DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

CARRERA DE INGENIERÍA EN SISTEMAS E INFORMÁTICA

TRABAJO DE TITULACIÓN, PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN SISTEMAS E INFORMATICA

"MINERÍA DE DATOS STREAMS APLICADA A PARÁMETROS ABIÓTICOS; CASO PRÁCTICO: INVERNADERO DE ROSAS ESPE-IASA I"

AUTORES: JAMI FERNÁNDEZ, JHONNY ALONSO;

MACHÁNGARA QUILCA, KLÉVER GEOVANY

DIRECTOR: ING. DÍAZ ZÚÑIGA, MAGI PAÚL MSC, MBA

SANGOLQUÍ, 2019

CERTIFICADO

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

CARRERA DE INGENIERIA EN SISTEMAS

CERTIFICACIÓN

Certifico que el trabajo de titulación, *"MINERÍA DE DATOS STREAMS APLICADA A PARÁMETROS ABIÓTICOS; CASO PRÁCTICO: INVERNADERO DE ROSAS ESPE-IASA I"* fue realizado por los señores: ***Jhonny Alonso Jami Fernández y Kléver Geovany Machángara Quilca***, el mismo que ha sido revisado en su totalidad, analizado por la herramienta de verificación de similitud de contenido; por lo tanto, cumple con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de las Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente públicamente.

Sangolquí, 04 de Febrero de 2018

Ing. Paúl Díaz Zuñiga MsC.
C.C.: 1707210072

AUTORÍA DE RESPONSABILIDAD

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

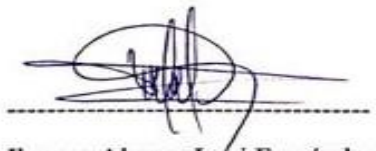
CARRERA DE INGENIERIA EN SISTEMAS E INFORMÁTICA

AUTORÍA DE RESPONSABILIDAD

Nosotros, ***Jhonny Alonso Jami Fernández y Kléver Geovany Machángara Quilca***, declaramos que el contenido, ideas y criterios del trabajo de titulación: ***"MINERÍA DE DATOS STREAMS APLICADA A PARÁMETROS ABIÓTICOS; CASO PRÁCTICO: INVERNADERO DE ROSAS ESPE-IASA I"*** es de nuestra autoría y responsabilidad, cumpliendo con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Consecuentemente el contenido de la investigación mencionada es veraz

Sangolquí, 5 de febrero de 2019



Jhonny Alonso Jami Fernández
C.C.: 1723506265
Telf: 0987814383

Kléver Geovany Machángara Quilca
C.C.: 1004138697
Telf: 0990025479

AUTORIZACIÓN

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

CARRERA DE INGENIERIA EN SISTEMAS E INFORMÁTICA

AUTORIZACIÓN

Nosotros, ***Jhonny Alonso Jami Fernández y Kléver Geovany Machángara Quilca,*** *autorizamos a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación:* ***"MINERÍA DE DATOS STREAMS APLICADA A PARÁMETROS ABIÓTICOS; CASO PRÁCTICO: INVERNADERO DE ROSAS ESPE-IASA I"*** *en el Repositorio Institucional, cuyo contenido, ideas y criterios son de nuestra responsabilidad.*

Sangolquí, 5 de febrero de 2019

Jhonny Alonso Jami Fernández
C.C.: 1723506265

Kléver Geovany Machángara Quilca
C.C.: 1004138697

DEDICATORIA

A Dios, por habernos dado la vida y bendecirnos cada día, por su infinita bondad, por ser una fuente de fe y fortaleza para no desfallecer durante toda nuestra carrera universitaria.

A nuestras familias, quienes han forjado nuestro carácter con valores y principios, por mostrarnos el camino hacia la superación, por su comprensión, consejos, paciencia, apoyo y ejemplo.

A nuestras amigas y amigos con los que hemos compartido momentos inolvidables dentro y fuera de la universidad, y que sin lugar a dudas seguirán formando parte de nuestras vidas, aunque nuestros caminos no sean los mismos después de culminar esta etapa.

Todo esto fue posible gracias a ustedes.

AGRADECIMIENTOS

Nuestro infinito agradecimiento al Ingeniero Paúl Díaz Zúñiga Msc, MBA, por brindarnos la confianza para desarrollar este proyecto, el mismo que con la calidad de Director de Tesis con su experiencia, amplio conocimiento y disponibilidad de tiempo ha sabido asesorarnos clara y acertadamente en el desarrollo de este trabajo. A la Dra. Elisabeth Urbano experta en floricultura quién nos ayudó a comprender y discernir conceptos en esta área fundamental, gracias por su guía, consejos y recomendaciones.

A nuestros padres y hermanos por ser los principales promotores de nuestros sueños, gracias por su paciencia y amor durante toda esta etapa, gracias por enseñarnos que el camino es duro pero que con constancia y esfuerzo todo es posible.

A las personas que conforman la UTIC (Unidad de Tecnologías de la Información y Comunicación) por su conocimiento y apoyo, ya que han sido parte fundamental en este proceso final.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

La generación de datos a gran velocidad y en grandes cantidades provenientes de repositorios, organizaciones, sistemas de monitoreo etc, deben ser tratados en tiempo real para poder actuar de forma rápida y oportuna. Bajo este contexto el campus IASA I de la Universidad de las Fuerzas Armadas ESPE cuenta con un invernadero para el cultivo de rosas, en el que se encuentra ya implementada una WSN “Wireless Sensor Network” que genera streams cada cierto tiempo y en el cual se necesita extraer conocimiento, para ello se ha propuesto mediante está presente investigación aplicar técnicas de minería de datos stream utilizando herramientas de software libre. Este proceso inició con la identificación y selección de las técnicas de DSM para detectar el comportamiento de los factores abióticos, seguidamente se realizó una inferencia con el experto del área. Para finalizar los resultados obtenidos fueron presentados a través de dashboards interactivos que ayudan abstraer conocimiento y por ende aportan al mejoramiento de cultivo de rosas bajo invernadero.

PALABRAS CLAVE:

- **MINERÍA DE DATOS**
- **FLUJOS DE DATOS**
- **ORANGE**
- **CLUSTERIZACIÓN**
- **GRANDES DATOS**
- **FLORICULTURA**

ABSTRACT

The generation of data at high speed and in large quantities from repositories, organizations, monitoring systems, etc., must be treated in real time in order to act quickly and in a timely manner. In this context, the IASA I campus of the University of the Armed Forces ESPE has a greenhouse for the cultivation of roses, in which a WSN "Wireless Sensor Network" is already implemented, which generates streams from time to time and in which It needs to extract knowledge, for it has been proposed by means of this present investigation to apply stream mining techniques using free software tools. This process began with the identification and selection of DSM techniques to detect the behavior of abiotic factors, followed by an inference with the area expert. To finalize the results obtained were presented through interactive dashboards that help to abstract knowledge and therefore contribute to the improvement of greenhouse cultivation of roses.

KEYWORDS:

- **DATA MINING**
- **STREAMS**
- **ORANGE**
- **CLUSTERIZATION**
- **BIG DATA**
- **FLORICULTURE**

CAPÍTULO 1

INTRODUCCIÓN

1.1. Antecedentes

La minería de datos también conocida como exploración de los datos surgió desde los años 60 en conjunto con los conceptos de “data fishing o data archeology”, pero no fue hasta los años 80 donde este término empezó a consolidarse con el único y fundamental objetivo de ayudarnos a comprender el manejo de la información de grandes volúmenes de datos que hoy en la actualidad nos resultan fáciles de generar o recopilar por medio de la tecnología y el internet.

No obstante, que esta información en forma de streams de datos generada por sensores, sistemas de vigilancia, transacciones bancarias, mercados de valores y sistemas Groupware representan un activo muy valioso para cualquier organización; “Un stream es una secuencia continua, potencialmente infinita, de datos que llegan en determinados instantes de tiempo a un sistema para su almacenamiento o procesamiento” (Nauman Chaudhry; Kevin Shaw; Mahdi Abdelguerfi, 2005), por ello se invierten considerables recursos para poder encontrar patrones que nos ayuden a interpretar fenómenos o sucesos que son imperceptibles a nuestros ojos.

“Un patrón es un punto en el espacio de representación de los patrones, espacio de dimensionalidad determinada por el número de variables consideradas”, por eso es

razonable que los patrones de una misma clase estén cercanos, mientras que los de clases diferentes deberían estar en otras regiones.

Dentro del proceso llamado Descubrimiento de Conocimiento en Base de Datos (Knowledge Discovery in Databases – KDD) la minería de datos puede considerarse como el núcleo de toda esta metodología, debido a que mediante la aplicación de técnicas, buscamos comprender o encontrar patrones significativos que sean válidos, novedosos, potencialmente útiles y comprensibles para un usuario dentro de todo este conjunto de datos en bruto, este proceso básicamente consiste en transformar información de bajo nivel en conocimiento de alto nivel.

1.2. Planteamiento del problema

El Ecuador es un país que se destaca por su producción y exportación de rosas debido a las características que éstas presentan, llegando a ser únicas y muy solicitadas a nivel mundial, sin embargo, los floricultores para conseguir este nivel de calidad deben realizar fuertes inversiones en su cuidado y desarrollo.

El proceso apropiado del cultivo de rosas, radica en ciertas variables ambientales tales como: temperatura ambiente, humedad relativa, humedad del suelo e intensidad luminosa. El sector de la floricultura en el país se encuentra principalmente en puntos geográficos remotos y muchos de ellos tienen dificultades económicas para el acceso a tecnología de punta. Por tal razón, los agricultores que se dedican al área florícola requieren de construir un invernadero para poder proteger el cultivo de las condiciones

climáticas desfavorables que pueden aparecer. Los invernaderos generalmente suelen ser pasivos y son estructuras básicas los cuales no están provistos de ventiladores, sistemas de calefacción ni sistemas de riego.

Existen algunas soluciones tecnológicas que ayudan en el monitoreo de áreas de cultivo, sin embargo, implementar esas soluciones implican altos costos que repercuten en el valor de producción, herramientas como un fitómetro que permite monitorear este tipo de ambientes en tiempo real puede superar los 300 mil dólares. Por esta razón, el sector florícola, puede ser beneficiado con el uso de otras tecnologías, especialmente con soluciones de bajo costo que puedan ser adquiridas e implementadas.

La minería de datos descriptiva descubre patrones en los datos disponibles y mediante la interpretación de expertos en el dominio se puede llegar a contribuir a la toma de decisiones tácticas y estratégicas para mejor el crecimiento de los cultivos.

Actualmente en el invernadero ESPE-IASA I sobre el cual se realizará la investigación, se encuentra ya implementada una red WSN (Wireless Sensor Network) la cual genera datos de temperatura, luz, humedad ambiental y humedad del suelo, mismos que son recibidos en una plataforma de manipulación de datos en tiempo real llamada Apache Kafka, pero, estos datos no son analizados por lo cual necesitan recibir un tratamiento para obtener conocimiento y de esta manera prever riesgos y aprovechar oportunidades.

Dentro de los invernaderos los trabajadores son completamente responsables de las plantas, por esta razón deben ser capacitados constantemente para que puedan estar preparados para cualquier problema, esta capacitación implica mayores costos de producción y la experiencia adquirida por los agricultores se la ha obtenido de forma empírica.

Uno de los objetivos de los invernaderos es mantener a las plantas en las condiciones óptimas para su desarrollo, por esta y los anteriores escenarios mencionados es necesario implementar prácticas para tener un análisis descriptivo en tiempo real del ambiente

1.3. Justificación

En base al gran crecimiento de los datos, el enfoque estándar con el que suele manejarse ya no es suficiente para tratar con big data analytics en las condiciones que actualmente se presentan, ya que el recoger, limpiar, modelar y realizar el despliegue de la información es un proceso considerado ya no muy adecuado ni por su escalabilidad ni por su velocidad de respuesta, entonces, continuar con este mecanismo puede considerarse como un equivalente a perder valor de los datos y, por eso, la minería de datos debe adaptarse a estos nuevos requisitos para poder sacar el máximo provecho.

Dentro de la minería de datos, es importante mencionar que existen dos grandes categorías la minería predictiva la cual predice el valor de un atributo en base un

conjunto de datos y la minería descriptiva sobre la cual este proyecto se enfocara la misma que ayuda a descubrir patrones, tendencias, anomalías o correlaciones entre variables de los datos disponibles, que luego mediante la interpretación de expertos en el dominio se puede llegar a contribuir a la toma de decisiones tácticas y estratégicas para mejorar el cultivo de las rosas.

Por tal motivo la aplicación de este tipo de técnicas de minería sobre streams de datos provenientes de cualquier medio pueden revelar el comportamiento de los individuos inmersos en el ambiente compartido y producir conocimiento que antes no era factible encontrar debido a la variedad y complejidad de la información, como por ejemplo: estrategias de inteligencia de negocios, modelos de predicción avanzada y propagación de historias (Valbuena, Cardona y Fernández, 2015).

Precisamente esta última forma de analizar los datos está ganando mucha importancia, ya que vamos a ser capaces de mantener modelos en línea, incorporar nuevos datos sobre la marcha y sobre todo, son efectivos a la hora de detectar cambios ya que por su naturaleza pueden ajustarse a ellos inmediatamente.

Obviamente hay que tomar muy en cuenta que este tipo de minería optimiza su valor mucho más cuando lo aplicamos a la búsqueda de patrones de comportamiento (modelo descriptivo), en este contexto y bajo nuestro caso práctico se planea analizar todos los datos provenientes de sensores instalados previamente como: temperatura, luz, humedad ambiental y humedad del suelo, para de esta manera determinar si existe interacciones entre estos factores abióticos que ayuden a mejorar el cultivo de rosas

1.4. Objetivos

a. Objetivo General

Identificar interacciones de factores abióticos mediante la aplicación de técnicas de minería de datos streams para el cultivo de rosas en el invernadero ESPE-IASA I.

b. Objetivos Específicos

i. Realizar el estudio y aplicación de las principales técnicas no supervisadas para manipulación de datos streams.

ii. Detección de cambios de comportamiento en base a los datos streams provenientes de los sensores para encontrar anomalías que impiden el crecimiento normal del cultivo.

iii. Interpretar y evaluar los patrones descubiertos junto al floricultor para la elaboración de una propuesta de mejoramiento del invernadero.

iv. Establecer las condiciones óptimas para el cultivo de rosas mediante la interpretación de los patrones de comportamiento encontrados.

1.5. Alcance

El proceso de extracción de conocimiento inicia con la obtención de datos, para este proyecto la primera fase ya se encuentra ejecutándose actualmente sobre Apache

Kafka, la cual recoge los datos y los presenta, mismos que son obtenidos por medio de los sensores de temperatura, luz, humedad ambiental y humedad del suelo.

Nuestra propuesta dentro de los tiempos establecidos abarcará los siguientes procesos de extracción de conocimiento:

i. Preparación de los datos e identificación de variables que serán considerados para la construcción del modelo de minería y eliminación de ruido e inconsistencias.

ii. Aplicar técnicas no supervisadas de minería sobre stream de datos identificando correlaciones, tendencias, grupos, trayectorias y anomalías.

iii. Analizar la información para realizar un proceso de inferencia y luego mostrar los resultados de los patrones obtenidos en la fase de minería de datos descriptiva.

iv. Presentar dashboards interactivos identificando los patrones que se expresaron en el análisis del invernadero.

1.6. Hipótesis

H_0: El descubrimiento de patrones ambientales presentarán aporte para mejorar el cultivo de rosas.

H_1: El descubrimiento de patrones ambientales no presentarán aporte para mejorar el cultivo de rosas.

CAPÍTULO 2

PRODUCCIÓN DE ROSAS

En sus inicios la rosa era considerada como un símbolo de belleza por babilonios, sirios, egipcios, romanos y griegos.

