UNIVERSIDAD NACIONAL JOSÉ MARÍA ARGUEDAS

FACULTAD DE INGENIERÍA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

Presentado por:

ANAI LUZ AMPARO LIRA FLORES

DETERMINACIÓN DE PATRONES DE COMPORTAMIENTO DE CONSUMO DE AGUA POTABLE CON ALGORITMOS DE CLUSTERIZACION EN LA PROVINCIA DE ANDAHUAYLAS

Asesor:

ING. JUAN JOSÉ ORE CERRÓN

TESIS PARA OPTAR EL TÍTULO PROFESIONAL DE INGENIERO DE SISTEMAS

ANDAHUAYLAS – APURÍMAC – PERÚ

2018

APROBACIÓN DEL ASESOR

Quien suscribe:
Ing. Juan José Ore Cerrón por la presente:

CERTIFICA,
Que, la Bachiller en **Ingeniería de Sistemas,** Anai Luz Amparo Lira Flores ha culminado satisfactoriamente el informe final de tesis intitulado: "DETERMINACIÓN DE PATRONES DE COMPORTAMIENTO DE CONSUMO DE AGUA POTABLE CON ALGORITMOS DE CLUSTERIZACION EN LA PROVINCIA DE ANDAHUAYLAS" para optar el Título Profesional de Ingeniero de Sistemas.

Andahuaylas, 17 de diciembre de 2018

_____
Ing. Juan José Ore Cerrón
Asesor

_____
Anai Luz Amparo Lira Flores
Tesista

APROBACIÓN DEL JURADO DICTAMINADOR

LA TESIS: "DETERMINACIÓN DE PATRONES DE COMPORTAMIENTO DE CONSUMO DE AGUA POTABLE CON ALGORITMOS DE CLUSTERIZACION EN LA PROVINCIA DE ANDAHUAYLAS"; para optar el Título Profesional de Ingeniero de Sistemas, ha sido evaluada por el Jurado Dictaminador conformado por:

PRESIDENTE: Dr. JULIO CÉSAR HUANCA MARÍN
PRIMER MIEMBRO: MSc. CARLOS YINMEL CASTRO BULEJE
SEGUNDO MIEMBRO: MSc. MAGALY ROXANA ARANGÜENA YLLANES

Habiendo sido aprobado por UNANIMIDAD/MAYORIA, en la ciudad de Andahuaylas el día 20 del mes de diciembre de 2018

Andahuaylas, 17 de diciembre de 2018.

_____
Dr. JULIO CÉSAR HUANCA MARÍN
PRESIDENTE DEL JURADO DICTAMINADOR

_____
MSc. CARLOS YINMEL CASTRO BULEJE
PRIMER MIEMBRO DEL JURADO DICTAMINADOR

_____
MSc. MAGALY ROXANA ARANGÜENA YLLANES
SEGUNDO MIEMBRO DEL JURADO DICTAMINADOR

ACTA DE SUSTENTACIÓN DE TESIS

Universidad Nacional José María Arguedas

Identidad y Excelencia para el Trabajo Productivo y el Desarrollo

FACULTAD DE INGENIERÍA

ACTA DE SUSTENTACIÓN DE TESIS

En la Av. José María Arguedas del Local Académico SL01 (Ccoyahuacho) en el auditorio de la Escuela Profesional de Ingeniería de Sistemas de la Universidad Nacional José María Arguedas ubicado en el distrito de San Jerónimo de la Provincia de Andahuaylas, siendo las 10:00 horas del día 20 de noviembre del año 2018, se reunieron los docentes: Dr. Julio César Huanca Marín, MSc. Carlos Yinmel Castro Buleje, MSc. Magaly Roxana Arangüena Yllanes, en condición de integrantes del Jurado Evaluador del Informe Final de Tesis intitulado: "DETERMINACIÓN DE PATRONES DE COMPORTAMIENTO DE CONSUMO DE AGUA POTABLE CON ALGORITMOS DE CLUSTERIZACION EN LA PROVINCIA DE ANDAHUAYLAS", cuyo autor es el Bachiller en Ingeniería de Sistemas **ANAI LUZ AMPARO LIRA FLORES,** el asesor Ing. Juan José Oré Cerrón, con el propósito de proceder a la sustentación y defensa de dicha tesis.

Luego de la sustentación y defensa de la tesis, el Jurado Evaluador **ACORDÓ**: APROBAR por UNANIMIDAD al Bachiller en Ingeniería de Sistemas **ANAI LUZ AMPARO LIRA FLORES,** obteniendo la siguiente calificación y mención:

Nota escala vigesimal		Mención
Números	Letras	
14	CATORCE	BUENO

En señal de conformidad, se procedió a la firma de la presente acta en 03 ejemplares.

.....
Dr. Julio César Huanca Marín
Presidente del Jurado Evaluador

.....
MSc. Carlos Yinmel Castro Buleje
Primer Miembro del Jurado Evaluador

.....
MSc. Magaly Roxana Arangüena Yllanes
Segundo Miembro del Jurado Evaluador

DEDICATORIA

Dedico este trabajo de tesis a Dios, a mis padres Juana Flores Mesares y Cesar Lira Serna por brindarme siempre su amor y apoyo en todo momento para superar cada obstáculo que se presenta en el día a día; a mi tía Cristina Flores Mesares que me cuida y guía en todo momento desde el cielo en cada paso que doy.

AGRADECIMIENTO

Agradezco a Dios; a mis padres y hermana por su apoyo incondicional en todo momento. Agradezco a mis docentes por haberme brindado sus enseñanzas; a mis amigos y compañeros con quienes compartí y aprendí cada día en la universidad; a Luis Miguel por su apoyo moral cada día.

Agradezco a mi asesor Ing. Juan José Ore Cerrón por su apoyo en la realización de este proyecto de tesis.

TABLA DE CONTENIDOS

Pág.

APROBACIÓN DEL ASESOR ..... **ii**
APROBACIÓN DEL JURADO DICTAMINADOR ..... **iii**
ACTA DE SUSTENTACIÓN DE TESIS ..... **iv**
DEDICATORIA ..... **v**
AGRADECIMIENTO ..... **vi**
RESUMEN ..... **xii**
ABSTRACT ..... **xiii**
CHUMASQA ..... **xiv**
CAPÍTULO 1: INTRODUCCIÓN **1**
1.1 Planteamiento del problema 3
1.1.1 Realidad problemática 3
1.2 Formulación del problema 6
1.2.1 Formulación del problema central 6
1.3 Objetivos 6
1.3.1 Objetivo General 6
1.3.2 Objetivos Específicos 6
1.4 Justificación 6
1.5 Viabilidad técnica, económica, operativa y social 7
1.5.1 Viabilidad técnica 7
1.5.2 Viabilidad Económica 7
1.5.3 Viabilidad Operativa 10
1.5.4 Viabilidad social 10
1.6 Limitación del estudio 10
CAPÍTULO 2: MARCO TEÓRICO **11**
2.1 Estado del arte 11
2.1.1 Antecedentes internacionales 11
2.1.2 Antecedentes nacionales 12
2.2 Marco conceptual 12
2.2.1 Tecnologías de la información y comunicación 12
2.2.2 Minería de datos 13

2.2.3 Proceso de descubrimiento de conocimiento en bases de datos 14
2.2.4 Metodología CRISP-DM 15
2.2.5 Técnicas 16
2.2.6 Clustering 18
2.2.7 Algoritmos de agrupamiento 19
2.2.8 Patrones de comportamiento 24
2.2.9 Comportamiento del consumidor 26
CAPÍTULO 3: MATERIALES Y MÉTODOS 27
3.1 Propuesta de solución 27
3.1.1 Modelo conceptual 27
3.2 Gestión de proyecto 29
3.2.1 Procesos de la dirección del proyecto 29
3.2.2 Áreas de conocimiento 29
3.3 Desarrollo de la determinación de patrones de comportamiento de consumo de agua 44
3.3.1 Comprender el dominio de aplicación 45
3.3.2 Extraer la base de datos objetivo 46
3.3.3 Preparar la base de datos objetivo 49
3.3.4 Minería de datos 49
CAPITULO 4: RESULTADOS Y DISCUSIÓN 53
4.1 Resultados año 2015 53
4.2 Resultados del 2016 56
4.3 Resultados del año 2017 59
4.4 Interpretación 62
4.4.1 Evaluar resultados 62
4.4.2 Utiliza el conocimiento descubierto 63
CONCLUSIONES 64
RECOMENDACIONES 65
REFERENCIAS BIBLIOGRÁFICAS 66
ANEXO 69

LISTA DE FIGURAS

Pág.

Figura 1: Proceso KDD 15
Figura 2*: Clustering particional* 20
Figura 3: Modelo conceptual para la determinación de patrones de comportamiento de consumo de agua potable con algoritmos de clusterización en la provincia de Andahuaylas. 28
Figura 4: Estructura de Descomposición de Trabajo (EDT) 33
Figura 5: Cronograma de Actividades 35
Figura 6: Organigrama del personal involucrado 38
Figura 7: Categorías de riesgo 40
Figura 8: Análisis de riesgo 42
Figura 9: Interface para ejecutar el algoritmo 46
Figura 10: Cuadro de datos 47
Figura 11: Cuadro de datos 48
Figura 12: Librerías instaladas 50
Figura 13: Librerías instalas 50
Figura 14: Lista de datos 51
Figura 15: Resultados de los clusters utilizando la ecuación distancia de Manhattan 52
Figura 16: Resultados de Centroides 52
Figura 17: Resultados de Centroides y cantidad de usuarios por categoría año 2015 53
Figura 18: Datos ubicados sin algoritmo k_means del año 2015 53
Figura 19: Resultados de clusters gráficamente del año 2015 54
Figura 20: Resultados del clusters 0 del año 2015 54
Figura 21: Resultados del clusters 1del año 2015 55
Figura 22: Resultados del clusters 2 del año 2015 55
Figura 23: Usuarios que son de los centroides del año 2015 55
Figura 24: Resultados de Centroides y cantidad de usuarios por categoría del año 2016 56
Figura 25: Datos ubicados sin algoritmo k_means del año 2016 56
Figura 26: Resultados de clusters gráficamente del año 2016 57
Figura 27: Resultados del clusters 0 del año 2016 57
Figura 28: Resultados del clusters 1 del año 2016 58
Figura 29: Resultados del clusters 2 del año 2016 58
Figura 30: Usuarios que son de los centroides del año 2016 58

Figura 31: Resultados de Centroides y cantidad de usuarios por categoría del año 2017.....59
Figura 32: Datos ubicados sin algoritmo k_means del año 2017.....59
Figura 33: Resultados de clusters gráficamente del año 2017.....60
Figura 34: Resultados del clusters 0 del año 2017.....60
Figura 35: Resultados del clusters 1 del año 2017.....61
Figura 36: Resultados del clusters 2 del año 2017.....61
Figura 37: Usuarios que son de los centroides del año 2017.....61

LISTA DE TABLAS

Tabla 1: Análisis de costo de Inversión 8
Tabla 2: Características de las NN.TT. de la Información y Comunicaciones 13
Tabla 3: Gestión de Integración del proyecto 30
Tabla 4: Gestión de Costo del proyecto 36
Tabla 5: Perfil de Recursos Humanos del Proyecto 38
Tabla 6: Asignación de responsabilidad al personal 39
Tabla 7: Reunión con el asesor 39
Tabla 8: Probabilidad de riesgo 41
Tabla 9: Gestión de las adquisiciones del proyecto 43
Tabla 10: Transformación de categoría 49
Tabla 11: Comparación de consumo 62

RESUMEN

El presente trabajo de tesis se realizó con el propósito de determinar patrones de comportamiento de consumo de agua potable en la Entidad Prestadora de Servicios de Saneamiento Municipal Chanka en la provincia de Andahuaylas, utilizando un algoritmo de clusterización para determinar grupos de consumo de agua potable.

El objetivo de esta tesis es obtener patrones de comportamiento de consumo de agua potable en la EPS EMSAP CHANKA, con el objeto de idear un modelo sencillo que obtenga clusters o grupos de consumo utilizando el algoritmo k_means, la cual busca encontrar patrones de comportamiento que distinga a cada uno de los clusters encontrados.

Asimismo, se desarrolló una herramienta utilizando el algoritmo k_means la cual usa el consumo de agua potable y la categoría a la que pertenece un determinado usuario, analizando y mostrando los clusters encontrados con los datos utilizados; para así evaluar cuáles son los patrones de comportamiento que posee cada cluster y de esta manera obteniendo los patrones de comportamiento realizar una mejor toma de decisiones para optimizar el consumo de agua potable en la ciudad de Andahuaylas.

Palaras Clave: Patrones de comportamiento, clustering, algoritmo, k_means, agua.

ABSTRACT

The present thesis work was carried out with the purpose of determining patterns of consumption behavior of drinking water in the Provider Entity of Chanka Municipal Sanitation Services in the province of Andahuaylas, using a clustering algorithm to determine groups of drinking water consumption.

The objective of this thesis is to obtain patterns of consumption behavior of drinking water in EPS EMSAP CHANKA, in order to devise a simple model that obtains clusters or consumption groups using the k_means algorithm, which seeks to find patterns of behavior that distinguish to each of the clusters found.

Likewise, a tool was developed using the k_means algorithm which uses the drinking water consumption and the category to which a specific user belongs, analyzing and showing the clusters found with the data used; in order to evaluate which are the behavior patterns that each cluster possesses and in this way obtaining the behavior patterns to make a better decision making to optimize the drinking water consumption in the city of Andahuaylas.

Keywords: Patterns of behavior, clustering, algorithm, k_means, water.

CHUMASQA

Kay kunan llankay rurasqay tesis rurakurqa propositokunata tukupaynin allin humalliq comportamientokunata sumaq hanpisqa yaku tumanapaq kay humalliq llanapakuq Saneamiento Municipal Chanka nisqa wuasimanta kay provincia Andahuaylas llactanchipi, huk ruaykuna algoritmo de clusterización kay chayay achka hanpisha yaku tupanachikpaq.

Kay ruraykunaqa allin punim tesis haypay kamachikuykuna hinaspa allin hasnpisha llaku tumanapaq kay umalliq llanapaq Emsap Chanka wuasimanta, kay huk yachaykunawuan sumaq yuyaykunahuan huk modelokunata haypay clusters huñulla tukuy chaninchay kay algoritmo k_means, maskay tarinchik patrunkunata purtakuykunata allin akllanapaq sapankamata kay clausters tarisqamanta.

Chaynallataqmi rurakun huk herramientakunapas allin ruraykunapaq kay algoritmo k_means tukuy chaninchay hampisqa yaku chayaqiy kay mitan determinaduta uk runa kuskiy chuskichikunapaq kay clusters tarisqanta yuyaykuna chaninchay; chaynallataq ima kaqta mayqin willay patrunkunata kay comportamientokunata ima kapay sapa cluster chay hina haypay patronkunata kay comportamientokunata uk allin ruraykuna qayka kamarikuy hinaspa allin hampisqa yaku tumanapaq kay Andahuaylas llaqtanchikpi.

Allin rimaykuna: Patrón de comportamiento nisqan, clustering nisqan, algoritmo nisqan, k_means nisqan, yacu.

CAPÍTULO 1: INTRODUCCIÓN

Hoy en día hay muchas decisiones que se tienen que tomar con tantos datos obtenidos en cada instante. Las Organizaciones a nivel mundial viven pensando en cual decisión será la correcta para su entidad ya que los datos crecen cada vez más y la capacidad de tomar decisiones correctas con datos grandes también llega a crecer. La minería de datos es una manera de buscar respuestas ya que analizando una gran cantidad de datos podemos llegar a encontrar patrones ocultos que se pueden utilizar para predecir comportamiento futuro y de esta manera transformar los datos en conocimiento.

El presente trabajo de tesis se realizó con el fin de obtener patrones de comportamiento del consumo de agua potable en la Empresa Prestadora de Servicios de Saneamiento Municipal Chanka (EMSAP CHANKA) en la ciudad de Andahuaylas con la finalidad de mejorar la toma de decisiones en cuanto al consumo de agua potable de los usuarios; con el desarrollo de un sistema que permite observar clusters o grupos de consumo de agua además de poder observar que usuarios se encuentran en cada cluster o grupo, utilizando el algoritmo K_means que se encarga del proceso de formar los clusters utilizando el consumo de agua y la categoría a la que pertenece cada usuario.

K_means es un algoritmo automático de clusterización sin supervisión que agrupa puntos de datos similares y de esta manera llegue a descubrir patrones. Busca un número fijo de grupos (K) en el conjunto de datos, para luego definir un numero de centroides que represente el centro de cada grupo establecido, realiza esta acción varias veces para optimizar las posiciones de los centroides.

En la actualidad las organizaciones que poseen gran cantidad de información, buscan evaluar los datos que tienen y buscar patrones ocultos que revelen el comportamiento futuro de los usuarios y de esta manera obtener mejores resultados. En la presente tesis se realizó un sencillo sistema que muestre grupos de consumo de agua utilizando el algoritmo k_means de clusterizacion con el fin de obtener patrones de comportamiento del consumo de agua potable en la EPS EMSAP CHANKA de Andahuaylas; este informe final de tesis se dividió en 4 capítulos donde se desarrolla lo siguiente:

1. CAPÍTULO 1: INTRODUCCIÓN; en este capítulo llegaremos a encontrar el planteamiento del problema se menciona la realidad problemática, la que nos llevó a formular el problema general, así mismo, los objetivos a alcanzar.

2. CAPÍTULO 2: MARCO TEÓRICO; en este capítulo encontraremos todos los conceptos relacionados a la minería de datos y algoritmos de clustering, para realizar un sistema utilizando un algoritmo de clustering para obtener patrones de comportamiento en la EPS EMSAP CHANKA en Andahuaylas, así como los antecedentes, que nos ayudaran a entender los diferentes términos utilizados en la presente tesis.

3. CAPÍTULO 3: MATERIALES Y MÉTODOS; en este capítulo llegaremos a encontrar propuesta de sola general, así mismo, la metodología desarrollada utilizando un algoritmo de clustering para obtener patrones de comportamiento.

4. CAPÍTULO 4: RESULTADOS Y DISCUSIÓN; en este capítulo encontraremos los resultados obtenidos utilizando el algoritmo de clustering, realizando una comparación entre las mismas para llegar a obtener los patrones de comportamiento de consumo de agua potable.

1.1 Planteamiento del problema

1.1.1 Realidad problemática

El agua es un recurso hídrico sumamente importante, esencial para la vida y el desarrollo de todos los seres vivos que habitan en la tierra, ya sea que se utilice para beber, para el uso doméstico, para producir alimentos, el agua puede impulsar el crecimiento económico de los países, según el Programa Conjunto OMS/UNICEF de Monitoreo menciona que “cada año podrían evitarse 842 000 muertes con la mejora de agua, saneamiento e higiene”(PCM; 2015), mientras el Informe de las Naciones Unidas sobre el Desarrollo de los Recursos Hídricos en el Mundo “a mediados del presente siglo, 7.000 millones de personas en 60 países sufrirán escasez de agua, en el peor de los casos, y en el mejor se tratara de 2.000 millones de personas en 48 países.”(UNESCO, 2003).

La Organización Mundial de la Salud, establece que el acceso óptimo en un consumo medio superior o igual a 100 litros por habitante al día. Alrededor de 3 de cada 10 personas (2100 millones de personas) carecen de acceso a agua potable y disponible en el hogar, por lo tanto, tienen que caminar grandes distancias hasta encontrar pozos, manantiales, estanques, lagos, ríos o arroyos no protegidos. Se calcula que 842 000 personas y 361 000 niños menores de cinco años mueren cada año de diarrea como consecuencia de insalubridad de agua, además 240 millones de personas se ven afectadas por lombrices parasitarias contraídas con agua infestada (OMS, 2017).

Suiza es a veces llamada la “torre de agua de Europa” y posee muchos lagos y ríos, como el Rin y el Ródano, que nacen en su territorio. Las estaciones de purificación que se construyeron en los años 60 y 70 garantizan una muy alta calidad del agua. Para eliminar los micro contaminantes causados por los productos fitosanitarios de la agricultura, Suiza decidió agregar un nuevo paso en el tratamiento de las aguas residuales en más de cien plantas potabilizadoras. El estado de las vías fluviales suizas mejorará para 2030 gracias a las nuevas construcciones en curso y la renovación de las plantas de tratamiento existentes. (House of Switzerland, 2017).

Según el INE (Instituto Nacional de Estadísticas), (2018) en las actividades diarias se llegan a consumir aproximadamente las siguientes cantidades de agua potable: un baño o wáter gasta entre 150 y 300 litros, tomar una ducha gasta 50-100 litros, lavar los platos a mano gasta 23 litros, utilizar un lavavajillas gasta entre 20 y 40 litros, una lavadora gasta entre 40 y 80 litros de agua, descongelar un alimento bajo el grifo gasta entre 15 y 25 litros, mantener el grifo abierto durante un minuto y medio, durante el lavado de dientes, puede gastar más de 18 litros y el lavado de coche con manguera gasta entre 200 y 500 litros.

Según el Índice de Desarrollo Humano Nacional, el Perú es el país con mayor disponibilidad de agua dulce por habitante en América Latina (IDM, 2013), ¿Qué pasaría si en el futuro no contaríamos con este elemento vital? aumentaría la tasa de mortalidad en el mundo.

En nuestro país el consumo promedio de agua potable por persona al día es de unos 250 litros por habitante al día (SEDAPAL, 2016). La Organización Panamericana de la Salud, menciona que la oferta de agua en el Perú y en el mundo está cada vez más limitada y la demanda sigue aumentando. Por ello, suministrar y manejar el agua en todos los sectores es importante mediante políticas públicas concertadas, además de instituciones fortalecidas, la gobernanza del sector del agua es esencial para la coordinación y articulación de acciones por ello permitirá la mejora en la cobertura efectiva del agua en particular en los sectores más vulnerables de la población mejorando su expectativa y calidad de vida (OPS, 2018).

“En nuestra provincia de Andahuaylas uno de los problemas principales en los tres últimos años fue disminuir el tiempo de consumo del agua potable a 9 horas diarias a comparación de hace 4 años atrás que se podía contar con este servicio de agua potable las 24 horas al día” (Sopanta, 2018); al inicio muchos tuvieron problemas como las amas de casa, los restaurantes, colegios, hospitales y otros, ya que los pobladores tenían una costumbre de contar todo el día con agua potable para realizar sus diferentes actividades uno de los problemas que afectó a todo poblador ya sea en su casa o lugar de trabajo fue que se necesitó recurso hídrico para

utilizar los servicios higiénicos, esto podía ocasionar enfermedades o infecciones por la falta de limpieza en los baños; hoy en día la mayoría de los andahuaylinos cuentan con tanques en sus casas para poder tener acceso de este elemento vital todo el día, pero no todas las familias cuentan con la posibilidad de tener un tanque en casa y recurren a empozar el agua en baldes o bidones, por otro lado esta solución de empozar agua en baldes o bidones puede ocasionar enfermedades como la diarrea, tifoidea, el cólera, dengue o infecciones, causando gastos mayores a las familias; las enfermedades diarreicas causadas por el agua sucia y el saneamiento deficiente cuesta la vida de 1,8 millones de niños anualmente a nivel mundial y perjudican gravemente la salud. Las sequías las inundaciones o sistemas de saneamiento inadecuados son causa de malnutrición y en muchos casos pueden evitarse.

