cumplir los plazos.

Código	Nombre	Responsable	Estimación (horas)
1.4	Reunión con el cliente	Jefe de Proyecto, Analista	8

Descripción

La última tarea antes de completar la fase de estudio del proyecto. Una vez se ha planificado de forma general el proyecto en los distintos niveles, y se ha estimado el coste y tiempos, se le presenta al cliente toda esta documentación para mantenerlo informado y para comprobar que el enfoque es el correcto.

Si todo está de acuerdo se sigue adelante, en caso contrario, se hacen las modificaciones pertinentes.

➢ Análisis

Código	Nombre	Responsable	Estimación (horas)
2.1	Análisis casos de uso	Analista	8

Descripción

Análisis de las funcionalidades de la aplicación a desarrollar.

Código	Nombre	Responsable	Estimación (horas)
2.2	Capturas de los indicadores (KPI)	Analista	8

Descripción

Se realiza un análisis para la captura de los indicadores más importantes que nos mostraran la situación de los datos.

Código	Nombre	Responsable	Estimación (horas)
2.3	Planificar entorno de desarrollo	Analista	8

Descripción

Se analiza y se decide el entorno de desarrollo que se va a emplear durante el proyecto. Es decir, se considera las tecnologías que se van a utilizar, los distintas herramientas o entornos en los que se va a trabajar.

- **Diseño**

Código	Nombre	Responsable	Estimación (horas)
3.1	Identificación de los datos de origen	Analista	16

Descripción

Se realizará un análisis del origen de los datos, analizando las diferentes fuentes que puedan provenir los datos.

Código	Nombre	Responsable	Estimación (horas)
3.2	Integración de la información	Analista	8

Descripción

Se realizará la unificación de la información mediante procesos ETL, con los que obtendremos unos datos más estándar.

Código	Nombre	Responsable	Estimación (horas)
3.3	Análisis de los datos	Analista	16

Descripción

Se realizará un análisis de los datos, donde podremos comprobar el estado de los datos actuales.

Código	Nombre	Responsable	Estimación (horas)
3.4	Elaboración de pruebas	Analista, Programador1, Programador2	16

Descripción

Diseño de los casos de prueba generales que necesitarán hacerse en la fase de implementación y en la de despliegue de la aplicación para comprobar que todo el proceso de análisis está integrado correctamente y como se espera.

- **Implementación de la solución**

Código	Nombre	Responsable	Estimación (horas)
4.1	Implementación de la solución	Analista, Programador1, Programador2	16

Descripción

Implementación de todos los procesos ETL y análisis en la infraestructura. Será necesario documentar cada funcionalidad implementada para que todo el equipo de trabajo tenga facilidad de compresión.

Código	Nombre	Responsable	Estimación (horas)
4.2	Pruebas unitarias	Programador1, Programador2	8

Descripción

Una vez realizada la implementación, se llevarán a cabo numerosas pruebas para comprobar que todas las funcionalidades funcionan como se han de esperar. En caso de error, se documenta y se procede a su solución.

Código	Nombre	Responsable	Estimación (horas)
4.3	Elaboración de documentación de uso	Analista, Programador1, Programador2	16

Descripción

Se realiza un documento de uso del sistema, principalmente dirigido a los usuarios finales, con el máximo detalle posible.

- **Despliegue**

Código	Nombre	Responsable	Estimación (horas)
5.1	Implantación de la solución	Jefe de Proyecto, Analista, Programador1, Programador 2	16

Descripción

Se pasa la aplicación de local a los servidores de los usuarios finales.

Código	Nombre	Responsable	Estimación (horas)
5.2	Pruebas de implantación	Analista, Programador1, Programador2	8

Descripción

Una vez subido la aplicación a los servidores, se realiza:

- Pruebas de integración a nivel de servidor, para comprobar si algo que antes funcionaba en local, puede dar problemas en los servidores por configuraciones o compatibilidades.
- Pruebas de integración a nivel de la arquitectura de la aplicación. Se vuelve a comprobar que todo en conjunto funciona completamente.
- Pruebas unitarias de funcionalidades pequeñas en concreto. Se vuelve a comprobar cada funcionalidad para que el resultado sea el esperado.

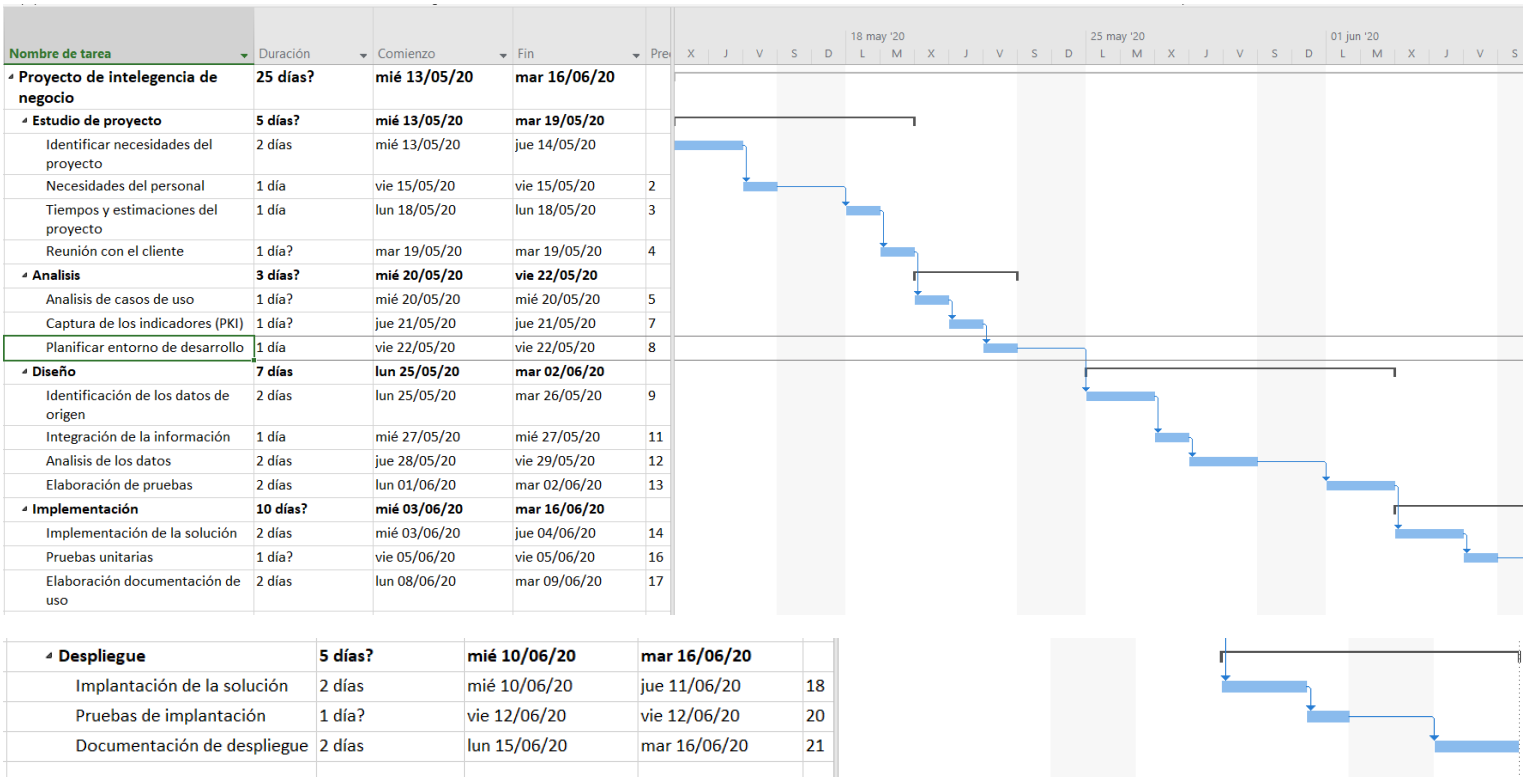
Código	Nombre	Responsable	Estimación (horas)
5.3	Documentación de despliegue	Analista	16

Descripción

Documentación de toda la fase de despliegue. Al mismo tiempo que se sube la aplicación y se realiza todas las tareas de pruebas, se tiene que ir documentando, tanto lo que funciona como lo que no.

2.2.4 Planificación temporal

Se presenta a continuación una planificación del proyecto, en lo que se refiere a los módulos de explotación estadística, con un diagrama de Gantt, que contiene las fases y tareas descritas en el EDT, que se ha presentado:



Nombre de tarea	Duración	Comienzo	Fin	Pre
Proyecto de intelegencia de negocio	**25 días?**	**mié 13/05/20**	**mar 16/06/20**	
Estudio de proyecto	**5 días?**	**mié 13/05/20**	**mar 19/05/20**	
Identificar necesidades del proyecto	2 días	mié 13/05/20	jue 14/05/20	
Necesidades del personal	1 día	vie 15/05/20	vie 15/05/20	2
Tiempos y estimaciones del proyecto	1 día	lun 18/05/20	lun 18/05/20	3
Reunión con el cliente	1 día?	mar 19/05/20	mar 19/05/20	4
Analisis	**3 días?**	**mié 20/05/20**	**vie 22/05/20**	
Analisis de casos de uso	1 día?	mié 20/05/20	mié 20/05/20	5
Captura de los indicadores (PKI)	1 día?	jue 21/05/20	jue 21/05/20	7
Planificar entorno de desarrollo	1 día	vie 22/05/20	vie 22/05/20	8
Diseño	**7 días**	**lun 25/05/20**	**mar 02/06/20**	
Identificación de los datos de origen	2 días	lun 25/05/20	mar 26/05/20	9
Integración de la información	1 día	mié 27/05/20	mié 27/05/20	11
Analisis de los datos	2 días	jue 28/05/20	vie 29/05/20	12
Elaboración de pruebas	2 días	lun 01/06/20	mar 02/06/20	13
Implementación	**10 días?**	**mié 03/06/20**	**mar 16/06/20**	
Implementación de la solución	2 días	mié 03/06/20	jue 04/06/20	14
Pruebas unitarias	1 día?	vie 05/06/20	vie 05/06/20	16
Elaboración documentación de uso	2 días	lun 08/06/20	mar 09/06/20	17
Despliegue	**5 días?**	**mié 10/06/20**	**mar 16/06/20**	
Implantación de la solución	2 días	mié 10/06/20	jue 11/06/20	18
Pruebas de implantación	1 día?	vie 12/06/20	vie 12/06/20	20
Documentación de despliegue	2 días	lun 15/06/20	mar 16/06/20	21

2.3 Análisis de la viabilidad

2.3.1 Beneficio

En esta parte se detalla una aproximación del tiempo que tardará en realizar la nueva herramienta de inteligencia de negocio los diferentes procesos.