Aproximadamente 200 especies botánicas de rosas son nativas del hemisferio norte, aunque no se conoce la cantidad real debido a la existencia de poblaciones híbridas en estado silvestre. (InfoAgro, 2012).

Las primeras rosas cultivadas eran de floración estival, hasta que posteriores trabajos de selección y mejora realizados en oriente sobre algunas especies, fundamentalmente Rosa gigantea y R. chinensis dieron como resultado la “rosa de té”; de carácter refloreciente (variedad que florece). Esta rosa fue introducida en occidente en el año 1793 sirviendo de base a numerosos híbridos creados desde esta fecha. (Infoagro, 2015).

2.1. Morfología

Actualmente, las variedades comerciales de rosa son híbridos de especies de rosa desaparecidas. Para flor cortada se utilizan los tipos de té híbrida y en menor medida los de floribunda. Los primeros presentan largos tallos y atractivas flores dispuestas individualmente o con algunos capullos laterales, de tamaño mediano o grande y numerosos pétalos que forman un cono central visible. Los rosales floribunda

presentan flores en racimos, de las cuales algunas pueden abrirse simultáneamente. (InfoAgro, 2012).

Las flores se presentan en una amplia gama de colores: rojo, blanco, rosa, amarillo, lavanda, etc., con diversos matices y sombras. Éstas nacen en tallos espinosos y verticales. (Infoagro, 2015).

2.2. Requerimientos Edafoclimáticos

Las flores más vendidas en el mundo son, en primer lugar, las rosas seguidas por los crisantemos, tercero los tulipanes, cuarto los claveles y en quinto lugar los lilium. Ninguna flor ornamental ha sido y es tan estimada como la rosa. A partir de la década de los 90 su liderazgo se ha consolidado debido principalmente a una mejora de las variedades, ampliación de la oferta durante todo el año y a su creciente demanda. (Infoagro, 2015).

Sus principales mercados de consumo son Europa, donde figura Alemania en cabeza, Estados Unidos y Japón. (Infoagro, 2015).

Se trata de un cultivo muy especializado que ocupa 1.000 ha de invernadero en Italia, 920 ha en Holanda, 540 ha en Francia, 250 en España, 220 en Israel y 200 ha en Alemania. (Infoagro, 2015).

Los países Sudamericanos han incrementado en los últimos años su producción, destacando, México, Colombia (cerca de 1.000 ha) y Ecuador, la producción se

desarrolla igualmente en África del Este: Zimbabwe con 200 ha y Kenia con 175 ha. (Infoagro, 2015).

En Japón, primer mercado de consumo en Asia, la superficie destinada al cultivo de rosas va en aumento y en la India, se cultivan en la actualidad 100 ha. (Infoagro, 2015).

2.2.1. Temperatura

Para la mayoría de los cultivares de rosa, las temperaturas óptimas de crecimiento son de 17ºC a 25ºC, con una mínima de 15ºC durante la noche y una máxima de 28ºC durante el día. Pueden mantenerse valores ligeramente inferiores o superiores durante períodos relativamente cortos sin que se produzcan serios daños, pero una temperatura nocturna continuamente por debajo de 15ºC retrasa el crecimiento de la planta, produce flores con gran número de pétalos y deformes, en el caso de que abran. Temperaturas excesivamente elevadas también dañan la producción, apareciendo flores más pequeñas de lo normal, con escasos pétalos y de color más cálido. (Infoagro, 2015).

2.2.2 Iluminación

El índice de crecimiento para la mayoría de los cultivares de rosa sigue la curva total de luz a lo largo del año. Así, en los meses de verano, cuando prevalecen elevadas intensidades luminosas y larga duración del día, la producción de flores es más alta que durante los meses de invierno. (Infoagro, 2015).

Una práctica muy utilizada en Holanda consiste en una irradiación durante 16 horas, con un nivel de iluminación de hasta 3.000 lux (lámparas de vapor de sodio), pues de este modo se mejora la producción invernal en calidad y cantidad. (Infoagro, 2015).

No obstante, a pesar de tratarse de una planta de día largo, es necesario el sombreo u oscurecimiento durante el verano e incluso la primavera y el otoño, dependiendo de la climatología del lugar, ya que elevadas intensidades luminosas van acompañadas de un calor intenso. La primera aplicación del oscurecimiento deberá ser ligera, de modo que el cambio de la intensidad luminosa sea progresivo. (Infoagro, 2015).

Se ha comprobado que en lugares con días nublados y nevadas durante el invierno, podría ser ventajosa la iluminación artificial de las rosas, debido a un aumento de la producción, aunque siempre hay que estudiar los aspectos económicos para determinar la rentabilidad. (Infoagro, 2015).

2.2.3 Ventilación y enriquecimiento en CO2

En muchas zonas las temperaturas durante las primeras horas del día son demasiado bajas para ventilar y, sin embargo, los niveles de CO2 son limitantes para el crecimiento de la planta, bajo condiciones de invierno en climas fríos donde la ventilación diurna no es económicamente rentable, es necesario aportar CO2 para el crecimiento óptimo de la planta, elevando los niveles a 1.000 ppm, asimismo, si el cierre

de la ventilación se efectúa antes del atardecer, a causa del descenso de la temperatura, los niveles de dióxido de carbono siguen reduciéndose debido a la actividad fotosintética de las plantas. (Infoagro, 2015).

Por otro lado, hay que tener en cuenta que las rosas requieren una humedad ambiental relativamente elevada, que se regula mediante la ventilación y la nebulización o el humedecimiento de los pasillos durante las horas más cálidas del día. (Infoagro, 2015).

La aireación debe poder regularse, de forma manual o automática, abriendo los laterales y las cumbreras, apoyándose en ocasiones con ventiladores interiores o incluso con extractores (de presión o sobrepresión). Ya que así se produce una bajada del grado higrométrico y el control de ciertas enfermedades. (Infoagro, 2015).

2.3 Cultivo en Invernadero

Con el cultivo de rosa bajo invernadero se consigue producir flor en épocas y lugares en los que de otra forma no sería posible, consiguiendo los mejores precios. Para ello, estos invernaderos deben cumplir unas condiciones mínimas: tener grandes dimensiones (50 x 20 y más), la transmisión de luz debe ser adecuada, la altura tiene que ser considerable y la ventilación en los meses calurosos debe ser buena. Además, es recomendable la calefacción durante el invierno, junto con la instalación de mantas térmicas para la conservación del calor durante la noche. (Infoagro, 2015).

2.3.1. Preparación del suelo

Para el cultivo de rosas el suelo debe estar bien drenado y aireado para evitar encharcamientos, por lo que los suelos que no cumplan estas condiciones deben mejorarse en este sentido, pudiendo emplear diversos materiales orgánicos. (Infoagro, 2015).

Las rosas toleran un suelo ácido, aunque el pH debe mantenerse en torno a 6. No toleran elevados niveles de calcio, desarrollándose rápidamente las clorosis debido al exceso de este elemento. Tampoco soportan elevados niveles de sales solubles, recomendando no superar el 0,15%. (Infoagro, 2015).

La desinfección del suelo puede llevarse a cabo con calor u otro tratamiento que cubra las exigencias del cultivo. En caso de realizarse fertilización de fondo, es necesario un análisis de suelo previo. (Infoagro, 2015).

2.3.2. Plantación

La época de plantación va de noviembre a marzo. Esta se realizará lo antes posible a fin de evitar el desecamiento de las plantas, que se recortan 20 cm; se darán riegos abundantes (100 l de agua/m2), manteniendo el punto de injerto a 5 cm por encima del suelo. (Infoagro, 2015).

En cuanto a la distancia de plantación la tendencia actual es la plantación en 4 filas (60 x 15 cm) o 2 filas (40 x 20 ó 60 x 12,5 cm) con pasillos al menos de 1 m, es

decir, una densidad de 6 a 8 plantas/m2 cubierto. De este modo se consigue un mantenimiento más sencillo y menores inversiones. (Orange, Data mining fruitful and fun open source machine learning, 2016).

2.3.3. Humedad

La humedad ambiental es el segundo factor ambiental que influye en el desarrollo y modifica el rendimiento de los cultivos. Es inversamente proporcional a la temperatura, es decir si la temperatura se incrementa la humedad disminuye, al aumentar la humedad ambiental los estomas permanecen abiertos y la fotorrespiración disminuye, obteniéndose mejores resultados en el desarrollo. Humedad óptima en Rosa: 60-70%. (Infoagro, 2015).

Tabla 1
Humedad de invernaderos

Humedad	Efecto	Recomendaciones
Baja	Aire se reseca, consumo de reservas muy alto. Nuevos tallos florales cortos, débiles y con botones pequeños. HR<40% presencia de enfermedades: Ácaros en zonas secas y con altas temperaturas.	Foog System, aspersores aéreos, aspersores en camas, riego de duchas (a primera hora de la mañana). Ductos para cerrar los cenitales y evitar la reducción de la HR. Pantallas de sombreo, malla sombreado, doble cubierta, etc.

CONTINÚA

Alta	Punto de saturación, punto de rocío. HR > 80%, presencia de enfermedades: Chancros en los tallos, botrytis en los botones y tallos, Peronospora en hojas, tallos y botones. Reduce la transpiración, por lo tanto tamaño de hoja y producción.	Movimiento de aire mediante ventiladores. Plásticos con menor transmitancia de infrarrojo. Doble cubierta. Cubiertas plásticas antigoteo. Pantallas térmicas (mejor distribución del calor). Riego por goteo (< humedad ambiental). Mulch (reducen humedad absoluta). Eliminar las hierbas (también transpiran y aportan humedad al ambiente). No regar después de las 10 AM, nunca en la tarde.

Fuente: (Infoagro, 2015).

Tabla 2
Niveles de referencia de nutrientes en hojas

Macroelementos	**Niveles deseables (%)**
Nitrógeno	3,00-4,00
Fósforo	0,20-0,30
Potasio	1,80-3,00
Calcio	1,00-1,50
Magnesio	0,25-0,35
Microelementos	**Niveles deseables (ppm)**
Zinc	15-50
Manganeso	30-250
Hierro	50-150
Cobre	5-15
Boro	30-60

Fuente: (Infoagro, 2015).

El pH puede regularse con la adición de ácido y teniendo en cuenta la naturaleza de los fertilizantes. Así, por ejemplo, las fuentes de nitrógeno como el nitrato de amonio y el sulfato de amonio, son altamente ácidas, mientras que el nitrato cálcico y el nitrato potásico son abonos de reacción alcalina. Si el pH del suelo tiende a aumentar, la aplicación de sulfato de hierro da buenos resultados. El potasio suele aplicarse como nitrato de potasio, el fósforo como ácido fosfórico o fosfato monopotásico y el magnesio como sulfato de magnesio. (Infoagro, 2015).

2.3.4. Formación de la planta y poda posterior

Los arbustos de dos años ya tienen formada la estructura principal de las ramas y su plantación debe realizarse de forma que el injerto de yema quede a nivel del suelo o enterrado cerca de la superficie. Las primeras floraciones tenderán a producirse sobre brotes relativamente cortos y lo que se buscará será la producción de ramas y más follaje antes de que se establezca la floración, para lo cual se separan las primeras yemas florales tan pronto como son visibles. Las ramas principales se acortan cuatro o seis yemas desde su base y se eliminan por completo los vástagos débiles. Puede dejarse un vástago florecer para confirmar la autenticidad de la variedad. (Infoagro, 2015).

Hay que tener en cuenta que los botones puntiagudos producirán flores de tallo corto y éstos se sitúan en la base de la hoja unifoliada, la de tres folíolos y la primera hoja de cinco folíolos por debajo del botón floral del tallo. En la mitad inferior del tallo las yemas son bastante planas y son las que darán lugar a flores con tallo largo, por lo que cuando un brote se despunta es necesario retirar toda la porción superior hasta un punto por debajo de la primera hoja de cinco folíolos. (Infoagro, 2015).

Posteriormente la poda se lleva a cabo cada vez que se cortan las flores, teniendo en cuenta los principios antes mencionados. (Infoagro, 2015).

2.3.5. Cultivo sin suelo

En los últimos años, el cultivo sin suelo se está convirtiendo en una alternativa muy aconsejable para el cultivo del rosal. Esta técnica se desarrolló como consecuencia de problemas patológicos y agronómicos (fatiga del suelo). (Infoagro, 2015).

De las 4 técnicas posibles (lana de roca, canalones, contenedores planos y contenedores), las dos primeras son actualmente las más utilizadas. Los canalones pueden recibir los siguientes sustratos: perlita, arena, cortezas y fibras vegetales. (Infoagro, 2015).

El suelo del invernadero debe estar nivelado para permitir una irrigación regular (pendiente del 0,5%). Se puede recubrir totalmente de una tela sin suelo que evita posibles contaminaciones a partir de la tierra. Los sacos de cultivo denominados comúnmente "salchichas" son enviados en módulos de 1 ó 2 m de longitud y de 7,5-10 cm de espesor y de 15 a 20 cm de ancho. Algunos productores instalan las "salchichas" sobre los caballones o los canales a 50-80 cm de altura para facilitar la recolección y los tratamientos y mejorar las condiciones sanitarias. (Infoagro, 2015).

Debido al débil espesor del sustrato (7,5-10 cm), las plantas serán más pequeñas que en el cultivo tradicional: miniesquejes y miniinjertos. Después de la plantación (normalmente en marzo) y durante 4 a 5 semanas, la planta crece naturalmente, y dependiendo de los cultivares hay que intervenir para formar una estructura arqueada que favorezca la formación de maderas sobre las cuales se suprimen los botones florales. Después de 1 ó 2 pinzamientos, la recolección de las primeras flores tiene lugar mes y medio más tarde. (Infoagro, 2015).

En el caso del cultivo en canalones sobre perlita, es diferente la elección de las plantas que son parecidas a las utilizadas para el cultivo tradicional en tierra. (Infoagro, 2015).

En la multiplicación se emplean 2 técnicas: el esquejado del cultivar (franco de pie) y el semiinjerto. En el primer caso se realiza con esquejes de trozos de brotes con hojas (con una hoja), dispuestos directamente en el contenedor de comercialización, por ejemplo cubo de lana de roca. En el segundo caso se realiza el injerto a la inglesa simple realizada sobre un trozo de brote del mismo tipo que el utilizado para el esquejado. El miniinjerto se aplica a ciertos cultivares que se multiplican difícilmente por esquejado, como es el caso del cultivar Dallas. (Infoagro, 2015).

Ventajas del cultivo sin suelo:

- La productividad es superior en relación con el cultivo tradicional (incremento entre el 10-30%, según cultivares).
- La calidad es comparable a la del cultivo continuo.
- El estado sanitario es excelente.

2.4. Recolección

Generalmente el corte de las flores se lleva a cabo en distintos estadios, dependiendo de la época de recolección, así, en condiciones de alta luminosidad durante el verano, la mayor parte de las variedades se cortan cuando los sépalos del

cáliz son reflejos y los pétalos aún no se han desplegado, y sin embargo, el corte de las flores durante el invierno se realiza cuando están más abiertas, aunque con los dos pétalos exteriores sin desplegarse, entonces si se cortan demasiado inmaduras, las cabezas pueden marchitarse y la flor no se endurece, ya que los vasos conductores del pedicelo aún no están suficientemente lignificados. (Infoagro, 2015).

En todo caso, siempre se debe dejar después del corte, el tallo con 2-3 yemas que correspondan a hojas completas. Si cortamos demasiado pronto, pueden aparecer problemas de cuello doblado, como consecuencia de una insuficiente lignificación de los tejidos vasculares del pedúnculo floral. (Infoagro, 2015).