Según la Empresa prestadora de servicios se agua potable Emsap Chanka (2018), el consumo de agua potable en la ciudad de Andahuaylas fue incrementando exponencialmente cada año, según las disposiciones finales de cada año diariamente se tiene los siguientes datos.

- En el año 2011 se consumió 871898 metros cúbicos de agua potable.
- En el año 2012 se consumió 955762 metros cúbicos de agua potable.
- En el año 2013 se consumió 1032148 metros cúbicos de agua potable.
- En el año 2014 se consumió 1075306 metros cúbicos de agua potable.
- En el año 2015 se consumió 1145135 metros cúbicos de agua potable.
- En el año 2016 se consumió 1149624 metros cúbicos de agua potable.
- En el año 2017 se consumió 1079840 metros cúbicos de agua potable.

Analizando datos históricos, de consumo de agua potable al mes de un sector con tarifas determinadas podremos obtener patrones de comportamiento de consumo de agua potable en los usuarios de la EPS Emsap Chanka.

1.2 Formulación del problema

1.2.1 Formulación del problema central

¿Cuáles son los patrones de comportamiento del consumo de agua potable con algoritmos de clusterizacion en la provincia de Andahuaylas?

1.3 Objetivos

1.3.1 Objetivo General

Obtener patrones de comportamiento del consumo de agua potable con algoritmos de clusterizacion en la provincia de Andahuaylas.

1.3.2 Objetivos Específicos

- Recopilar y preparar los datos históricos encontrados en la Empresa Prestadora de servicios Emsap Chanka.
- Implementar y clasificar los datos utilizando el algoritmo de clusterizacion.
- Obtener los patrones de comportamiento y evaluar los resultados.

1.4 Justificación

La crisis que se está viviendo a nivel mundial según muchos investigadores y organizaciones que tienen menciones en cuestión de salud y medio ambiente es la escases de agua por falta de una buena conservación de este líquido elemento, en la actualidad la preservación de este líquido elemento es fundamental ya que en la provincia Andahuaylas los pilares fundamentales por los cuales gira la economía son la agricultura y la ganadería.

La presente investigación pretende obtener patrones de comportamiento logrando localizar diferentes grupos de usuarios como los usuarios que consumen una cantidad alta de agua potable.

Los problemas localizados como un consumo alto de agua o la falta de agua para otros usuarios podrán llegar a encontrarlas con mayor facilidad ya que se obtendrá grupos de usuarios que consumen más o menos elemento líquido y de esta manera dar a conocer a la empresa prestadora de este servicio qué tipo de factores lo ocasionan para que luego se pueda tomar decisiones y realizar acciones que mejoren el servicio de agua para los usuarios que no llegan a tener un buen acceso, así como concientizar a los

usuarios sobre el uso adecuado que se debe realizar para cada actividad que se hace en el día a día cuando se utiliza el agua potable, los beneficiarios directos en esta investigación serán los usuarios además de la empresa prestadora de este servicio.

1.5 Viabilidad técnica, económica, operativa y social

1.5.1 Viabilidad técnica

En la actualidad, el uso de tecnologías de información y comunicación como sistemas inteligentes, redes TIC entre otros ayudan a reducir la contaminación ambiental y recursos naturales. Para el presente proyecto de investigación se necesita información histórica la cual ya fue adquirida del Sistema Comercial con Base de Datos Relacional (SICI-BDR) que posee la empresa prestadora de servicios de agua potable EMSAP CHANKA de la ciudad de Andahuaylas además, además de herramientas software como MySQL, Erwin data modeler entre otros para realizar los procesos de limpiar, transformar y cargar teniendo como resultado un Data Werehaouse que nos permita realizar el análisis de datos para luego obtener los patrones de comportamiento y realizar diferentes tomas de decisiones.

1.5.2 Viabilidad Económica

El costo de las herramientas hardware y demás materiales al realizar la investigación son asumidos por quien presenta el proyecto, cubriendo el costo total del proyecto que asciende a S/. 3.884.70 soles.

Tabla 1: Análisis de costo de Inversión

ÍTEM	DESCRIPCIÓN	Cantidad	UNID. MEDIDA	Precio Unit	Precio parcial
1.00.00	**BIENES**				**1608,00**
1.01.00	**Hardware y accesorios**				**1.540,00**
1.01.01	Computadoras	1	Unid	1.500,00	1.500,00
1.01.02	Memoria USB 8GB	2	Unid	20,00	40,00
1.02.00	**Materiales de escritorio**				**68,00**
1.02.01	Papel Bond A4	2	Millar	30,00	60,00
1.02.02	Lapiceros	3	Unid	1,50	4,50
1.02.03	Cuaderno de apuntes	1	Unid	3,50	3,50
2.00.00	**SERVICIOS**				**830,00**
2.01.00	**Servicios diversos**				**730,00**
2.01.01	Internet	300	Horas	1,00	300,00
2.01.02	Fotocopias	1000	Unid	0,10	100,00
2.01.03	Anillado	6	Unid.	5,00	30,00
2.01.04	Empastado de Tesis	5	Unid	30,00	150,00
2.01.05	Impresión	1500	Unid.	0,10	150,00

2.02.00	**Servicios personales**				**100,00**
2.02.01	Revisión ortográfica, de estilo, redacción y diseño	1	Evento	100,00	100,00
3.00.00	**MATERIAL BIBLIOGRÁFICO**				**380,00**
3.01.00	**Libros de Especialidad**				**280,00**
3.01.01	Libros algoritmos de clustering	2	Unid	100,00	200,00
3.01.02	Libros de minería de datos	1	Unid	80,00	80,00
3.02.00	**Libros Metodológicos**				**100,00**
3.01.03	Libros de Metodología de Investigación	2	Unid	50,00	100,00
4.00.00	**RECURSOS HUMANOS**				**400,00**
4.01.00	**Asesorías externas**	**4**	**Sesión**	**100,00**	**400,00**
5.00.00	**GASTOS GENERALES**				**666,70**
5.01.00	Transporte	4	mes	40,00	160,00
5.02.00	**Imprevistos**				
5.02.01	Imprevistos 15% del costo total	1	Unid		506,70
TOTAL					**3.884,70**

Fuente: Elaboración propia

1.5.3 Viabilidad Operativa

Al realizar el análisis de los datos y obtener los patrones de comportamiento, la empresa prestadora del servicio de agua podrá seguir utilizando el método realizado ya que cuentan con personal capacitado además de hardware disponible y así de esta manera puedan tomar decisiones o acciones para mejorar el consumo de agua.

1.5.4 Viabilidad social

La sociedad se fue acoplando al uso de la tecnología a medida que pasa el tiempo, mejora la calidad de vida además de automatizar algunos procesos que ayudan a optimizar tiempo y recursos; es viable socialmente porque beneficia aproximadamente a 7347 usuarios de la empresa prestadora del servicio de agua potable en Andahuaylas, de esta manera se podrán implementar planes de concientización hacia la población y mejorar el uso cotidiano que se da al agua potable.

1.6 Limitación del estudio

El presente trabajo de investigación está limitado a que la institución prestadora del servicio de agua potable Emsap Chanka, al obtener la información de los patrones de comportamiento resultantes de la investigación no tomen decisiones ni realicen acciones para mejorar el modo de consumo del agua potable y se pueda preservar para el futuro.

Solo se obtuvo y analizó los datos de los usuarios que poseían un medidor en su vivienda, por lo cual los datos de los usuarios que no tuvieron un medidor y no recibieron el servicio de la empresa prestadora de servicios de agua potable Emsap Chanka no fueron analizados.

Solo se analizó a los 1575 usuarios que se encontraron en el sector 4 que se encontró dentro de la jurisdicción que posee la empresa prestadora de servicios de agua potable Emsap Chanka.

CAPÍTULO 2: MARCO TEÓRICO

2.1 Estado del arte

2.1.1 Antecedentes internacionales

Salamó, Gras, & Font-Mayolas. (2010), realizaron una investigación titulada "Patrones De Consumo De Alcohol En La Adolescencia" desarrollada en la Universidad de Girona en España, en este estudio analiza los patrones de consumo de alcohol y el papel de la influencia social y la percepción de peligrosidad de esta sustancia en dicho consumo. Asimismo, confirma la influencia del entorno (amigos y familiares) y de la percepción de peligrosidad del alcohol tanto en el consumo de esta sustancia como en su uso de alto riesgo, con la conclusión de que el modelo predictivo del consumo de alcohol incluye variables como la edad, el consumo del mejor amigo y del hermano y la percepción de peligrosidad de esta sustancia. Tienden a consumir los adolescentes de más edad, que tienen amigos o algún hermano consumidor y que perciben el alcohol como poco peligroso. Las variables que mejor predicen el consumo de alto riesgo son el sexo, la edad y el consumo de alcohol por parte de algún hermano. Así, la probabilidad de realizar un consumo de riesgo es mayor entre las chicas, los adolescentes mayores y aquellos que tienen algún hermano consumidor de alcohol. No obstante, la poca variabilidad explicada por ambos modelos sugiere la existencia de otras variables predictivas relevantes que no se han considerado en el presente estudio.

Diez, Lacruz. (2008), realizaron una investigación titulada "Patrones de comportamiento relacionados con la salud en una muestra española de población general" en la Universidad de Valencia en España. Realizaron un análisis en una muestra de población general sobre la presencia de conductas de riesgo para la salud como: consumo de tabaco, consumo de alcohol, consumo de café, falta de higiene bucodental, ausencia de control del peso corporal e inactividad física realizados análisis de correlación para determinar relaciones entre las conductas medidas, así como análisis de varianza para establecer comparaciones en función de las variables género y edad, con la conclusión que en general, los hábitos no saludables han aparecido más presentes entre los varones, los hallazgos de la presente investigación muestran que los hábitos de cuidado de la

salud muestran, en todos los casos, una incidencia diferencial en los distintos rangos de edad, aunque no existe una pauta genérica que explique las relaciones entre ambas variables, sino que el sentido de estas relaciones está en función del tipo de comportamiento del que se trate.

2.1.2 Antecedentes nacionales

Grandéz. (2017), realizo una investigación titulada “Aplicación de minería de datos para determinar patrones de consumo futuro en clientes de una distribuidora de suplementos nutricionales” en la Universidad San Ignacio de Loyola. Ha planteado encontrar reglas que determinen el patrón de consumo de clientes en una distribuidora de suplementos nutricionales, se aplica técnicas de Minería de datos a través del software Visual Studio 2015 que contiene la extensión Service Analysis y como motor de Base de datos hacemos uso de SQL server 2014, con la conclusión que el algoritmo de asociación podemos ver que se forman reglas con cierta probabilidad de ocurrencia, así como también con una importancia, estas reglas lo que hacen es encontrar patrones sólidos dentro de la base de datos que relacione las características de los consumidores con los productos que suelen consumir, mientras tanto en el algoritmo neuronal se indica que tan determinante es una variable en la compra de un determinado producto y en el algoritmo Clúster se agrupa las variables de acuerdo al valor que puede tomar y asigna a cada una un nivel de ocurrencia.

2.2 Marco conceptual

2.2.1 Tecnologías de la información y comunicación

En líneas generales podríamos decir que las nuevas tecnologías de la información y comunicación son las que giran en torno a tres medios básicos: la informática, la microelectrónica y las telecomunicaciones; pero giran, no sólo de forma aislada, sino lo que es más significativo de manera interactiva e Inter conexionadas, lo que permite conseguir nuevas realidades comunicativas. (Cabero, 1998)

Las TIC se desarrollan a partir de los avances científicos producidos en los ámbitos de la informática y las telecomunicaciones, son el conjunto de tecnologías que permiten el acceso, producción, tratamiento y comunicación de

información presentada en diferentes códigos (texto, imagen, sonido,...). El elemento más representativo de las nuevas tecnologías es sin duda el ordenador y más específicamente, Internet. Como indican diferentes autores, Internet supone un salto cualitativo de gran magnitud, cambiando y redefiniendo los modos de conocer y relacionarse del hombre. (Ortí, 2011)

Tabla 2: Características de las NN.TT. de la Información y Comunicaciones

Características de las NN.TT. de la Información y Comunicaciones.
Inmaterialidad.
Interactividad e interconexión.
Instantaneidad.
Elevados parámetros de calidad de imagen y sonido.
Digitalización.
Influencia más sobre los procesos que sobre los productos.
Penetración en todos los sectores (culturales, económicos, educativos, industriales…).
Innovación.
Tendencia hacia automatización.
Diversidad.

Fuente: Cabero, 1998: Características de las NN.TT

2.2.2 Minería de datos

Fuera del ámbito informático, el término “minería” hace referencia al arte de extraer minerales de la corteza terrestre. Gracias a la minería se obtienen materiales preciados como, por ejemplo, el oro o la plata. El término “dato” se define formalmente como el valor que toma una variable, parámetro, atributo, característica, etc. De forma más general un dato es la representación mediante algún símbolo (número, letra, etc.), de un atributo de una determinada entidad. La minería de datos es una disciplina de la informática que estudia el análisis de grandes cantidades de datos con el objetivo de obtener conocimiento a partir de ellos. (Lara, 2015)

La disponibilidad de grandes volúmenes de información y el uso generalizado de herramientas informáticas ha transformado el análisis de datos orientándolo hacia determinadas técnicas especializadas englobadas bajo el nombre de

minería de datos o data mining. Las técnicas de minería de datos persiguen el descubrimiento automático del conocimiento contenido en la información almacenada de modo ordenado en grandes bases de datos. (López, 2004, p.1)

2.2.3 Proceso de descubrimiento de conocimiento en bases de datos

El proceso KDD es interactivo e iterativo conteniendo diferentes pasos. Riquelme, Ruiz, & Gilbert (2006) mencionan los siguientes pasos.

- Comprender el dominio de aplicación: este paso incluye el conocimiento relevante previo y las metas de la aplicación.
- Extraer la base de datos objetivo: recogida de los datos, evaluar la calidad de los datos para familiarizarse con ellos.
- Preparar los datos: incluye limpieza, transformación integración y reducción de datos a la vez que disminuir el tiempo requerido por el algoritmo de aprendizaje aplicado posteriormente.
- Minería de datos: es la fase fundamental del proceso. Está constituido por una o más de las siguientes funciones de clasificación, regresión, clustering, resumen, recuperación de imágenes, extracción de reglas, etc.
- Interpretación: explicar los patrones descubiertos, así como la posibilidad de visualizarlos.
- Utiliza el conocimiento descubierto: hacer uso del modelo creado.

"El proceso KDD persigue la extracción automatizada de conocimiento no trivial, implícito, previamente desconocido y potencialmente útil a partir de grandes volúmenes de datos" (Lara, 2014, p.14).

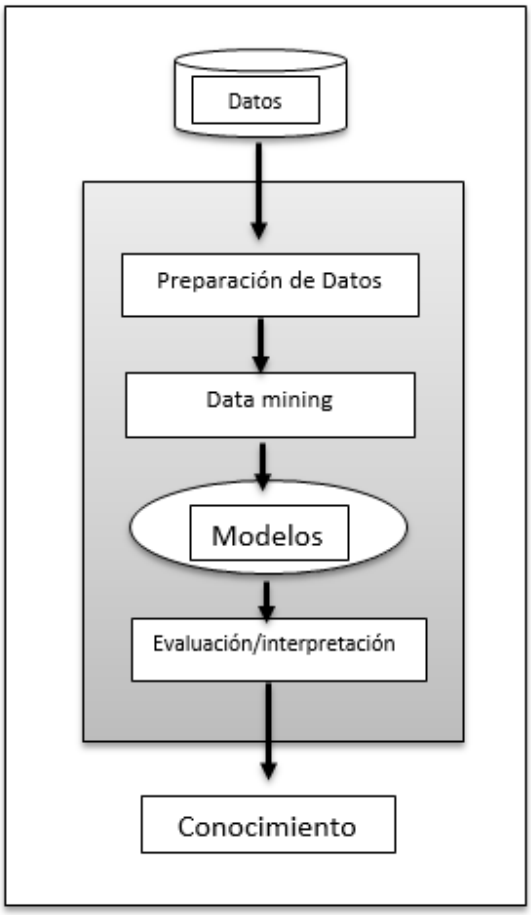


Figura 1: Proceso KDD
Fuente: Lara (2014), Visión general del proceso KDD

2.2.4 Metodología CRISP-DM

Grimes(como se citó en Grandéz, 2017), El proceso CRISP-DM se desarrolló por medio del esfuerzo de un consorcio inicialmente compuesto por Daimler Chryrler, SPSS y NCR. CRISP-DM (CROSS-Industry Standard Process for Data Mining). Consiste en un ciclo que consta de seis etapas:

1. Comprensión del negocio. - Esta fase inicial se centra en la comprensión de los objetivos del proyecto y los requisitos desde una perspectiva de negocio, a continuación, convertir este conocimiento en una definición de problema de minería de datos y un plan preliminar diseñado para lograr los objetivos.
2. Comprensión de los datos. - La fase de comprensión de los datos comienza con una recopilación inicial de datos y prosigue con las actividades para familiarizarse con los datos, identificar problemas de calidad de los datos, descubrir los primeros datos o detectar subconjuntos interesantes para

formar hipótesis para Información oculta.

3. Preparación de datos. - La fase de preparación de datos abarca todas las actividades para construir el conjunto de datos final a partir de los datos iniciales sin procesar.
4. Modelado. - En esta fase, se seleccionan y aplican diversas técnicas de modelado y se calibran sus parámetros a valores óptimos.
5. Evaluación. - En esta etapa el modelo o modelos obtenidos se evalúan más a fondo y se revisan los pasos ejecutados para construir el modelo para asegurarse de que alcanza adecuadamente los objetivos de negocio.
6. Implementación. - La creación del modelo generalmente no es el final del proyecto. Incluso si el propósito del modelo es aumentar el conocimiento de los datos, el conocimiento adquirido deberá ser organizado y presentado de manera que el cliente pueda utilizarlo.

2.2.5 Técnicas

"Estas técnicas tienen como objetivo descubrir patrones, perfiles y tendencias a través del análisis de los datos utilizando tecnologías de reconocimiento de patrones, redes neuronales, lógica difusa, algoritmos genéticos y otras técnicas avanzadas de análisis de datos" (López, 2007, p.1).

En general, cualquiera que sea el problema a resolver, no existe una única técnica para solucionarlo, sino que puede ser abordado siguiendo aproximaciones distintas. Aluja (2001) menciona las siguientes técnicas.

- Análisis Factoriales Descriptivos. Permiten hacer visualizaciones de realidades multivariantes complejas y, por ende, manifestar las regularidades estadísticas, así como eventuales discrepancias respecto de aquella y sugerir hipótesis de explicación.
- Market Basket Analysis o análisis de la cesta de compra. Permite detectar que productos se adquieren conjuntamente, permite incorporar variables técnicas que ayudan en la interpretación, como el día de la semana, localización, forma de pago. También puede aplicarse en contextos diferentes de las grandes superficies, en particular el e-comercio, e incorporar el factor temporal.

- Técnicas de "clustering". Son técnicas que parten de una medida de

proximidad entre individuos y a partir de ahí, buscar los grupos de individuos más preciados entre sí, según una serie de variables mesuradas.

- Series Temporales. A partir de la serie de comportamiento histórica, permite modelizar las componentes básicas de la serie, tendencia, ciclo y estacionalidad y así poder hacer predicciones para el futuro, tales como cifra de ventas, previsión de consumo de un producto o servicio, etc.
- Redes bayesianas. Consiste en representar todos los posibles sucesos en que estamos interesados mediante un grafo de probabilidades condicionales de transición entre sucesos. Puede codificarse a partir del conocimiento de un experto o puede ser inferido a partir de los datos. Permite establecer relaciones causales y efectuar predicciones.
- Modelos Lineales Generalizados. Son modelos que permiten tratar diferentes tipos de variables de respuesta, por ejemplo, la preferencia entre productos concurrentes en el mercado. Al mismo tiempo, los modelos estadísticos se enriquecen cada vez más y se hacen más flexibles y adaptativos, permitiendo abordar problemas cada vez más complejos: (GAM, Projection Pursuit, PLS, MARS).
- Previsión local. La idea de base es que individuos parecidos tendrán comportamientos similares respecto de una cierta variable de respuesta. La técnica consiste en situar los individuos en un espacio euclídeo y hacer predicciones de su comportamiento a partir del comportamiento observado en sus vecinos.
- Redes neuronales. Inspiradas en el modelo biológico, son generalizaciones de modelos estadísticos clásicos. Su novedad radica en el aprendizaje secuencial, el hecho de utilizar transformaciones de las variables originales para la predicción y la no linealidad del modelo. Permite aprender en contextos difíciles, sin precisar la formulación de un modelo concreto. Su principal inconveniente es que para el usuario son una caja negra.
- Arboles de decisión. Permiten obtener de forma visual las reglas de decisión bajo las cuales operan los consumidores, a partir de datos históricos almacenados. Su principal ventaja es la facilidad de interpretación.

- Algoritmos genéticos. También aquí se simula el modelo biológico de la evolución de las especias, solo que a una velocidad infinitamente mayor. Es una técnica muy prometedora. En principio cualquier problema que se plantee, como la optimización de una combinación entre distintas componentes, estando estas componentes sujetas a restricciones, puede resolverse mediante algoritmos genéticos.

2.2.6 Clustering

La agrupación de datos es un proceso de extracción de patrones previamente desconocidos, válidos, posicionalmente útiles y ocultos de grandes conjuntos de datos. El Clustering de Datos es una técnica estadística de análisis de datos sin supervisión. Se utiliza para la clasificación en grupos homogéneos descubriendo patrones y la relación oculta ayuda a tomar decisiones rápidas y eficientes. Con esta técnica se segmenta un gran conjunto de datos en subconjuntos llamados clusters (grupos). Cada cluster es una colección de objetos que son similares entre sí y se colocan dentro del mismo grupo, pero son diferentes a los objetos en otros clusters. Montoya (2016)

Según la forma en la que, aplicando dicha medida de similaridad, se construyen los clusters, los algoritmos de clustering se pueden dividir en cuatro grandes tipos. Lara (2014) menciona los siguientes:

- Clustering particional. Los algoritmos de este tipo intentan buscar una división del conjunto de datos en subconjuntos con intersección vacía. Todos ellos realizan una asignación de los objetos a los diferentes clusters en función de la proximidad de dichos objetos a un representante elegido para cada cluster. El número de clusters se indica inicialmente y, tras una serie de iteraciones, se alcanza una partición óptima de los datos. El algoritmo K-means (K-medias) es el representante más conocido.
- Clustering jerárquico. Las técnicas de clustering jerárquico (también llamado clustering basado en jerarquías) se basan en la generación de sucesiones ordenadas (jerarquías) de clusters. La estructura jerárquica se representa en forma de árbol y se llama dendograma. AGNES (AGlomerative NESting) y DIANA (DIvisive ANAlysis) son dos de los algoritmos de clustering jerárquico más conocidos.