FUNCIÓN	TIEMPO
Extracción de los datos	2 minuto Aproximado.
Transformación y manipulación de los datos	2 minuto Aproximado
Análisis de la información	8 minutos Aproximados
Realización de gráficos y tablas.	8 minutos Aproximados
Tiempo total	**20 minutos Aproximados**

Como podemos comprobar el tiempo aproximado que puede tardar nuestra propuesta de inteligencia de negocio, es un tiempo muy bajo.

Por lo que obtendremos un beneficio de bastante información, en un pequeño periodo de tiempo.

2.3.2 Coste Económico

En este apartado podemos ver los gastos económicos que tendrá la solución de inteligencia de negocio.

Software necesario.

Programa	Precio	Licencias
Pentaho Data Integration	0€	4
Weka	0€	4
R-Studio	0€	4
Total	**0€**	**12**

Podemos comprobar que el coste económico empleado en software es de 0€, debido a que los 3 software que utilizamos son open source.

Personal

Trabajador	Precio por día	Días Trabajados	Total
Jefe de Proyecto	80€	25	2000€
Analista	70€	25	1750€
Programador 1	60€	15	900€
Programador 2	60€	15	900€
		Total	**5550€**

El coste aproximado del personal seria de 5550€, este coste abarcaría el proceso completo del desarrollo de la implantación del proceso de inteligencia de negocio.

2.4 Riesgos y tecnología usada

2.4.1 Riesgos

Los riesgos que se han identificado inicialmente se muestran a continuación, ya evaluados y con alguna traza de las acciones propuestas en caso de que se produzcan las incidencias.

Id. Del Riesgo	Origen del riesgo	Descripción del riesgo	Probabilidad	Severidad	Plan contingencia	Prioridad
R01	Riesgos originados por aspectos tecnológicos	Entorno de desarrollo mal planificado (mala elección de tecnologías, equipo no preparado para trabajar, etc)	Media	Alta	Contactar con el personal encargado de dicha tarea para que haga una evaluación del caso	Moderado
R02	Riesgos que provienen de la gestión del proyecto	No se aplican estándares de diseño y desarrollo en el proyecto	Baja	Baja	Definir unos estándares mínimos de diseño y construcción para el proyecto	Muy leve
R03	Riesgos que provienen de los aspectos conceptuales	Estimación de plazos afectada por baja de algún empleado	Baja	Media	Aumentar número de horas de empleados con tareas similares para cumplir los plazos previstos	Leve
R04	Riesgos que provienen de la gestión del proyecto	Problemas para cerrar el proyecto (normalmente a causa del cliente)	Baja	Baja	Negociar con el cliente para cerrar el proyecto	Leve
R05	Riesgos originados por aspectos tecnológicos	Ataque a la seguridad del sistema	Baja	Alta	Aplicar plan de seguridad según el caso concreto	Moderado
R06	Riesgos originados por terceros	Los datos suministradores por terceras partes no están correctos	Media	Media	Aplicar plan para investigar cómo se obtienen los datos	Moderado
R07	Riesgos originados por aspectos tecnológicos	Posibilidad de que destruyan, dañen, borren, deterioren, alteren o supriman datos informáticos	Media	Media	Realizar Backups de la información	Moderado
R08	Riesgos que provienen de los aspectos conceptuales	Definición incorrecta de los indicadores para medir el desempeño	Media	Alta	Realizar una redefinición de los indicadores	Alta

2.4.2 Tecnologías usadas

La tecnología que hemos usado en nuestro proyecto de inteligencia de negocio han sido los programas Pentaho Data Integration y Weka, los 2 son de software libre

- **Pentaho Data Integration**
  Esta aplicación se ha utilizado para realizar las técnicas ETL, con las que hemos podido implementar los procesos de extracción, transformación y carga de datos, en concreto hemos utilizado la versión 9.0.

- **Weka**
  Waikato Environment for Knowledge Analysis (WEKA) es una librería Java de machine learning desarrollada en la Universidad de Waikato, Nueva Zelanda. Con la aplicación Weka hemos resuelto las tareas de preprocesado, en concreto hemos utilizado la versión 3.9.4.

- **R Studio**
  R Studio es la plataforma grafica de desarrollo de R, con esta herramienta hemos aplicado algoritmos de clustering a los datos y hemos realizado visualizaciones de los datos.

3. Análisis

3.1 Establecimiento de los requisitos del sistema

- Definir requisitos funcionales: los KPI (Indicador de rendimiento Clave) son el centro de toda aplicación de inteligencia de negocio. Se debe determinar qué información se debe suministrar, cuándo, y en qué formato

- Involucrar a los Usuarios: se deberían crear prototipos de solución inicial. Así, se pueden llevar a cabo revisiones y verificar si los requerimientos se están cumpliendo desde el principio.

- Obtener apoyo de la Dirección: para incorporar objetivos a largo y corto plazo. Esto a través de un monitoreo permanente de KPIs (Indicador de rendimiento Clave) más importantes.

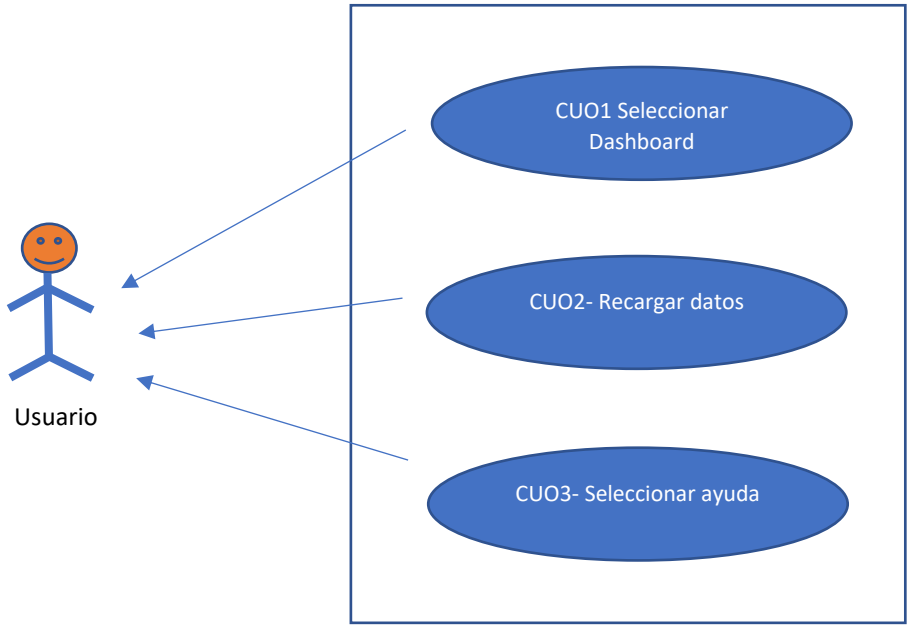
- Asegurar Integridad y Calidad de Datos: El equipo debe identificar los sistemas operacionales en los que la información requerida está disponible y cómo se debe acceder a esos datos.

- Conocer Herramientas de inteligencia de negocio ya disponibles: Cuando se empieza un proyecto nuevo, hay que decidir si las herramientas existentes de usuario final deben continuar siendo utilizadas o si debiesen ser totalmente reemplazadas. En la mayoría de los casos, es preferible estandarizar en un único sistema para asegurar la consistencia del suministro de información a través de la empresa.

- Elegir la aplicación, software o herramienta de inteligencia de negocio adecuada. Con una Prueba de Concepto, el equipo del proyecto toma la decisión final sobre el software más adecuado basándose en un informe específico.

- Poseer evolución constante del proyecto: Las necesidades de las empresas varían constantemente. Todas las soluciones de inteligencia de negocio deben desarrollarse y optimizarse continua y permanentemente, para poder satisfacer los requerimientos de la compañía.

3.2 Análisis de casos de uso



➢ CU01 – Seleccionar Dashboard

CU-01	Seleccionar Dashboard
Descripción	El sistema mostrara un panel con diversos tipos de gráficas, que mostraran los indicadores con toda la información.
Dependencias	-
Actores	Usuario
Precondición	-
Postcondición	-
Flujo Normal	1. El usuario accede a seleccionar dashboard. 2. El sistema muestra las gráficas y la información.
Flujos Alternativos	2. El sistema no tiene información que mostrar. 3. El usuario accede al CU02 – Recargar datos. 4. El sistema realiza la recarga de datos. 5. El usuario accede a seleccionar dashboard. 6. El sistema muestra las gráficas y la información.

➢ CU02 – Recargar Datos

CU-02	Recargar Datos
Descripción	El sistema realizara una carga de los datos en la aplicación.
Dependencias	-
Actores	Usuario
Precondición	-
Postcondición	-
Flujo Normal	1. El usuario accede a recargar datos. 2. El sistema solicita que indique los nuevos archivos de datos. 3. El usuario indica los archivos a procesar. 4. El sistema notifica la correcta subida de los archivos.
Flujos Alternativos	-

➢ CU03 – Seleccionar ayuda

CU-03	Seleccionar ayuda
Descripción	El sistema mostrara el manual de la aplicación.
Dependencias	-
Actores	Usuario
Precondición	-
Postcondición	-
Flujo Normal	1. El usuario accede a seleccionar ayuda. 2. El sistema mostrara el manual de ayuda de la aplicación.
Flujos Alternativos	-

3.3 Especificación del plan de pruebas

Pruebas Para Realizar	Descripción
Prueba de Carga	Se verificará que los datos del dataset son cargados correctamente
Prueba de Modelos de Carga	Se verificará que los modelos de carga estén tomando los datos de buena manera
Pruebas de Rendimiento	Se realizara una muestra completa con las herramientas pentaho ,weka y Rstudio , para tomar los tiempos respectivos de ejecución de los procesos.
Pruebas de Velocidad	Se testeará la velocidad de ejecución, carga y desempeño del transcurso de los procesos.

4. Diseño

4.1 Identificación de los datos

Los datos han sido suministrados por el ministerio de agua de Tanzania y la plataforma open-source Taarifa, plataforma que sirve para la generación de informes y el triaging de problemas relacionados con la infraestructura.