Las cualidades deseadas de las rosas para corte, según los gustos y exigencias del mercado en cada momento, son:

- Tallo largo y rígido: 50-70 cm, según zonas de cultivo.
- Follaje verde brillante.
- Flores: apertura lenta, buena conservación en florero.
- Buena floración (igual a rendimiento por pie o por m2).
- Buena resistencia a las enfermedades.
- Posibilidad de ser cultivados a temperaturas más bajas, en invierno.
- Aptitud para el cultivo sin suelo.

CAPÍTULO 3

ESTADO DE CUESTIÓN

Para desarrollar el presente trabajo de titulación se revisará los principales conceptos de minería de datos streams, herramientas, definiciones y sus respectivas técnicas para el tratamiento de datos continuos, por ende también es importante realizar un análisis de las características que nos ofrecen cada herramienta y determinar cuál de ellas tienen los métodos competentes para nuestro proyecto.

3.1 MINERÍA DE DATOS

Es la etapa de análisis de Knowledge Discovery in Databases o KDD es un campo de la estadística y las ciencias de la computación referente al proceso de intentar descubrir patrones en grandes volúmenes de conjuntos de datos, este se basa en los métodos de la inteligencia artificial, aprendizaje automático, estadística y sistemas de bases de datos, como objetivo general del proceso de minería de datos consiste en extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior. (Camejo, 2012).

3.1.1 Técnicas de minería de datos

Las técnicas de minería de datos se clasifican en dos grupos los que permiten realizar predicciones y los de descubrimiento de conocimiento.

- **Algoritmos supervisados:** Los algoritmos trabajan con datos “etiquetados” (labeled data), intentado encontrar una función que, dadas las variables de entrada (input data), les asigne la etiqueta de salida adecuada, El algoritmo se entrena con un “histórico” de datos y así “aprende” a asignar la etiqueta de salida adecuada a un nuevo valor, es decir, predice el valor da salida. (Luca, 2017).
- **Algoritmos no supervisados:** El aprendizaje no supervisado tiene lugar cuando no se dispone de datos “etiquetados” para el entrenamiento, sólo conocemos los datos de entrada, pero no existen datos de salida que correspondan a un determinado input, por tanto, sólo podemos describir la estructura de los datos, para intentar encontrar algún tipo de organización que simplifique el análisis, por ello, tienen un carácter exploratorio. (Luca, 2017).

Tabla 3
Clasificación de técnicas de DM

SUPERVISADOS	NO SUPERVISADOS
Árboles de decisión	Segmentación
Red neuronal	Agrupamiento (clustering)
Regresión	Reglas de asociación
Series temporales	Patrones secuenciales

Fuente: (Santos, 2015).

a) Árboles de Decisión

Es un modelo de predicción utilizado en diversos ámbitos que van desde la inteligencia artificial hasta la Economía, dado un conjunto de datos se fabrican diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema, un área donde son aplicados los árboles de decisión es la teoría de juegos, esta utiliza modelos para estudiar interacciones de estructuras formalizadas de incentivos, esta técnica es muy utilizada en diversos campos: economía, política, biología, sociología, psicología, filosofía y ciencias de la computación entre otras. (Camejo, 2012).

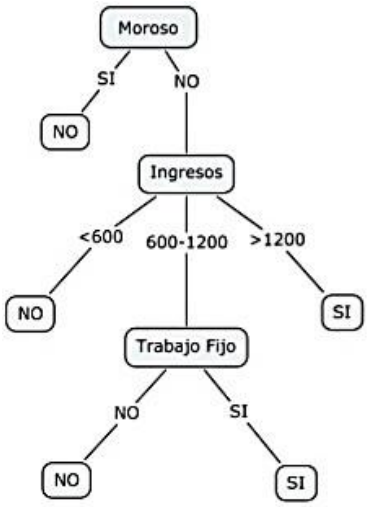


Figura 1. Árbol de decisión para préstamo.
Fuente: (Xataka, 2014)

b) Red Neuronal

Las redes neuronales se basan en una serie de pequeños nodos (neuronas artificiales) que se encuentran conectados entre sí conformando una serie de capas. Su funcionamiento pretende simular el comportamiento de un cerebro humano a la hora de resolver un problema. A medida que la señal avanza, ésta irá tomando un camino u otro en función a unos parámetros establecidos previamente hasta ofrecer un resultado. Las redes neuronales se han utilizado durante años para realizar varias tareas como, por ejemplo el reconocimiento de voz o la creación de sistema expertos con conocimiento incorporado, algo que marcó un cambio en el rumbo de la inteligencia artificial en los años 80. (Luca, 2017).

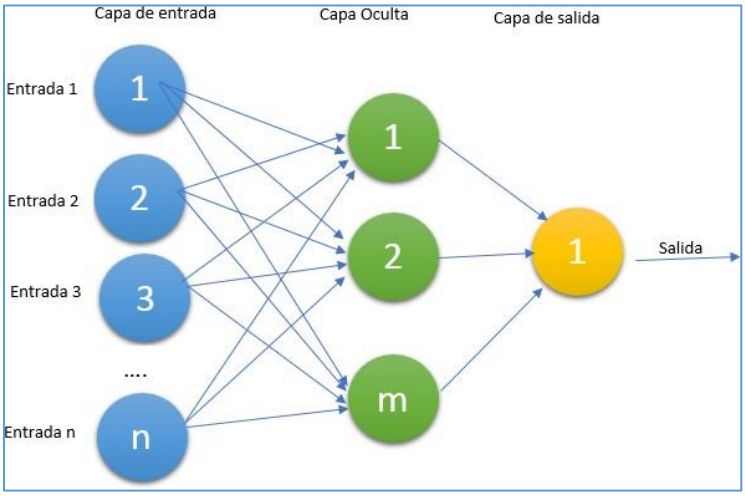


Figura 2. Ejemplo de Red Neuronal
Fuente: (Xataka, 2014)

c) Regresión Lineal

Permite determinar el grado de dependencia de las series de valores X e Y, prediciendo el valor y estimado que se obtendría para un valor x que no esté en la distribución, existen dos tipos de regresión.

La regresión lineal simple se basa en estudiar los cambios en una variable, no aleatoria, que afectan a una variable aleatoria, en el caso de existir una relación funcional entre ambas variables la cual puede ser establecida por una expresión lineal, lo cual su representación gráfica es una línea recta, es decir, que estamos en presencia de una regresión lineal simple cuando una variable independiente ejerce influencia sobre otra variable dependiente.

La regresión lineal permite trabajar con una variable a nivel de intervalo o razón, así también se puede comprender la relación de dos o más variables y permitirá relacionar mediante ecuaciones, una variable en relación a otras variables llamándose Regresión múltiple. O sea, la regresión lineal múltiple es cuando dos o más variables independientes influyen sobre una variable dependiente, uno de los campos donde se aplica la regresión lineal ha sido en la Medicina ya que las primeras evidencias son estudios relacionados con la mortalidad y el fumar tabaco. (Lizarraga, 2015).

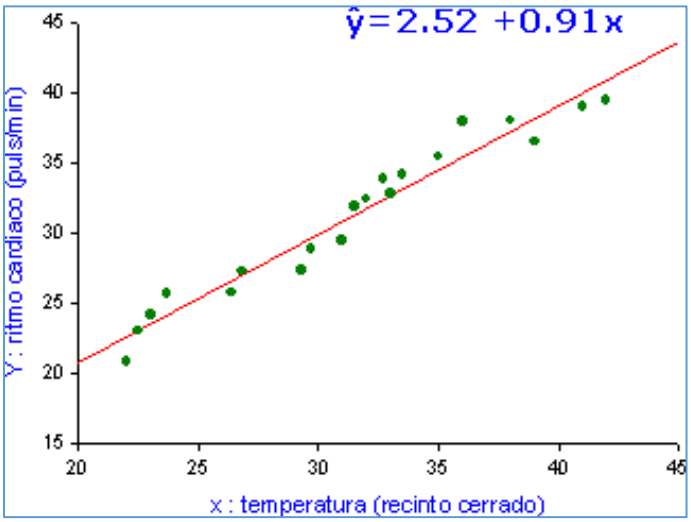


Figura 3. Regresión Lineal
Fuente: (Weka, Practica machine learning tools and techniques, 2015).

d) Series Temporal

Una serie temporal o cronológica es una secuencia de datos, observaciones o valores, medidos en determinados momentos y ordenados cronológicamente, los datos pueden estar espaciados a intervalos iguales (como la temperatura en un observatorio meteorológico en días sucesivos al mediodía) o desiguales (como el peso de una persona en sucesivas mediciones en el consultorio médico, la farmacia, etc.), para el análisis de las series temporales se usan métodos que ayudan a interpretarlas y que permiten extraer información representativa sobre las relaciones subyacentes entre los datos de la serie o de diversas series y que permiten en diferente medida y con distinta confianza extrapolar o interpolar los datos y así predecir el comportamiento de la serie en momentos no observados, que pueden existir en el futuro (extrapolación pronostica),

en el pasado (extrapolación retrógrada) o en momentos intermedios (interpolación). (Martinez de Pinson Ascacibar, 2015).

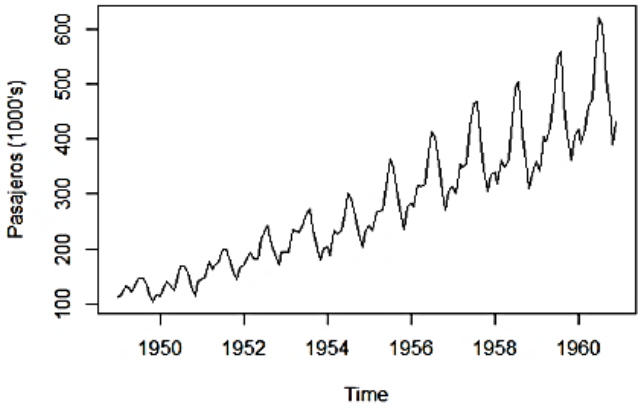


Figura 4. Series Temporales
Fuente: (Xataka, 2014)

e) Segmentación

Cuando se tiene información en una empresa, no es suficiente contar solamente con ella, se requiere analizarla, explorarla y descubrir los patrones que proporcionan la información, por tanto el objetivo es clasificar la información que se desea evaluar y analizar, y así determinar las variables continuas para segmentar o clasificar en grupos y finalmente determinar que ocurre en una base de datos de gran tamaño, la detección de segmentos sería el objetivo principal de la minería de datos. La clasificación es un modo de segmentar datos asignándolos a grupos que están previamente definidos, por tanto, clustering divide la base de datos en diferentes grupos, su objetivo será encontrar grupos que son diferentes entre si. (Conexionesan, 2014).

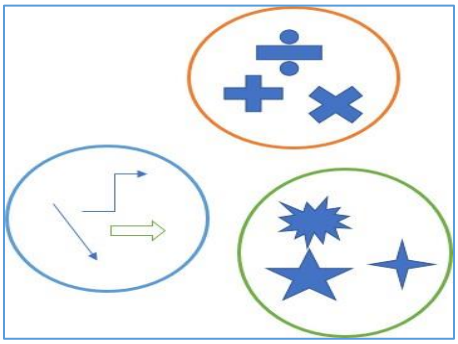


Figura 5. Segmentación
Fuente: (Xataka, 2014)

f) Agrupamiento (Clustering)

Un algoritmo de agrupamiento (en inglés, clustering) es un procedimiento de agrupación de una serie de vectores de acuerdo con un criterio, esos criterios son por lo general distancia o similitud, lo cual la cercanía se define en términos de una determinada función de distancia, como la euclídea, aunque existen otras más robustas o que permiten extenderla a variables discretas, la medida más utilizada para medir la similitud entre los casos es la matriz de correlación entre los nxn casos. Sin embargo, también existen muchos algoritmos que se basan en la maximización de una propiedad estadística llamada verosimilitud, generalmente, los vectores de un mismo grupo (o clústers) comparten propiedades comunes, el conocimiento de los grupos puede permitir una descripción sintética de un conjunto de datos multidimensional complejo. (Minitab, Interpretación de los resultados clave, 2012).

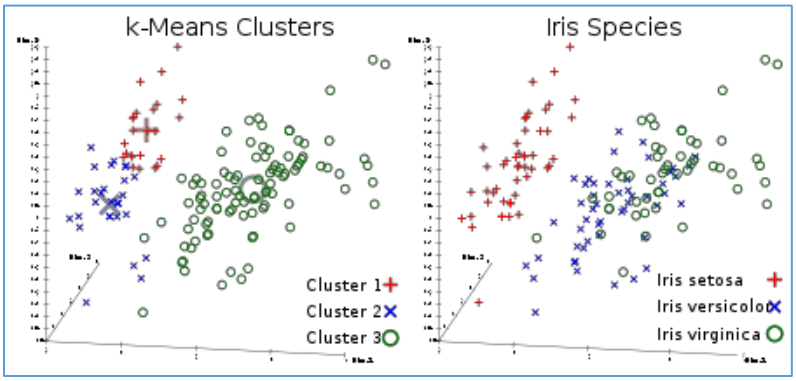


Figura 6. Ejemplo Clustering
Fuente: (Xataka, 2014)

g) Patrones Secuenciales

El campo de la minería de datos y extracción de conocimiento, la minería de secuencias es un caso particular de la minería de datos estructurados. Consiste en encontrar patrones estadísticamente relevantes en colecciones de datos que están representados de forma secuencial, debido a la frecuencia con que aparecen este tipo de datos en escenarios de aplicaciones reales, esta técnica constituye uno de los métodos más populares de descubrimiento de patrones, los patrones frecuentes obtenidos durante el minado de secuencias, se usan en tareas de detección de dependencias funcionales, predicción de tendencias, interpretación de fenómenos y como soporte de decisiones en estrategias de producción. (Solsoft, 2014).

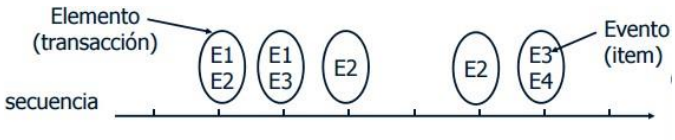


Figura 7. Secuencia
Fuente: (Weka, Practica machine learning tools and techniques, 2015)

3.2 MINERÍA DE DATOS STREAMS

La minería de datos también conocida como exploración de los datos surgió desde los años 60 en conjunto con los conceptos de “data fishing o data archeology”, pero no fue hasta los años 80 donde este término empezó a consolidarse con el único y fundamental objetivo de ayudarnos a comprender el manejo de la información de grandes volúmenes de datos que hoy en la actualidad nos resultan fáciles de generar o recopilar por medio de la tecnología y el internet.

Esta evolución ha hecho que se vaya relegando en un segundo plano la tradicional forma de minería de datos que se basaba en la creación de modelos a través del análisis de muestras de datos o registros históricos.

Sin embargo, y pese a lo profundo de sus raíces, la minería de datos sigue siendo tendencia y a través de distintas aplicaciones, entre las cuales destacan: Redes neuronales (secuenciación, clasificación), Clustering (segmentación), Clasificación por segmentación (filtros spam y análisis sentimental de Twitter) y Minería de Datos Streams.

Cabe mencionar que esta última está ganando importancia, debido a que es necesario tomar acciones en tiempo real y, para eso, hace falta contar con un buen conocimiento de la situación, característica fundamental que provee este tipo de minerías.

3.3 STREAMS DE DATOS

En un contexto general se utiliza el término “stream” para representar el flujo continuo de los datos, un caso muy común de “streams” son las señales de video o audio en directo que pueden ser transmitidas por un sin número de aplicaciones móviles disponibles, pero, en nuestro caso los “streams” son generados por la red WSN (Wireless Sensor Network) ya implementada en un invernadero del campus IASA I.

La problemática que se generan en estos escenarios es que existe paradigma antiguo llamado store-then-process donde, como su nombre lo dice, primero se almacenan los datos y luego se procesan. Por tal motivo y en consecuencia no es posible satisfacer necesidades en procesamiento rápido”. (Callau Zori, 2012).