- Clustering Basado en densidad. Son algoritmos basados en el concepto de densidad de un punto, que mide el número de puntos que son alcanzables desde el considerando un determinado radio. El algoritmo más conocido de clustering basado en densidad es DBSCAN.
- Clustering basado en grid. Estos algoritmos se basan en una división del espacio en un número finito de celdas que constituyen una estructura en forma de rejilla en la que se llevan a cabo las operaciones de clustering. Los objetos que se encuentran en cada celda. El agrupamiento se lleva a cabo utilizando la información estadística en cada celda, en vez de usar todo el conjunto de datos. STING es un ejemplo de algoritmo de clustering basado en grid.

2.2.7 Algoritmos de agrupamiento

El problema de formar grupos en un conjunto de datos es muy importante el conocimiento del comportamiento de una población de la cual solo se tiene una cantidad n de sus elementos. La solución de estos problemas se realiza mediante la creación de algoritmos de agrupamiento. (Pascual, Pla y Sánchez, 2007) afirman tres métodos de agrupamiento:

- Los métodos de agrupamiento no paramétricos pueden dividirse en tres grupos fundamentales: jerárquicos, particionales y basados en densidad.
- Los algoritmos jerárquicos son aquellos en los que se va particionando el conjunto de datos por niveles, de modo tal que en cada nivel generalmente, se unen o se dividen dos grupos del nivel anterior, según si es un algoritmo aglomerativo o divisivo.
- Los algoritmos particionales son los que realizan una división inicial de los datos en grupos y luego mueven los objetos de un grupo a otro según se optimice alguna función objetivo.

2.2.7.1 Algoritmo K-MEANS

Dicho informalmente, el procedimiento k-means consiste simplemente en comenzar con k grupos, cada uno de los cuales consiste en un solo punto aleatorio, y luego agregar cada nuevo punto al grupo cuya media es el nuevo punto más cercano. Después de agregar un punto a un grupo, la media de ese grupo se ajusta para tener en cuenta el nuevo

punto. Así, en cada etapa los k-means son, de hecho, los medios de los grupos que representan (de ahí el término k-means) (MacQueen, 1967, p. 283).

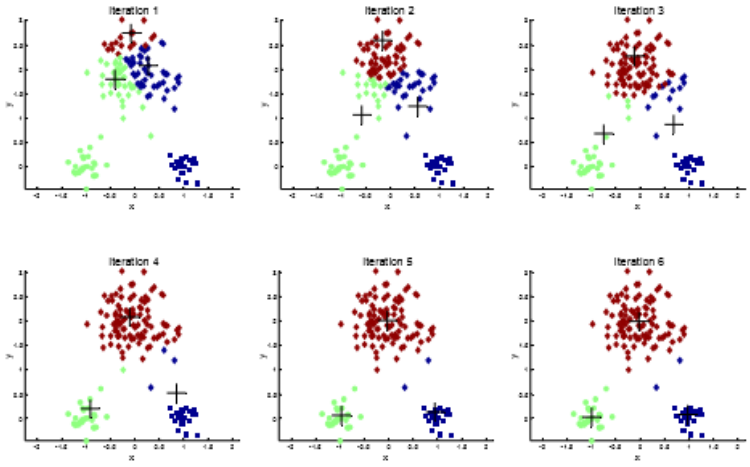


Figura 2*: Clustering particional*
Fuente: Fernando Berzal, Clustering particional

Para calcular los centroides se utilizará la distancia euclídea y luego se minimiza utilizando la media aritmética para cada variable:

- $d(t_i, t_j) = \sqrt{\sum_{m=1}^{p} W_m (t_{im} - t_{jm})}$

K-Means comienza particionando los datos en k subconjuntos no vacíos, aleatoriamente o usando alguna heurística. Luego calcula el centroide de cada partición como el punto medio del clúster y asigna cada dato al clúster cuyo centroide sea el más próximo. Luego los centroides son recalculados para los grupos nuevos y el algoritmo se repite hasta la convergencia, la cual es obtenida cuando no haya más datos que cambien de grupo de una iteración a otra. Para calcular el centroide más cercano a cada punto se debe utilizar una función de distancia. Para datos reales se suele utilizar la distancia euclídea. Para datos categóricos se debe establecer una función específica de distancia para ese conjunto de datos. Algunas de las opciones son utilizar una matriz de distancias predefinidas o una función heurística. (Valencia, 2017, p. 32)

El algoritmo K-means contempla cuatro pasos y estas son las siguientes:

a. Primero se escoge los objetos en K subconjuntos aleatoriamente, asignando un punto al centroide que se encuentre más cerca.
b. Mientras obtenemos el centroide más cercano, se calcula los puntos a los k centroides.
c. Asignar cada objeto al clúster cuyo centroide se encuentren más cercano.
d. Se vuelve a recalcular los centroides hasta que no se encuentre un centroide más cercano.

2.2.7.2 Algoritmo COBWEB

Fisher (como se citó en Garre, Cuadrado, Sicilia, Rodríguez & Rejas, 2007), se trata de un algoritmo jerárquico, se caracteriza por que utiliza aprendizaje incremental, esto es, realiza agrupaciones instancia a instancia. Durante su ejecución del algoritmo se forma un árbol (árbol de clasificación) donde las hojas representan los segmentos y el nodo raíz engloba por completo el conjunto de datos de entrada. Al principio el árbol consiste en un único nodo raíz. Las instancias se van añadiendo una a una y el árbol se va actualizando en cada paso. La actualización consiste en encontrar el mejor sitio donde incluir la nueva instancia, operación que puede necesitar de la reestructuración de todo árbol (incluyendo la generación de un nuevo nodo anfitrión para la instancia y/o la fusión/partición de nodos existentes.

Según Fisher (como se citó en Garre et al., 2007), El algoritmo es muy sensible a otros dos parámetros:

- Acuity: este parámetro es muy necesario, ya que la utilidad de categoría se basa en una estimación de la media y la desviación estándar del valor de los atributos, pero cuando se estima la desviación estándar del valor de un atributo para un nodo en particular, el resultado es cero si dicho nodo sólo contiene una instancia.

- Cut-off: este valor se utiliza para evitar el crecimiento desmesurado del número de segmentos. Indica el grado de mejoría que se debe producir en la utilidad de categoría para que la instancia sea tenida en cuenta de manera individual.

2.2.7.3 Algoritmo K-NEAREST NEIGHBORS

K-nearest neighbors es un algoritmo simple que almacena todos los casos disponibles y clasifica los casos nuevos en función de una medida de similitud (por ejemplo, funciones de distancia). KNN se ha utilizado en la estimación estadística y reconocimiento de patrones ya en el comienzo de la década de 1970 como una técnica no paramétrica. Un caso se clasifica por el voto mayoritario de sus vecinos, asignándose el caso a la clase más común entre sus vecinos K más cercanos medida por una función de distancia. Si K = 1, entonces el caso simplemente se asigna a la clase de su vecino más cercano. (Data minig map, 2018)

FUNCIÓN DE DISTANCIA

EUCLIDEANA	$\sqrt{\sum_{i=1}^{k}(X_i - Y_i)^2}$
MANHATTAN	$\Sigma_{i=1}^{k}\lvert X_i - Y_i^2\rvert$
MINKOWSKI	$\left[\Sigma_{i=1}^{k}(\lvert X_i - Y_i^2\rvert)^q\right]^{1/q}$

Elegir el valor óptimo para K se logra mejor inspeccionando primero los datos. En general, un gran valor de K es más preciso ya que reduce el ruido general, pero no hay garantía. La validación cruzada es otra forma de determinar retrospectivamente un buen valor K mediante el uso de un conjunto de datos independiente para validar el valor K. Históricamente, la K óptima para la mayoría de los conjuntos de datos ha estado entre 3-10. Eso produce resultados mucho mejores que 1NN. (Data minig map, 2018)

2.2.7.4 Algoritmo CURE

Emplea un novedoso algoritmo de agrupamiento jerárquico que adopta un término medio entre los extremos basados en centroides y puntos extremos. En CURE, primero se elige un número constante de puntos bien dispersos en un clúster. Los puntos dispersos capturan la forma y la extensión del clúster. Los puntos dispersos elegidos se reducen luego hacia el centroide del conglomerado por una fracción cr. Estos puntos dispersos después de la reducción se utilizan como representantes del clúster. Los clústeres con el par más cercano de puntos representativos son los clústeres que se fusionan en cada paso del algoritmo de agrupación jerárquica de CURE. El enfoque de puntos dispersos empleado por CURE alivia las deficiencias de los enfoques tanto de puntos como de centroides. Para el algoritmo basado en centroide, el espacio que constituye la vecindad del centroide único para un clúster es esférico. Por lo tanto, favorece los cúmulos esféricos, divide los cúmulos alargados. (Guha, Rastogi, y Shim, 1998)

2.2.7.5 Algoritmo BFR

El algoritmo BFR asume que cada cluster está normalmente distribuido en torno a un centroide en un espacio euclídeo. Se cuantifica la probabilidad de que un punto se encuentre en un cluster dada la distancia del punto al centroide, dimensión por dimensión. La suposición de normalidad implica que los clusters son elipsoides alineados con los ejes (las desviaciones en distintas dimensiones pueden ser diferentes).

Tres tipos de puntos

- DS [Discard Set]: Puntos cercanos al centroide (resumidos)
- CS [Compression Set]: Puntos cercanos entre sí pero lejos de centroides.
- RS [Retained Set]: Puntos aislados, a la espera de ser asignados a un CS.

2.2.8 Patrones de comportamiento

Según Gonzales (2016), se denomina patrón de comportamiento cuando ciertas reacciones de la persona, se hacen muy frecuentes en determinados ambientes o situaciones, constituyen lo que llamamos un patrón de comportamiento. Un patrón de comportamiento es una forma constante que tiene una persona, de pensar, sentir, reaccionar físicamente y actuar en determinada situación.

Gonzales (2016), menciona los siguientes tipos de patrones:

A. Patrones Constructivos

   a. Realista o Racional-Objetivo:

      Conductas apropiadas a la situación, sentimientos de paz y seguridad, actitud firme, palabras claras, con un tono de voz y volumen apropiado. Sus afirmaciones y juicios se basan en hechos comprobados. Mira las cosas como realmente son.

   b. Crítico Positivo:

      Se esfuerza porque haya y se respeten límites y reglas razonables. Es firme y consistente en sus juicios, que se basan en hechos comprobados. Es protector y justo cuando aplica disciplina. No maltrata, ni humilla.

   c. Adaptado Sano:

      Cumple las reglas y normas, aunque le cause malestar. Hace lo que haya que hacer, aunque no le guste, lo hace de buen ánimo y manera correcta, como si fuera para Dios. Se deja proteger. Es responsable y cumplidor de sus deberes. Cuida todo lo que es suyo y participa en el cuidado de los otros y del ambiente.

   d. Nutritivo Positivo:

      Está pendiente y ayuda en la satisfacción de las necesidades de las otras personas y también de sus propias necesidades. Apoya, estimula y alienta a los demás y se da a sí mismo aliento, estímulo y reconocimiento. Identifica sus recursos y los utiliza con talento y sabiduría.

   e. Natural y Espontáneo:

      Tiene conciencia de lo que está sintiendo y lo expresa sanamente. Siente y expresa las emociones y sentimientos agradables (amor, y alegría) y las que son desagradables como rabia, miedo y tristeza. Está familiarizado con su cuerpo y con sus necesidades. Disfruta sin arriesgarse, ni dañar a otros.

B. Patrones Destructivos

a. Crítico Negativo:

Para los críticos negativos, los otros están mal y sólo ellos están bien. Acostumbran hacer juicios o críticas de valor sobre los otros. Resaltan constantemente las fallas y faltas de los demás. Se inventan cualquier razón para no dar reconocimiento positivo a los demás.

b. Sobreprotector:

Atiende exageradamente las necesidades de los demás y desatiende las propias. No pregunta, ni oye para conocer las necesidades de los otros, las asume. Funciona como un adivino o una especie de dios, que cree saber lo que los demás necesitan y se jacta de eso.

c. Rebelde:

Se opone a las autoridades, a las reglas, normas y regulaciones. No se deja proteger, porque no acepta consejos, ni sugerencias. Mantiene la actitud de ser sabio de su propia opinión. Sus conductas predominantes son la agresividad, la violencia y el aislamiento. Su emoción preferida es la ira.

d. Servil o Fácil de Dominar:

Actúa según lo que cree que son las expectativas de los demás, para complacerles y ganarse así el afecto, aceptación y respeto de ellos. Es asustadizo, inseguro, débil y depresivo. Mantiene creencias y actitudes de pesimismo y derrota. Es constantemente negativo. No utiliza los recursos que Dios le ha dado para su protección y desarrollo.

e. Vengativo o Rencoroso:

Le cuesta perdonar porque recuerda constantemente lo que una vez alguien le dijo, le hizo o dejó de hacerle o de decirle. Cultiva la venganza hacia los otros y la lleva a cabo a cualquier costo y de cualquier manera. Predomina en estas personas el resentimiento, la rabia, el rencor y la actitud vengativa.

2.2.9 Comportamiento del consumidor

La definición del comportamiento del consumidor es aquel “proceso de decisión y actividades físicas individuales en el cual se comprometen la evaluación, adquisición y uso económico de los bienes y servicios”. De modo que se puede intuir el concepto en cuestión, con la finalidad de comprender, explicar e intentar pronosticar las acciones humanas relativas al consumo. De las posiciones anteriores se desprenden que el comportamiento de los consumidores es un proceso complejo, que comprende distintas actividades y funciones desde que aquellos deciden comprar hasta que finalmente lo hacen. Y un poco más allá, si se diferencia entre comprador y consumidor (siendo este último que finalmente compra y “destruye” o disfruta del bien o servicio), el proceso sería una decisión compartida o individual, si es el caso en el que el comprador y consumidor son personas u hogares. Los consumidores pueden ser individuos u hogares, o indirectamente negocios privados, organismos o agencias de gobierno otras instituciones. (Quispe, 2012)

El comportamiento del consumidor se define como el comportamiento que los consumidores muestran al buscar, comprar, utilizar, evaluar y desechar los productos y servicios que, consideran, satisfarán sus necesidades. El comportamiento del consumidor se enfoca en la forma en que los individuos toman decisiones para gastar sus recursos disponibles (tiempo, dinero y esfuerzo). (Schiffman y Kanuk, 2005).

Según Schiffman y Kanuk. (2005), el comportamiento del consumidor se divide en dos:

a) Consumidor personal. - Compra bienes y servicios para su consumo propio, para uso de su familia o como obsequio para un amigo. En cada uno de tales contextos, los productos son comprados para consumo final por parte de individuos, a quienes se conoce como usuarios finales o consumidores últimos.
b) Consumidor organizacional. - Incluye empresas con propósitos de lucro o sin ellos, dependencias gubernamentales (locales, estatales y nacionales) e instituciones (por ejemplo, escuelas, hospitales y prisiones), todos los cuales deben comprar productos, equipo y servicios para mantener en marcha sus organizaciones.

CAPÍTULO 3: MATERIALES Y MÉTODOS

3.1 Propuesta de solución

3.1.1 Modelo conceptual

Para la determinación de patrones de comportamiento de consumo de agua potable con algoritmos de clusterización en la provincia de Andahuaylas se tomó como modelo el siguiente esquema.

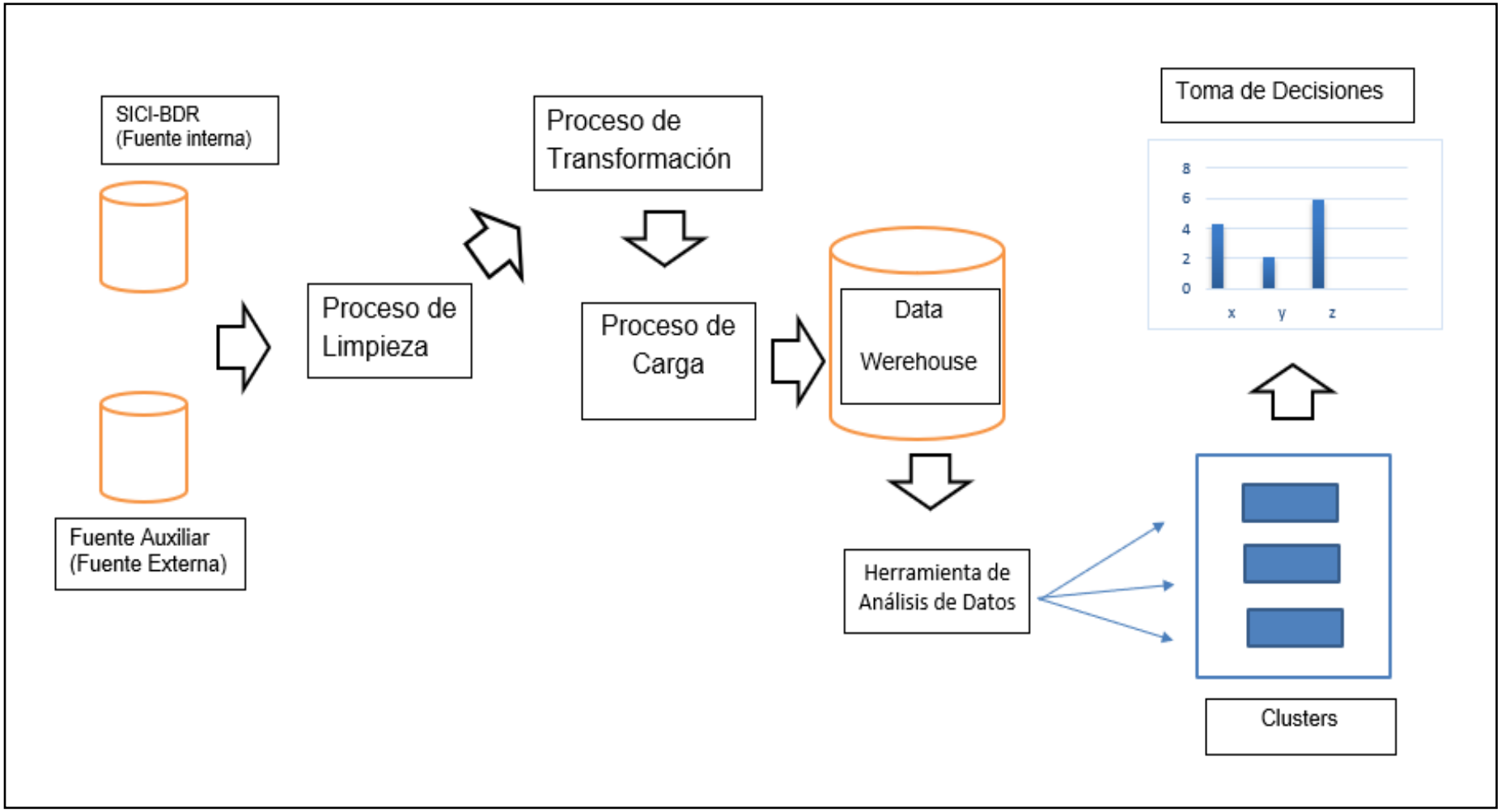


Figura 3: Modelo conceptual para la determinación de patrones de comportamiento de consumo de agua potable con algoritmos de clusterización en la provincia de Andahuaylas.

Fuente: Elaboración propia.

3.2 Gestión de proyecto

Para la elaboración del plan del Informe Final de tesis se ha tomado como referencia los conocimientos, técnicas y prácticas vigentes, para la gestión exitosa de proyectos reunidas por el PMI (Project Management Institute) en el documento llamado PMBOK (Project Management Body of Knowledge) quinta edición.

A continuación, se muestra las áreas de conocimiento y los procesos que se tomaran en cuenta para el presente proyecto de tesis:

3.2.1 Procesos de la dirección del proyecto

Se agrupan en 5 procesos los cuales son:

- **El grupo de procesos de iniciación**: Aquí se encuentran los procesos que definieron el proyecto de investigación en el Emsap Chanka de Andahuaylas.
- **El grupo de procesos de planificación:** En este grupo de procesos se definió los procesos que establecieron el alcance del proyecto, definieron objetivos y las acciones que se tomaron para alcanzar los objetivos.
- **El grupo de procesos de ejecución:** en este grupo se encuentran los procesos que se realizaron para completar el proyecto.
- **El grupo de procesos de seguimiento y control:** en este grupo se encuentran los procesos que se usaron para dar seguimiento, analizar y regular el progreso y el desempeño del proyecto, además que permitió identificar áreas en las que el proyecto requiere cambios.
- **El grupo de procesos de cierre:** En este grupo están los procesos que permitieron finalizar todas las actividades a fin de cerrar formalmente el proyecto de investigación en la entidad prestadora de servicio municipal de saneamiento municipal chanka en Andahuaylas**.**

3.2.2 Áreas de conocimiento

En la versión actual de PMBOK (5ta. Edición), cuenta con 10 áreas de conocimiento, con las cuales se desarrolló el presente proyecto:

A. **Gestión de la Integración del Proyecto**: El proyecto está basado en la implementación del algoritmo K-means para determinar patrones de comportamiento de consumo de agua potable en la ciudad de Andahuaylas.

La gestión del proyecto está dada por la integración de los conocimientos, técnicas y prácticas vigentes en el PMBook.