El dataset facilitado consta de 26 variables y unas 14400 instancias, en la siguiente tabla podemos ver la descripción y el tipo de cada variable:

ATRIBUTO	TIPO DE ATRIBUTO	DESCRIPCIÓN
Fecha	Nominal	Indica la fecha de instalación de la bomba.
Financiador	Nominal	Indica la empresa u organización que fue el financiador del proyecto.
Gps	Numérico	Indica las coordenadas GPS del pozo.
Instalador	Nominal	Indique el nombre del instalador de la bomba.
longitud	Numérico	Indica la longitud geográfica del pozo.
latitud	Numérico	Indicada la latitud geográfica del pozo.
Nombre_wpt	Nominal	Nombre del punto de agua
cuenca	Nominal	Nombre de la cuenca donde está situado el pozo.
subpoblacion	Nominal	Comunidad a la que pertenece el pozo.
region	Nominal	Región de Tanzania en la que está situado el pozo.
Código_region	Numérico	Código numérico que identifica a cada región.
Código_distrito	Numérico	Código del distrito al que pertenece dicho pozo.
poblacion	Numérico	Número de habitantes que se abastecen de dicho pozo.
organización	Nominal	Organización que se encarga del mantenimiento del pozo.
Nombre_esq	Nominal	Nombre de la organización
permiso	Nominal	Variable que indica si obtuvieron permiso o no para la instalación del pozo.
Ano_construcion	Numérico	Año de construcción del pozo.
Tipo_extracion	Nominal	Tipo de extracción del agua del pozo.
Grupo_tipo_extracion	Nominal	Nombre del grupo que se encarga de la extración
Grupo_gestion	Nominal	Nombre del grupo que gestiona el mantenimiento del pozo.
pago	Nominal	Método de pago
Calidad_agua	Nominal	Calidad del agua extraída del pozo.
Calidad_grupo	Nominal	Calidad del grupo que lleva el mantenimiento del pozo.
cantidad	Nominal	Caudal de agua que suministra el pozo.
fuente	Nominal	Tipo de fuente
Tipo_PuntoAgua	Nominal	Tipo de punto de agua

4.1.1 Procesos ETL

Los procesos ETL han sido desarrollados en la aplicación de Pentaho Data Integratión, en la que hemos repartido el trabajo en 2 transformaciones que se unificaran en un trabajo.

- Transformación 1

En esta primera transformación realizaremos un renombre de los diferentes atributos del dataset, también se realiza una eliminación de los atributos innecesarios, ya que más de uno nos proporcionaba información redundante, por últimos eliminaremos las instancias duplicadas que podamos encontrar, con lo que conseguiremos un proceso de limpieza de nuestro dataset.

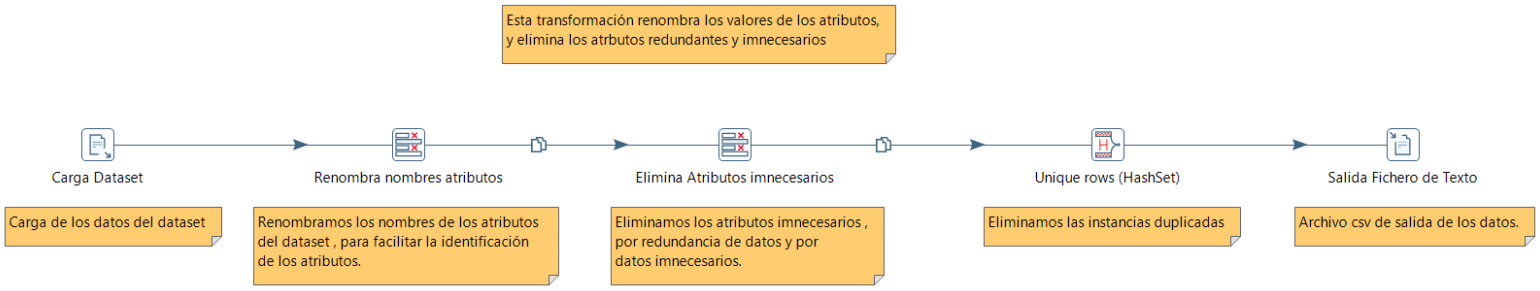


Imagen 1. Proceso de Pentaho Data Integration

- Transformación 2

En esta segunda transformación realizaremos diversos reemplazos de diferentes valores en las instancias, con lo que conseguiremos unos valores más unificados en cada variable.

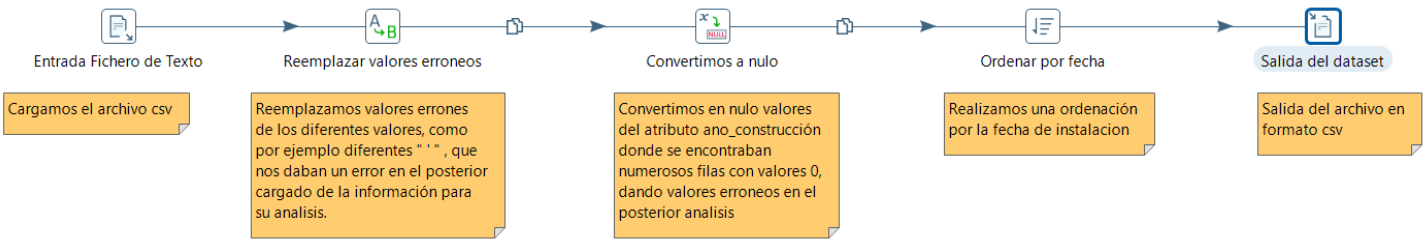


Imagen 2. Proceso de Pentaho Data Integration

- Trabajo

Trabajo general donde realizaremos las 2 transformaciones anteriormente especificadas.

Imagen 3. Trabajo general de Pentaho Data Integration

4.2 Análisis de datos

4.2.1 Preprocesamiento

Una vez hemos realizado la extracción y unificación de la información, pasaremos a la etapa de preprocesamiento de los datos, con lo que conseguiremos eliminar el ruido e inconsistencias de los datos originales, dejando unos datos más puros.

Para la realización del preprocesamiento utilizaremos la aplicación Weka en su versión 3.9.4.

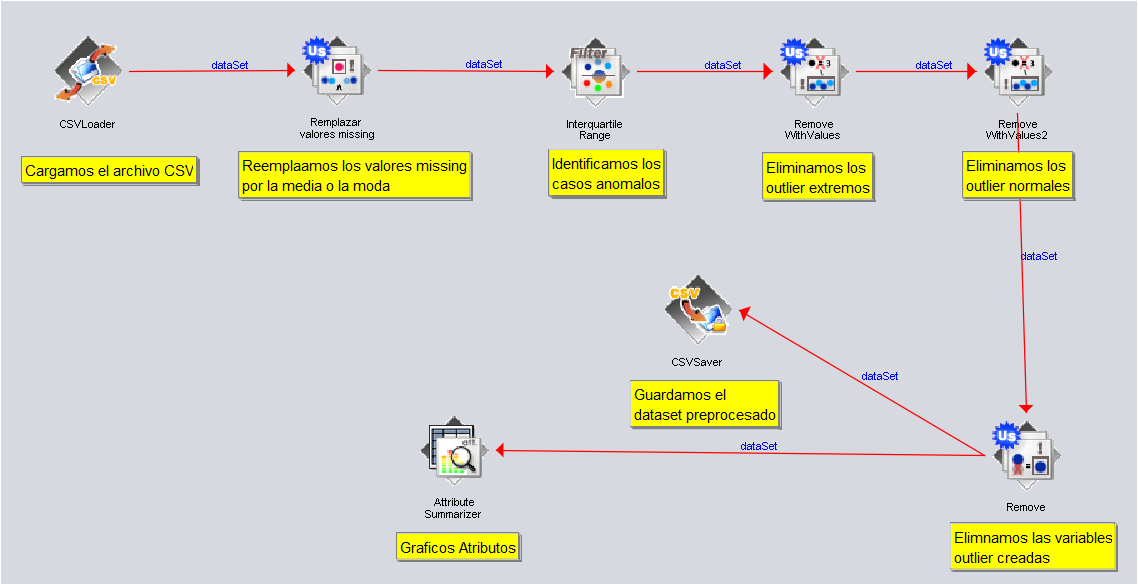


Imagen 4. Proceso de preprocesamiento realizado con la herramienta de Weka knowledgeFlow

Donde podemos ver los diferentes procesos que se llevan a cabo:

1. Realizamos un reemplazo de los valores missing, donde cada valor será reemplazado por la media o su valor modal.

2. Identificamos los valores anómalos, creando 2 variables en las que se indicara los valores outlier normales y los valores outlier extremos.

3. Eliminación de los casos anómalos detectados en el proceso anterior, este proceso es dividido en 2 pasos.

Una vez realizado el preprocesado de los datos, podemos ver la situación actual de los datos, aplicando métodos de visualización, donde podremos ver diferentes gráficos estadísticos:

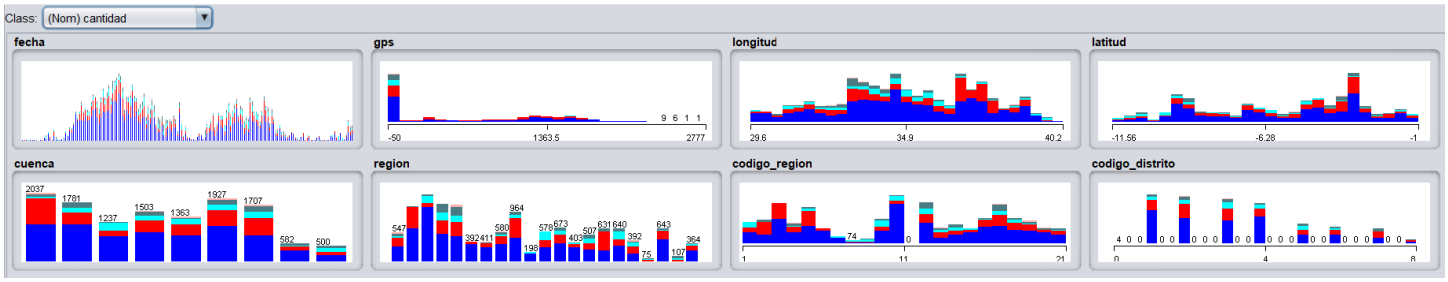


Imagen 5. Imagen extraída del módulo Attribute Summarizer de Weka

4.2.1 Visualización y clustering de los datos

En este apartado realizaremos la visualización de los datos sobre el mapa geográfico de Tanzania, y aplicaremos una técnica de clustering sobre el conjunto de datos, donde comprobaremos las relaciones de las diferentes instancias.

Este proceso será realizado completamente en la plataforma de RStudio.

Visualización

Para la visualización de los datos utilizaremos dos API de Google:

- Maps Static API.
- Geocoding API.

Con las que mostraremos la distribución de las diferentes instancias de los datos sobre el mapa geográfico de Tanzania.

Se mostrarán los atributos más relevantes, aunque se podrán mostrar cualquier atributo de los datos realizando un pequeño cambio de variable, por defecto se mostraran las variables:

- Cantidad de agua.
- Región.
- Cuenca.
- Calidad de agua.