Entonces, para satisfacer y solucionar los problemas anteriores nacen los flujos de datos (data stream) como un modelo o paradigma en donde se procesan los datos a su llegada mediante algoritmos, los mismos que una vez procesados pueden ser visualizados en el mismo instantes.

Los dos principales modelos para trabajar sobre streams de datos son el evolutivo y el de ventana deslizante. El primero ellos considera todo el histórico de los

datos. El segundo, y más utilizado, hace uso de ventanas deslizantes, que permite considerar los datos más recientes. Una ventana deslizante, se comporta de forma similar a una cola (Primero en entrar primero en salir), almacenando y procesando únicamente el sub string definido por el tamaño de la ventana (Jaramillo, Cardona & Fernández, 2015).

Además, se debe tomar en cuenta que cuando se presentan flujos de datos a éstos se los piensa como “transitorios”, y por tanto las fuentes de su procedencia puede ser muy variada.

Por tal motivo al terminar de procesarse un elemento de un flujo de datos, éste automáticamente se descarta o archiva, generalmente no es fácil de almacenar en memoria debido a que su tamaño es pequeño en relación al tamaño del flujo de datos. (Babcock, Babu, Datar, Motwani, & Widom, 2002)

3.4 HERRAMIENTAS PARA BIG DATA

En la actualidad existe una gran variedad de herramientas para solventar las complicaciones generadas por los “grandes volúmenes de datos”. No obstante, cabe mencionar que la diferencia entre una u otra es precisamente el modo en que realizan el procesamiento de los datos.

Bajo la naturaleza de nuestro proyecto, las herramientas escogidas para un análisis preliminar son aquellas enfocadas a procesamiento en streaming y de código abierto. Entre las cuales y hasta la fecha tenemos Weka, Spark Streaming, Apache

Storm, Orange (la más popular de las herramientas por su interfaz interactiva), Kafka Streaming. A continuación, se proveerá de un breve resumen de las herramientas encontradas.

3.4.1 Weka

Weka es una herramienta de tipo software para el aprendizaje automático y minería de datos diseñado a base de Java y desarrollado en la universidad de Waikato en Nueva Zelanda en el año 1993, esta herramienta por su nombre en inglés (Waikato Environment for Knowledge Analysis) es de distribución de licencia GNU-GLP o software libre. (Weka, Practica machine learning tools and techniques, 2015).

Esta contiene una colección de algoritmos para realizar análisis de datos descriptivo como modelado predictivo, herramientas para la visualización de datos y provee una interfaz gráfica que unifica todas las herramientas para que estén a una mejor disposición. (Weka, Practica machine learning tools and techniques, 2015).

Weka está diseñado como una herramienta orientada a la extensibilidad por lo que una de las propiedades más interesantes de este software, es su facilidad para añadir extensiones, modificar métodos etc. (Weka, Practica machine learning tools and techniques, 2015).

Ventajas

- Software de licencia publica GNU que acceso a cualquier usuario.

- Contiene un sin número de técnicas de procesamiento modelado y procesamiento de datos.
- Interfaz fácil de usar.
- Multiplataforma ya que es una implementación en java.

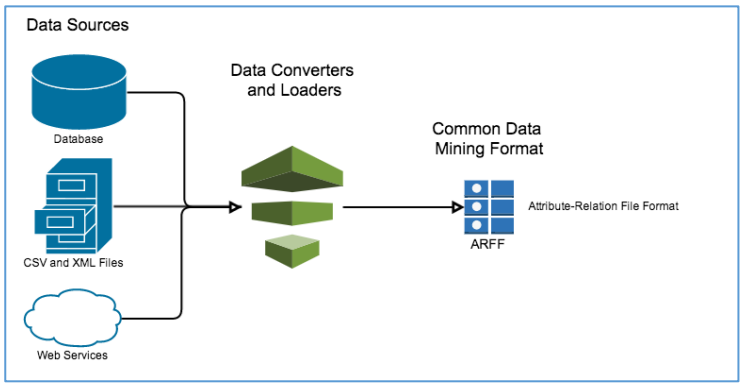


Figura 8. Arquitectura Weka
Fuente: (Weka, Weka 3 machine learning software in java, 2017)

3.4.2 Orange

Es un software informático de código abierto para realizar minería de datos o análisis predictivo, provee todas las herramientas necesarias para machine learning y visualización de datos, ideal para usuarios novatos como expertos en el área.

Esta herramienta posee un entorno gráfico llamado Orange Canvas, en el cual los usuarios pueden colocar widgets sobre esta y luego conectarlos a un esquema, cada widget realiza alguna función básica, pero a diferencia de un sin número de visualizaciones de datos y modelos incluye búsqueda inteligente para una mejor

visualización. (Orange, Data mining fruitful and fun open source machine learning, 2016).

Orange consta de muchas características que pueden ser añadidas para una mejor solución en base a tu proyecto, como puede ser el análisis con datos en flujo, por ejemplo: Flux Vision, una solución iniciada por Orange Labs, analiza los flujos de población en tiempo real utilizando los datos de la red móvil de Orange. Convierte millones de elementos de información técnica de la red móvil en indicadores estadísticos para analizar con qué frecuencia se visitan diferentes áreas geográficas y cómo se mueven las personas. Además de los indicadores de ubicación (como la densidad, de dónde provienen y van los flujos de usuarios) proporcionados por la red móvil, este servicio también ofrece datos sociodemográficos anónimos adicionales (como edad, género y categoría socio profesional), que brindan información local. Las autoridades y las empresas tienen una mejor visión de los perfiles de sus clientes y usuarios. Flux Vision utiliza un sistema innovador patentado para garantizar el anonimato irreversible de los datos que respeta todos los requisitos de la vigilancia de la privacidad de Francia, la Comisión Nacional de Informática y Libertad (CNIL). (Astellia, 2015).

Figura 9. *Flux Vision Real Time*
Fuente: (Orange, Fluz vision real time statistics on mobility patterns, 2015)

Orange permite incorporar varios widget como Continuize el cual permite ver transformaciones en tiempo real. Este widget esencialmente crea nuevos atributos a partir de sus discretos. Si tiene, por ejemplo, un atributo con el color de ojos de las personas, donde los valores pueden ser azul, marrón o verde, probablemente querría tener tres atributos separados 'azul', 'verde' y 'marrón' con 0 o 1 si una persona tiene ese color de ojos Algunos alumnos se desempeñan mucho mejor si los datos se transforman de esa manera. También puede tener solo atributos en los que supondría que 0 es una condición normal y solo le gustaría tener las desviaciones del estado normal registrado ('objetivo o primer valor como base') o el estado normal sería el valor más común ('la mayoría valor frecuente como base '). El widget Continuize le ofrece mucho espacio para jugar. Lo mejor es seleccionar un pequeño conjunto de datos con valores discretos, conectarlo a Continuize y luego a la Tabla de datos y cambiar los

parámetros. Así es como puedes observar las transformaciones en tiempo real. Es útil para proyectar puntos de datos discretos en Proyección Lineal. (Biolab, 2016).

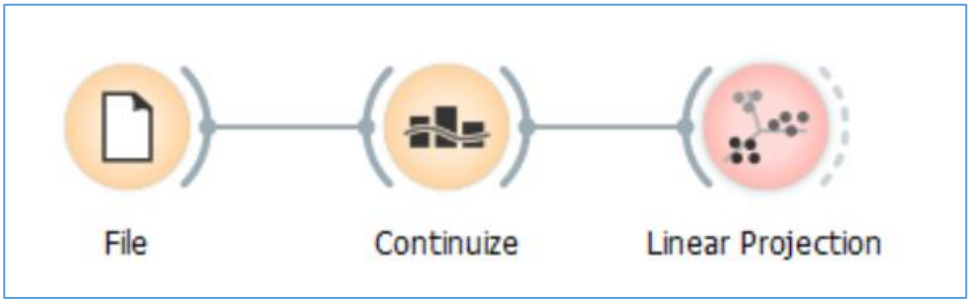


Figura 10. Widget Continuize
Fuente: (Orange, Fluz vision real time statistics on mobility patterns, 2015)

3.4.3 Spark Streaming

Es un sistema de computación distribuida en clústers de computadores. Spark streaming tiene como base Apache Spark, y éste a su vez parte de los conceptos básicos de Hadoop como MapReduce; es decir que Apache Spark tiene a su disposición a DAG (Direct Acyclic Graph) y RDD (Resilent Distributed Dataset). Con estas dos funcionalidades a su disposición, Spark Streaming mejora la tolerancia a fallos de MapReduce y a la vez reduce costos de procesamientos al explotar al máximo la memoria caché y no escribir en disco cada parte del proceso.

Spark Streaming tiene la característica de poder recibir datos de otras fuentes “como Apache Kafka, Flume, Twitter, ZeroMQ, Kinesis o sockets TCP, los cuales

pueden ser procesados utilizando complejos algoritmos a los que se accede mediante funciones de alto nivel como map, reduce o join". (Pérez Esteso, 2015).

Figura 11. Spark streaming componentes
Fuente: (Hortonworks, 2015)

Una vez recibidos dichos datos el procesamiento se realiza por lotes (batchs), y es cuando se puede configurar el tiempo de procesamiento en cada lote y de cuántos datos constan cada uno. Debido a la naturaleza de procesamiento en lotes, siempre suele existir un grado de latencia y esto puede representar una desventaja para Spark Streaming.

3.4.4 Apache Storm

Esta herramienta es un sistema de computación distribuida de código abierto. Su mayor ventaja es el funcionamiento en tiempo real y el hecho de que el sistema se basa en construir topologías de Big data para transformar y analizar los datos que entran constantemente en un proceso continuo.

Una de las características es que “Apache Storm garantiza el procesamiento de un dato de entrada al menos una vez, lo que puede causar inconsistencia debido a que un dato de entrada puede ser procesado dos veces” (Pérez Esteso, 2015)

Pero de todos modos es una herramienta fiable y sencilla que maneja eventos complejos (Complex Event Processing, CEP) en un sistema de procesamiento. Lo ideal, desde la visión empresarial, es que Apache Storm permite responder de manera inmediata a eventos repentinos, y por ende ayuda en la toma de decisiones.

3.4.5 Kafka Streaming

Es una librería cliente para la construcción de aplicaciones y micro servicios, donde los datos de entrada y salida se almacenan en los clusters de Kafka. Es decir combina la simplicidad de escribir e implementar aplicaciones Java y Scala estándar en el lado del cliente con los beneficios de la tecnología cluster del lado del servidor Kafka.

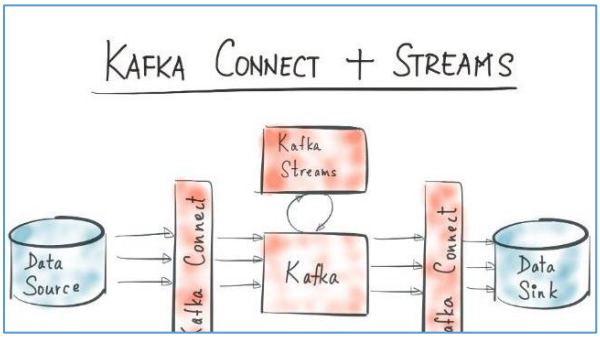


Figura 12. Infraestructura
Fuente: (Confluent, 2015)

3.5 METODOLOGIAS DE MINERIA DE DATOS

Una metodología ayuda hace referencia al conjunto de procedimientos racionales utilizados para alcanzar el objetivo, por este motivo usar una metodología en el campo de minería de datos como en cualquiera otro campo es muy relevante, son diversos los modelos de proceso que han sido propuestos para el desarrollo de proyectos de Data Mining tales como SEMMA, DMAMC, o CRISP-DM sin embargo uno de los modelos principalmente utilizados en los ambientes académico e industrial es el modelo CRISP-DM, esta es la guía de más referencia utilizada en el desarrollo de proyectos de Data Mining, como se puede observar en la Figura 13, la cual fue publicada el año 2007 por kdnuggets.com, el cual representa el resultado obtenido en sucesivas encuestas efectuadas durante los últimos años. (Rodriguez, 2014).

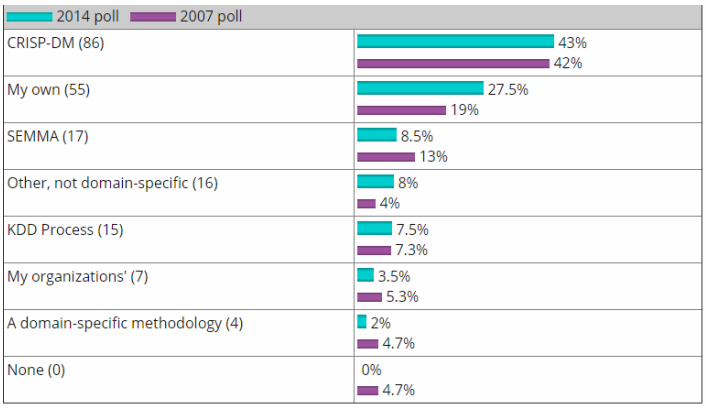


Figura 13. Comparación de metodologías más usadas
Fuente: (Weka, Practica machine learning tools and techniques, 2015)

3.5.1 Metodología Semma

El nombre SEMMA es el acrónimo a las cinco fases:(Sample, Explore, Modify, Model, Assess), esta es una metodología es propuesta por SAS Institute Inc, y lo definen como: proceso de selección, exploración y modelamiento de grandes cantidades de datos para descubrir patrones de negocios desconocidos. (Wikipedia, 2015). En la Figura 14 se muestran sus fases.

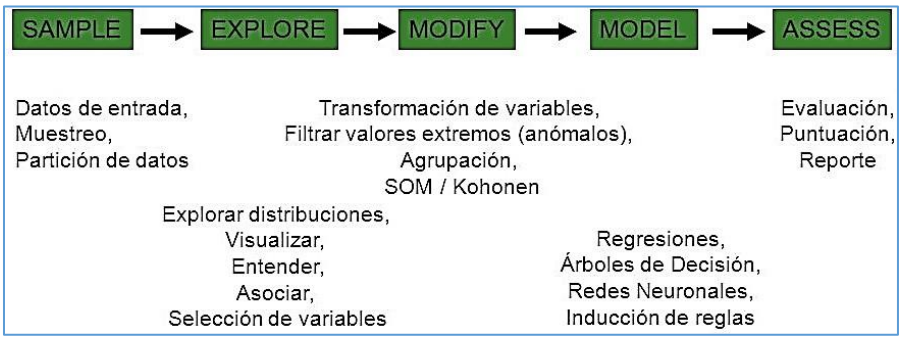


Figura 14. Fases de la metodología
Fuente: (Weka, Weka 3 machine learning software in java, 2017)

- **Muestreo (Sample)**: Extracción de una muestra representativa, en la primera fase de la metodología, se realiza la extracción de un conjunto de datos (población maestral) sobre la que se va a llevar a cabo el análisis, en esta metodología, para cada una de las muestras escogidas se debe asociar un determinado nivel de confianza. (Lizarraga, 2015).
- **Exploración (Explore)**: En esta fase, se realiza un análisis de los datos extraídos en la muestra, para lo cual se propone el uso de herramientas de visualización o de diferentes técnicas estadísticas para la exploración de la

información seleccionada, que contribuyan a poner de manifiesto relaciones entre variables. (Lizarraga, 2015).