Tabla 3: Gestión de Integración del proyecto

PROYECTO	**DETERMINACIÓN DE PATRONES DE COMPORTAMIENTO DE CONSUMO DE AGUA POTABLE CON ALGORITMOS DE CLUSTERIZACIÓN EN LA PROVINCIA DE ANDAHUAYLAS.**		
Patrocinador	**Jefe de proyecto:** Anai Luz Amparo Lira Flores		
Preparado por	**Jefe de Proyecto:** Anai Luz Amparo Lira Flores	**Fecha**	08/05/2018
Revisado por	**Asesor de Proyecto:** Ing. Juan José Ore Cerrón	**Fecha**	10/06/2018
Aprobado por	**Universidad: Facultad** de ingeniería	**Fecha**	16/07/2018
REVISIÓN	**DESCRIPCIÓN**		**FECHA**
1	Preparación de acta de constitución (Anai Luz Amparo Lira Flores)		16/07/2018
BREVE DESCRIPCIÓN DEL SISTEMA DE PROYECTO			
El proyecto incluirá el desarrollo de un Sistema que ayude a determinar patrones de comportamiento que beneficiará a la entidad prestadora de servicio de saneamiento municipal chanka en Andahuaylas garantizando una mejor toma de decisiones para optimizar el consumo de agua potable.			
ALINEAMIENTO DEL PROYECTO			
OBJETIVOS ESTRATÉGICOS DE LA ORGANIZACIÓN	**PROPÓSITO DEL PROYECTO**		
Obtener patrones de comportamiento del consumo de agua potable con algoritmos de clusterizacion en la provincia de Andahuaylas.	Optimizar el consumo de agua potable en la ciudad de Andahuaylas.		
OBJETIVOS DEL PROYECTO			
1. Recopilar y preparar los datos históricos encontrados en la Empresa Prestadora de servicios Emsap Chanka.			
2. Implementar y clasificar los datos utilizando el algoritmo de clusterizacion.			
3. Obtener los patrones de comportamiento y evaluar los resultados.			

FACTORES CRÍTICOS DE ÉXITO DEL PROYECTO	
4. Disposición de las herramientas para el desarrollo del sistema dentro de los plazos establecidos.	
5. Diseño del sistema acorde a los requerimientos de los interesados	
REQUERIMIENTO DE ALTO NIVEL	
1. El sistema debe ser desarrollada exactamente de acuerdo a los requerimientos funcionales planteados.	
EXTENSIÓN Y ALCANCE DEL PROYECTO	
FASES DEL PROYECTO	**PRINCIPALES ACTIVIDADES**
FASE I: COMPRENDER EL DOMINIO DE APLICACIÓN	Análisis de problemática
FASE II: EXTRAER LA BASE DE DATOS OBJETIVO	Recojo de datos
	Evaluar calidad de datos
FASE III: PREPARAR LOS DATOS	Limpieza y preprocesamiento de datos
	Transformación de datos
FASE IV: MINERÍA DE DATOS	Selección de la tarea de minería de datos
	Elección y aplicación del algoritmo de clustering
FASE V: INTERPRETACIÓN	Evaluar resultados
FASE VI: UTILIZA EL CONOCIMIENTO DESCUBIERTO	Aplicar el conocimiento encontrado a partir de los resultados.
INTERESADOS CON EL PROYECTO	
1. Personal la entidad prestadora de servicio de saneamiento municipal chanka.	
2. Programador o desarrollador	
RIESGOS	
1. Demora en la adquisición de la laptop y las herramientas de desarrollo.	
2. Poca información brindada y tiempo disponible por los interesados.	
3. Retraso en la programación	
HITOS PRINCIPALES DEL PROYECTO	
1. Aprobación del acta de constitución del proyecto.	
2. Alcanzar todos los objetivos trazados.	
3. Pruebas y correcciones de la aplicación.	
PRESUPUESTO DEL PROYECTO	

El costo del proyecto que haciende a la suma total cuatro mil novecientos sesenta y nueve 0/100 soles S/ 4, 969.00, el cual es asumido en su totalidad por el jefe del proyecto.

Fuente: Elaboración propia

B. Gestión del Alcance del Proyecto: El alcance del proyecto que tiene la DETERMINACIÓN DE PATRONES DE COMPORTAMIENTO DE CONSUMO DE AGUA POTABLE CON ALGORITMOS DE CLUSTERIZACION EN LA PROVINCIA DE ANDAHUAYLAS, está dirigido directamente al personal de la Entidad prestadora de Servicio de Saneamiento Municipal Chanka, para lo cual se utiliza la Estructura de Descomposición de Trabajo (EDT).

La Estructura de Descomposición de Trabajo Determinación De Patrones De Comportamiento De Consumo De Agua Potable Con Algoritmos De Clusterizacion En La Provincia De Andahuaylas, está compuesta por cuatro (06) fases, en las cuales se describe cada una de las actividades realizadas, tal como se muestra en la figura 23.

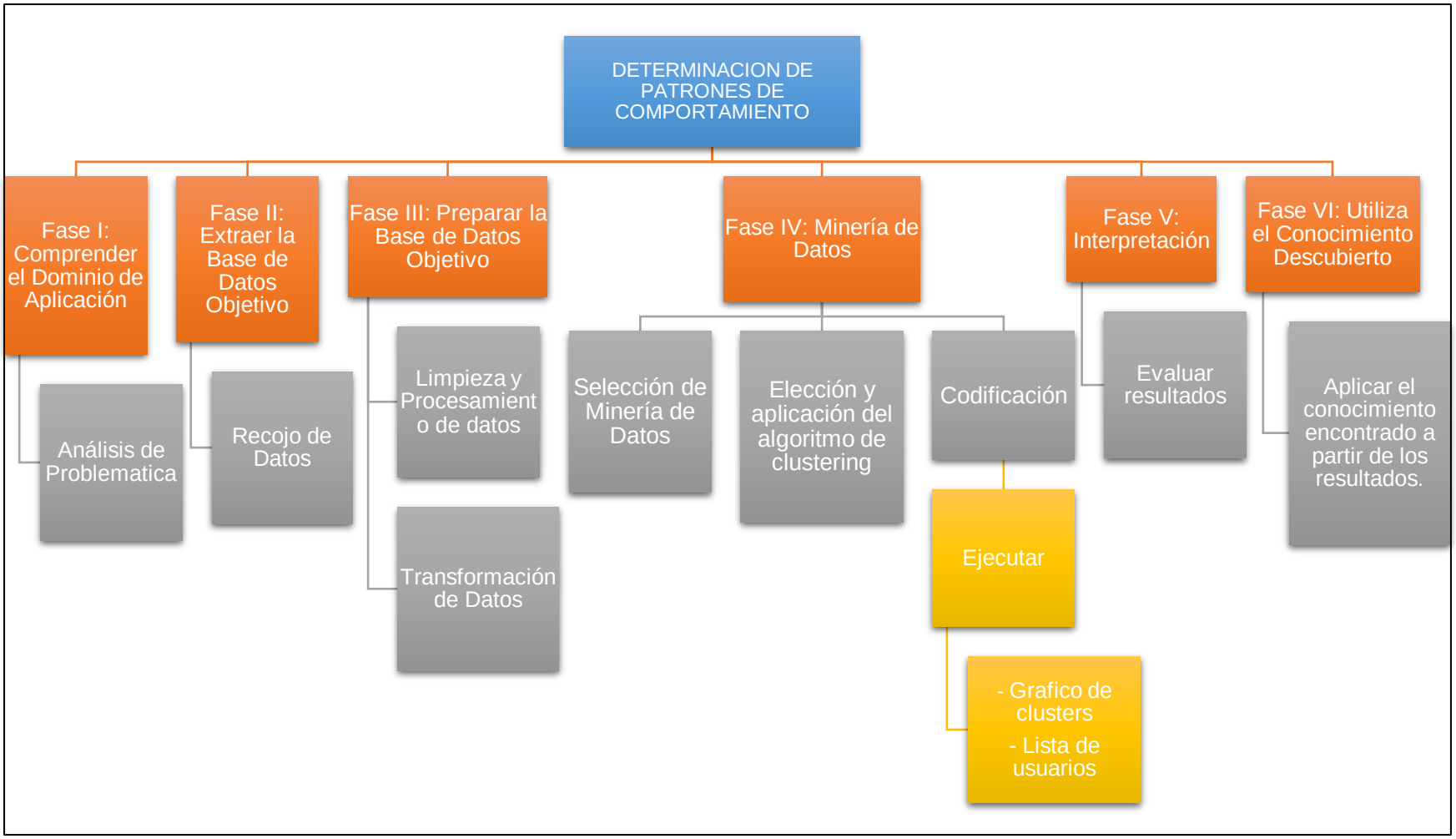


Figura 4: Estructura de Descomposición de Trabajo (EDT)

Fuente: Elaboración propia

C. **Gestión del Tiempo del Proyecto**: la gestión de tiempo del proyecto se llevó a cabo a través de un diagrama GANT teniendo una ruta crítica de 9 actividades que a continuación se detalla:

- **Documentación de proyecto de tesis:** La elaboración del proyecto de investigación, tiene un tiempo de duración de 50 días dando inicio el martes 08 de mayo del 2018 hasta el día lunes 16 de julio del 2018 sin tomar en cuenta los días sábados y domingos.

- **Recopilación y análisis de datos:** La recopilación de datos y análisis de los mismos tienen un tiempo de duración de 23 días dando inicio el día martes 17 de julio hasta el jueves 16 de agosto de 2016.

- **Preparación de los datos:** La preparación de datos, tiene una duración de 15 días, dando inicio el día martes 17 de julio hasta el día lunes 06 de agosto del 2018.

- **Implementación del algoritmo de clusterización:** La implementación del algoritmo, tiene un tiempo de duración de 40 días iniciando el martes 07 de agosto hasta el lunes 01 de octubre del 2018.

- **Análisis de los resultados obtenidos**: El análisis de los resultados que se obtuvo con la implementación del algoritmo, tuvo una duración de 07 días, iniciando el martes 02 de octubre hasta el 09 de octubre del 2018.

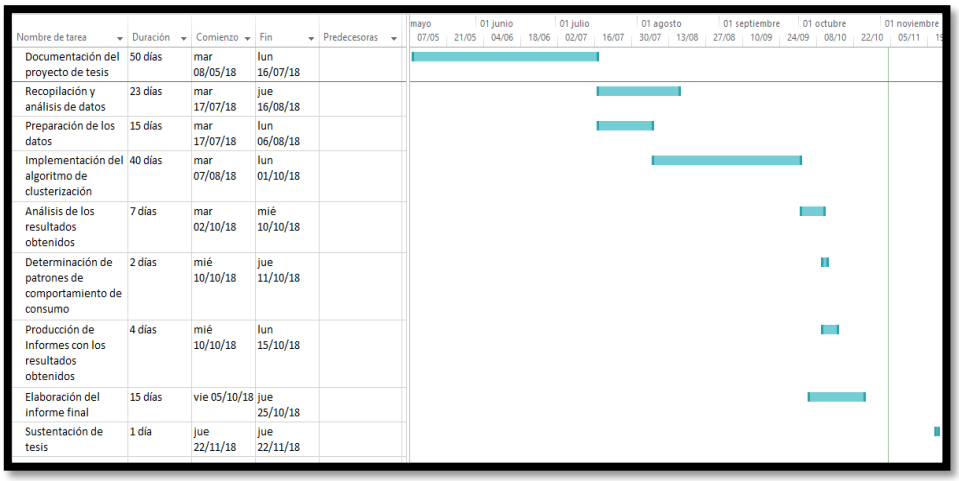
- **Determinación de patrones de comportamiento de consumo:** La determinación de patrones de comportamiento tuvo una duración de 02 días, iniciando desde el 10 de octubre hasta el 11 de octubre del 2018.

- **Producción de informes con los resultados obtenidos:** La producción de los informes con los resultados obtenidos después de la implementación tuvo una duración de 04 días, iniciando desde miércoles 10 de octubre hasta el lunes 15 de octubre del 2018.

- **Elaboración del informe final:** La elaboración del informe final consta de 15 días, iniciando el viernes 05 de octubre hasta el viernes 25 de octubre del 2018.

- **Sustentación y defensa del informe final:** La Sustentación y defensa de la tesis, tiene un tiempo de duración de 14 días dando inicio el día viernes 26 de noviembre hasta el 30 de noviembre de 2018.

El proyecto está estimado en un tiempo de duración de 140 días, dando inicio el día martes 08 de mayo del 2018 con la recopilación de información y se sustentará el Informe Final de tesis final el día jueves 22 de noviembre del 2018, el mismo que se realizó en un diagrama de Gantt, tal como se muestra en figura 5:



Nombre de tarea	Duración	Comienzo	Fin	Predecesoras
Documentación del proyecto de tesis	50 días	mar 08/05/18	lun 16/07/18	
Recopilación y análisis de datos	23 días	mar 17/07/18	jue 16/08/18	
Preparación de los datos	15 días	mar 17/07/18	lun 06/08/18	
Implementación del algoritmo de clusterización	40 días	mar 07/08/18	lun 01/10/18	
Análisis de los resultados obtenidos	7 días	mar 02/10/18	mié 10/10/18	
Determinación de patrones de comportamiento de consumo	2 días	mié 10/10/18	jue 11/10/18	
Producción de Informes con los resultados obtenidos	4 días	mié 10/10/18	lun 15/10/18	
Elaboración del informe final	15 días	vie 05/10/18	jue 25/10/18	
Sustentación de tesis	1 día	jue 22/11/18	jue 22/11/18	

Figura 5: Cronograma de Actividades
Fuente: Elaboración propia.

D. **Gestión de los costos del proyecto:** la gestión de costos del proyecto se basará el 5 Ítems que se mencionado en la tabla 10 Presupuesto, que se detalla a continuación.

Ítem 1: bienes, dividido en Materiales de Escritorios y Materiales respectivamente haciendo un total de S/ 1 608.00soles.

Ítem 2: Servicios, dividido en Servicios de Comunicación, Servicios de Movilidad y Servicios de Impresión que haciende la suma respectivamente haciendo un total de S/. 830.00 soles.

Ítem 3: Material Bibliográfico, para el desarrollo del sistema se utilizó diverso material bibliográfico respectivamente haciendo un total de S/380.00 soles para el desarrollo del proyecto.

Ítem 4: Recursos Humanos, dividido en el Asesor, que hacienden a la suma de S/400.00 soles.

Ítem 5: Gastos Generales, En este Ítem se consideran los gastos imprevistos que es el 15% de la suma total del proyecto haciendo un total de S/ 666.70 soles.
El costo de proyecto en los 5 ítems antes mencionados haciende a un total de S/ 3 884.70 soles.

Tabla 4: Gestión de Costo del proyecto

ÍTEM	DESCRIPCIÓN	Cantidad	UNID. MEDIDA	Precio Unit.	Precio parcial
1.00.00	BIENES				1608,00
1.01.00	Hardware y accesorios				1.540,00
1.01.01	Computadoras	1	Unid	1.500,00	1.500,00
1.01.02	Memoria USB 8GB	2	Unid	20,00	40,00
1.02.00	Materiales de escritorio				68,00
1.02.01	Papel Bond A4	2	Millar	30,00	60,00

1.02.02	Lapiceros	3	Unid	1,50	4,50
1.02.03	Cuaderno de apuntes	1	Unid	3,50	3,50
2.00.00	**SERVICIOS**				**830,00**
2.01.00	**Servicios diversos**				**730,00**
2.01.01	Internet	300	Horas	1,00	300,00
2.01.02	Fotocopias	1000	Unid	0,10	100,00
2.01.03	Anillado	6	Unid.	5,00	30,00
2.01.04	Empastado de Tesis	5	Unid	30,00	150,00
2.01.05	Impresión	1500	Unid.	0,10	150,00
2.02.00	**Servicios personales**				**100,00**
2.02.01	Revisión ortográfica, de estilo, redacción y diseño	1	Evento	100,00	100,00
3.00.00	**MATERIAL BIBLIOGRÁFICO**				**380,00**
3.01.00	**Libros de Especialidad**				**280,00**
3.01.01	Libros algoritmos de clustering	2	Unid	100,00	200,00
3.01.02	Libros de minería de datos	1	Unid	80,00	80,00
3.02.00	**Libros Metodológicos**				**100,00**
3.01.03	Libros de Metodología de Investigación	2	Unid	50,00	100,00
4.00.00	**RECURSOS HUMANOS**				**400,00**
4.01.00	**Asesorías externas**	**4**	**Sesión**	**100,00**	**400,00**
5.00.00	**GASTOS GENERALES**				**666,70**
5.01.00	Transporte	4	mes	40,00	160,00
5.02.00	**Imprevistos**				
5.02.01	Imprevistos 15% del costo total	1	Unid		506,70
TOTAL					3.884,70

Fuente: Elaboración propia

E. Gestión de los Recursos Humanos del Proyecto: La gestión de Recursos Humanos se desarrolló de acuerdo al planificador organizacional del personal involucrado en el proyecto tal como se muestra en la figura 6.

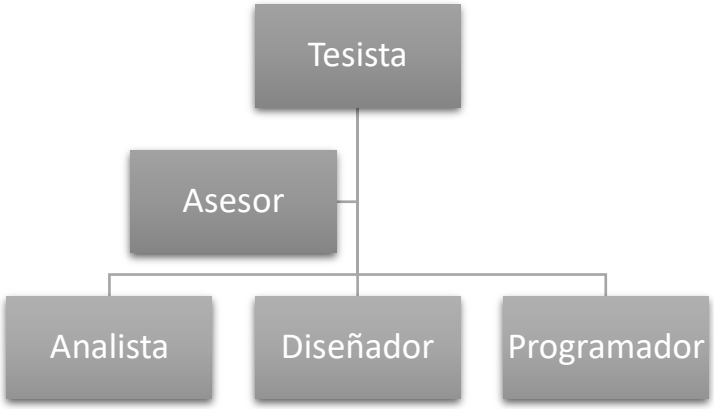


Figura 6: Organigrama del personal involucrado

Fuente: Elaboración propia

Tabla 5: Perfil de Recursos Humanos del Proyecto

PERSONAL	PERFIL
Tesista	Bachiller en Ingeniería de Sistemas.
Asesor	Ingeniero de Sistemas
Analista	• Conocimiento en minería de datos, clustering y patrones de comportamiento. • Modelado funcional: Diagrama de flujo de datos. • Modelado de datos y sus técnicas • Conocimiento de la tecnología: arquitectura de software, bases de datos.
Diseñador	***Dominio*** de Python.
Programador	• Entre sus conocimientos destacan ser gestor de base de datos.

Fuente: Elaboración propia

Tabla 6: Asignación de responsabilidad al personal

ÍTEM	ROL	PERFIL	FUNCIONES	RESULTADOS
1	Asesor	Asesoría	Asesorar	Cumplido
			Revisar	Cumplido
2	Programador	Python	Analizar	Cumplido
			Diseñar	Cumplido
			Implementar	Cumplido
			Realizar pruebas	Cumplido

Fuente: Elaboración propia

F. Gestión de las Comunicaciones del Proyecto: la Gestión de comunicación del proyecto se realizó mediante reuniones programadas con el asesor durante el desarrollo del proyecto e informe final de acuerdo a la siguiente manera:

Tabla 7: Reunión con el asesor

REUNIÓN CON EL ASESOR DE TESIS	
Frecuencia	Una vez a la semana
Día de la semana	Los días viernes
Duración	2 horas
Agenda	Revisión del avance del informe Observaciones del informe Correcciones del informe

Fuente: Elaboración propia

G. Gestión de los riesgos del proyecto: La Gestión de riesgos de proyecto se identificó de acuerdo a las categorías de riesgo de la ISO 31000 tal como se muestra en la figura 7.

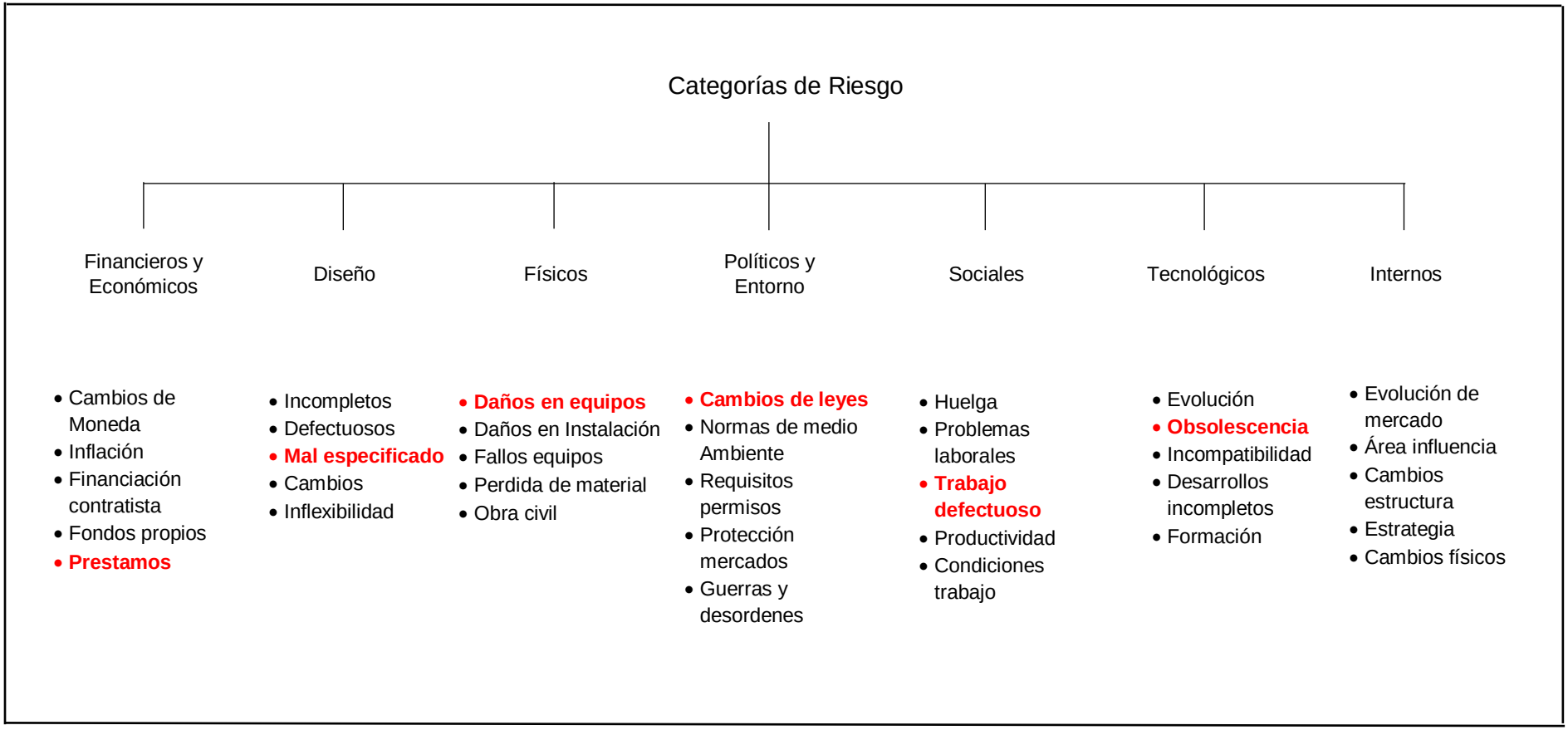


Figura 7: Categorías de riesgo

Fuente: www.iso31000.com
Fuente: Elaboración propia

Prestamos: será considerado medio por la necesidad de ellos y un eventual recorte de estos en el futuro podría representar un riesgo para el proyecto.