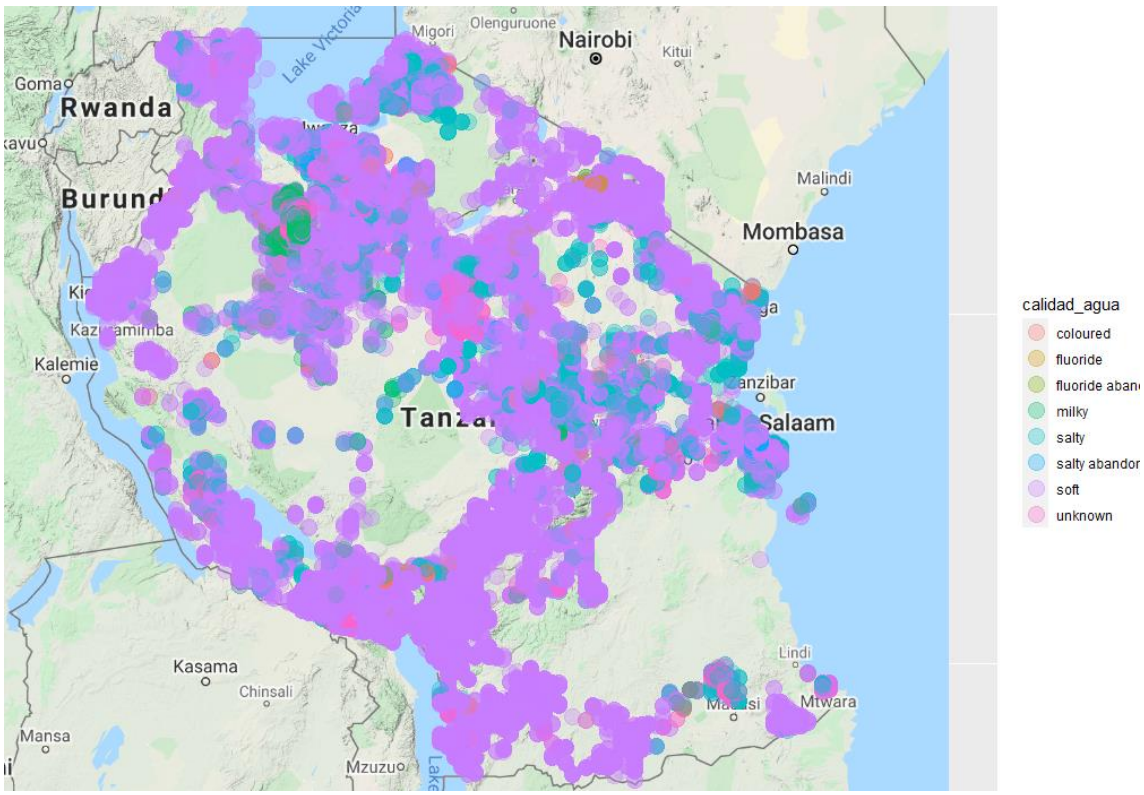


Imagen 6. Imagen que muestra la calidad del agua de cada bomba.

En la anterior imagen, podemos comprobar el resultado obtenido cuando aplicamos a nuestra vista la variable de calidad de agua. Donde podemos observar a simple vista que el tipo de agua más común es el soft (suave), por lo que podemos comprobar que estas imágenes pueden proporcionar bastante información con un simple vistazo.

Otra vista que pueda resultar muy interesante es la que nos proporciona la información de la cuenca hidrográfica de la que se alimenta cada bomba de agua.

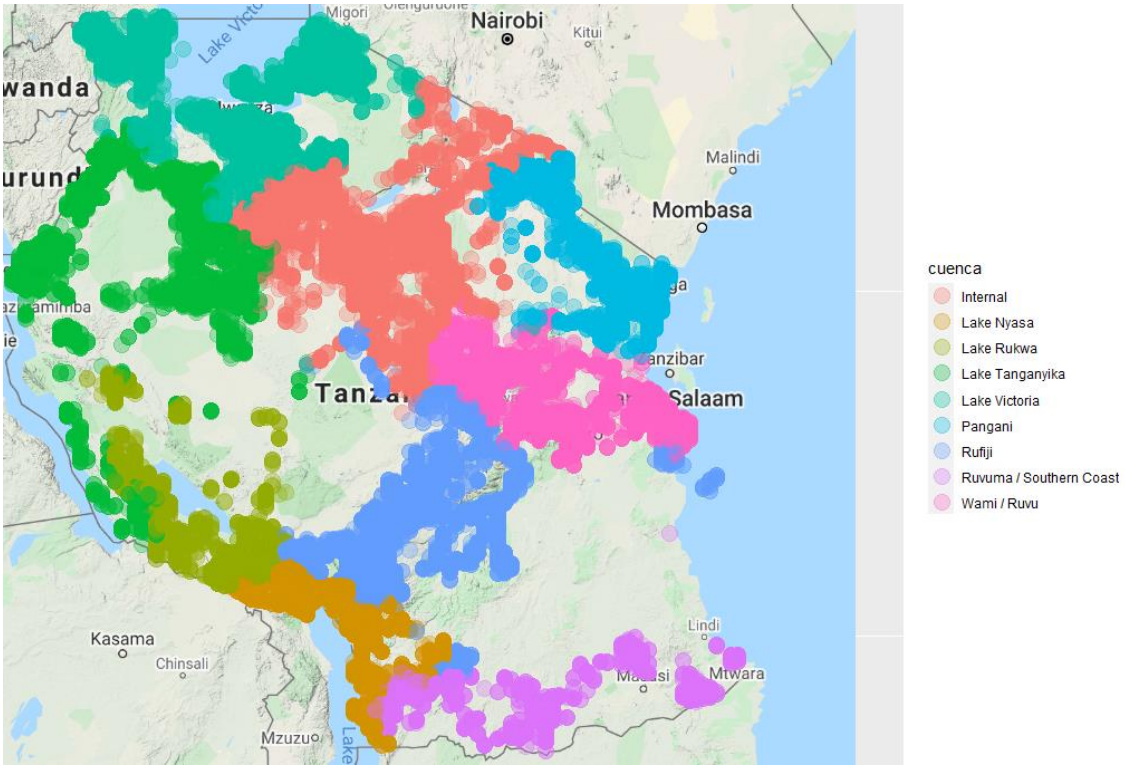


Imagen 7. Imagen que muestra la cuenca hidrográfica de cada bomba.

Para las presentaciones de estos mapas, se podrán exportar a archivos PDF o imágenes de diversos formatos (PNG, JPEG, TIF, BMP, Metafile, SVG, EPS) desde la aplicación RStudio.

Clustering

El clustering nos permite comprender mejor como una muestra de datos, puede estar compuesta por subgrupos distintos, que tienen similitudes entre sus variables.

Cuando se quiere aplicar un algoritmo de clustering, hay tres decisiones que deben tomarse:

- Elegir el medidor de distancia que utilizaremos.
- Elegir un algoritmo de agrupamiento.
- Elegir el número de clústeres idóneo.

En nuestro caso, el dataset de estudio está formado por variables de tipo numérico y nominal, por lo que para realizar las medidas de distancias de este tipo de dataset hemos elegido la distancia de Gower, que es utilizada para medir las distancias entre variables mixtas.

Una vez elegido el medidor de distancias, elegiremos el algoritmo de clustering, en nuestro caso hemos elegido el algoritmo de PAM (Partición Alrededor de Medoids), con el que tendremos mas resistencia al ruido y a los valores atípicos, aunque por el contrario el tiempo y el uso de memoria en ejecución es cuadrático. Por último, tendremos que elegir la métrica, que será la que nos recomendara el número de clústeres a utilizar, para ello hemos elegido la métrica del ancho de silueta, esta técnica proporciona una representación gráfica sucinta de qué tan bien se ha clasificado cada objeto.

Lo primero que realizaremos será el cálculo de la matriz de distancias con el método de Gower, una vez calculado podremos comprobar si el resultado es el idóneo, comprobando las instancias que más se acercan:

```
> data[ which(mat_distancias == min(mat_distancias[mat_distancias != min(mat_distancias)]),arr.ind = TRUE)[1, ], ]
          fecha gps longitud   latitud                cuenca        region codigo_region codigo_distrito poblacion     organizacion permiso ano_construcion tipo_extracion grupo_tipo_extracion grupo_gestion
23209 2011-08-03   0     33.4 -2.936435 Lake Victoria          Mwanza                19               4         0 VWC                     Y         1996.815  mono                mono               user-group
23268 2011-08-03   0     33.4 -2.936443 Lake Victoria          Mwanza                19               4         0 VWC                     Y         1996.815  mono                mono               user-group
                    pago    calidad_agua calida_grupo     cantidad             fuente             tipo_PuntoAgua cluster
23209 pay per bucket        soft                 good    enough       machine dbh          communal standpipe multiple       2
23268 pay per bucket        soft                 good    enough       machine dbh          communal standpipe multiple       2
> |
```

Imagen 8. Instancias más cercanas, tras cálculo de la matriz de distancias.

Y las que más se alejan:

```
> data[which(mat_distancias == max(mat_distancias[mat_distancias != max(mat_distancias)]), arr.ind = TRUE)[1, ], ]
          fecha  gps longitud    latitud                cuenca        region codigo_region codigo_distrito poblacion     organizacion permiso ano_construcion tipo_extracion grupo_tipo_extracion grupo_gestion
36205 2013-02-03  -32     40.2 -10.273411 Ruvuma / Southern Coast Mtwara                 9               5        50 Private operator       N            2011  ksb                 submersible        commercial
33207 2013-01-20 1352     30.3  -4.480916 Lake Tanganyika       Kigoma                16               2       755 Water authority        Y            1974  gravity             gravity            user-group
                    pago    calidad_agua calida_grupo     cantidad             fuente             tipo_PuntoAgua cluster
36205 pay per bucket        salty                salty   enough       machine dbh          communal standpipe                3
33207 pay annually          soft                 good    insufficient river                communal standpipe multiple       1
> |
```

Imagen 9. Instancias más lejanas, tras cálculo de la matriz de distancias.

Donde podemos comprobar en la primera imagen las diversas similitudes que existen entre las dos instancias, y en la segunda las numerosas diferencias. Por lo que podemos comprobar que la realización de las distancias sea efectuada correctamente.

Lo siguiente que se realizará será la comprobación del número óptimo de clústeres a utilizar, que se realiza con la métrica del ancho de la silueta.