- **Modifica (Modify)**: Modificación de los datos, en esta tercera fase de la metodología, involucra la transformación de los datos que van a ser ingresados al modelo para que tengan el formato adecuado, mejorando la definición de los mismos. (Lizarraga, 2015).
- **Modela (Model)**: Se procede a modelar el conjunto de datos, permitiendo al software realizar una búsqueda completa de combinaciones de datos que ayudarán a predecir los resultados esperados de manera confiable, el objetivo de esta fase es establecer una relación entre las variables objeto del estudio y las variables explicativas, de manera tal que posibiliten inferir el valor de las mismas con un nivel de confianza determinado. (Lizarraga, 2015).
- **Evalúa (Assess)**: Esta fase consiste en la valoración de los resultados mediante el análisis de bondad del modelo o modelos, contrastado con otros métodos estadísticos o con nuevas poblaciones muéstrales y entregando un reporte de lo que pudo encontrar con los resultados obtenidos de todos los análisis que puedo hacer. (Rodriguez Montequín & Alvarez Cabal, 2014).

3.5.2 Proceso de extracción de conocimiento (KDD)

Es un proceso iterativo que exhaustivamente explora volúmenes muy grandes de datos para determinar relaciones, es un proceso que extrae información de calidad que puede usarse para dibujar conclusiones basadas en relaciones o

modelos dentro de los datos. (Webmining, 2014). Se compone de cinco fases como lo muestra la Figura 15.

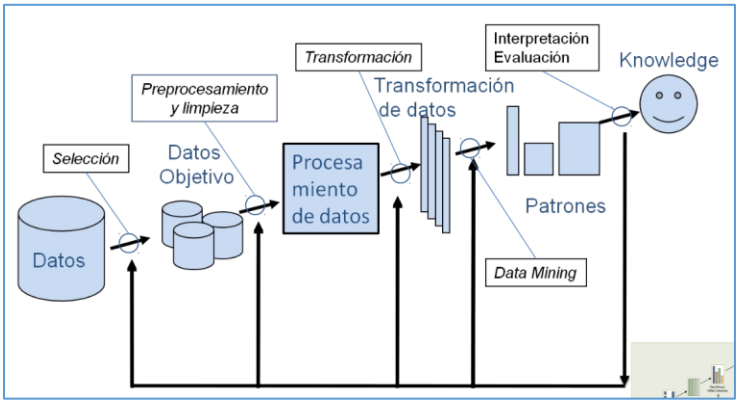


Figura 15. Fases del KDD
Fuente: (Weka, Weka 3 machine learning software in java, 2017)

1. **Selección de datos**. En esta etapa se determinan las fuentes de datos y el tipo de información a utilizar, es la etapa donde los datos relevantes para el análisis son extraídos desde la o las fuentes de datos. (Webmining, 2014).
2. **Pre procesamiento**. Esta etapa consiste en la preparación y limpieza de los datos extraídos desde las distintas fuentes de datos en una forma manejable, necesaria para las fases posteriores. (Webmining, 2014).
3. **Transformación**. Consiste en el tratamiento preliminar de los datos, transformación y generación de nuevas variables a partir de las ya existentes con una estructura de datos adecuada, aquí se realizan operaciones de agregación o normalización, consolidando los datos de una forma necesaria para la fase siguiente. (Webmining, 2014).

4. **Data Mining**. Es la fase de modelamiento propiamente tal, en donde métodos inteligentes son aplicados con el objetivo de extraer patrones previamente desconocidos, válidos, nuevos, potencialmente útiles y comprensibles y que están contenidos u ocultos en los datos. (Webmining, 2014)**.**
5. **Interpretación y Evaluación.** Se identifican los patrones obtenidos y que son realmente interesantes, basándose en algunas medidas y se realiza una evaluación de los resultados obtenidos. (Webmining, 2014).

3.5.3 Metodología CRISP-DM

Proporciona una descripción normalizada del ciclo de vida de un proyecto estándar de análisis de datos, de forma análoga a como se hace en la ingeniería del software, en este contexto tiene en cuenta la existencia de un cliente que no es parte del equipo de desarrollo, así como el hecho de que el proyecto no termina una vez se halla el modelo idóneo (se requiere un despliegue y mantenimiento), sino que está relacionado con otros proyectos, y es preciso documentarlo de forma exhaustiva para que otros equipos de desarrollo utilicen el conocimiento adquirido y trabajen a partir de él. (Villena Román, 2016). Como se muestra en la Figura 16.

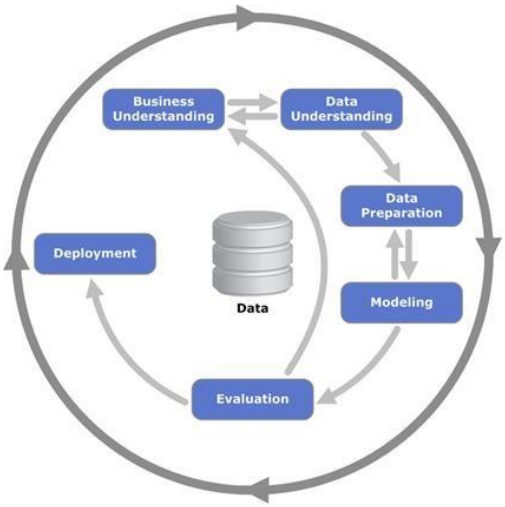


Figura 16. Metodología CRISP-DM
Fuente: (Villena Román, 2016)

A continuación se describen los pasos de la metodología el cual servirá de referencia para el desarrollo de este proyecto de titulación.

a) Comprensión del negocio.

La primera fase es la fase de comprensión del negocio o problema, es probablemente la más importante y aglutina las tareas de comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial, con el fin de convertirlos en objetivos técnicos y en un plan de proyecto DM sin lograr comprender dichos objetivos, ningún algoritmo por muy sofisticado que sea, permitirá obtener resultados fiables. (Oldemar, 2015).

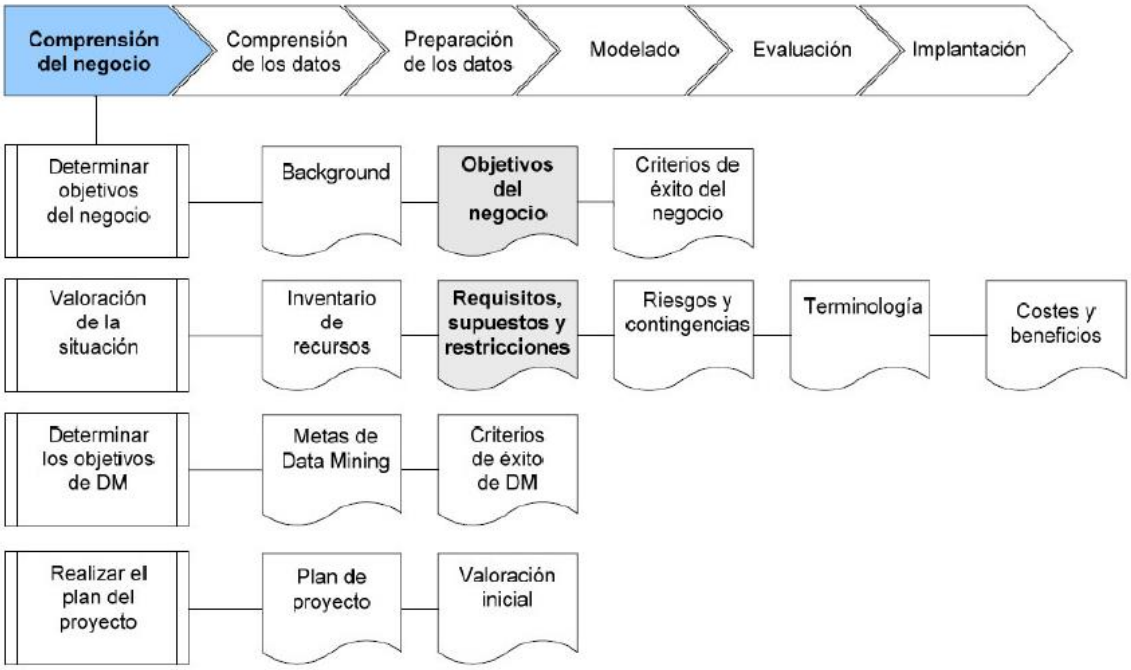


Figura 17. Fase Comprensión del negocio
Fuente: (Rodriguez, 2014)

1. Determinar objetivos del negocio.

Esta es la primera tarea a desarrollar y tiene como metas, determinar cuál es el problema que se desea resolver, por qué es necesario utilizar Data Mining y especificar los criterios de éxito, los problemas pueden ser diversos como por ejemplo, detectar fraude en el uso de tarjetas de crédito, detección de intentos de ingreso indebido a un sistema, asegurar el éxito de una determinada campaña publicitaria, etc. (Oldemar, 2015).

2. **Valoración de la situación.**

En esta tarea se debe calificar el estado de la situación antes de iniciar el proceso de DM, considerando aspectos tales como ¿cuál es el conocimiento previo disponible acerca del problema?, ¿se cuenta con la cantidad de datos requerida para resolver el problema?, ¿cuál es la relación coste beneficio de la aplicación de DM?, entre otras cosas, esta fase define los requisitos del problema, tanto en términos de negocio como en términos de Data Mining. (Oldemar, 2015).

3. **Determinar los objetivos de la minería de datos.**

Esta tarea tiene como objetivo representar los objetivos del negocio en términos de las metas del proyecto de DM, como por ejemplo, si el objetivo del negocio es el desarrollo de una campaña publicitaria para incrementar la asignación de créditos hipotecarios, la meta de DM será por ejemplo, determinar el perfil de los clientes respecto de su capacidad de endeudamiento. (Oldemar, 2015).

4. **Realizar el plan del proyecto.**

Finalmente esta última tarea de la primera fase de CRISP-DM, tiene como fin desarrollar un plan para el proyecto, que describa los pasos a seguir y las técnicas que se emplearan en cada paso. (Oldemar, 2015).

b) Comprensión de los datos.

Esta fase de comprensión de los datos, comprende la recolección inicial de datos, con el objetivo de establecer un primer contacto con el problema, familiarizándose con ellos, identificar su calidad y establecer las relaciones más evidentes que permitan definir las primeras hipótesis esta fase junto a las próximas dos fases, son las que demandan el mayor esfuerzo y tiempo en un proyecto de DM, por lo general si la organización cuenta con una base de datos corporativa, es deseable crear una nueva base de datos ad-hoc al proyecto de DM, pues durante el desarrollo del proyecto, es posible que se generen frecuentes y abundantes accesos a la base de datos a objeto de realizar consultas y probablemente modificaciones, lo cual podría generar muchos problemas. (Oldemar, 2015). La fase tiene cuatro tareas que se muestran en la Figura 18.

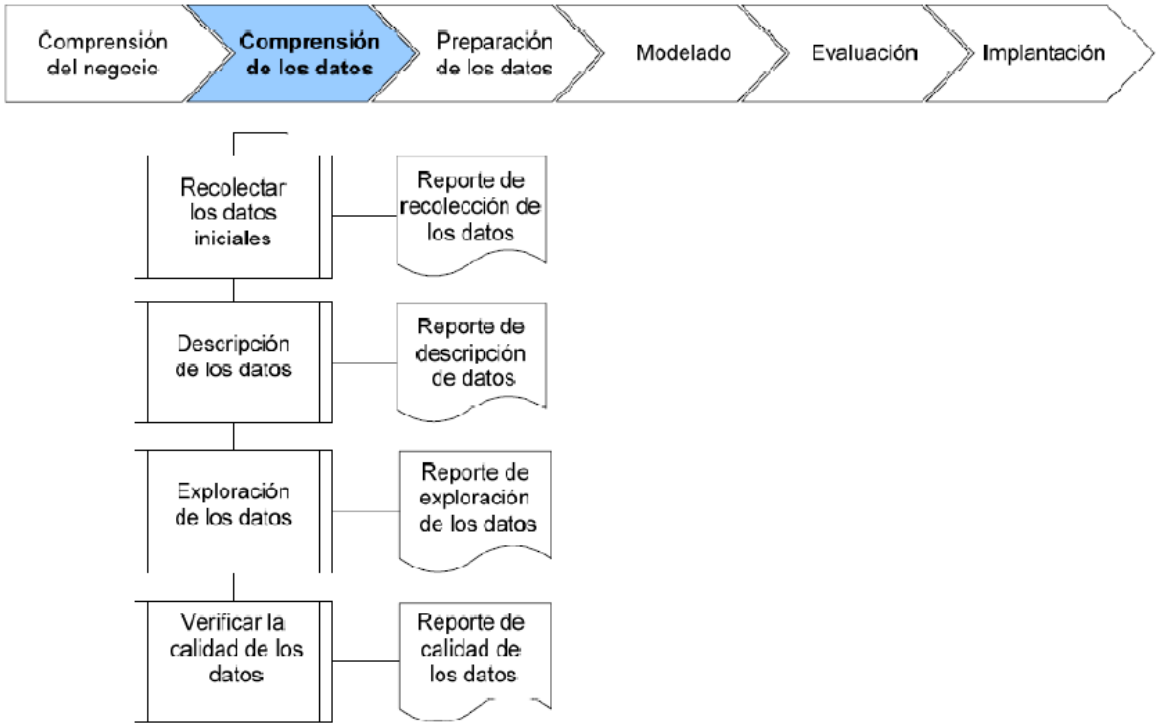


Figura 18. Fase comprensión de los datos
Fuente: (Rodriguez, 2014)

1. Recolectar los datos.

La primera tarea en esta segunda fase del proceso de CRISP DM, es la recolección de los datos iniciales y su adecuación para el futuro procesamiento, como se dice a continuación, esta tarea tiene como objetivo, elaborar informes con una lista de los datos adquiridos, su localización, las técnicas utilizadas en su recolección y los problemas y soluciones inherentes a este proceso. (Oldemar, 2015).

2. Descripción de datos.

Después de adquiridos los datos iniciales, estos deben ser descritos, este proceso involucra establecer volúmenes de datos (número de registros y campos por registro), su identificación, el significado de cada campo y la descripción del formato inicial. (Oldemar, 2015).

3. Exploración de datos.

A continuación, se procede a su exploración, cuyo fin es encontrar una estructura general para los datos, esto involucra la aplicación de pruebas estadísticas básicas, que revelen propiedades en los datos recién adquiridos, se crean tablas de frecuencia y se construyen gráficos de distribución, la salida de esta tarea es un informe de exploración de los datos. (Oldemar, 2015).

4. Verificar la calidad de datos.

En esta tarea, se efectúan verificaciones sobre los datos, para determinar la consistencia de los valores individuales de los campos, la cantidad y distribución de los valores nulos, y para encontrar valores fuera de rango, los cuales pueden constituirse en ruido para el proceso. La idea en este punto, es asegurar la completitud y corrección de los datos. (Oldemar, 2015).

c) Fase preparación de datos

En esta fase y al finalizar la recolección inicial de datos, se procede a su preparación para adaptarlos a las técnicas de Data Mining que se utilicen

posteriormente, tales como técnicas de visualización de datos, de búsqueda de relaciones entre variables u otras medidas para exploración de los datos, la preparación de datos incluye las tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato. (Oldemar, 2015). Las cinco tareas que contiene esta fase se muestran en la Figura 19.