Mal especificados: será considerado bajo por que los profesionales que ejecutan el proyecto tienen un amplio conocimiento.

Daños de Equipos: será considerado bajo ya que se cuenta con garantía de los equipos por su reciente adquisición.

Cambio de Leyes: Podrían ser más exigentes y por ende involucrar mayor inversión en un determinado momento, en tanto es considerado medio.

Trabajos defectuosos: es considerado bajo ya que se trabaja con una adecuad planificación y diseño.

Obsolescencia: Los quipos e instalaciones serán nuevas con un mínimo de 3 años de vida útil siendo el horizonte del proyecto de 1.5 años, en tanto se considera bajo.

Tabla 8: Probabilidad de riesgo

	PROBABILIDAD DE QUE SUCEDA	IMPACTO	RIESGO
Préstamo	**2**	**3**	**5**
Mal especificado	**1**	**3**	**4**
Daño de equipos	**1**	**3**	**4**
Cambios de leyes	**2**	**3**	**5**
Trabajos defectuosos	**1**	**3**	**4**
Obsolescencia	**1**	**3**	**4**

Fuente: Elaboración propia

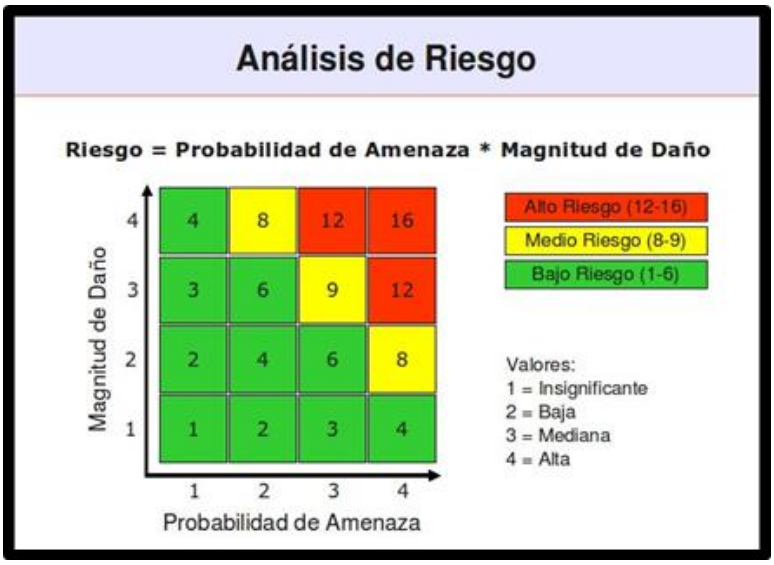


Figura 8: Análisis de riesgo

Fuente: Elaboración propia

El proyecto alcanza valores de 5, 4, 4, 5, 4 y 4 considerando como de Bajo riesgo.

H. Gestión de las Adquisiciones del Proyecto: La gestión de las adquisiciones del proyecto se desarrollará de la siguiente manera:

Tabla 9: Gestión de las adquisiciones del proyecto

ÍTEM	DESCRIPCIÓN	Cantidad	UNID. MEDIDA	Precio Unit	Precio parcial	Forma de Adquisición	Meses – 2018				
							julio	agosto	setiembre	octubre	Noviembre
1.00.00	**BIENES**				**1608,00**						
1.01.00	**Hardware y accesorios**				**1.540,00**						
1.01.01	Computadoras	1	Unid	1.500,00	1.500,00	Compra	1.500,00				
1.01.02	Memoria USB 8GB	2	Unid	20,00	40,00	Compra	20,00	20,00			
1.02.00	**Materiales de escritorio**				**68,00**						
1.02.01	Papel Bond A4	2	Millar	30,00	60,00	**Compra**	**30,00**			**30,00**	
1.02.02	Lapiceros	3	Unid	1,50	4,50	Compra	1,50	3,00			
1.02.03	Cuaderno de apuntes	1	Unid	3,50	3,50	Compra	3,50				
2.00.00	**SERVICIOS**				**830,00**						
2.01.00	**Servicios diversos**				**730,00**						
2.01.01	Internet	300	Horas	1,00	300,00	**Alquiler**	**75,00**	75,00	75,00	75,00	
2.01.02	Fotocopias	1000	Unid	0,10	100,00	Compra	40,00			60,00	
2.01.03	Anillado	6	Unid.	5,00	30,00	Compra	15,00			15,00	
2.01.04	Empastado de Tesis	5	Unid	30,00	150,00	Compra				150,00	
2.01.05	Impresión	1500	Unid.	0,10	150,00	Compra	50,00			100,00	
2.02.00	**Servicios personales**				**100,00**						
2.02.01	Revisión ortográfica, de estilo, redacción y diseño	1	Evento	100,00	100,00	**Contrato**	**25,00**	**25,00**	**25,00**	**25,00**	
3.00.00	**MATERIAL BIBLIOGRÁFICO**				**380,00**						

3.01.00	**Libros de Especialidad**				**280,00**						
3.01.01	Libros algoritmos de clustering	2	Unid	100,00	200,00	**Compra**	**200,00**				
3.01.02	Libros de minería de datos	1	Unid	80,00	80,00	Compra	80,00				
3.02.00	**Libros Metodológicos**				**100,00**						
3.01.03	Libros de Metodología de Investigación	2	Unid	50,00	100,00	**Compra**	**100,00**				
4.00.00	**RECURSOS HUMANOS**				**400,00**						
4.01.00	**Asesorías externas**	**4**	**Sesión**	**100,00**	**400,00**	**Contrato**	**100,00**	**100,00**	**100,00**	**100,00**	
5.00.00	**GASTOS GENERALES**				**666,70**						
5.01.00	Transporte	4	mes	40,00	160,00	Compra	40,00	40,00	20,00	40,00	20,00
5.02.00	**Imprevistos**										
5.02.01	Imprevistos 15% del costo total	1	Unid		506,70	Compra	101,34	101,34	101,34	101,34	101,34
FLUJO DE CAJA					**3.884,70**		**2381,34**	**364,34**	**321,34**	**696,34**	**121,34**

Fuente: Elaboración Propia

3.3 Desarrollo de la determinación de patrones de comportamiento de consumo de agua

El desarrollo de la determinación de patrones de comportamiento de consumo de agua, se basó en seis fases de la metodología KDD (Knowledge Discovery in Databases), debido a que permite encontrar un modelo entendible que describa patrones de acuerdo a la información utilizada.

3.3.1 Comprender el dominio de aplicación

3.3.1.1 Análisis de la problemática

Uno de los principales problemas que se observó en la ciudad de Andahuaylas fue disminuir el consumo de agua 9 horas al día a comparación de 4 años atrás que era 24 horas al día, las instituciones educativas u otras instituciones públicas que no tenían un tanque de agua fueron afectadas ya que los servicios higiénicos podían ser un comienzo a enfermedades o infecciones hacia los niños y adolescentes, en los hogares las amas de casa tenían que empozar agua en recipientes para realizar sus actividades con normalidad con el paso del tiempo tener agua empozada muchos días puede causar algunas enfermedades como el dengue o infecciones estomacales; la mejor solución que se presenta es utilizar los datos de consumo y categoría de cada usuario de la EPS Emsap Chanka para analizarlos utilizando un algoritmo de clusterización para obtener las características que poseen cada grupo y de esta manera determinar los patrones de comportamiento de consumo de agua.

3.3.2 Extraer la base de datos objetivo

3.3.2.1 Selección de datos

Para poder llegar a acceder a los datos se fue a la Empresa Prestadora de Servicios Emsap Chanka y solicitar datos que poseen, se logró recolectar datos de 1570 usuarios todos del sector cuatro, los datos obtenidos se encontraban en documentos .csv, se obtuvo dos documentos dentro de ellos los datos más resaltantes que se encontró fueron los siguientes: nombre del usuario, dirección del usuario, consumo de agua por mes, la categoría a la que pertenece el usuario, el año, el número id de usuario.

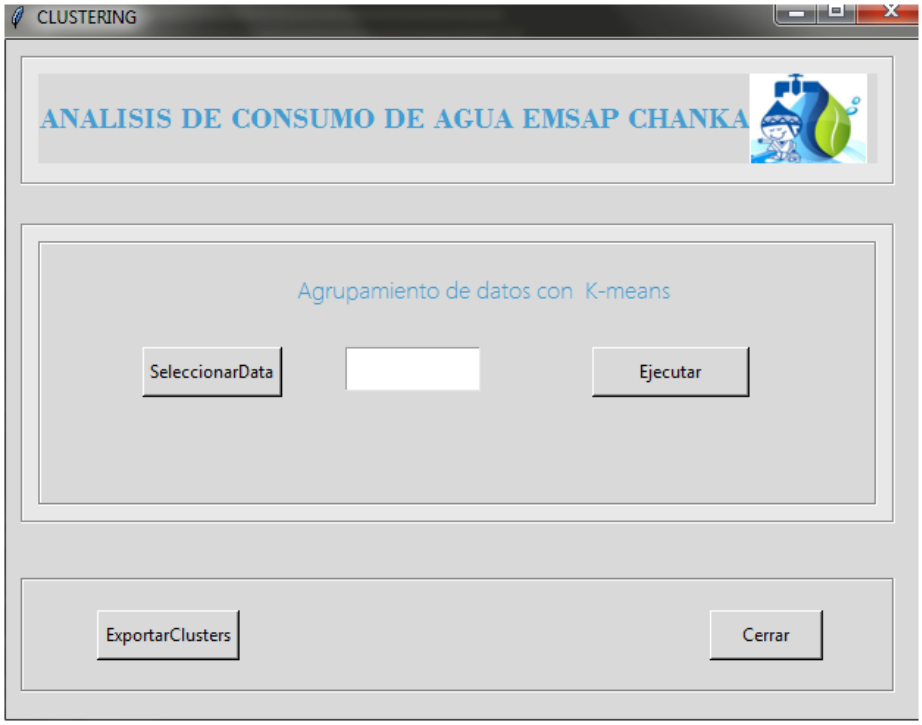
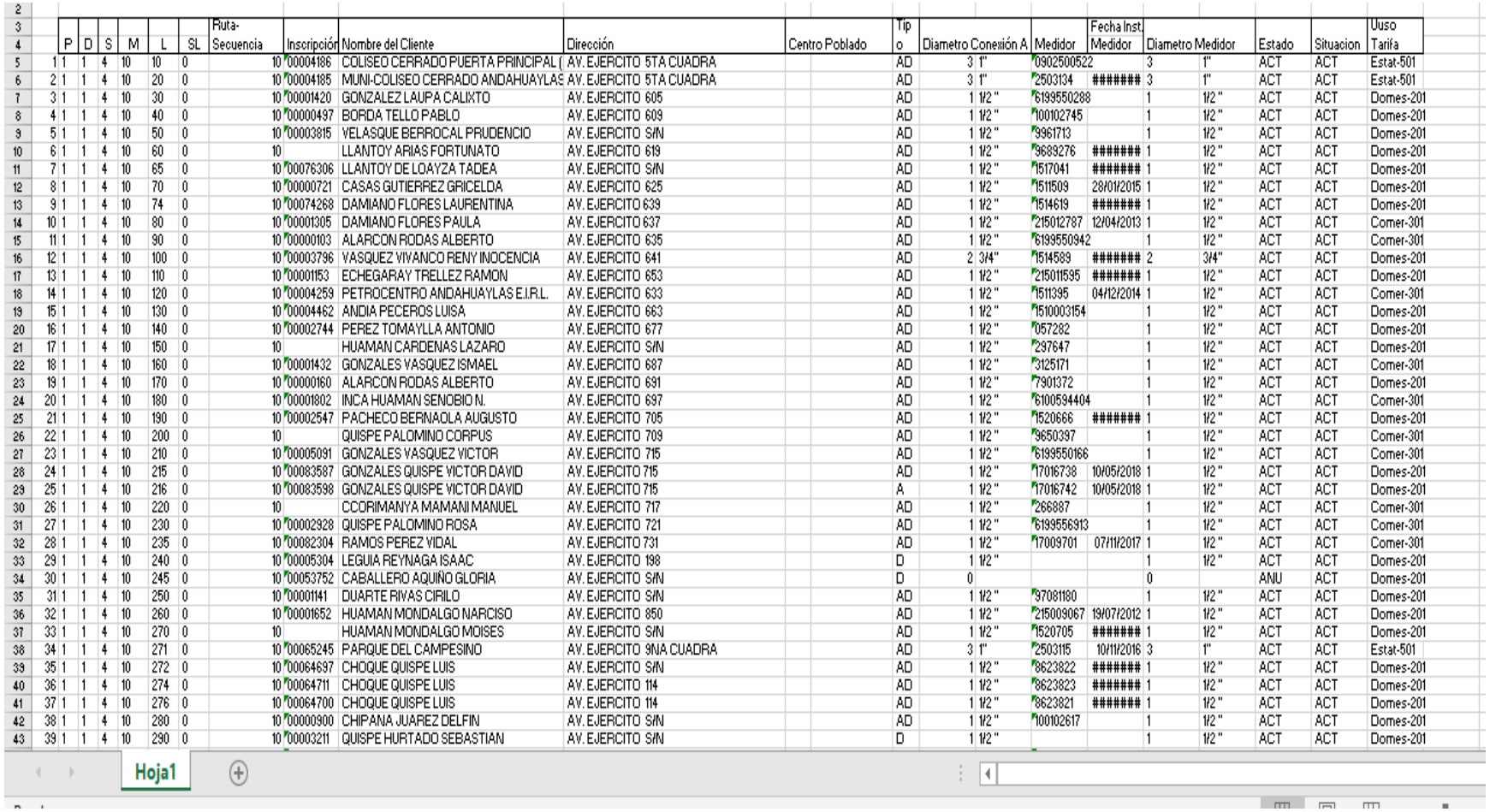


Figura 9: Interface para ejecutar el algoritmo
Fuente: Elaboración Propia

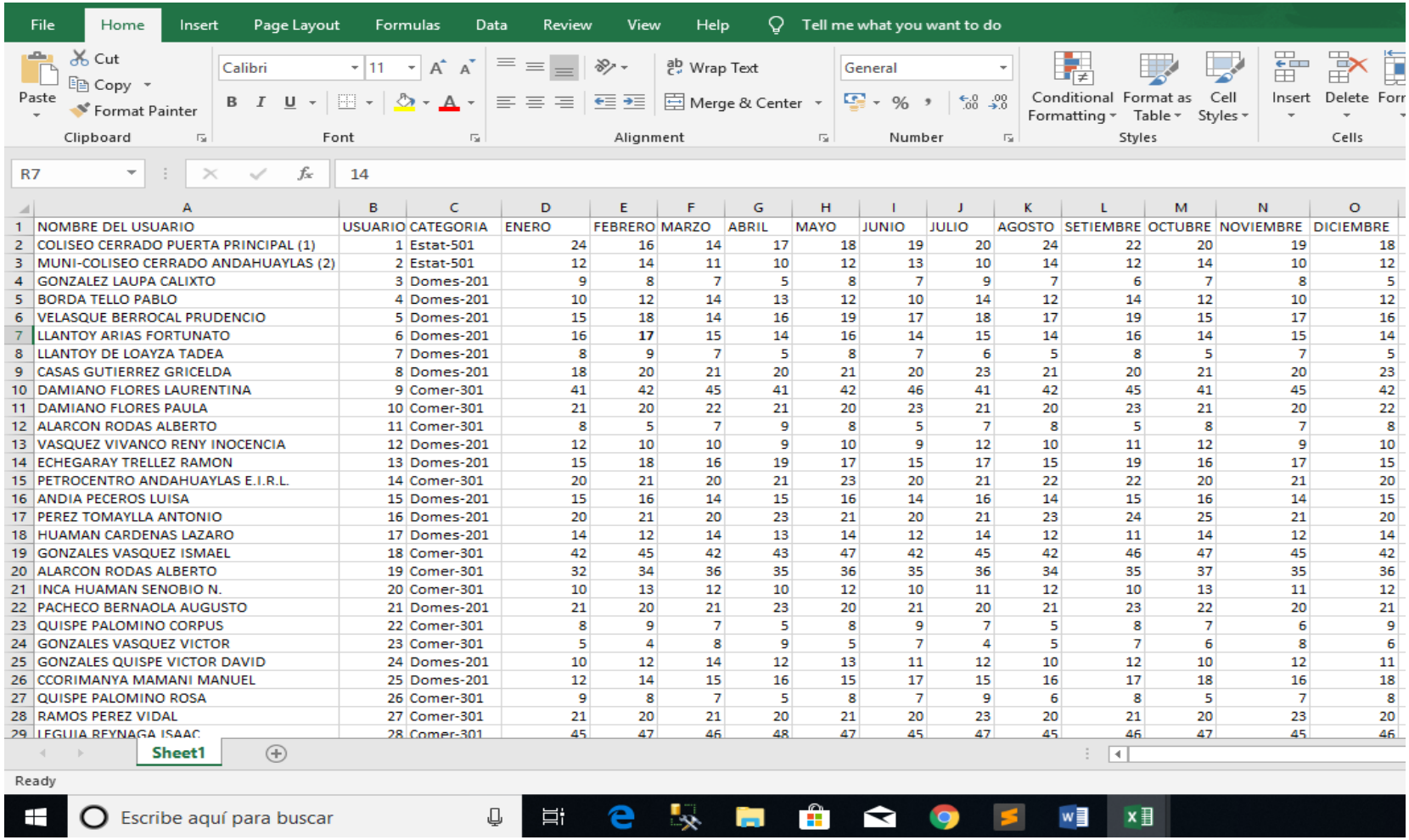


	P	D	S	M	L	SL	Ruta-Secuencia	Inscripción	Nombre del Cliente	Dirección	Centro Poblado	Tipo	Diametro Conexión A	Medidor	Fecha Inst. Medidor	Diametro Medidor	Estado	Situacion	Uuso Tarifa
1	1	1	4	10	10	0	10	00004186	COLISEO CERRADO PUERTA PRINCIPAL (	AV. EJERCITO 5TA CUADRA		AD	3 1"	0902500522		3 1"	ACT	ACT	Estat-501
2	1	1	4	10	20	0	10	00004185	MUNI-COLISEO CERRADO ANDAHUAYLAS	AV. EJERCITO 5TA CUADRA		AD	3 1"	2503134	########	3 1"	ACT	ACT	Estat-501
3	1	1	4	10	30	0	10	00001420	GONZALEZ LAUPA CALIXTO	AV. EJERCITO 605		AD	1 1/2"	6199550288		1 1/2"	ACT	ACT	Domes-201
4	1	1	4	10	40	0	10	00000497	BORDA TELLO PABLO	AV. EJERCITO 609		AD	1 1/2"	100102745		1 1/2"	ACT	ACT	Domes-201
5	1	1	4	10	50	0	10	00003815	VELASQUE BERROCAL PRUDENCIO	AV. EJERCITO S/N		AD	1 1/2"	3961713		1 1/2"	ACT	ACT	Domes-201
6	1	1	4	10	60	0	10		LLANTOY ARIAS FORTUNATO	AV. EJERCITO 619		AD	1 1/2"	9689276	########	1 1/2"	ACT	ACT	Domes-201
7	1	1	4	10	65	0	10	00076306	LLANTOY DE LOAYZA TADEA	AV. EJERCITO S/N		AD	1 1/2"	1517041	########	1 1/2"	ACT	ACT	Domes-201
8	1	1	4	10	70	0	10	00000721	CASAS GUTIERREZ GRICELDA	AV. EJERCITO 625		AD	1 1/2"	1511509	28/01/2015	1 1/2"	ACT	ACT	Domes-201
9	1	1	4	10	74	0	10	00074268	DAMIANO FLORES LAURENTINA	AV. EJERCITO 639		AD	1 1/2"	1514619	########	1 1/2"	ACT	ACT	Domes-201
10	1	1	4	10	80	0	10	00001305	DAMIANO FLORES PAULA	AV. EJERCITO 637		AD	1 1/2"	215012787	12/04/2013	1 1/2"	ACT	ACT	Comer-301
11	1	1	4	10	90	0	10	00000103	ALARCON RODAS ALBERTO	AV. EJERCITO 635		AD	1 1/2"	6199550942		1 1/2"	ACT	ACT	Comer-301
12	1	1	4	10	100	0	10	00003796	VASQUEZ VIVANCO RENY INOCENCIA	AV. EJERCITO 641		AD	2 3/4"	1514589	########	2 3/4"	ACT	ACT	Domes-201
13	1	1	4	10	110	0	10	00001153	ECHEGARAY TRELLEZ RAMON	AV. EJERCITO 653		AD	1 1/2"	215011595	########	1 1/2"	ACT	ACT	Domes-201
14	1	1	4	10	120	0	10	00004259	PETROCENTRO ANDAHUAYLAS E.I.R.L.	AV. EJERCITO 633		AD	1 1/2"	1511395	04/12/2014	1 1/2"	ACT	ACT	Comer-301
15	1	1	4	10	130	0	10	00004462	ANDIA PECEROS LUISA	AV. EJERCITO 663		AD	1 1/2"	1510003154		1 1/2"	ACT	ACT	Domes-201
16	1	1	4	10	140	0	10	00002744	PEREZ TOMAYLLA ANTONIO	AV. EJERCITO 677		AD	1 1/2"	057282		1 1/2"	ACT	ACT	Domes-201
17	1	1	4	10	150	0	10		HUAMAN CARDENAS LAZARO	AV. EJERCITO S/N		AD	1 1/2"	297647		1 1/2"	ACT	ACT	Domes-201
18	1	1	4	10	160	0	10	00001432	GONZALES VASQUEZ ISMAEL	AV. EJERCITO 687		AD	1 1/2"	3125171		1 1/2"	ACT	ACT	Comer-301
19	1	1	4	10	170	0	10	00000160	ALARCON RODAS ALBERTO	AV. EJERCITO 691		AD	1 1/2"	7901372		1 1/2"	ACT	ACT	Domes-201
20	1	1	4	10	180	0	10	00001802	INCA HUAMAN SENOBIO N.	AV. EJERCITO 697		AD	1 1/2"	6100594404		1 1/2"	ACT	ACT	Comer-301
21	1	1	4	10	190	0	10	00002547	PACHECO BERNAOLA AUGUSTO	AV. EJERCITO 705		AD	1 1/2"	1520666	########	1 1/2"	ACT	ACT	Domes-201
22	1	1	4	10	200	0	10		QUISPE PALOMINO CORPUS	AV. EJERCITO 709		AD	1 1/2"	9650397		1 1/2"	ACT	ACT	Comer-301
23	1	1	4	10	210	0	10	00005091	GONZALES VASQUEZ VICTOR	AV. EJERCITO 715		AD	1 1/2"	6199550166		1 1/2"	ACT	ACT	Comer-301
24	1	1	4	10	215	0	10	00083587	GONZALES QUISPE VICTOR DAVID	AV. EJERCITO 715		AD	1 1/2"	17016738	10/05/2018	1 1/2"	ACT	ACT	Domes-201
25	1	1	4	10	216	0	10	00083598	GONZALES QUISPE VICTOR DAVID	AV. EJERCITO 715		A	1 1/2"	17016742	10/05/2018	1 1/2"	ACT	ACT	Domes-201
26	1	1	4	10	220	0	10		CCORIMANYA MAMANI MANUEL	AV. EJERCITO 717		AD	1 1/2"	266887		1 1/2"	ACT	ACT	Comer-301
27	1	1	4	10	230	0	10	00002928	QUISPE PALOMINO ROSA	AV. EJERCITO 721		AD	1 1/2"	6199556913		1 1/2"	ACT	ACT	Comer-301
28	1	1	4	10	235	0	10	00082304	RAMOS PEREZ VIDAL	AV. EJERCITO 731		AD	1 1/2"	17009701	07/11/2017	1 1/2"	ACT	ACT	Comer-301
29	1	1	4	10	240	0	10	00005304	LEGUIA REYNAGA ISAAC	AV. EJERCITO 198		D	1 1/2"			1 1/2"	ACT	ACT	Domes-201
30	1	1	4	10	245	0	10	00053752	CABALLERO AQUIÑO GLORIA	AV. EJERCITO S/N		D	0			0	ANU	ACT	Domes-201
31	1	1	4	10	250	0	10	00001141	DUARTE RIVAS CIRILO	AV. EJERCITO S/N		AD	1 1/2"	97081180		1 1/2"	ACT	ACT	Domes-201
32	1	1	4	10	260	0	10	00001652	HUAMAN MONDALGO NARCISO	AV. EJERCITO 850		AD	1 1/2"	215009067	19/07/2012	1 1/2"	ACT	ACT	Domes-201
33	1	1	4	10	270	0	10		HUAMAN MONDALGO MOISES	AV. EJERCITO S/N		AD	1 1/2"	1520705	########	1 1/2"	ACT	ACT	Domes-201
34	1	1	4	10	271	0	10	00065245	PARQUE DEL CAMPESINO	AV. EJERCITO 9NA CUADRA		AD	3 1"	2503115	10/11/2016	3 1"	ACT	ACT	Estat-501
35	1	1	4	10	272	0	10	00064697	CHOQUE QUISPE LUIS	AV. EJERCITO S/N		AD	1 1/2"	8623822	########	1 1/2"	ACT	ACT	Domes-201
36	1	1	4	10	274	0	10	00064711	CHOQUE QUISPE LUIS	AV. EJERCITO 114		AD	1 1/2"	8623823	########	1 1/2"	ACT	ACT	Domes-201
37	1	1	4	10	276	0	10	00064700	CHOQUE QUISPE LUIS	AV. EJERCITO 114		AD	1 1/2"	8623821	########	1 1/2"	ACT	ACT	Domes-201
38	1	1	4	10	280	0	10	00000900	CHIPANA JUAREZ DELFIN	AV. EJERCITO S/N		AD	1 1/2"	100102617		1 1/2"	ACT	ACT	Domes-201
39	1	1	4	10	290	0	10	00003211	QUISPE HURTADO SEBASTIAN	AV. EJERCITO S/N		D	1 1/2"			1 1/2"	ACT	ACT	Domes-201