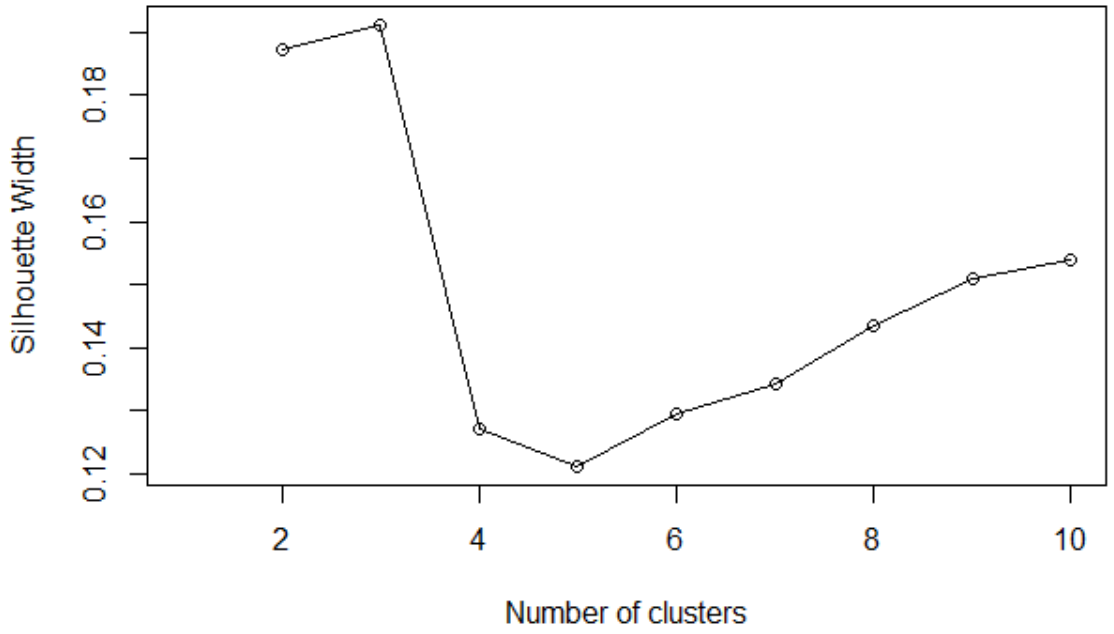


Imagen 10. Numero óptimo de clústeres a utilizar.

Como podemos comprobar el proceso nos mostrara una gráfica, donde el valor más alto nos indicara el valor óptimo de clústeres a utilizar, en este caso el número optimo seria de tres.

Una vez se sabe el número óptimos de clústeres, podremos ejecutar nuestro algoritmo de clustering PAM, siempre que antes de su ejecución lo configuremos correctamente con el número de clústeres que se nos recomendó en el proceso anteriormente.

Por último, mostraremos el resultado de nuestro clustering, con el que mostraremos los grupos formados tras el proceso de clustering.

Imagen 11. Gráfico de los diferentes clústeres formados.

Para terminar, también podemos realizar la visualización de los clústeres formados sobre el mapa geográfico de Tanzania, donde podemos ver a que clúster pertenece cada bomba de agua.

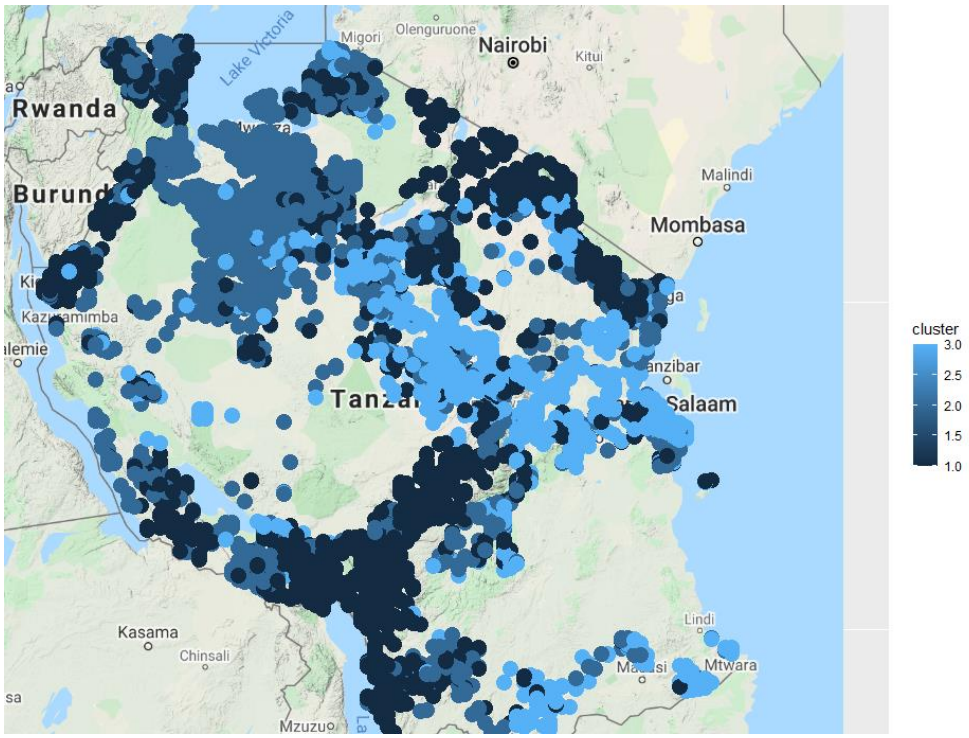


Imagen 12. Distribución del clúster en el mapa de Tanzania.

5. Conclusión

Podemos determinar que con esta solución de inteligencia de negocio el gobierno de Tanzania tendrá una visualización geográfica de las diferentes bombas de agua que tiene repartidas por el territorio nacional, pudiendo realizar múltiples visualizaciones dependiendo del atributo que se desee visualizar. Con lo que se espera que, con la ayuda de esta información, el gobierno tenga más información a la hora de tomas de decisiones en un futuro.

Ventajas que nos aporta la solución:

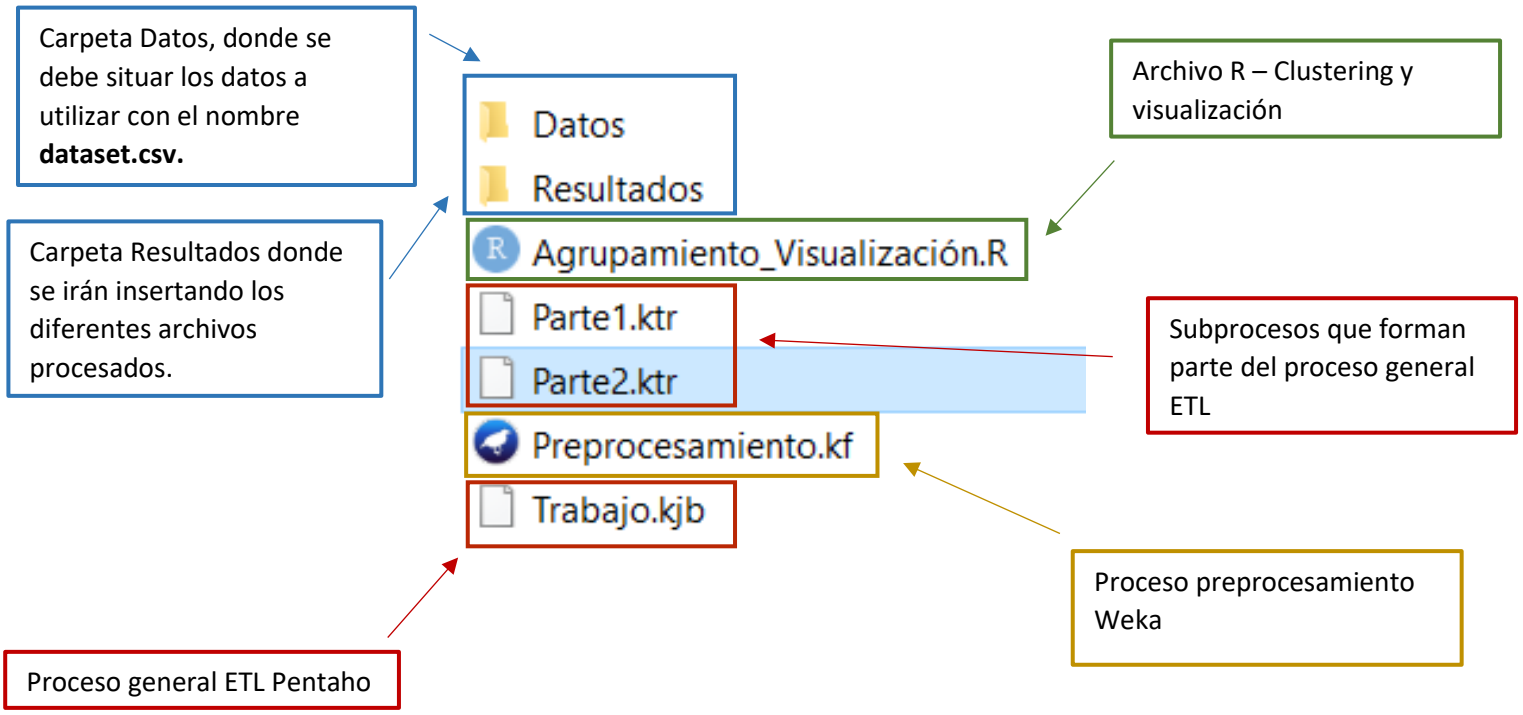
- Visualización de la información sobre el mapa geográfico del país.
- Visualización de graficas.
- Claridad en la información.
- Extracción de la información en archivo PDF e imágenes.
- Tomas de decisiones futuras más acertadas.

Anexo – Documentación del uso de la aplicación.

En este documento se explicará el proceso de instalación y ejecución de nuestro proceso de inteligencia de negocio, antes de todo lo primero que deberemos realizar es la instalación del software necesario:

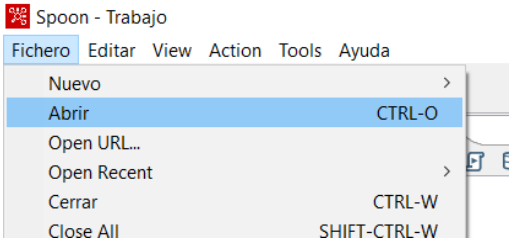
- WEKA 3.9.4
- RStudio
- Pentaho Data Integration

Una vez descargado e instalado todo el software necesario, deberá descomprimir la carpeta del proyecto que estará estructurada de la siguiente forma:

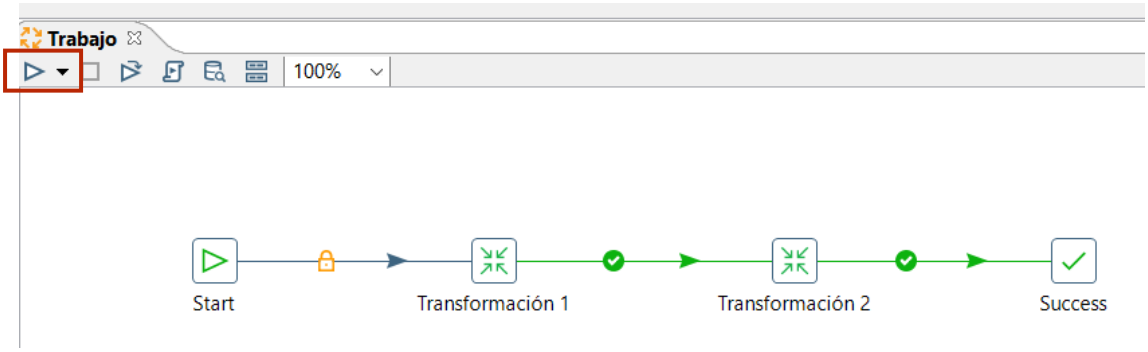


Los datos que vamos a utilizar los encontraremos en la carpeta **Datos** , estos han sido descargados de la plataforma Driven Data a partir del siguiente enlace. En el caso de que se lleve a cabo la descarga del archivo desde el enlace será necesario renombrar dicho archivo al nombre de **dataset.csv** e introducirlo en el interior de la carpeta **Datos** que se encuentra junto a los archivos de ejecución.