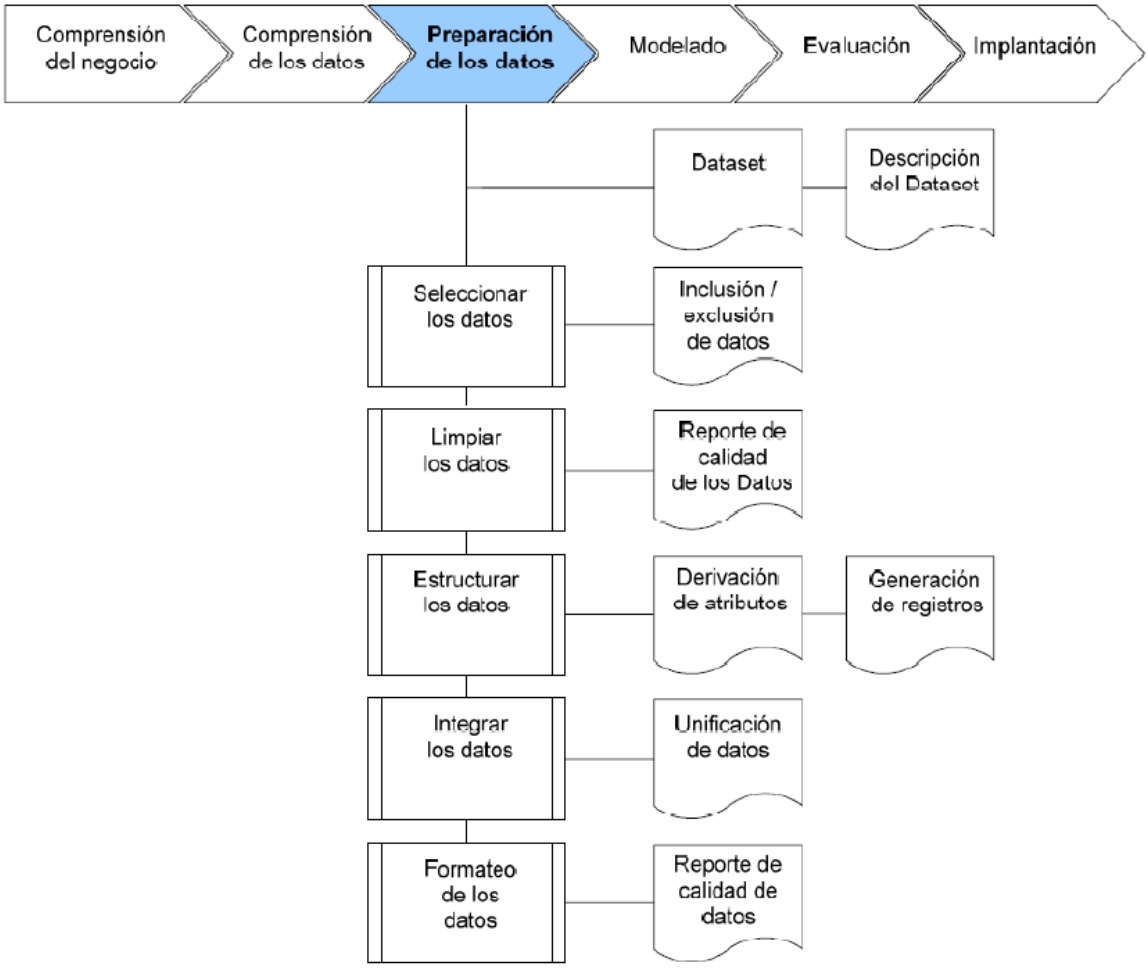


Figura 19. Fase preparación de los datos
Fuente: (Rodriguez, 2014)

1. Seleccionar los datos.

En esta etapa, se selecciona un subconjunto de los datos adquiridos en la fase anterior, apoyándose en criterios previamente establecidos en las fases anteriores: calidad de los datos en cuanto a completitud y corrección de los datos

y limitaciones en el volumen o en los tipos de datos que están relacionadas con las técnicas de DM seleccionadas. (Oldemar, 2015).

2. **Limpiar los datos.**

Esta complementa a la anterior, y es una de las que más esfuerzo y tiempo conlleva debido a que existe una gran diversidad de técnicas orientadas a optimizar la calidad de los datos, tomando en cuenta que queden listos para la fase posterior, entre las técnicas que encontramos podemos mencionar las siguientes: descretización de los campos numéricos, normalización de los datos, tratamiento de valores ausentes o faltantes, reducción de la cantidad de los datos, entre otros. (Oldemar, 2015).

3. **Estructurar los datos.**

Esta fase incluye operaciones de la fase anterior como: generación de nuevos atributos a partir de los ya existentes, integración de los nuevos o transformación de valores para los existentes. (Oldemar, 2015).

4. **Integrar los datos.**

Esta tarea involucra las siguientes sub tareas: generación de nuevos campos a partir de otros existentes, creación de nuevos registros, fusión de tablas campo o nuevas tablas, en las cuales se resumen características de múltiples campos o registros en nuevas tablas resumen. (Oldemar, 2015).

5. Formatear de los datos.

El objetivo de esta tarea es la realización de transformaciones a nivel sintáctico de los datos tomando en cuenta que no se debe modificar su significado, ya que debe permitir la aplicación de cualquier técnica de minería de datos como: ajuste de valores (eliminar tabuladores, comas, caracteres especiales, máximos y mínimos para las cadenas de caracteres, reordenación de campos, etc.). (Oldemar, 2015).

d) Modelado

En esta fase de CRISP-DM, se seleccionan las técnicas de modelado más apropiadas para el proyecto de Data Mining específico, las técnicas a utilizar en esta fase se eligen en función de los siguientes criterios: ser apropiada al problema, disponer de datos adecuados, cumplir los requisitos del problema, tiempo adecuado para obtener un modelo, conocimiento de la técnica, previamente al modelado de los datos, se debe determinar un método de evaluación de los modelos que permita establecer el grado de bondad de ellos. (Oldemar, 2015). Las cuatro tareas que tiene esta fase son los presentados en la Figura 20.

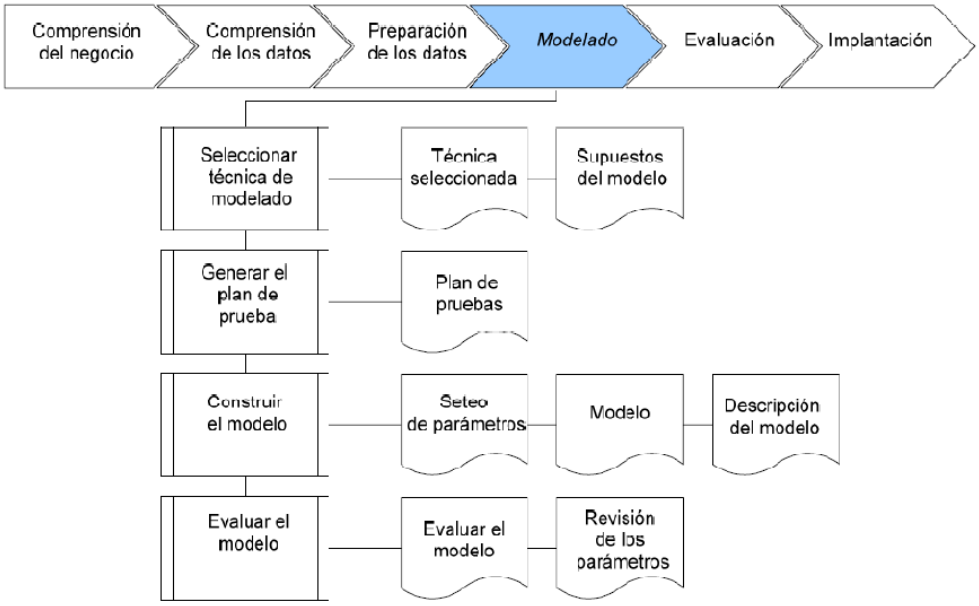


Figura 20. Fase modelado
Fuente: (Rodriguez, 2014)

1. Selección de técnica de modelado.

Esta tarea consiste en la selección de la técnica de DM más apropiada al tipo de problema a resolver, para esta selección, se debe considerar el objetivo principal del proyecto y la relación con las herramientas de DM existentes, por ejemplo, si el problema es de clasificación, se podrá elegir de entre árboles de decisión, k-nearest neighbour o razonamiento basado en casos, si el problema es de predicción, análisis de regresión, redes neuronales; o si el problema es de segmentación, redes neuronales, técnicas de visualización, entre otras. (Oldemar, 2015).

2. **Generación el plan de prueba.**

Una vez construido un modelo, se debe generar un procedimiento destinado a probar la calidad y validez del mismo, por ejemplo, en una tarea supervisada de DM como la clasificación, es común usar la razón de error como medida de la calidad, entonces, típicamente se separan los datos en dos conjuntos, uno de entrenamiento y otro de prueba, para luego construir el modelo basado en el conjunto de entrenamiento y medir la calidad del modelo generado con el conjunto de prueba. (Oldemar, 2015).

3. **Construir modelo.**

Después de seleccionar la técnica, se ejecuta sobre los datos previamente preparados para generar uno o más modelos, todas las técnicas de modelado tienen un conjunto de parámetros que determinan las características del modelo a generar, la selección de los mejores parámetros es un proceso iterativo y se basa exclusivamente en los resultados generados, estos deben ser interpretados y su rendimiento justificado. (Oldemar, 2015).

4. **Evaluar modelo.**

En esta tarea, los ingenieros de DM interpretan los modelos de acuerdo al conocimiento preexistente del dominio y los criterios de éxito preestablecidos, expertos en el dominio del problema juzgan los modelos dentro del contexto del dominio y expertos en Data Mining aplican sus propios criterios (seguridad del conjunto de prueba, pérdida o ganancia de tablas, entre otras). (Oldemar, 2015)..

e) Evaluación

En esta fase se va a evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito establecidos, debe considerarse además, que la fiabilidad calculada para el modelo se aplica solamente para los datos sobre los que se realizó el análisis, es preciso examinar el proceso, teniendo en cuenta los resultados adquiridos, para poder repetir algún paso anterior, en el que se haya eventualmente cometido algún error, considerar que se pueden utilizar múltiples herramientas para la definición de los resultados. (Oldemar, 2015). Las tres tareas con las que cuenta se muestran en la Figura 21.

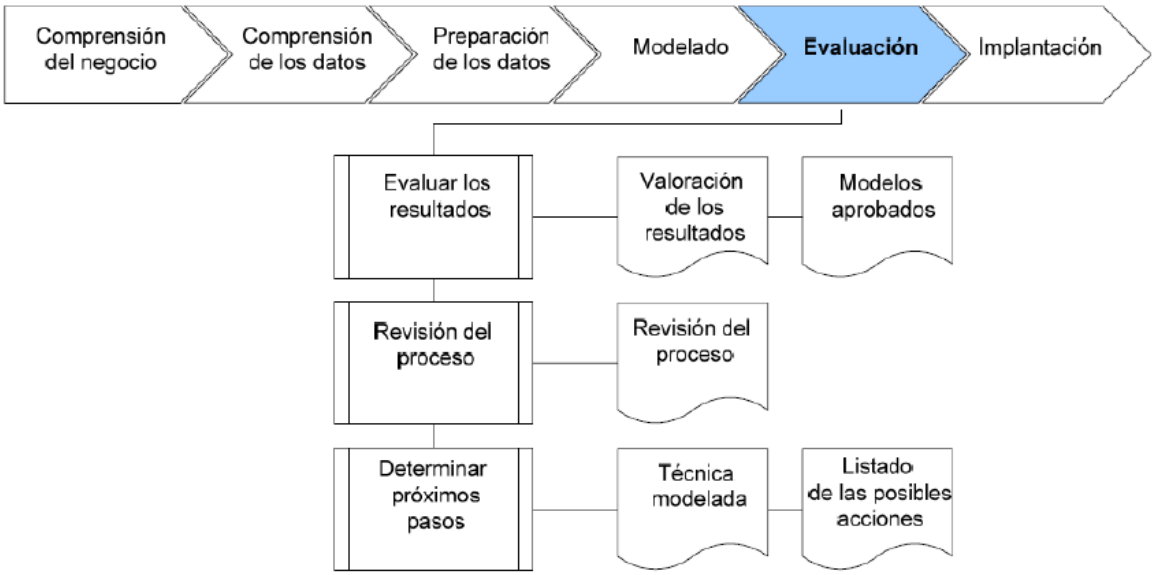


Figura 21. Fase Evaluación
Fuente: (Rodriguez, 2014)

1. Evaluación de resultados.

En los pasos de evaluación anteriores, se trataron factores tales como la precisión y generalización del modelo generado, esta tarea involucra la

evaluación del modelo en proporción a los objetivos del negocio y se busca determinar si hay alguna razón de negocio para la cual, el modelo sea deficiente, o si es aconsejable probar el modelo, en un problema real con datos reales si el tiempo y restricciones lo permiten. (Oldemar, 2015).

2. **Revisión el proceso.**

En esta parte del proceso de revisión, se refiere a calificar al proceso entero de DM, a objeto de identificar elementos que pudieran ser mejorados o pudieran tener algún problema. (Oldemar, 2015).

3. **Determinación de próximos pasos.**

Si se ha definitivo que las fases hasta este momento han generado resultados satisfactorios, podría continuar a la fase siguiente, en caso contrario podría decidirse por otra iteración desde la fase de preparación de datos o de modelación con otros parámetros, podría ser incluso que en esta fase se concluya partir desde cero con un nuevo proyecto de DM. (Oldemar, 2015).

f) Implantación

El objetivo de esta tarea es la realización de transformaciones a nivel sintáctico de los datos tomando en cuenta que no se debe modificar su significado, ya que debe permitir la aplicación de cualquier técnica de minería de datos como: ajuste

de valores (eliminar tabuladores, comas, caracteres especiales, máximos y mínimos para las cadenas de caracteres, reordenación de campos, etc.). (Oldemar, 2015). Las tareas con la que cuenta esta fase se muestran en la Figura 22.

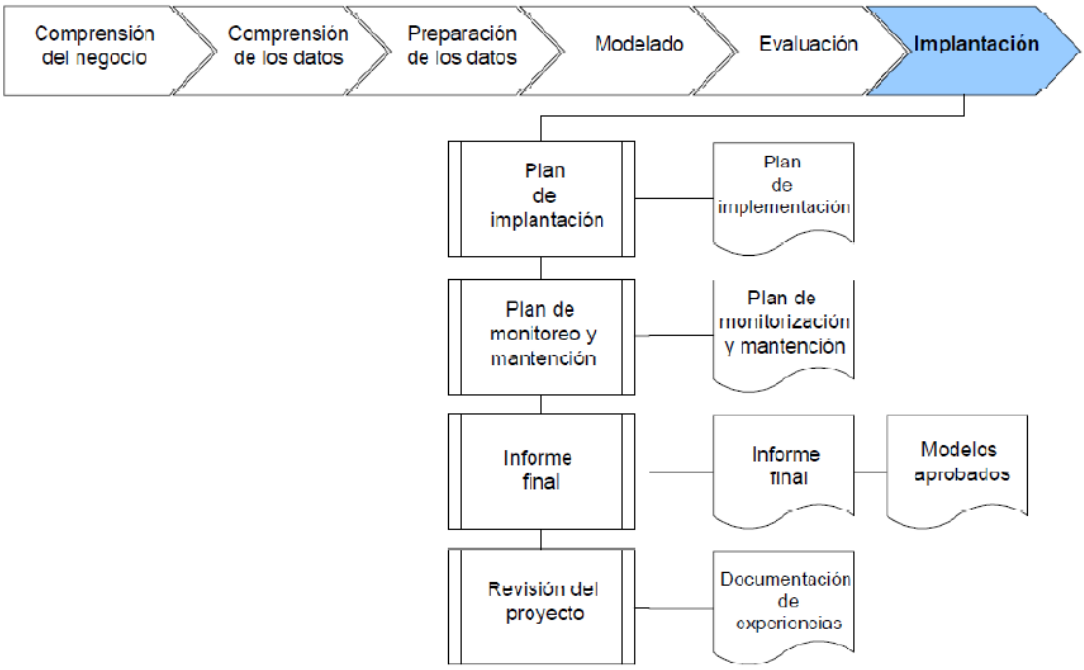


Figura 22. Fase Implantación
Fuente: (Rodriguez, 2014)

1. **Plan de la implementación.**

En esta tarea se toma los resultados de la evaluación realizada y se procede a elaborar una estrategia para su implementación en el caso de que se haya planeado un modo general para crear el modelo, este debe ser

documentado para su aplicación y tomar las medidas para uso por terceros. (Oldemar, 2015).