Hoja1

Figura 10: Cuadro de datos

Fuente: EPS Emsap Chanka



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	NOMBRE DEL USUARIO	USUARIO	CATEGORIA	ENERO	FEBRERO	MARZO	ABRIL	MAYO	JUNIO	JULIO	AGOSTO	SETIEMBRE	OCTUBRE	NOVIEMBRE	DICIEMBRE
2	COLISEO CERRADO PUERTA PRINCIPAL (1)	1	Estat-501	24	16	14	17	18	19	20	24	22	20	19	18
3	MUNI-COLISEO CERRADO ANDAHUAYLAS (2)	2	Estat-501	12	14	11	10	12	13	10	14	12	14	10	12
4	GONZALEZ LAUPA CALIXTO	3	Domes-201	9	8	7	5	8	7	9	7	6	7	8	5
5	BORDA TELLO PABLO	4	Domes-201	10	12	14	13	12	10	14	12	14	12	10	12
6	VELASQUE BERROCAL PRUDENCIO	5	Domes-201	15	18	14	16	19	17	18	17	19	15	17	16
7	LLANTOY ARIAS FORTUNATO	6	Domes-201	16	17	15	14	16	14	15	14	16	14	15	14
8	LLANTOY DE LOAYZA TADEA	7	Domes-201	8	9	7	5	8	7	6	5	8	5	7	5
9	CASAS GUTIERREZ GRICELDA	8	Domes-201	18	20	21	20	21	20	23	21	20	21	20	23
10	DAMIANO FLORES LAURENTINA	9	Comer-301	41	42	45	41	42	46	41	42	45	41	45	42
11	DAMIANO FLORES PAULA	10	Comer-301	21	20	22	21	20	23	21	20	23	21	20	22
12	ALARCON RODAS ALBERTO	11	Comer-301	8	5	7	9	8	5	7	8	5	8	7	8
13	VASQUEZ VIVANCO RENY INOCENCIA	12	Domes-201	12	10	10	9	10	9	12	10	11	12	9	10
14	ECHEGARAY TRELLEZ RAMON	13	Domes-201	15	18	16	19	17	15	17	15	19	16	17	15
15	PETROCENTRO ANDAHUAYLAS E.I.R.L.	14	Comer-301	20	21	20	21	23	20	21	22	22	20	21	20
16	ANDIA PECEROS LUISA	15	Domes-201	15	16	14	15	16	14	16	14	15	16	14	15
17	PEREZ TOMAYLLA ANTONIO	16	Domes-201	20	21	20	23	21	20	21	23	24	25	21	20
18	HUAMAN CARDENAS LAZARO	17	Domes-201	14	12	14	13	14	12	14	12	11	14	12	14
19	GONZALES VASQUEZ ISMAEL	18	Comer-301	42	45	42	43	47	42	45	42	46	47	45	42
20	ALARCON RODAS ALBERTO	19	Comer-301	32	34	36	35	36	35	36	34	35	37	35	36
21	INCA HUAMAN SENOBIO N.	20	Comer-301	10	13	12	10	12	10	11	12	10	13	11	12
22	PACHECO BERNAOLA AUGUSTO	21	Domes-201	21	20	21	23	20	21	20	21	23	22	20	21
23	QUISPE PALOMINO CORPUS	22	Comer-301	8	9	7	5	8	9	7	5	8	7	6	9
24	GONZALES VASQUEZ VICTOR	23	Comer-301	5	4	8	9	5	7	4	5	7	6	8	6
25	GONZALES QUISPE VICTOR DAVID	24	Domes-201	10	12	14	12	13	11	12	10	12	10	12	11
26	CCORIMANYA MAMANI MANUEL	25	Domes-201	12	14	15	16	15	17	15	16	17	18	16	18
27	QUISPE PALOMINO ROSA	26	Comer-301	9	8	7	5	8	7	9	6	8	5	7	8
28	RAMOS PEREZ VIDAL	27	Comer-301	21	20	21	20	21	20	23	20	21	20	23	20
29	LEGUIA REYNAGA ISAAC	28	Comer-301	45	47	46	48	47	45	47	45	46	47	45	46

Figura 11: Cuadro de datos
Fuente: EPS Emsap Chanka

3.3.3 Preparar la base de datos objetivo

3.3.3.1 limpieza y procesamiento de datos

Al seleccionar los datos se tomó en cuenta que el id de algunos usuarios estaban incompletos entonces se decidió no utilizarlo y eliminarlo de la lista de variables que serán utilizadas para el análisis.

3.3.3.2 Transformación de datos

En la siguiente actividad se vio por conveniente la transformación de los datos de la variable categoría de los usuarios y convertirlos en datos simples con el que pueda trabajar el algoritmo implementado.

Tabla 10: Transformación de categoría

Nombre de Categoría	Código de Categoría	Id de Categoría
Social	SOC-101	1
Doméstica	DOM-201	2
Comercial	COM-301	3
Industrial	IND-401	4
Estatal	EST-501	5

Fuente: Elaboración Propia

3.3.4 Minería de datos

3.3.4.1 selección de minería de datos

En el presente trabajo se eligió realizar la agrupación de los datos utilizando el consumo mensual de cada usuario y la categoría a la que pertenece cada uno, ya que se busca encontrar patrones de comportamiento de consumo de agua potable y de esta manera optimizar el consumo de agua potable en la ciudad de Andahuaylas.

3.3.4.2 Elección y aplicación del algoritmo de Clustering

En esta etapa se eligió utilizar el algoritmo de clustering k_means ya que es fácil de entender e implementar para obtener la agrupación de los datos y de esta manera obtener los patrones de comportamiento de consumo, asimismo tenemos que instalar el lenguaje de programación Python, este lenguaje posee diversas librerías que nos ayudaran a obtener los resultados esperados.

- **Paso 1:**

  Primero se tuvo que instalar las siguientes librerías que nos ayudaran; la primera librería que es pandas nos ayuda analizar los datos y nos permite trabajar de manera eficiente; la segunda librería que es numpy contiene funciones matemáticas que nos ayuda a trabajar con vectores o matrices; la tercera librería matplotlib nos ayuda a generar gráficos a partir de los datos que tenemos para trabajar; la cuarta librería sklearn nos sirve para implementar el algoritmo que estamos utilizando, la quinta librería seaborn nos sirve para realizar gráficos estadísticos.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sb
from sklearn.cluster import KMeans
from sklearn.metrics import pairwise_distances_argmin

from mpl_toolkits.mplot3d import Axes3D
```

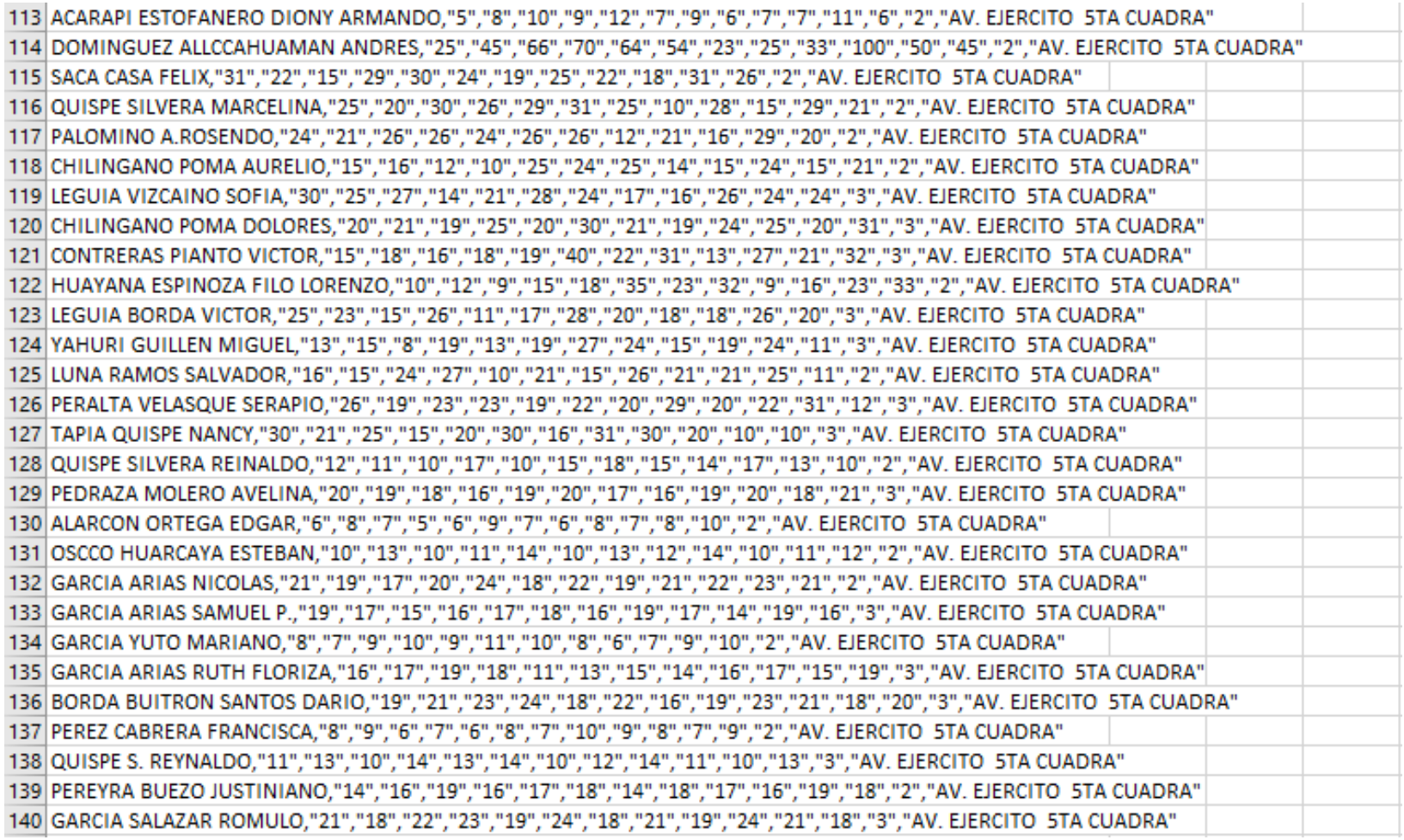
Figura 12: Librerías instaladas
Fuente: imagen propia, realizada con recortes de pantalla.

- **Paso 2:**

  Después de realizar la transformación de datos y tenerlos en un formato que el sistema lo cargue fácilmente pasamos a que el algoritmo lea los datos que estamos utilizando.

```
dataframe = pd.read_csv(r"D:\data2016.csv")
```

Figura 13: Librerías instalas
Fuente: imagen propia, realizada con recortes de pantalla



113	ACARAPI ESTOFANERO DIONY ARMANDO,"5","8","10","9","12","7","9","6","7","7","11","6","2","AV. EJERCITO 5TA CUADRA"
114	DOMINGUEZ ALLCCAHUAMAN ANDRES,"25","45","66","70","64","54","23","25","33","100","50","45","2","AV. EJERCITO 5TA CUADRA"
115	SACA CASA FELIX,"31","22","15","29","30","24","19","25","22","18","31","26","2","AV. EJERCITO 5TA CUADRA"
116	QUISPE SILVERA MARCELINA,"25","20","30","26","29","31","25","10","28","15","29","21","2","AV. EJERCITO 5TA CUADRA"
117	PALOMINO A.ROSENDO,"24","21","26","26","24","26","26","12","21","16","29","20","2","AV. EJERCITO 5TA CUADRA"
118	CHILINGANO POMA AURELIO,"15","16","12","10","25","24","25","14","15","24","15","21","2","AV. EJERCITO 5TA CUADRA"
119	LEGUIA VIZCAINO SOFIA,"30","25","27","14","21","28","24","17","16","26","24","24","3","AV. EJERCITO 5TA CUADRA"
120	CHILINGANO POMA DOLORES,"20","21","19","25","20","30","21","19","24","25","20","31","3","AV. EJERCITO 5TA CUADRA"
121	CONTRERAS PIANTO VICTOR,"15","18","16","18","19","40","22","31","13","27","21","32","3","AV. EJERCITO 5TA CUADRA"
122	HUAYANA ESPINOZA FILO LORENZO,"10","12","9","15","18","35","23","32","9","16","23","33","2","AV. EJERCITO 5TA CUADRA"
123	LEGUIA BORDA VICTOR,"25","23","15","26","11","17","28","20","18","18","26","20","3","AV. EJERCITO 5TA CUADRA"
124	YAHURI GUILLEN MIGUEL,"13","15","8","19","13","19","27","24","15","19","24","11","3","AV. EJERCITO 5TA CUADRA"
125	LUNA RAMOS SALVADOR,"16","15","24","27","10","21","15","26","21","21","25","11","2","AV. EJERCITO 5TA CUADRA"
126	PERALTA VELASQUE SERAPIO,"26","19","23","23","19","22","20","29","20","22","31","12","3","AV. EJERCITO 5TA CUADRA"
127	TAPIA QUISPE NANCY,"30","21","25","15","20","30","16","31","30","20","10","10","3","AV. EJERCITO 5TA CUADRA"
128	QUISPE SILVERA REINALDO,"12","11","10","17","10","15","18","15","14","17","13","10","2","AV. EJERCITO 5TA CUADRA"
129	PEDRAZA MOLERO AVELINA,"20","19","18","16","19","20","17","16","19","20","18","21","3","AV. EJERCITO 5TA CUADRA"
130	ALARCON ORTEGA EDGAR,"6","8","7","5","6","9","7","6","8","7","8","10","2","AV. EJERCITO 5TA CUADRA"
131	OSCCO HUARCAYA ESTEBAN,"10","13","10","11","14","10","13","12","14","10","11","12","2","AV. EJERCITO 5TA CUADRA"
132	GARCIA ARIAS NICOLAS,"21","19","17","20","24","18","22","19","21","22","23","21","2","AV. EJERCITO 5TA CUADRA"
133	GARCIA ARIAS SAMUEL P.,"19","17","15","16","17","18","16","19","17","14","19","16","3","AV. EJERCITO 5TA CUADRA"
134	GARCIA YUTO MARIANO,"8","7","9","10","9","11","10","8","6","7","9","10","2","AV. EJERCITO 5TA CUADRA"
135	GARCIA ARIAS RUTH FLORIZA,"16","17","19","18","11","13","15","14","16","17","15","19","3","AV. EJERCITO 5TA CUADRA"
136	BORDA BUITRON SANTOS DARIO,"19","21","23","24","18","22","16","19","23","21","18","20","3","AV. EJERCITO 5TA CUADRA"
137	PEREZ CABRERA FRANCISCA,"8","9","6","7","6","8","7","10","9","8","7","9","2","AV. EJERCITO 5TA CUADRA"
138	QUISPE S. REYNALDO,"11","13","10","14","13","14","10","12","14","11","10","13","3","AV. EJERCITO 5TA CUADRA"
139	PEREYRA BUEZO JUSTINIANO,"14","16","19","16","17","18","14","18","17","16","19","18","2","AV. EJERCITO 5TA CUADRA"
140	GARCIA SALAZAR ROMULO,"21","18","22","23","19","24","18","21","19","24","21","18","3","AV. EJERCITO 5TA CUADRA"

Figura 14: Lista de datos

Fuente: imagen propia, realizada con recortes de pantalla

- **Paso 3:**

  Luego de cargar los datos el algoritmo calculó los centroides en este caso tuvo tres grupos por lo que el algoritmo escoge al azar tres datos que serán los centroides de cada grupo por un momento, por lo tanto, el algoritmo calculara tres centroides, luego el algoritmo k_means utiliza la función de distancia manhattan ya que mediante una serie de pruebas se observó que según a los datos obtenidos la agrupación de los clusters se definen de mejor manera; para la obtención de los centroides fijos se realizó diferentes interacciones.

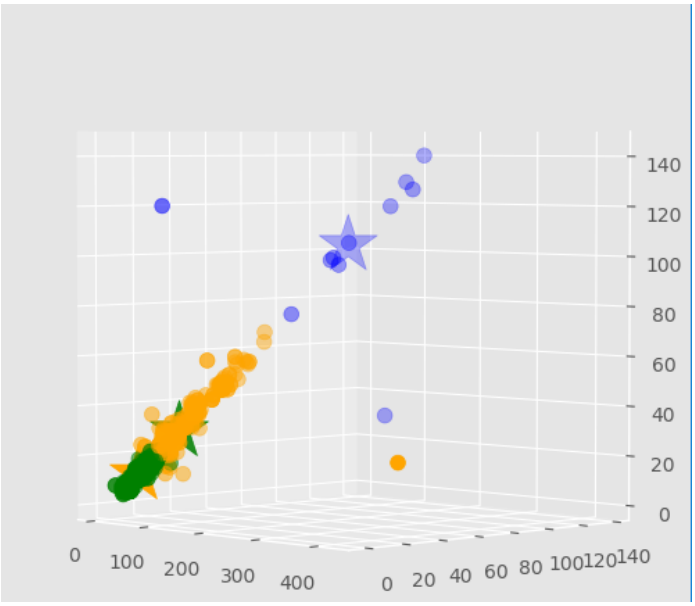


Figura 15: Resultados de los clusters utilizando la ecuación distancia de Manhattan

Fuente: Imagen propia, realizada con recortes de plantillas

```
==========================================CENTROIDES==========================================
centroide 0 [15.20923913 15.36141304 15.25543478 15.24796196 15.32608696 15.21127717]
centroide 1 [48.76470588 49.68627451 50.16666667 51.19607843 50.05882353 63.08823529]
centroide 2 [  1. 817.  17.  18.  18.  19.]
==============================================================================================
```

Figura 16: Resultados de Centroides
Fuente: Imagen propia, realizada con recortes de plantillas

CAPITULO 4: RESULTADOS Y DISCUSIÓN

4.1 Resultados año 2015

El algoritmo nos muestra cuantos usuarios existen en cada categoría más nos muestra los centroides calculados utilizando los datos del 2015.