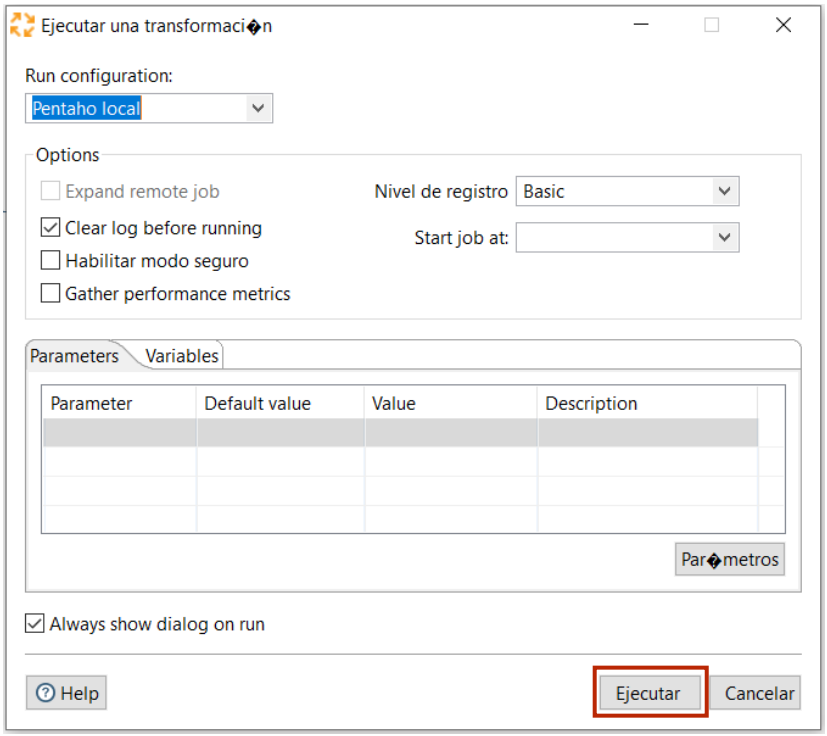
Una vez tenemos nuestros datos listos, deberemos de iniciar nuestra aplicación de pentaho, en la que deberemos de abrir nuestro proceso **Trabajo.kjb**



Una vez abramos nuestro trabajo, lo único que tendremos que realizar es ejecutar nuestro trabajo.



Pulsamos también sobre ejecutar en la ventana emergente que nos aparecera.



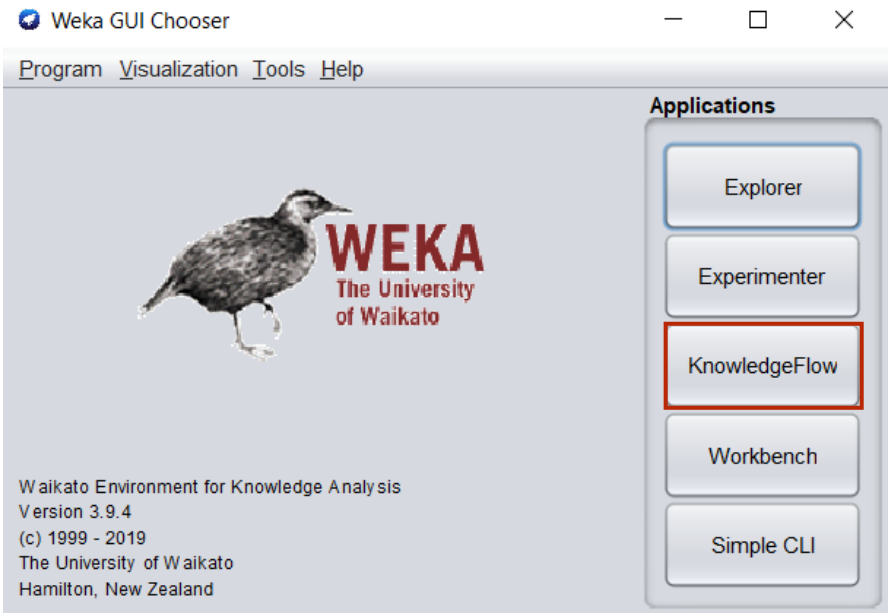
Una vez se concluye el proceso, podemos comprobar como en la carpeta **Resultados** encontramos 2 nuevos archivos:

- datasetRenombrado.csv
- datasetListo.csv

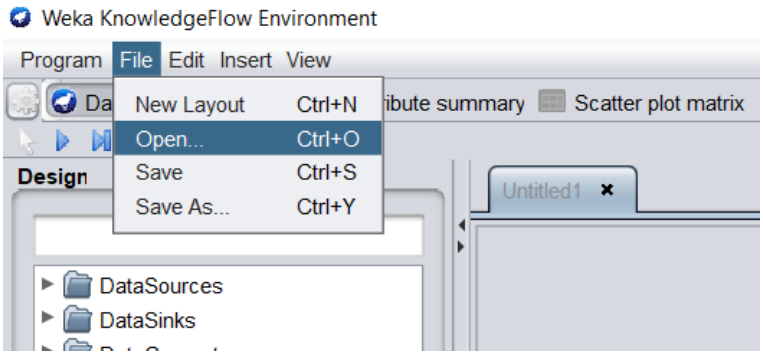
El archivo que nos servirá será el segundo **datasetListo.csv.**

Con este proceso tendremos concluido nuestro proceso ETL, ahora pasaremos a la fase de preprocesamiento, que la realizaremos con la aplicación Weka.

Lo primero que debemos realizar será abrir nuestra aplicación Weka, y acceder al entorno de trabajo knowledgeFlow.

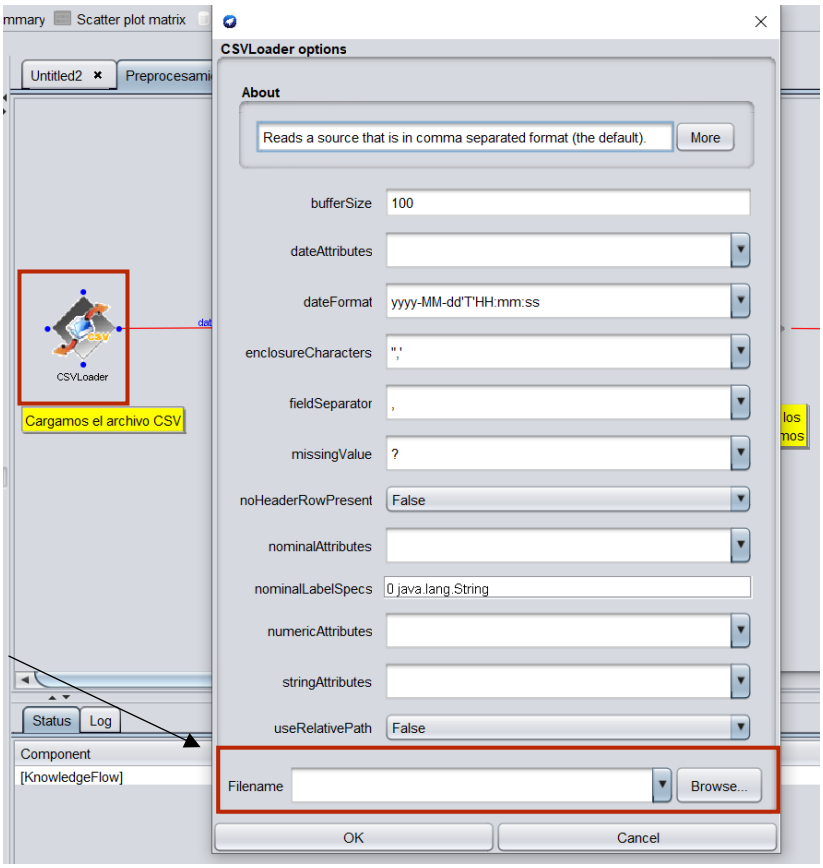


Lo siguiente a realizar será abrir nuestro proceso de Weka.



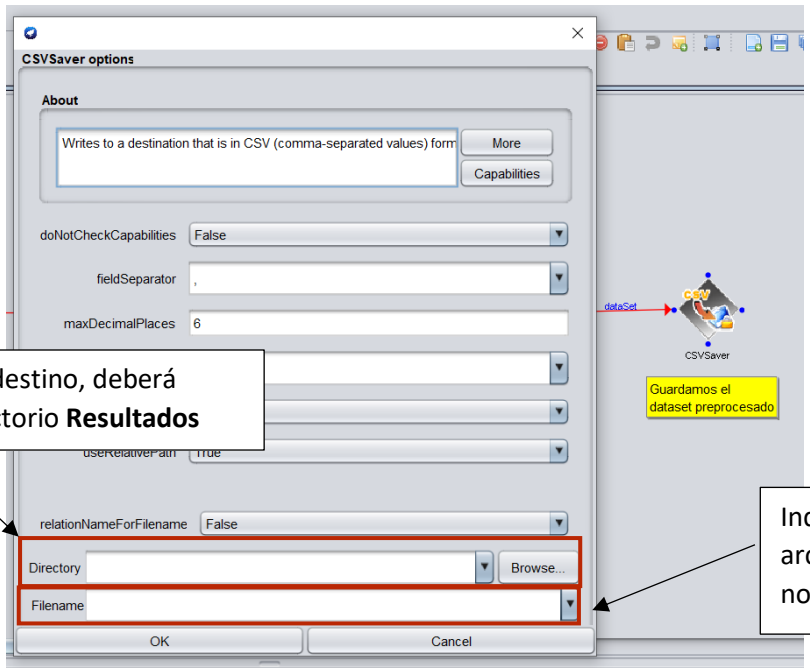
Una vez lo tenemos abierto deberemos configurar los módulos de lectura y escritura de los archivos, esto es debido a que weka knowledgeFlow tiene varios conflictos a la hora de dejar los módulos de lectura y escritura configurados con rutas, por lo que deberemos indicarle las ruta en los 2 módulos de forma manual.