2. **Planeación de la supervisión y mantenimiento.**

En esta tarea es aconsejable preparar las diferentes estrategias de monitorización y mantenimiento para su posterior aplicación en los modelos en este punto la retroalimentación generada por esta tarea puede mostrar si dicho modelo se lo está utilizando adecuadamente. (Oldemar, 2015)..

3. **Informe final y revisión de proyecto.**

Estas son las últimas tareas de esta metodología, por ende dependiendo del plan de implementación elaborado y aplicado, este solo debe contener un resumen de aquellos puntos importantes del proyecto, también se debe considerar realizar una presentación final que muestre toda la experiencia lograda o indique los resultados logrados, finalmente, en la revisión del proyecto se lleva a cabo una evaluación total para encontrar que fue lo correcto e incorrecto que se realizó, y también determinar lo que se requiere mejorar. (Oldemar, 2015).

CAPÍTULO 4

DESARROLLO

4.1. Arquitectura de minería de datos stream

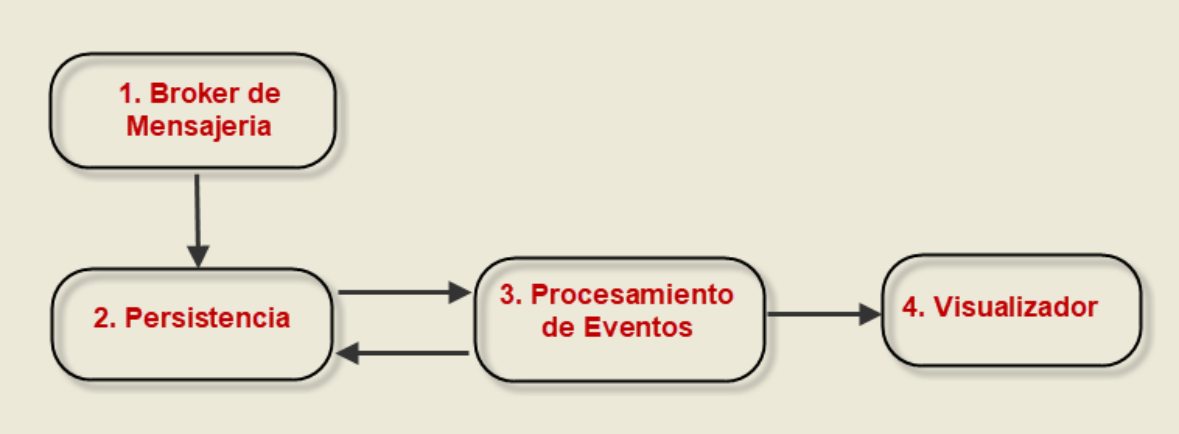


Figura 23. Arquitectura de minería datos stream

Bróker de mensajería: Utilizado para centralizar la recepción de información sobre los eventos que se produzcan, Ejemplo Apache Kafka.

Persistencia: Mantener los datos para un análisis histórico posterior después de recibirlos y analizarlos en streams. Herramienta: PostgreSQL

Procesamiento de Eventos: Sistema de procesamiento de eventos en stream, capaz de definir los “caminos” y “transformaciones” que sufren los eventos para poder extraer datos de interés para la organización. Herramienta: Orange

Visualizador: Se convierte en una herramienta poderosa para el Sistema de análisis e interpretación de datos grandes y complejos, volviéndose un medio eficiente en la transmisión de conceptos en un formato universal. Herramienta: Orange.

El bróker de mensajería captura los datos streams provenientes de los sensores y los envía a una base de datos, el procesador de eventos captura los datos registrados en la base de datos y los analiza cada 0.001 seg., con las técnicas de minería de datos seleccionados para luego ser presentados en dashboards interactivos.

En este proyecto los datos ya se encuentran almacenados en la base de datos y por consiguiente se va a proceder a analizarlos en forma de streams.

Diseño del procesamiento de datos con Orange

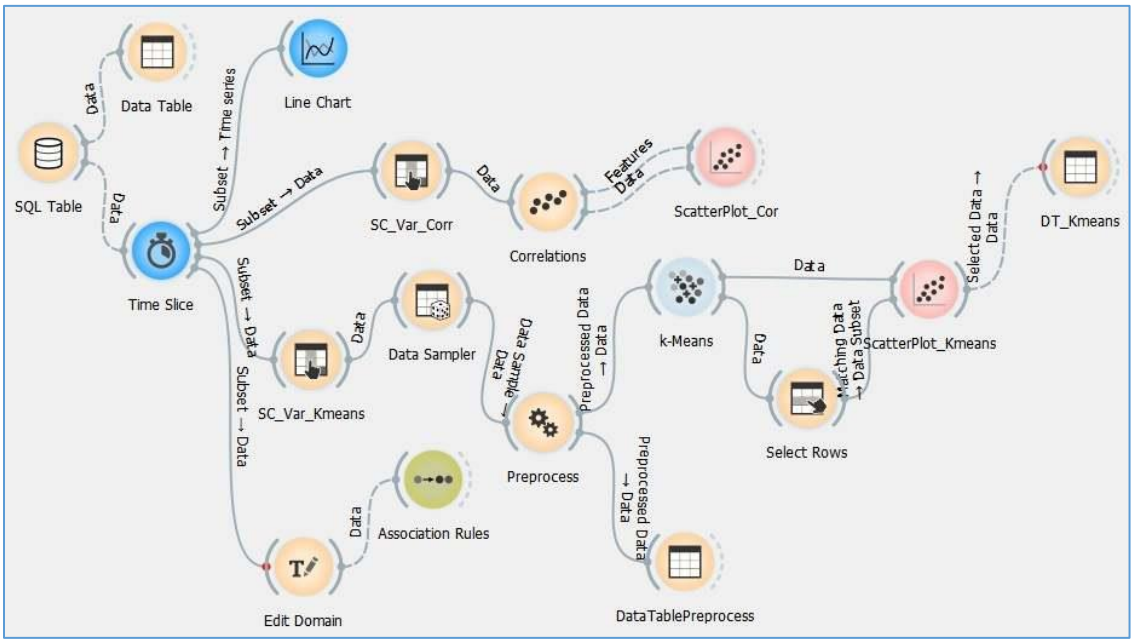


Figura 24. Diseño en Orange

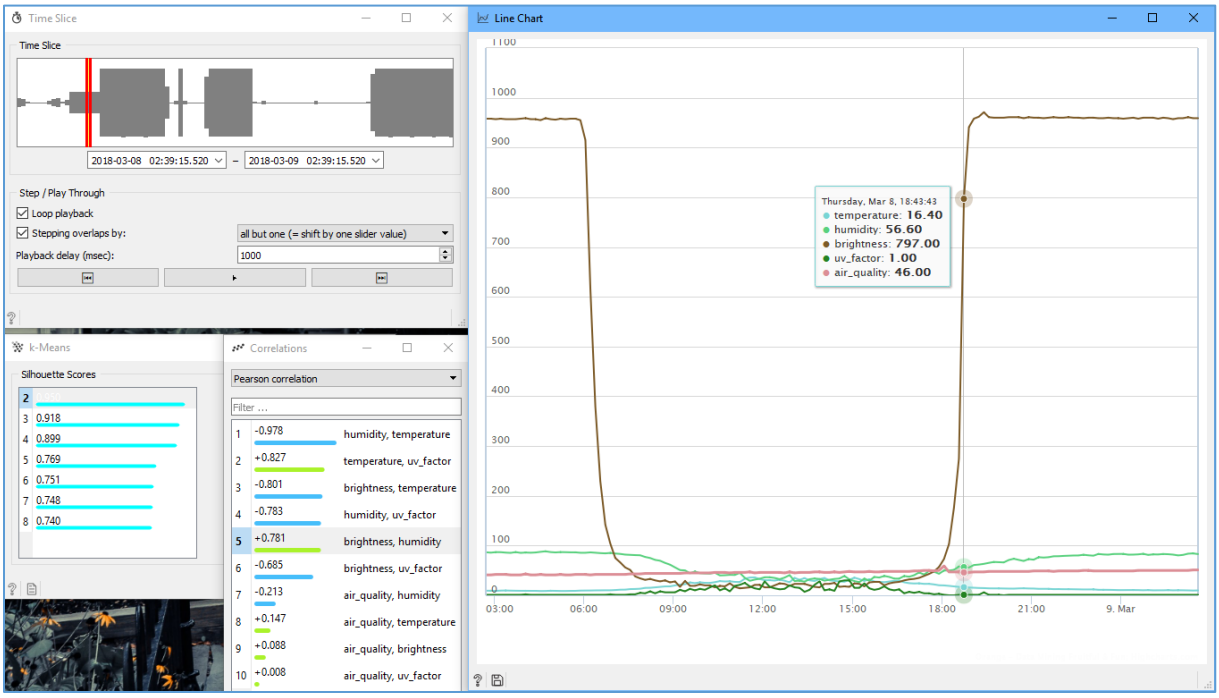


Figura 25. Procesamiento de datos stream

4.2. Comprensión del negocio

La comprensión del negocio es la primera fase de la metodología de minería de datos CRISP-DM, la cual se desarrolla realizando una serie de tareas esenciales para definir el marco en que se elaborará la minería estableciendo metas y objetivos que le permitan a la empresa trascender, dichas tareas se describen a continuación.

4.2.1. Objetivos de negocio

Desde el enfoque de la minería de datos tenemos como criterio de éxito la probabilidad de identificar patrones de comportamiento que ayuden a mejorar el cultivo

de las rosas bajo invernadero, y como segundo criterio de éxito del negocio se tiene la probabilidad de elaborar una propuesta de mejoramiento para dichas instalaciones.

4.2.2. Situación Actual

Los recursos disponibles con los que se dispone para el desarrollo del proyecto puede sub dividirse en los siguientes que se detallan a continuación:

Personal: en este punto podemos decir que el proyecto es viable operativamente debido a que se encuentra conformado por dos alumnos, los señores Jami Fernández Jhonny Alonso y Machángara Quilca Kléver Geovany, mismos que presentan las capacidades y destrezas necesarias para llevar a cabo esta investigación.

Datos: El Campus IASA I de la Universidad de las Fuerzas Armadas – ESPE y en específico el área de Floricultura a cargo de la Dra. Elizabeth Urbano dispone de un invernadero que ya tiene implementada una red WSN (Wireless Sensor Network) la cual genera datos streams de: temperatura, humedad relativa, luminosidad, factor UV, calidad de aire, mismos que han sido almacenados en un archivo “datos_sensores.csv” para su tratamiento y que son necesarios para poder cumplir con los objetivos planteados de esta investigación.

Recursos Hardware y Software: Este proyecto es viable técnicamente debido a que los recursos de TI para el desarrollo de esta investigación están a disposición del grupo de proyecto, tal como se puede evidenciar en la siguiente Tabla 4:

Tabla 4
Factibilidad Técnica

	Herramienta	Licencia
SOFTWARE	ORANGE 3.18.0	GPL
	PostgreSQL	GPL
	Apache Kafka	GPL
HARDWARE	Lenovo core i7	NA
	Sony VAIO core i3 VPCEA45FL	NA

- Recursos, supuestos y restricciones

  Requisitos

  - Disponer de la autorización emitida por la Dra. Elizabeth Urbano encargada del invernadero en el IASA I.
  - Disponer de una amplia información para la aplicación de este tipo de minería de datos.
  - Contar con las tutorías de expertos en el área para llevar a cabo un correcto tratamiento de la información.

  Restricciones

- Los datos para el desarrollo de esta investigación se limitan a la información registrada y almacenada dentro del invernadero por la red WSN (Wireless Sensor Network).

4.2.3. Objetivos de la Minería de Datos

Los objetivos en términos de minería de datos son:

- Refinar, normalizar o discretizar los datos recopilados y prepararlos para el análisis.
- Identificar la existencia de relaciones, correlaciones entre los factores abióticos a investigar.
- Identificar mediante agrupamiento (clustering) las condiciones con las que el invernadero funcionaba.
- Presentar mediante dashboards los resultados obtenidos.

4.2.4. Generación del proyecto

Ver Anexo 1: Cronograma del proyecto.

4.3. Comprensión de los datos

En esta fase el objetivo principal es recolectar los datos, identificar la calidad de los mismos con los cuales se podrán establecer las primeras hipótesis que serán validadas posteriormente con el análisis, cabe recalcar que esto nos ayudara a evitar problemas en la fase de preparación de los datos.

4.3.1. Recolectar los Datos

Los datos han sido recolectados a través de una WSN (Wireless Sensor Network) ya implementada dentro del invernadero por la Dra. Elizabeth Urbano encargada de la floricultura en el IASA I e Ingeniero Paúl Díaz Director del Departamento de Ciencias de la Computación.

1. La recolección de datos se realizó a través de Apache Kafka, mismos que posteriormente fueron almacenados en una bases datos.
2. Toda esta información que se obtuvo fue exportada a un archivo de formato csv.
3. Este archivo csv contiene los datos capturados por el: Sensor Temperatura y Humedad, Sensor de Luminosidad, Sensor de Índice Ultravioleta, Sensor de Calidad de Aire.
4. Los datos obtenidos pertenecen al período Febrero 2018 – Junio 2018.

4.3.2. Descripción de los datos

En la Tabla 5 que se presenta a continuación se detallan los atributos para la tabla Datos_Sensores.

Tabla 5
Variables de la tabla Datos_Sensores

TABLA	**DESCRIPCIÓN**	**N° REGISTROS**
Datos Sensores	Esta tabla contiene los datos capturados por los sensores mencionados anteriormente.	25962
CAMPO	**DESCRIPCIÓN**	
Temperatura	Rango de medición de temperatura: -40°C a 80 °C Precisión de medición de temperatura: <±0.5 °C	
Humedad	Rango de medición de humedad: De 0 a 100% RH Precisión de medición de humedad: 2% RH	
Luminosidad	Rango de medición de luminosidad: 0 “muy claro” a 1024 “muy obscuro”	
Factor UV	Rango de medición índice Factor UV: 0 a 10.	
Calidad Aire	Detección de partes por millón: 10ppm~1000ppm. Concentración detectable: Amoniaco, sulfuro, benceno, humo.	

Nota: los datos capturados en los campos de luminosidad, Factor UV y calidad de aire presentan valores muy generales, es decir, no son los más adecuados o están en las

unidades de medida requeridas por el experto de la agrónoma, pero que desde otro contexto o perspectiva resultan ser muy útiles para esta investigación.

4.3.3. Exploración de los Datos.

A continuación se realizó el análisis estadístico de los datos descritos en el apartado anterior para revelar las características tales como: fecha y períodos de tiempo de captura de los datos, rangos máximos y mínimos, etc.

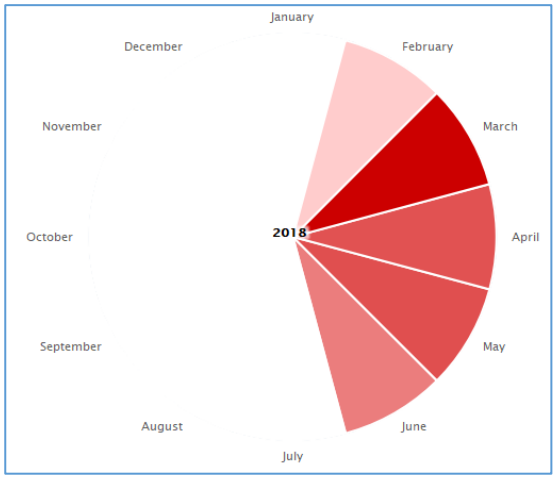


Figura 26. Porcentaje de datos por Mes

En la Figura se observa que los datos pertenecen al periodo Febrero 2018 – Junio 2018, de los cuales el 0,82 % corresponden al mes de Febrero, el 37.29 % al mes de Marzo, el 22,80 % al mes de Abril, el 23,09 % al mes de Mayo y el 15,20 % al mes de Junio respectivamente.