```
categoria
1       9
2    1383
3     153
4       2
5      28
dtype: int64
====================================CENTROIDES====================================
centroide 0 [11.71246588 11.73339399 11.86988171 11.84531392 11.76979072 11.96906278
 11.88717015]
centroide 1 [28.51187905 28.88336933 28.89200864 28.79049676 29.44276458 28.90280778
 28.64794816]
centroide 2 [ 85.92307692  82.15384615 115.76923077 100.23076923 102.07692308
 144.92307692 137.38461538]
==================================================================================
```

Figura 17: Resultados de Centroides y cantidad de usuarios por categoría año 2015
Fuente: Imagen propia, realizada con recortes de plantillas

Luego de sacar los centroides podemos observar una grafica que plasma la ubicación de los datos sin utilizar el algoritmo k_means.

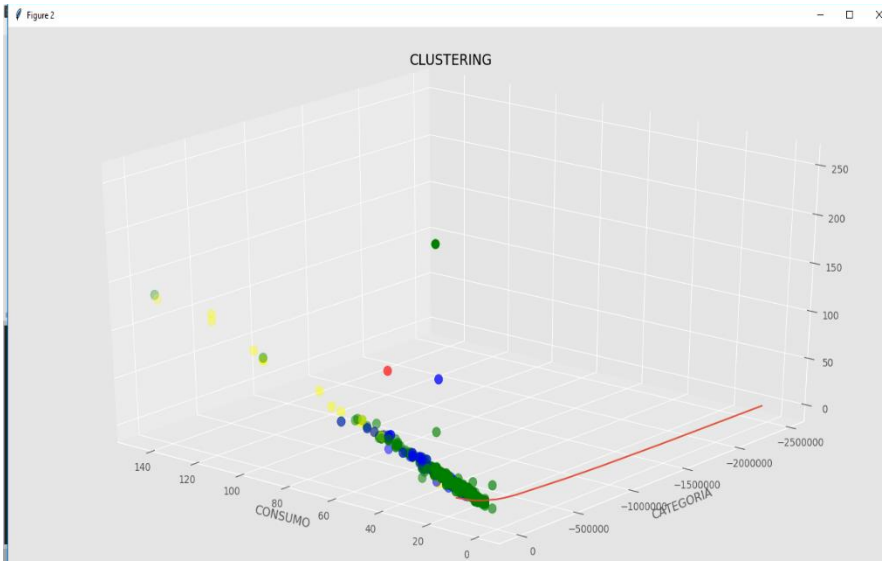


Figura 18: Datos ubicados sin algoritmo k_means del año 2015
Fuente: Imagen propia, realizada con recortes de plantillas

Después de obtener una imagen con los datos en el plano observaremos una gráfica donde se utiliza el algoritmo k-means

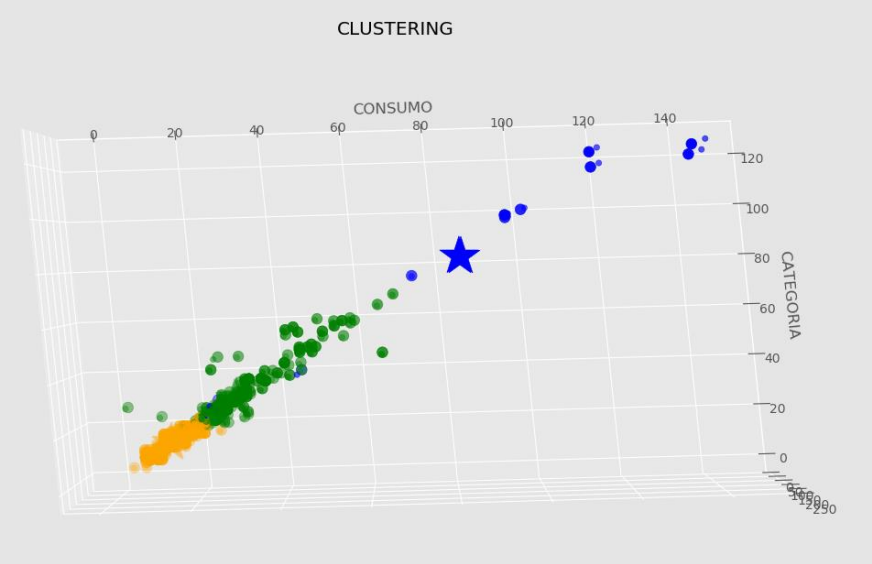


Figura 19: Resultados de clusters gráficamente del año 2015
Fuente: Imagen propia, realizada con recortes de plantillas

```
    color  cantidad
0  orange       463
1   green      1099
2    blue        13
imprimierndo lista ['orange', 'green', 'blue']
imprimierndo lista [463, 1099, 13]
                                  usuarios   ...   cluster
7                    CASAS GUTIERREZ GRICELDA   ...         0
8                   DAMIANO FLORES LAURENTINA   ...         0
9                       DAMIANO FLORES PAULA   ...         0
13           PETROCENTRO ANDAHUAYLAS E.I.R.L.   ...         0
15                    PEREZ TOMAYLLA ANTONIO   ...         0
17                   GONZALES VASQUEZ ISMAEL   ...         0
18                     ALARCON RODAS ALBERTO   ...         0
20                  PACHECO BERNAOLA AUGUSTO   ...         0
26                         RAMOS PEREZ VIDAL   ...         0
27                      LEGUIA REYNAGA ISAAC   ...         0
29                      DUARTE RIVAS CIRILO   ...         0
33                       CHOQUE QUISPE LUIS   ...         0
36                              PILETA - 12   ...         0
38                  CONCEJO PROV. - PILETA 11   ...         0
39                         MUNI BAÑO PUBLICO   ...         0
40                    CONCEJO PROV. PILETA 9   ...         0
41                                PILETA - 8   ...         0
42                  BAÑOS P. AGUSTINA PASTOR   ...         0
43                    CONCEJO PROV. PILETA 7   ...         0
49                  LOAYZA GUZMAN ANTONIO I.   ...         0
```

Figura 20: Resultados del clusters 0 del año 2015
Fuente: Imagen propia, realizada con recortes de plantillas

	usuarios	...	cluster
0	COLISEO CERRADO PUERTA PRINCIPAL (1)	...	1
1	MUNI-COLISEO CERRADO ANDAHUAYLAS (2)	...	1
2	GONZALEZ LAUPA CALIXTO	...	1
3	BORDA TELLO PABLO	...	1
4	VELASQUE BERROCAL PRUDENCIO	...	1
5	LLANTOY ARIAS FORTUNATO	...	1
6	LLANTOY DE LOAYZA TADEA	...	1
10	ALARCON RODAS ALBERTO	...	1
11	VASQUEZ VIVANCO RENY INOCENCIA	...	1
12	ECHEGARAY TRELLEZ RAMON	...	1
14	ANDIA PECEROS LUISA	...	1
16	HUAMAN CARDENAS LAZARO	...	1
19	INCA HUAMAN SENOBIO N.	...	1
21	QUISPE PALOMINO CORPUS	...	1
22	GONZALES VASQUEZ VICTOR	...	1
23	GONZALES QUISPE VICTOR DAVID	...	1
24	CCORIMANYA MAMANI MANUEL	...	1
25	QUISPE PALOMINO ROSA	...	1
28	CABALLERO AQUIÑO GLORIA	...	1
30	HUAMAN MONDALGO NARCISO	...	1
31	HUAMAN MONDALGO MOISES	...	1
34	CHIPANA JUAREZ DELFIN	...	1
35	QUISPE HURTADO SEBASTIAN	...	1
45	CONTRERAS G.BUENAVENTURA	...	1
46	ALVITES GARCIA CESAR	...	1
47	ALVITES GARCIA CORINA BETTY	...	1
48	ALVITES GARCIA SILVIA	...	1

Figura 21: Resultados del clusters 1del año 2015
Fuente: Imagen propia, realizada con recortes de plantillas

	usuarios	...	cluster
32	PARQUE DEL CAMPESINO	...	2
37	C.P.MUNIC.FERIA BAÑOS 6TA.CUADRA	...	2
44	MUNI-C.P. ESTADIO LOS CHANKAS	...	2
62	COLEGIO BELEN DE OSMA	...	2
79	CUARTEL LOS CHANKAS	...	2
80	MINISTERIO DE AGRICULTURA	...	2
176	COLEGIO JUAN ESPINOZA M	...	2
663	MUNI-C.P. ESTADIO LOS CHANKAS	...	2
769	GONZALES GONZALON JOSE L.	...	2
847	UNSACC	...	2
1088	CLUB CAMPESTRE PUMA DE PIEDRA S.A.C.	...	2
1089	MEZA DURAN GREGORIA RICARDINA	...	2
1393	ANDREU NECOCHEA WALDO	...	2

Figura 22: Resultados del clusters 2 del año 2015
Fuente: Imagen propia, realizada con recortes de plantillas

```
ALTAMIRANO GUZMAN MARCIAL
LOA MENDOZA REMIGIO
CUARTEL LOS CHANKAS
```

Figura 23: Usuarios que son de los centroides del año 2015
Fuente: Imagen propia, realizada con recortes de plantillas

4.2 Resultados del 2016

El algoritmo nos mostró cuantos usuarios existen en cada categoría más nos muestra los centroides calculados utilizando los datos del 2016.

```
categoria
1       9
2    1383
3     153
4       2
5      28
dtype: int64
=====================================CENTROIDES=====================================
centroide 0 [11.90036232 12.07880435 11.96648551 11.88134058 12.00996377 11.92028986
 11.98097826]
centroide 1 [28.76315789 29.04166667 28.6622807  29.13815789 28.82017544 28.61184211
 28.70175439]
centroide 2 [ 75.06666667  74.6         87.26666667  85.4         85.53333333
 175.86666667  86.6       ]
====================================================================================
```

Figura 24: Resultados de Centroides y cantidad de usuarios por categoría del año 2016
Fuente: Imagen propia, realizada con recortes de plantillas

Después de obtener una imagen con los datos en el plano observaremos una gráfica donde se utiliza el algoritmo k-means

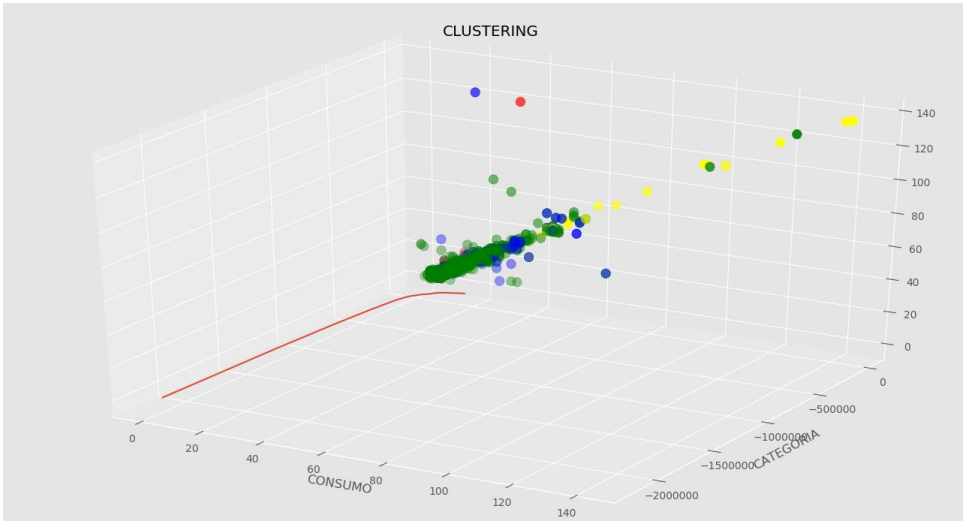


Figura 25: Datos ubicados sin algoritmo k_means del año 2016
Fuente: Imagen propia, realizada con recortes de plantillas

Después de obtener una imagen con los datos en el plano observaremos una gráfica donde se utiliza el algoritmo k-means

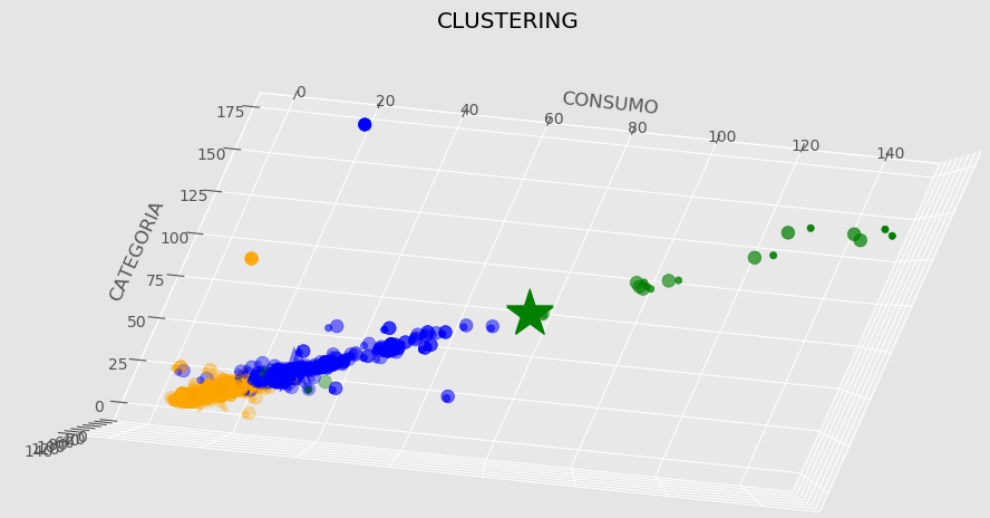


Figura 26: Resultados de clusters gráficamente del año 2016
Fuente: Imagen propia, realizada con recortes de plantillas

```
    color  cantidad
0  orange      1104
1   green       456
2    blue        15
imprimierndo lista ['orange', 'green', 'blue']
imprimierndo lista [1104, 456, 15]
                                usuarios   ...   cluster
0     COLISEO CERRADO PUERTA PRINCIPAL (1)   ...         0
1     MUNI-COLISEO CERRADO ANDAHUAYLAS (2)   ...         0
2                   GONZALEZ LAUPA CALIXTO   ...         0
3                        BORDA TELLO PABLO   ...         0
4              VELASQUE BERROCAL PRUDENCIO   ...         0
5                  LLANTOY ARIAS FORTUNATO   ...         0
6                  LLANTOY DE LOAYZA TADEA   ...         0
10                   ALARCON RODAS ALBERTO   ...         0
11          VASQUEZ VIVANCO RENY INOCENCIA   ...         0
12                 ECHEGARAY TRELLEZ RAMON   ...         0
14                     ANDIA PECEROS LUISA   ...         0
16                  HUAMAN CARDENAS LAZARO   ...         0
19                  INCA HUAMAN SENOBIO N.   ...         0
21                  QUISPE PALOMINO CORPUS   ...         0
22                 GONZALES VASQUEZ VICTOR   ...         0
23             GONZALES QUISPE VICTOR DAVID   ...         0
24                 CCORIMANYA MAMANI MANUEL   ...         0
25                     QUISPE PALOMINO ROSA   ...         0
28                 CABALLERO AQUIÑO GLORIA   ...         0
30                 HUAMAN MONDALGO NARCISO   ...         0
```

Figura 27: Resultados del clusters 0 del año 2016
Fuente: Imagen propia, realizada con recortes de plantillas

	usuarios	...	cluster
7	CASAS GUTIERREZ GRICELDA	...	1
8	DAMIANO FLORES LAURENTINA	...	1
9	DAMIANO FLORES PAULA	...	1
13	PETROCENTRO ANDAHUAYLAS E.I.R.L.	...	1
15	PEREZ TOMAYLLA ANTONIO	...	1
17	GONZALES VASQUEZ ISMAEL	...	1
18	ALARCON RODAS ALBERTO	...	1
20	PACHECO BERNAOLA AUGUSTO	...	1
26	RAMOS PEREZ VIDAL	...	1
27	LEGUIA REYNAGA ISAAC	...	1
29	DUARTE RIVAS CIRILO	...	1
33	CHOQUE QUISPE LUIS	...	1
36	PILETA - 12	...	1
38	CONCEJO PROV. - PILETA 11	...	1
39	MUNI BAÑO PUBLICO	...	1
40	CONCEJO PROV. PILETA 9	...	1
41	PILETA - 8	...	1
42	BAÑOS P. AGUSTINA PASTOR	...	1
43	CONCEJO PROV. PILETA 7	...	1
49	LOAYZA GUZMAN ANTONIO I.	...	1
51	SILVERA PALOMINO CLAUDIO-COOP.LOS ANDES	...	1
61	RAMIREZ CARDENAS EDGAR	...	1
64	VILLANO VASQUEZ JOSE L.	...	1
65	BARRIENTOS AYALA NIDIA	...	1
66	GARCIA ARIAS MAXIMILIANA	...	1

Figura 28: Resultados del clusters 1 del año 2016
Fuente: Imagen propia, realizada con recortes de plantillas

	usuarios	...	cluster
32	PARQUE DEL CAMPESINO	...	2
37	C.P.MUNIC.FERIA BAÑOS 6TA.CUADRA	...	2
44	MUNI-C.P. ESTADIO LOS CHANKAS	...	2
62	COLEGIO BELEN DE OSMA	...	2
79	CUARTEL LOS CHANKAS	...	2
80	MINISTERIO DE AGRICULTURA	...	2
161	PALOMINO ALTAMIRANO JUAN	...	2
176	COLEGIO JUAN ESPINOZA M	...	2
228	OSCCO HUARCAYA ESTEBAN	...	2
663	MUNI-C.P. ESTADIO LOS CHANKAS	...	2
847	UNSACC	...	2
1088	CLUB CAMPESTRE PUMA DE PIEDRA S.A.C.	...	2
1393	ANDREU NECOCHEA WALDO	...	2
1401	ROJAS CCENTE JORGE	...	2
1468	HERHUAY BAUTISTA SANTIAGO	...	2

Figura 29: Resultados del clusters 2 del año 2016
Fuente: Imagen propia, realizada con recortes de plantillas

```
FLORES BERTI HECTOR
LIZUNDI CCOICCA BENEDICTA
MINISTERIO DE AGRICULTURA
```

Figura 30: Usuarios que son de los centroides del año 2016
Fuente: Imagen propia, realizada con recortes de plantillas

4.3 Resultados del año 2017

El algoritmo nos mostró cuantos usuarios existen en cada categoría más nos mostró los centroides calculados utilizando los datos del 2016.

```
=====================================CENTROIDES=====================================
centroide 0 [11.93727273 11.83363636 11.87       11.88       11.89090909 12.02181818
 11.83363636]
centroide 1 [28.54545455 28.63203463 28.74242424 28.74242424 28.495671   28.53246753
 28.70779221]
centroide 2 [124.61538462  93.15384615  94.38461538  92.61538462 101.61538462
 101.38461538 100.38461538]
====================================================================================
```

Figura 31: Resultados de Centroides y cantidad de usuarios por categoría del año 2017
Fuente: Imagen propia, realizada con recortes de plantillas

Figura 32: Datos ubicados sin algoritmo k_means del año 2017
Fuente: Imagen propia, realizada con recortes de plantillas

Figura 33: Resultados de clusters gráficamente del año 2017
Fuente: Imagen propia, realizada con recortes de plantillas

	usuarios	...	cluster
0	COLISEO CERRADO PUERTA PRINCIPAL (1)	...	0
1	MUNI-COLISEO CERRADO ANDAHUAYLAS (2)	...	0
2	GONZALEZ LAUPA CALIXTO	...	0
3	BORDA TELLO PABLO	...	0
4	VELASQUE BERROCAL PRUDENCIO	...	0
5	LLANTOY ARIAS FORTUNATO	...	0
6	LLANTOY DE LOAYZA TADEA	...	0
10	ALARCON RODAS ALBERTO	...	0
11	VASQUEZ VIVANCO RENY INOCENCIA	...	0
12	ECHEGARAY TRELLEZ RAMON	...	0
14	ANDIA PECEROS LUISA	...	0
16	HUAMAN CARDENAS LAZARO	...	0
19	INCA HUAMAN SENOBIO N.	...	0
21	QUISPE PALOMINO CORPUS	...	0
22	GONZALES VASQUEZ VICTOR	...	0
23	GONZALES QUISPE VICTOR DAVID	...	0
24	CCORIMANYA MAMANI MANUEL	...	0
25	QUISPE PALOMINO ROSA	...	0
28	CABALLERO AQUIÑO GLORIA	...	0
30	HUAMAN MONDALGO NARCISO	...	0
31	HUAMAN MONDALGO MOISES	...	0
34	CHIPANA JUAREZ DELFIN	...	0
35	QUISPE HURTADO SEBASTIAN	...	0
45	CONTRERAS G.BUENAVENTURA	...	0
46	ALVITES GARCIA CESAR	...	0
47	ALVITES GARCIA CORINA BETTY	...	0
48	ALVITES GARCIA SILVIA	...	0

Figura 34: Resultados del clusters 0 del año 2017
Fuente: Imagen propia, realizada con recortes de plantillas

	usuarios	...	cluster
7	CASAS GUTIERREZ GRICELDA	...	1
8	DAMIANO FLORES LAURENTINA	...	1
9	DAMIANO FLORES PAULA	...	1
13	PETROCENTRO ANDAHUAYLAS E.I.R.L.	...	1
15	PEREZ TOMAYLLA ANTONIO	...	1
17	GONZALES VASQUEZ ISMAEL	...	1
18	ALARCON RODAS ALBERTO	...	1
20	PACHECO BERNAOLA AUGUSTO	...	1
26	RAMOS PEREZ VIDAL	...	1
27	LEGUIA REYNAGA ISAAC	...	1
29	DUARTE RIVAS CIRILO	...	1
33	CHOQUE QUISPE LUIS	...	1
36	PILETA - 12	...	1
38	CONCEJO PROV. - PILETA 11	...	1
40	CONCEJO PROV. PILETA 9	...	1
41	PILETA - 8	...	1
42	BAÑOS P. AGUSTINA PASTOR	...	1
43	CONCEJO PROV. PILETA 7	...	1
49	LOAYZA GUZMAN ANTONIO I.	...	1
51	SILVERA PALOMINO CLAUDIO-COOP.LOS ANDES	...	1
61	RAMIREZ CARDENAS EDGAR	...	1
64	VILLANO VASQUEZ JOSE L.	...	1
65	BARRIENTOS AYALA NIDIA	...	1
66	GARCIA ARIAS MAXIMILIANA	...	1
67	QUINTANA LEGUIA TULIO	...	1
74	ALARCON ZEDANO JUVENAL	...	1
78	GUTIERREZ CASTILLO ALCIDES	...	1

Figura 35: Resultados del clusters 1 del año 2017
Fuente: Imagen propia, realizada con recortes de plantillas

	usuarios	...	cluster
32	PARQUE DEL CAMPESINO	...	2
37	C.P.MUNIC.FERIA BAÑOS 6TA.CUADRA	...	2
39	MUNI BAÑO PUBLICO	...	2
44	MUNI-C.P. ESTADIO LOS CHANKAS	...	2
62	COLEGIO BELEN DE OSMA	...	2
79	CUARTEL LOS CHANKAS	...	2
80	MINISTERIO DE AGRICULTURA	...	2
176	COLEGIO JUAN ESPINOZA M	...	2
663	MUNI-C.P. ESTADIO LOS CHANKAS	...	2
769	GONZALES GONZALON JOSE L.	...	2
847	UNSACC	...	2
1088	CLUB CAMPESTRE PUMA DE PIEDRA S.A.C.	...	2
1393	ANDREU NECOCHEA WALDO	...	2

Figura 36: Resultados del clusters 2 del año 2017
Fuente: Imagen propia, realizada con recortes de plantillas