Por lo que para configurarlo deberemos hacer doble click sobre el módulo de lectura del archivo csv, donde seguidamente nos saldrá una ventana emergente donde deberemos seleccionar nuestro archivo datasetListo.csv.



Indicar archivo **datasetListo.csv**

Realizamos el mismo proceso, pero esta vez con el módulo de escritura, donde especificaremos el directorio donde queremos que nos lo guarde, que en nuestro caso será dentro de la carpeta **Resultados**, seguidamente indicaremos el nombre que tendrá dicho archivo, deberemos indicarle el nombre de **datasetPreprocesado**

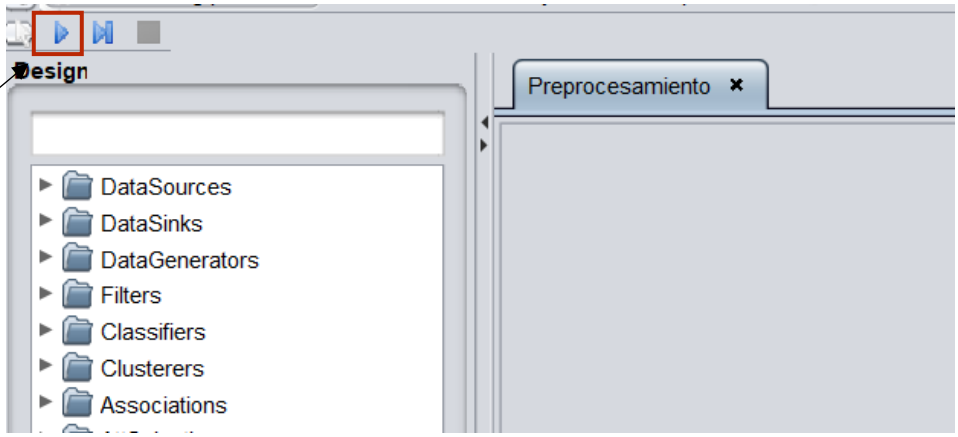


Directorio de destino, deberá indicar el directorio **Resultados**

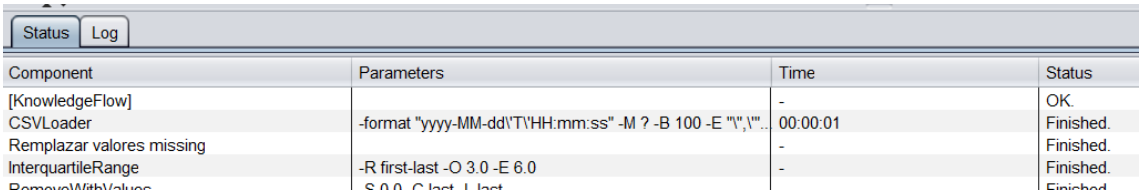
Indicamos el nombre que tendrá el archivo, deberemos ponerle el nombre de **datasetPreprocesado**

Una vez los hemos configurado pasaremos a la ejecución del proceso de preprocesamiento.

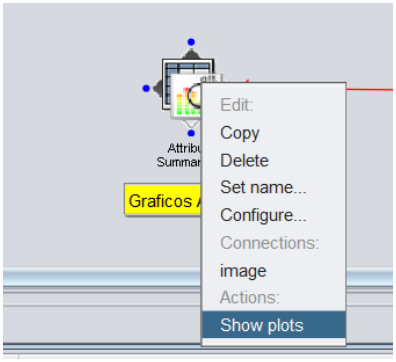
Ejecutamos proceso



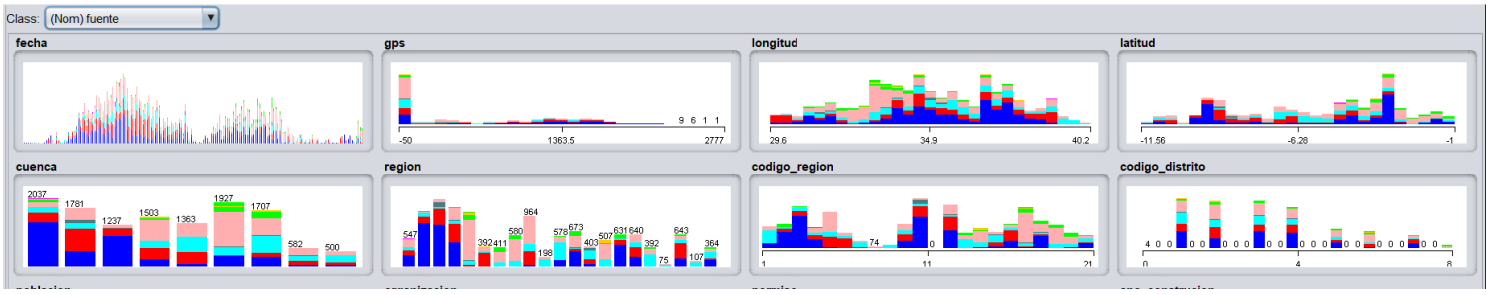
Podemos comprobar que el proceso se ha ejecutado correctamente, observando si se ha creado el archivo **datasetPreprocesado.csv** dentro de la carpeta **Resultados**, también podemos hacerlo fijándonos en el apartado inferior de la aplicación Weka , sobre la barra de Status.



Podremos también visualizar las diferentes graficas de los datos, una vez realizado el preprocesamiento, para ello debemos hacer click derecho con el ratón sobre el módulo de gráficos, y seleccionar *Show plots*

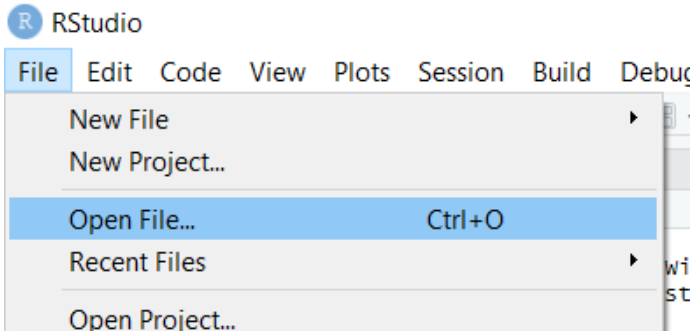


Donde podremos visualizar los diferentes gráficos.



Tras realizar el preprocesado, pasaremos a la fase de clustering y visualización, que realizaremos con RStudio.

Para ello deberemos de abrir la aplicación de RStudio y abrir nuestro archivo **Agrupamiento_Visualizacion.r.**



Al abrirlo nos encontraremos con un fichero de código que deberemos ejecutar línea por línea para evitar posibles errores.

Lo primero que deberemos realizar será la instalación de los paquetes que nos serán necesarios y la importación de esas librerías.

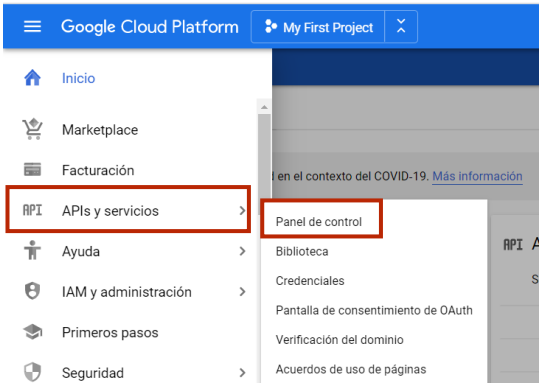
```
# PAQUETES NECESARIOS
install.packages("reshape")
install.packages("devtools")
install.packages("Rtools")
devtools::install_github ("dkahle/ggmap")
devtools::install_github ("hadley/ggplot2")
install.packages("dplyr")
install.packages("ISLR")
install.packages("cluster")
install.packages("Rtsne")
install.packages("ggplot2")

# IMPORTAMOS LIBRERIAS
library(dplyr)
library(ISLR)
library(cluster)
library(Rtsne)
library(ggplot2)
```

Una vez importadas pasaremos a la solicitud de una licencia de la API de Google, ya que necesitaremos las API de Maps Static y Geocoding, **en nuestro caso se le asignará una clave ya configurada de nuestra cuenta para la prueba del proyecto**, aun así, a continuación, se explicará el proceso para la solicitud de la licencia.

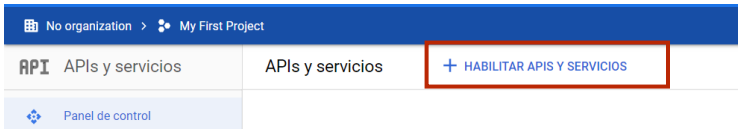
Para solicitar la licencia deberemos de tener una cuenta de Google activa y acceder a Google cloud, desde hace poco Google ha cambiado sus directivas de uso y para hacer uso de Google cloud es necesario crear una cuenta de facturación, con lo que implica que nos pedirá los datos de nuestra tarjeta de crédito, aunque el servicio es totalmente **GRATUITO**, aun así, al crear nuestra cuenta de facturación se nos dará un crédito de 300€ para gasto en Google cloud.

Una vez activemos nuestra cuenta de facturación, pasaremos a la activación de las API que necesitamos.

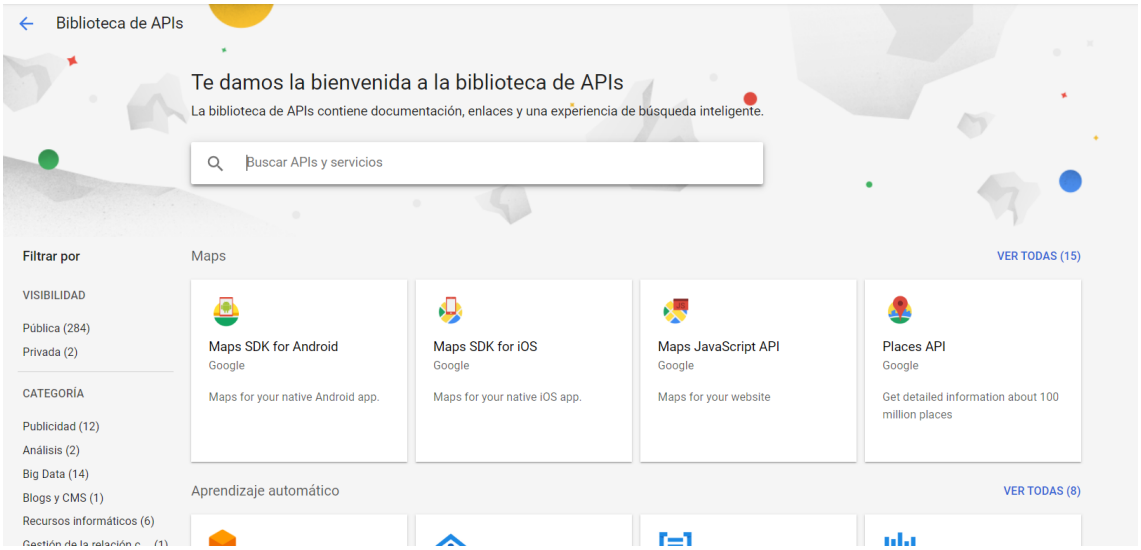
Accedemos al panel de control.