```
LOA MENDOZA REMIGIO
CARBAJAL GALINDO ALCIDES
PARQUE DEL CAMPESINO
```

Figura 37: Usuarios que son de los centroides del año 2017
Fuente: Imagen propia, realizada con recortes de plantillas

4.4 Interpretación

4.4.1 Evaluar resultados

En los diferentes clusters se obtuvo tres grupos que se clasificó en consumo bajo, consumo medio y consumo elevado, dentro de ello los consumos más elevados fueron en el clusters 2 en los tres años.

Tabla 11: Comparación de consumo

Año	Cluster 0	Cluster 1	Cluster 2
2015	11.80 m3	28.70 m3	100.23 m3
2016	11.88 m3	29.13 m3	85.4 m3
2017	11.83 m3	28.74 m3	92.61 m3

Fuente: Elaboración Propia

Los usuarios que se encontraron en el cluster 0 de los tres años 2015, 2016 y 2017 pertenecen al sector 04 de la ciudad de Andahuaylas, son los que tuvieron un consumo bajo y siguen un patrón de comportamiento; los usuarios fueron de la categoría Domestica (DOM-201), correspondieron a las direcciones: av. el ejército y jr. Túpac Amaru donde existe menor consumo de agua potable.

Los usuarios que se encontraron en el cluster 1 tuvieron un consumo medio y siguen un patrón de comportamiento; los usuarios fueron en su mayoría de la categoría comercial (COM-301), corresponden a las direcciones av. Perú, av. los cedros y jr. los cactus.

Los usuarios que se encontraron en el cluster 2 fueron aquellos que tuvieron un consumo elevado y lo conformaron en la mayoría los usuarios que pertenecen a la categoría estatal (EST-501) y comercial (COM-301), además, en este grupo la mayoría son instituciones públicas y tiendas comerciales grandes donde la concurrencia de personas es en grandes masas por ende son los mayores consumidores de agua potable.

4.4.2 Utiliza el conocimiento descubierto

- **Aplicar el conocimiento encontrado a partir de los resultados**

  Obteniendo los resultados de los patrones de comportamiento de consumo de agua potable podemos utilizarlos para realizar una mejor toma de decisiones respecto al cuidado y manejo del agua en la provincia de Andahuaylas; en los diferentes datos de los años analizados se pudo observar que el mayor índice de consumo se encontró en el cluster 2 y pudo notar que este grupo se encontró conformado por instituciones públicas, instituciones educativas y recreos por ende una manera de reducir el consumo de agua seria realizar una campaña de concientización sobre el manejo adecuado y el cuidado del agua en dichos lugares.

CONCLUSIONES

CONCLUSIÓN GENERAL

Se determinó los patrones de comportamiento de consumo de agua potable utilizando los datos de la Entidad Prestadora de Servicios de Saneamiento Municipal Chanka (Emsap Chanka) en la ciudad de Andahuaylas implementando el algoritmo de clusterización k_means.

CONCLUSIONES ESPECÍFICOS

- Se obtuvieron datos históricos de los años 2015, 2016 y 2017 de la Entidad Prestadora de Servicios de Saneamiento Municipal (Emsap Chanka) de los cuales se utilizó el nombre de usuario, consumo mensual de agua potable por metros cúbicos, la categoría a la que pertenece el usuario y la dirección del usuario; la transformación de datos solo se realizó en la variable categoría para trabajar mejor con el algoritmo.
- Se logró implementar el algoritmo k_means sin ningún problema; se realizó la clasificación de los clusters obteniendo tres grupos que se clasifican en consumo bajo, consumo medio y consumo elevado.
- Se obtuvo los patrones de comportamiento analizando los clusters encontrados teniendo como resultado:
  - En el cluster 0 tienen un consumo bajo los usuarios son de la categoría Domestica (DOM-201), corresponden a las direcciones: av. el ejército y jr. Tupac Amaru donde existe menor consumo de agua potable y consumen un promedio de 12 metros cúbicos de agua potable.
  - En el cluster 1 tienen un consumo medio los usuarios son de la categoría comercial (COM-301), correspondían a las direcciones av. Perú, av. los cedros y jr. los cactus donde existe un consumo medio de agua potable, además consumen un promedio de 29 metros cúbicos de agua potable.
  - En el cluster 2 tienen un consumo elevado y lo conforman en la mayoría los usuarios que pertenecen a la categoría estatal (EST-501) y comercial (COM-301), además, en este grupo la mayoría son instituciones públicas y tiendas comerciales grandes donde la concurrencia de personas es en grandes masas, su consumo promedio es de 100 metros cúbicos de agua potable.

RECOMENDACIONES

- El archivo que se analice debe contener los nombres de los usuarios, el consumo de agua desde el mes de enero hasta el mes de diciembre por metro cubico, la categoría a la que pertenece y la dirección del usuario.
- Para realizar el análisis de otro año el archivo debe estar con la extensión .csv para que el algoritmo lea los datos sin ningún inconveniente.
- Para realizar un buen análisis y obtener los cluster que nos den los resultados buscados se debe contar con datos reales autorizados por la entidad que posee los datos necesarios.

REFERENCIAS BIBLIOGRÁFICAS

Aluja, T. (2001). La minería de datos, entre la estadística y la inteligencia artificial. *Qüestiió: quaderns d'estadística i investigació operativa*, *25*(3), 479-498.

Berzal, F. Clustering basado por particiones. *Línea. Available: http://elvex. ugr. es/idbis/dm/slides/41% 20Clustering*.

Cabero, J. C. (1998). Impacto de las nuevas tecnologías de la información y la comunicación en las organizaciones educativas. *Grupo Editorial Universitaria*.

Data minig map. (2018). K Nearest Neighbors - Classification. Obtenido de http://www.saedsayad.com/k_nearest_neighbors.htm

Diez, X. P. y Lacruz, M. G. (2008). Patrones de comportamiento relacionados con la salud en una muestra española de población general. Acta Colombiana de Psicología. *11*(1).97-106.

Estadistica, I. (. (2018). *Ferrovial*. Obtenido de https://blog.ferrovial.com/es/2015/03/sabes-cuantos-litros-de-agua-necesita-una-persona-al-dia/

Garre, M., Cuadrado, J. J., Sicilia, M. A., Rodríguez, D., & Rejas, R. (2007). Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software. *REICIS. Revista Española de Innovación, Calidad e Ingeniería del Software*, *3*(1).

Grandéz, M. (2017). *Aplicación de mineria de datos para determinar patrones de consumo en clientes de una distribuidora de suplementos nutricionales.* Lima-Perú: Universidad San Ignacio de Loyola. Obtenido de http://repositorio.usil.edu.pe/bitstream/USIL/2763/1/2017_Granda_Aplicacion-de-mineria-datos.pdf

González, M. S. (2016). Patrones de comportamiento. *El campamento de Dios*.

Gutiérrez, J. (2016). Líneas de investigación en minería de datos en aplicaciones en ciencia e ingeniería: Estado del arte y perspectivas. *Pdfs. Semanticscholar. Org*, *1*, 1-17.

Guha, S., Rastogi, R., & Shim, K. (1998). CURE: an efficient clustering algorithm for large databases. In *ACM Sigmod Record* (Vol. 27, No. 2, pp. 73-84). ACM.

Hernandez, J. (2016). Clustering basado en el algoritmo K-means para la identificación de grupos de pacientes quirúrgicos.

House of Switzerland (2017). Aspectos clave de la protección ambiental en Suiza. Environment. Recuperado de https://houseofswitzerland.org/es/node/279

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 14, pp. 281-297).

Montoya, J. R. A. (2016). Evaluación de la performance de estudiantes mediante técnicas de clustering. *Revista de las Facultades de Arquitectura e Ingeniería*, (1).

López, C. P. (2007). *Minería de datos: técnicas y herramientas*. Editorial Paraninfo

Riquelme Santos, J. C., Ruiz, R., & Gilbert, K. (2006). Minería de datos: Conceptos y tendencias. *Inteligencia artificial: Revista Iberoamericana de Inteligencia Artificial*, *10*(29), 11-18.

Valencia Bacca, F. A. (2017). Plataforma para la agrupación de modelos de procesos de negocio bajo el algoritmo de clustering K-MENS.

Ortí, C. B. (2011). Las tecnologías de la información y comunicación (TIC). *Univ. Val., Unidad Tecnol. Educ*, (951), 1-7.

Organización Mundial de la Salud (2015). Informe 2015 del PCM sobre el acceso a agua potable y saneamiento: datos esenciales. Recuperado de http://www.who.int/water_sanitation_health/monitoring/jmp-2015-key-facts/es/

Organización Mundial de la Salud (2017). *Progresos en materia de agua potable, saneamiento e higiene: informe de actualización de 2017 y evaluación de los ODS.* Recuperado de http://www.who.int/es/news-room/detail/12-07-2017-2-1-billion-people-lack-safe-drinking-water-at-home-more-than-twice-as-many-lack-safe-sanitation

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (2003). Informe de las Naciones Unidas sobre el Desarrollo de los Recursos Hídricos en el Mundo. Recuperado de http://unesdoc.unesco.org/images/0012/001295/129556s.pdfOrganizacion

Pascual, D., Pla, F., & Sánchez, S. (2007). Algoritmos de agrupamiento. *Método Informáticos Avanzados*, 164-174.

Quispe, I. R. (2012). Cuantificación de la demanda insatisfecha de agua potable en las áreas rurales del departamento de la paz durante el periodo 2006 – 2011 (Tesis de Grado). Universidad Mayor de San Andrés. La Paz, Bolivia.

Salamó, A., Gras, E., & Font-Mayolas, S. (2010). Patrones de consumo de alcohol en la adolescencia. *Psicolothema*.

Sopanta, W. (25 de Abril de 2018). Horarios establecidos de consumo. (A. Lira, Entrevistador)

Schiffman, L. G., & Kanuk, L. L. (2005). *Comportamiento del consumidor*. Pearson Educación.

ANEXO

ANEXO 1: Código de la implementación del algoritmo k_means

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sb
from sklearn.cluster import KMeans
from sklearn.metrics import pairwise_distances_argmin_min

import xlwt
from tkinter import Tk
from tkinter import filedialog as k
from tkinter.filedialog import askopenfilename



def seleccionar():
    return askopenfilename(filetypes = (("txt files","*.csv"),("todos los archivos","*.*")))


from mpl_toolkits.mplot3d import Axes3D
plt.rcParams['figure.figsize'] = (5,5)
plt.style.use('ggplot')



dataframe = pd.read_csv(seleccionar(), header=0)



def inprimirCategoria():
    print(dataframe.groupby('categoria').size())
inprimirCategoria()


```

```
35
36
37
38
39  def obtenerX():
40      X = np.array(dataframe[["enero","febrero","marzo","abril","mayo","junio","julio","agosto",
41      return X
42  def ObtenerY():
43      y = np.array(dataframe['categoria'])
44      return y
45
46
47
48  def graficarClusters():
49
50      fig = plt.figure("DATA NO AGRUPADO REPRESENTADO POR CATEGORIA")
51      ax = Axes3D(fig)
52      colores=['red','blue','green','cyan','black',"orange"]
53      asignar=[]
54      cont=0
55
56      for row in ObtenerY():
57          asignar.append(colores[row-1])
58
59      ax.scatter(obtenerX()[:, 0], obtenerX()[:, 1], obtenerX()[:, 2], c=asignar,s=120 )
60
61      plt.title('CLUSTERING')
62      plt.show()
63
64
```

```
66  def numerodeCluster():
67      kmeans = KMeans(n_clusters=3).fit(obtenerX())
68      return kmeans
69
70
71  def sacarCentroides():
72
73      print("=====================================CENTROIDES=====================================
74      centroids = numerodeCluster().cluster_centers_
75      #print(centroids)
76      for i in range(len(centroids)):
77          print("centroide",i,centroids[i])
78
79      print("=====================================================================================
80
81  sacarCentroides()
82
83  def labels():
84
85      labels = numerodeCluster().predict(obtenerX())
86      return labels
87
88  colores1=['orange','green','blue']
89
90  def graficarClustersAgrupados():
91
92      C = numerodeCluster().cluster_centers_
93      colores=['orange','green','blue']
94      asignar=[]
95      for row in labels():
96          asignar.append(colores[row])
97
```

```
    fig = plt.figure("AGRUPADO")
    ax = Axes3D(fig)
    ax.scatter(obtenerX()[:, 0], obtenerX()[:, 1], obtenerX()[:, 2], c=asignar,s=60)
    ax.scatter(C[:, 0], C[:, 1], C[:, 2], marker='*', c=colores, s=1000)


graficarClustersAgrupados()

copy =  pd.DataFrame()
copy['usuarios']=dataframe['usuarios'].values
copy['categoria']=dataframe['categoria'].values
copy['label'] = labels;
cantidadGrupo =  pd.DataFrame()
cantidadGrupo['color']=colores1
cantidadGrupo['cantidad']=copy.groupby('label').size()







def guerdarClusters():

    cluster_map = pd.DataFrame()
    cluster_map['usuarios'] = dataframe['usuarios'].values

    cluster_map['direccion']=dataframe['direccion'].values
    cluster_map['categoria']=dataframe['categoria']
    cluster_map['cluster'] = numerodeCluster().labels_

```

```
    list1 = cluster_map['usuarios']
    list2 = cluster_map['direccion']
    lista3 =cluster_map['cluster']
    lista4=cluster_map['categoria']
    wb = xlwt.Workbook()
    w1 = wb.add_sheet('clusters')


    w1.write(0, 0, 'USUARIOS')
    w1.write(0, 1, 'DIRECCION')
    w1.write(0, 2, 'CLUSTER')
    w1.write(0, 3, 'CATEGORIA')


    nuevalista0=[]
    nuevalista2=[]
    nuevalista3=[]
    cate1=[]

    nuevalista4=[]
    nuevalista5=[]
    nuevalista6=[]
    cate2=[]

    nuevalista7=[]
    nuevalista8=[]
    nuevalista9=[]
    cate3=[]
```

```
    i = 1
    for x, y, z, a in zip(list1, list2,lista3, lista4):
      if z==0:
        nuevalista0.append(x)
        nuevalista2.append(y)
        nuevalista3.append(z)
        cate1.append(a)

      elif z==1:
        nuevalista4.append(x)
        nuevalista5.append(y)
        nuevalista6.append(z)
        cate2.append(a)
      elif z==2:
        nuevalista7.append(x)
        nuevalista8.append(y)
        nuevalista9.append(z)
        cate3.append(a)

      k=1

    for x,y,z ,a in zip(nuevalista0,nuevalista2,nuevalista3,cate2):
        w1.write(k, 0, x)
        w1.write(k, 1, y)
        w1.write(k, 2, z)
        w1.write(k, 3, a)
        k=k+1
    j=k
```

```
    for x,y,z,a in zip(nuevalista4,nuevalista5,nuevalista6,cate2):
        w1.write(j, 0, x)
        w1.write(j, 1, y)
        w1.write(j, 2, z)
        w1.write(j, 3, a)
        j=j+1
    s=j
    for x,y,z,a in zip(nuevalista7,nuevalista8,nuevalista9,cate3):
        w1.write(s, 0, x)
        w1.write(s, 1, y)
        w1.write(s, 2, z)
        w1.write(s, 3, a)
        s=s+1

    wb.save("clusters2.xls")

import os
import sys

try:
    import Tkinter as tk
except ImportError:
    import tkinter as tk

try:
    import ttk
    py3 = False
except ImportError:
    import tkinter.ttk as ttk
    py3 = True
import unknown_support
```

```
def vp_start_gui():


    global val, w, root
    root = tk.Tk()
    top = Toplevel1 (root)
    unknown_support.init(root, top)
    root.mainloop()

w = None
def cerrar():
    root.destroy()
def create_Toplevel1(root, *args, **kwargs):
    '''Starting point when module is imported by another program.'''
    global w, w_win, rt
    rt = root
    w = tk.Toplevel (root)
    top = Toplevel1 (w)
    unknown_support.init(w, top, *args, **kwargs)
    return (w, top)

def destroy_Toplevel1():
    global w
    w.destroy()
    w = None

class Toplevel1:
    def imprimir(self,etiqueta):
        valor=etiqueta.get()
        print("ggg",valor)
```

```
    def __init__(self, top=None):
        _bgcolor = '#d9d9d9'  # X11 color: 'gray85'
        _fgcolor = '#000000'  # X11 color: 'black'
        _compcolor = '#d9d9d9' # X11 color: 'gray85'
        _ana1color = '#d9d9d9' # X11 color: 'gray85'
        _ana2color = '#d9d9d9' # X11 color: 'gray85'
        font13 = "-family {Bodoni Bd BT} -size 15 -weight normal "  \
            "-slant roman -underline 0 -overstrike 0"
        font14 = "-family {Segoe UI Light} -size 12 -weight normal "  \
            "-slant roman -underline 0 -overstrike 0"

        top.geometry("625x442+231+114")
        top.title("CLUSTERING")
        top.configure(background="#d9d9d9")
        top.configure(highlightbackground="#f0f0f0")

        self.Frame1 = tk.Frame(top)
        self.Frame1.place(relx=0.016, rely=0.023, relheight=0.192
                , relwidth=0.952)
        self.Frame1.configure(relief='groove')
        self.Frame1.configure(borderwidth="2")
        self.Frame1.configure(relief='groove')
        self.Frame1.configure(background="#e8e8e8")
        self.Frame1.configure(width=595)
        self.Label1 = tk.Label(self.Frame1)
        self.Label1.place(relx=0.824, rely=0.118, height=61, width=94)
        self.Label1.configure(background="#d9d9d9")
        self.Label1.configure(disabledforeground="#a3a3a3")
        self.Label1.configure(foreground="#000000")
```

```
288         self.Label1.configure(width=94)
289
290         self.Label2 = tk.Label(self.Frame1)
291         self.Label2.place(relx=0.017, rely=0.118, height=61, width=484)
292         self.Label2.configure(background="#d9d9d9")
293         self.Label2.configure(disabledforeground="#a3a3a3")
294         self.Label2.configure(font=font13)
295         self.Label2.configure(foreground="#439dd1")
296         self.Label2.configure(text='''ANALISIS DE CONSUMO DE AGUA EMSAP CHANKA''')
297         self.Label2.configure(width=484)
298
299         self.Frame3 = tk.Frame(top)
300         self.Frame3.place(relx=0.016, rely=0.271, relheight=0.441
301                 , relwidth=0.952)
302         self.Frame3.configure(relief='groove')
303         self.Frame3.configure(borderwidth="2")
304         self.Frame3.configure(relief='groove')
305         self.Frame3.configure(background="#e8e8e8")
306         self.Frame3.configure(width=595)
307
308         self.Frame4 = tk.Frame(self.Frame3)
309         self.Frame4.place(relx=0.017, rely=0.051, relheight=0.897
310                 , relwidth=0.966)
311         self.Frame4.configure(relief='groove')
312         self.Frame4.configure(borderwidth="2")
313         self.Frame4.configure(relief='groove')
314         self.Frame4.configure(background="#d9d9d9")
315         self.Frame4.configure(width=575)
```

```
317         self.Button1 = tk.Button(self.Frame4)
318         self.Button1.place(relx=0.122, rely=0.4, height=34, width=95)
319         self.Button1.configure(activebackground="#d9d9d9")
320         self.Button1.configure(activeforeground="#000000")
321         self.Button1.configure(background="#d9d9d9")
322         self.Button1.configure(disabledforeground="#a3a3a3")
323         self.Button1.configure(foreground="#000000")
324         self.Button1.configure(highlightbackground="#d9d9d9")
325         self.Button1.configure(highlightcolor="black")
326         self.Button1.configure(pady="0")
327         self.Button1.configure(text='''SeleccionarData''')
328         self.Button1.configure(command=seleccionar)
329         self.Button1.configure(width=95)
330
331         self.Entry1 = tk.Entry(self.Frame4)
332         #self.valor=tk.StringVar()
333         #self.Entry1.configure(txtvariable=self.valor)
334         self.Entry1.place(relx=0.365, rely=0.4,height=30, relwidth=0.163)
335         self.Entry1.configure(background="white")
336         self.Entry1.configure(disabledforeground="#a3a3a3")
337         self.Entry1.configure(font="TkFixedFont")
338         self.Entry1.configure(foreground="#000000")
339         self.Entry1.configure(insertbackground="black")
340         self.Entry1.configure(width=94)
```

```
343         self.Button2 = tk.Button(self.Frame4)
344         self.Button2.place(relx=0.661, rely=0.4, height=34, width=107)
345         self.Button2.configure(activebackground="#d9d9d9")
346         self.Button2.configure(activeforeground="#000000")
347         self.Button2.configure(background="#d9d9d9")
348         self.Button2.configure(disabledforeground="#a3a3a3")
349         self.Button2.configure(foreground="#000000")
350         self.Button2.configure(highlightbackground="#d9d9d9")
351         self.Button2.configure(highlightcolor="black")
352         self.Button2.configure(pady="0")
353         self.Button2.configure(text='''Ejecutar''')
354         self.Button2.configure(command=graficarClusters)
355         self.Button2.configure(width=107)
356 
357         self.Label3 = tk.Label(self.Frame4)
358         self.Label3.place(relx=0.104, rely=0.114, height=21, width=484)
359         self.Label3.configure(background="#d9d9d9")
360         self.Label3.configure(disabledforeground="#a3a3a3")
361         self.Label3.configure(font=font14)
362         self.Label3.configure(foreground="#439dd1")
363         self.Label3.configure(text='''Agrupamiento de datos con  K-means''')
364         self.Label3.configure(width=484)
365 
366         self.Frame5 = tk.Frame(top)
367         self.Frame5.place(relx=0.016, rely=0.792, relheight=0.17, relwidth=0.952)
368 
369         self.Frame5.configure(relief='groove')
370         self.Frame5.configure(borderwidth="2")
371         self.Frame5.configure(relief='groove')
372         self.Frame5.configure(background="#d9d9d9")
373         self.Frame5.configure(width=595)
```

```
375         self.Button3 = tk.Button(self.Frame5)
376         self.Button3.place(relx=0.084, rely=0.267, height=34, width=97)
377         self.Button3.configure(activebackground="#d9d9d9")
378         self.Button3.configure(activeforeground="#000000")
379         self.Button3.configure(background="#d9d9d9")
380         self.Button3.configure(disabledforeground="#a3a3a3")
381         self.Button3.configure(foreground="#000000")
382         self.Button3.configure(highlightbackground="#d9d9d9")
383         self.Button3.configure(highlightcolor="black")
384         self.Button3.configure(pady="0")
385         self.Button3.configure(text='''ExportarClusters''')
386         self.Button3.configure(command=guerdarClusters)
387         self.Button3.configure(width=97)
388 
389         self.Button4 = tk.Button(self.Frame5)
390         self.Button4.place(relx=0.79, rely=0.267, height=34, width=77)
391         self.Button4.configure(activebackground="#d9d9d9")
392         self.Button4.configure(activeforeground="#000000")
393         self.Button4.configure(background="#d9d9d9")
394         self.Button4.configure(disabledforeground="#a3a3a3")
395         self.Button4.configure(foreground="#000000")
396         self.Button4.configure(highlightbackground="#d9d9d9")
397         self.Button4.configure(highlightcolor="black")
398         self.Button4.configure(pady="0")
399         self.Button4.configure(text='''Cerrar''')
```

```
402             self.Button4.configure(command=cerrar)
403             self.Button4.configure(width=77)
404 
405             self.menubar = tk.Menu(top,font="TkMenuFont",bg=_bgcolor,fg=_fgcolor)
406             top.configure(menu = self.menubar)
407 
408 if __name__ == '__main__':
409     vp_start_gui()
410 
```