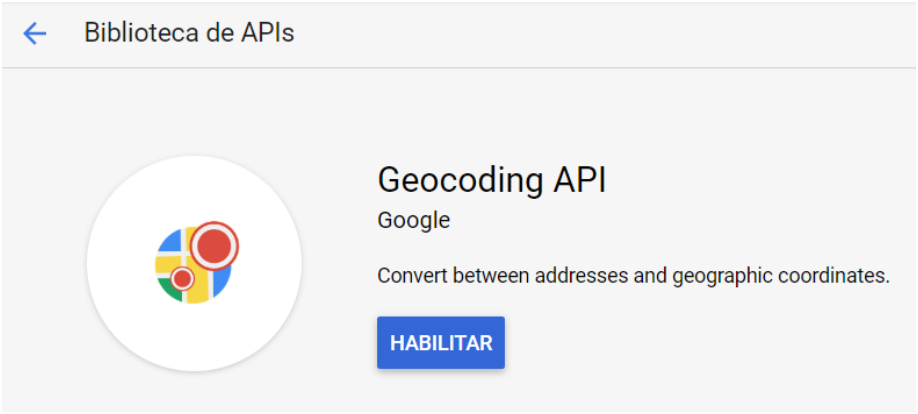
Accedemos a habilitar APIS y servicios



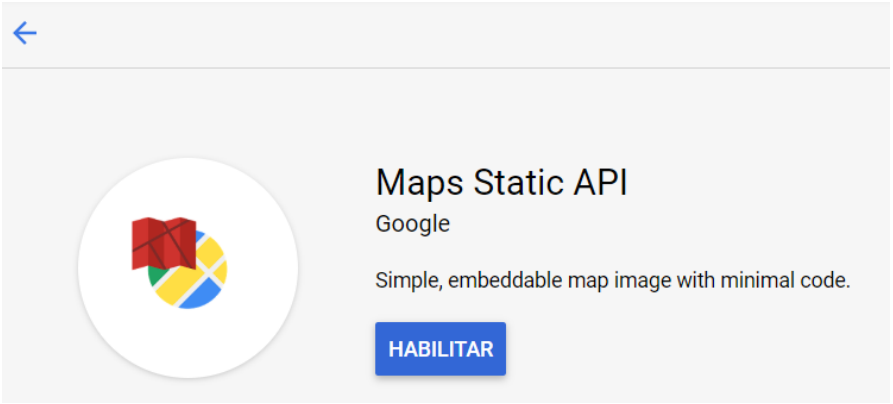
Se nos mostraran todas las API de Google disponibles, donde deberemos de buscar las API que deseamos utilizar.



Seleccionamos y habilitamos Geocoding.

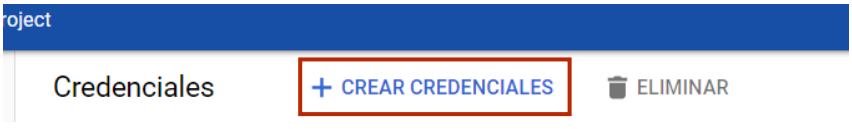


Seleccionamos y habilitamos Maps Static

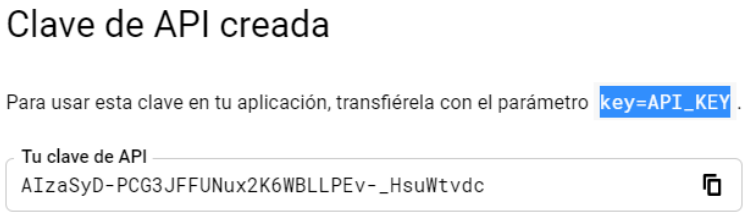


Una vez las tenemos activadas, deberemos de obtener la clave de API de nuestra cuenta, para ello accedemos a credenciales.

Creamos credenciales.



Se nos creará la clave, esta clave será la que tendremos que registrar en RStudio.



Una vez obtenida la clave, ya podremos volver nuevamente a nuestra aplicación RStudio, donde registraremos la clave con la siguiente instrucción.

```
## CLAVE DE LA API DE GOOGLE, NECESARIA PARA MOSTRAR MAPAS
register_google(key="AIzaSyD-PCG3JFFUNux2K6WBLLPEv-_HsuWtvdc")
```

Cargamos nuestro dataset de datos, recordamos que debemos carga el archivo **datasetPreprocesado.csv**, deberemos especificar la ruta donde se encuentra el archivo.

```
## CARGA DEL DATASET
dataset <- read.table("2º Cuatrimestre/Inteligencia Negocio/ProyectoFinal/Resultados/datasetPreprocesado.csv",sep=",",header=TRUE)
```

Ruta que debemos cambiar

Una vez tenemos cargados los datos, ejecutaremos la siguiendo orden, en las que pediremos al servidor de Google el mapa geográfico de Tanzania.

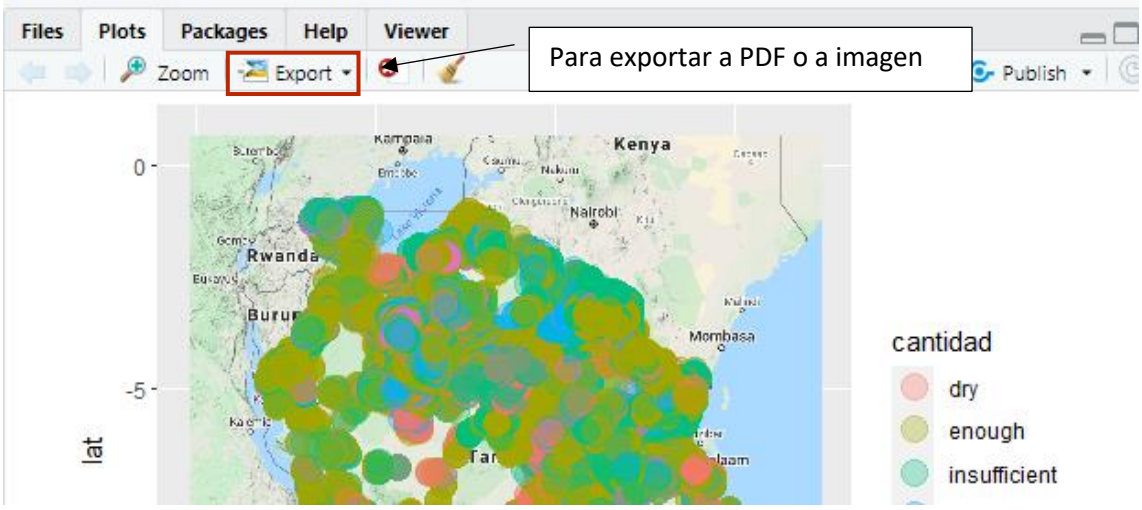
```
# BUSQUEDA DEL MAPA QUE QUEREMOS MOSTRAR
map <- get_map(location = "tanzania", zoom = 6, source = "google", maptype = "terrain")
```

Una vez tenemos cargado el mapa, podremos mostrar este con cualquier atributo de nuestros datos. Lo único que debemos cambiar es el nombre del atributo a mostrar, en la siguiente imagen mostramos por el atributo cantidad.

```
# MOSTRAR MAPA -- CANTIDAD DE CADA BOMBA DE AGUA
ggmap(map, extent  = TRUE) +  geom_point(aes(longitud, latitud, colour=cantidad), alpha=0.3, size=5, shape=19, data=dataset)
```

Opción que cambiar, para poder ver al atributo deseado

Tras ejecutarlo y esperar unos segundos, podemos comprobar como en la esquina inferior derecha se nos ha creado el mapa con la distribución del atributo seleccionado, desde este apartado de la aplicación podremos exportar ese mapa a un archivo PDF o a una imagen.



Una vez realizados los mapas, lo siguiente que realizaremos será realizar el clustering.

Ejecutamos las siguientes líneas para obtener la matriz de distancias.

```
# ============================================================
# ===== CALCULAMOS DISTANCIAS CON EL METODO DE GOWER =====
# ============================================================

distancias <- daisy(data,metric = "gower",type = list(logratio = 3))

summary(distancias)

# MATRIZ DE DISTANCIAS
mat_distancias <- as.matrix(distancias)
```

Una vez tenemos la matriz de distancias, ejecutamos las líneas para obtener el número óptimo de clústeres.

```
# ============================================================================================
# ===== COMPROBAR EL NUMERO OPTIMO DE CLUSTERES -- METRICA DEL ANCHO DE SILUETA =====
# ============================================================================================

sil_width <- c(NA)

for(i in 2:10){

  pam_fit <- pam(distancias, diss = TRUE, k = i)
  sil_width[i] <- pam_fit$silinfo$avg.width

}


# PINTAMOS GRAFICA PARA COMPOBAR EL NUMERO OPTIMO DE CLUSTERES
plot(1:10, sil_width, xlab = "Number of clusters",ylab = "Silhouette Width")
lines(1:10, sil_width)
```

Lo siguiente a realizar será la realización del clustering PAM, donde tendremos que configurarlo con el numero óptimo de clústeres.

```
# ========================================================================
# ===== EJECUTAMOS EL ALGORITMO PAM, CON EL NUMERO DE CLUSTERES OBTENIDO =====
# ========================================================================

pam_fit <- pam(distancias, diss = TRUE, k = 2)

pam_results <- data %>%
  dplyr::select(-tipo_PuntoAgua) %>%
  mutate(cluster = pam_fit$clustering) %>%
  group_by(cluster) %>%
  do(the_summary = summary(.))

pam_results$the_summary
```

Numero de clústeres a modificar

Seguidamente, podremos realizar una visualización de la distribución de las instancias en los clústeres creados, ejecutando las siguientes líneas.

```
# VISUALIZACIÓN DE LOS DATOS
tsne_obj <- Rtsne(distancias, is_distance = TRUE)

tsne_data <- tsne_obj$Y %>%
  data.frame() %>%
  setNames(c("X", "Y")) %>%
  mutate(cluster = factor(pam_fit$clustering),
         name = data$tipo_PuntoAgua)

ggplot(aes(x = X, y = Y), data = tsne_data) + geom_point(aes(color = cluster))
```

Para terminar para poder ver la distribución de los clústeres sobre el mapa, ejecutaremos la siguiente línea:

```
# ============================================================
# ===== MOSTRAMOS MAPA -- DISTRIBUCION CLUSTERES OBTENIDO =====
# ============================================================

ggmap(map, extent = TRUE) +  geom_point(aes(longitud, latitud, colour=cluster), alpha=1, size=5, shape=19, data=data)
```