4.3.4. Verificación de los datos

Para realización de esta tarea se utilizaron consultas SQL para verificar los campos que toman parte en los análisis, mismos en los que no se encontraron datos nulos o basura en los campos de temperatura, humedad relativa, luminosidad, factor UV y calidad del aire.

4.4. Preparación de los datos

4.4.1. Selección de los datos

En este punto y para cumplir con los objetivos propuestos con respecto al descubrimiento de conocimiento en datos stream, se propuso seleccionar aquellos registros completos de un día, una semana y un mes por medio del campo datetime de la base de datos.

4.4.2. Limpieza de los datos

La limpieza de datos no fue necesaria debido a que en el apartado “Verificación de los datos” se evidencio que no existen datos nulos o basura en los valores de cada uno de los campos.

- Transformación de Datos: Este método se utilizó para normalizar la base de datos debido a que si las variables de aglomeración están en escalas muy diferentes será necesario estandarizar previamente estas o trabajar con desviaciones con respecto a la media.

4.4.3. Estructura de los datos e Integración de los datos.

Para este caso no fue necesario que se aplique este tipo de medidas debido a que Apache Kafka unificaba los datos generados por los sensores en un solo registro y los almacenaba en una base de datos

4.5. Modelado

a) Escoger la técnica de modelado

Se utilizó la herramienta Orange para generar los modelos de minería de datos, para lo cual se usaron las técnicas de minería de datos no supervisados, que ofrece esta herramienta, las cuales van acorde a los objetivos de la minería de datos, porque los datos que se tienen están sin etiquetar ni ordenar de ninguna manera y el objetivo de estas técnicas no es predecir nuevos datos sino describir los existentes.

De las técnicas no supervisadas se utilizaron Clustering, Reglas de asociación y correlaciones porque son técnicas descriptivas que generan conocimiento a partir de no conocer el resultado de las características presentadas en los registros.

b) Generar el plan de prueba

Para verificar si la descripción presentada por parte de los métodos de minería no supervisados es la más correcta para los datos analizados, se los realizara

por medio de la aplicación Orange la cual nos provee de varios valores para seleccionar la solución más adecuada.

Validación de reglas de asociación

Para la generación de Reglas de asociación a continuación se muestran todas las medidas que se especifican en la generación de este método junto con los algoritmos que la calculan debido a que dependiendo del algoritmo que se utilice para las reglas de asociación, las medidas son distintas.

- **Soporte**: Calculada y presentada por el Algoritmo Apriori.
- **Confianza**: Calculada y presentada por el Algoritmo Apriori.
- **Medida Predictiva**: Calculada y presentada por el Algoritmo PredictiveApriori.
- **Mejora (Lift)**: Calculada y presentada por el Algoritmo Apriori.

En el procedimiento de selección de las Reglas de asociación para este proyecto no solo incluye las medidas más relevantes de confianza de cada regla (como son el soporte y la confianza) sino unos criterios que hacen que las reglas sean de utilidad para analizar los resultados obtenidos.

Cada regla de asociación fue revisada con el fin de no encontrar los siguientes problemas:

- **Reglas Redundantes:** Son reglas que contienen 2 o más atributos antecedentes similares o que tienden a una deducción lógica simple. Por

ejemplo una regla que tenga atributos antecedentes como: “Horas” y “Minutos”.

- **Reglas Innecesarias:** Son las reglas que no cumplen el objetivo de este proyecto ya que contienen consecuentes diferentes, por ejemplo dos variables con correlación negativa fuerte Temperatura Humedad.
- **Medida de Mejora (lift):** Si el resultado es mayor e igual a uno, la regla sirve, de lo contrario la regla se descarta. Esta medida es exclusiva para el algoritmo A priori.
- **Reglas con una confianza baja:** Son reglas que contienen la medida probabilística de confianza por debajo del 0.65.

Validación de Clustering

Para validar el Clustering, existen algunos tipos de validación. La validación externa y la validación interna son las dos categorías más importantes para la validación de clustering. La principal diferencia es si se usa o no información externa para la validación, es decir, información que no es producto de la técnica de agrupación utilizada, por ejemplo saber que se establecen dos condiciones ambientales en las plantaciones, día y noche.

A diferencia de técnicas de validación externas, las de validación interna miden el clustering únicamente basadas en información de los datos. Evalúan que tan buena es la estructura del clustering sin necesidad de información ajena al propio algoritmo y su resultado.

Como la validación externa mide la calidad del agrupamiento conociendo información externa de antemano, es principalmente usada para escoger un algoritmo de clustering óptimo sobre un data set específico.

- Las métricas de validación interna pueden usarse para escoger el mejor algoritmo de clustering, así como el número de clúster óptimo sin ningún tipo de información adicional.
- En la práctica, la información externa, como los labels de las clases, por lo general no se encuentra disponible en muchos escenarios de aplicación.

Como el objetivo del clustering es agrupar objetos similares en el mismo clúster y objetos diferentes ubicarlos en diferentes clúster, las métricas de validación interna están basadas usualmente en los dos siguientes criterios:

- **Cohesión**: El miembro de cada clúster debe ser lo más cercano posible a los otros miembros del mismo clúster.
- **Separación**: Los clúster deben estar ampliamente separados entre ellos. Existen varios enfoques para medir esta distancia entre clúster: distancia entre el miembro más cercano, distancia entre los miembros más distantes o la distancia entre los centroides. (León Guzmán, 2015)

Validación de correlaciones

Para validar las Correlaciones Pearson, se debe examinar la fuerza y la dirección de la relación lineal entre dos variables continuas.

Fuerza

El valor del coeficiente de correlación puede variar de −1 a +1. Mientras mayor sea el valor absoluto del coeficiente, más fuerte será la relación entre las variables.

Para la correlación de Pearson, un valor absoluto de 1 indica una relación lineal perfecta. Una correlación cercana a 0 indica que no existe relación lineal entre las variables.

Dirección

El signo del coeficiente indica la dirección de la relación. Si ambas variables tienden a aumentar o disminuir a la vez, el coeficiente es positivo y la línea que representa la correlación forma una pendiente hacia arriba. Si una variable tiende a incrementarse mientras la otra disminuye, el coeficiente es negativo y la línea que representa la correlación forma una pendiente hacia abajo. (Minitab, Interpretación de los resultados clave, 2012).

Las siguientes gráficas muestran datos con valores específicos del coeficiente de correlación para ilustrar diferentes patrones en la fuerza y la dirección de las relaciones entre las variables.

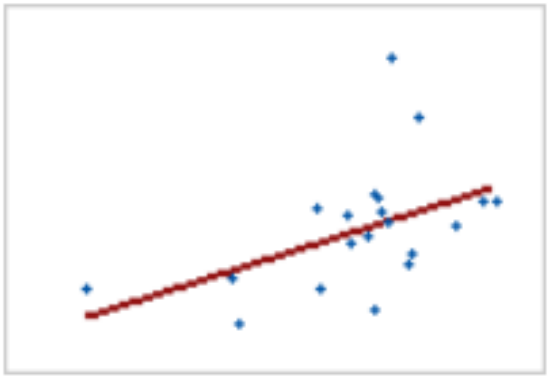


Figura 27. Ninguna relación: Pearson r = 0
Fuente: (Minitab, Interpretación de resultados para correlación, 2014)

Los puntos se ubican de forma aleatoria en la gráfica, lo que significa que no existe relación lineal entre las variables.

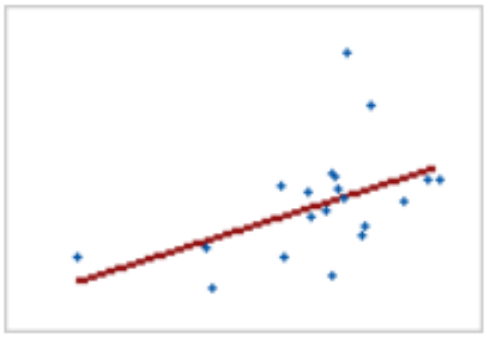


Figura 28. Relación positiva moderada: Pearson r = 0.476
Fuente: (Minitab, Interpretación de resultados para correlación, 2014)

Algunos puntos están cerca de la línea, pero otros puntos están lejos de ella, lo que indica que solo existe una relación lineal moderada entre las variables.

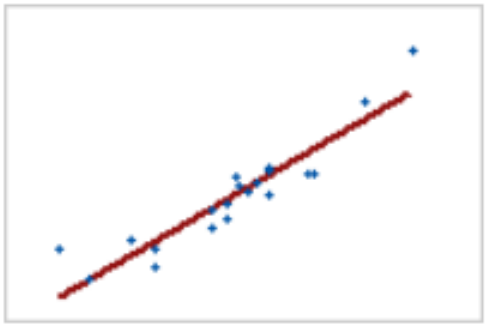


Figura 29. Relación positiva grande: Pearson r = 0.93
Fuente: (Minitab, Interpretación de resultados para correlación, 2014)

Los puntos se ubican cerca de la línea, lo que indica que existe una relación lineal fuerte entre las variables. La relación es positiva porque a medida que una variable aumenta, la otra variable también aumenta. (Minitab, Interpretación de los resultados clave, 2012)..

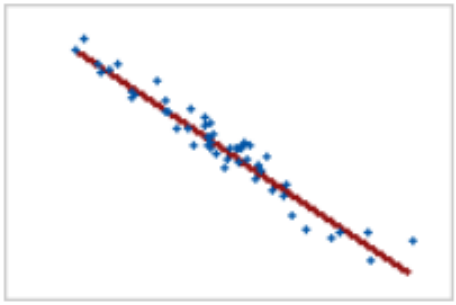


Figura 30. Relación negativa grande: Pearson r = −0.96
Fuente: (Minitab, Interpretación de resultados para correlación, 2014)

Los puntos se ubican cerca de la línea, lo que indica que existe una relación negativa fuerte entre las variables. La relación es negativa porque a medida que una variable aumenta, la otra variable disminuye.

Considere los siguientes puntos cuando interprete el coeficiente de correlación:

- Nunca se debe concluir que los cambios en una variable causan cambios en otra basándose solamente en la correlación. Solo los experimentos controlados adecuadamente permiten determinar si una relación es causal.
- El coeficiente de correlación de Pearson es muy sensible a valores de datos extremos. Un solo valor que sea muy diferente de los otros valores en un conjunto de datos puede cambiar considerablemente el valor del coeficiente. Usted debe tratar de identificar la causa de cualquier valor extremo. Corrija cualquier error de entrada de datos o de medición. Considere eliminar los valores de datos que estén asociados con eventos anormales y únicos (causas especiales). Luego, repita el análisis.
- Un coeficiente de correlación de Pearson bajo no significa que no exista relación entre las variables. Las variables pueden tener una relación no lineal. Para verificar gráficamente relaciones no lineales, cree una Gráfica de dispersión o utilice Gráfica de línea ajustada. (Minitab, Interpretación de los resultados clave, 2012).

c) Construir el modelo

Diseño en Orange para evaluar Kmeans

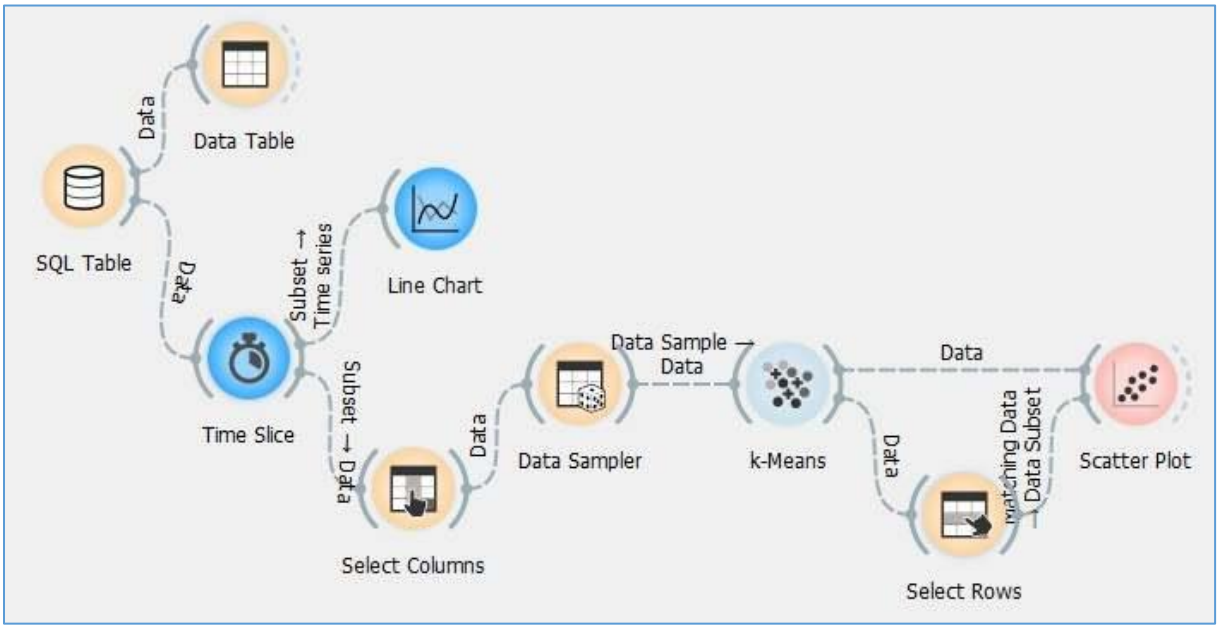


Figura 31. Diseño para evaluar Kmeans

Diseño en Orange para evaluar Correlaciones

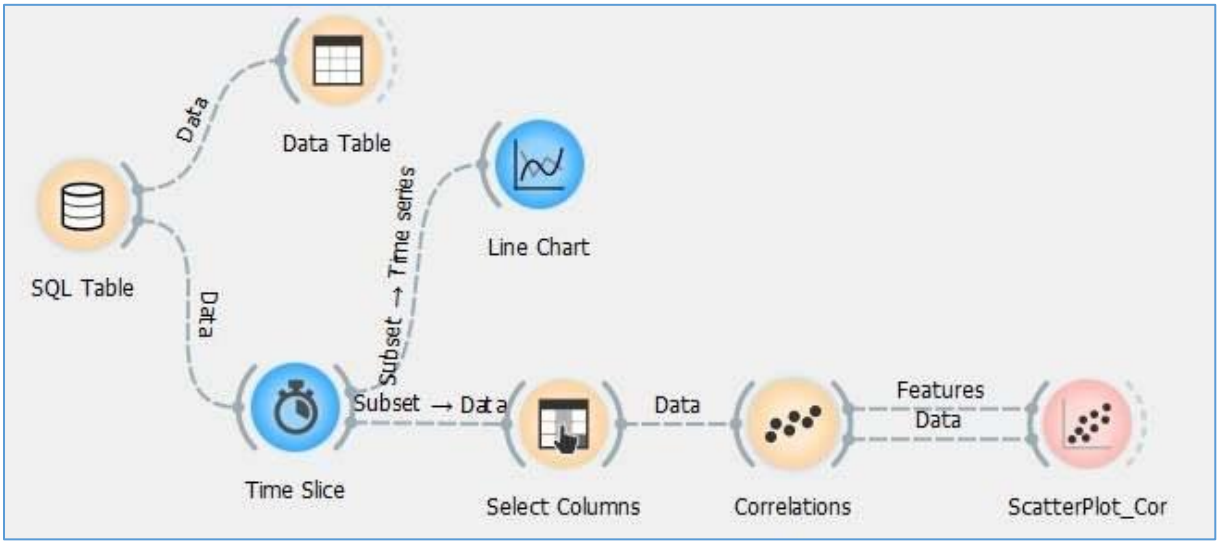


Figura 32. Diseño para evaluar Correlaciones

Diseño en Orange para evaluar Asociaciones

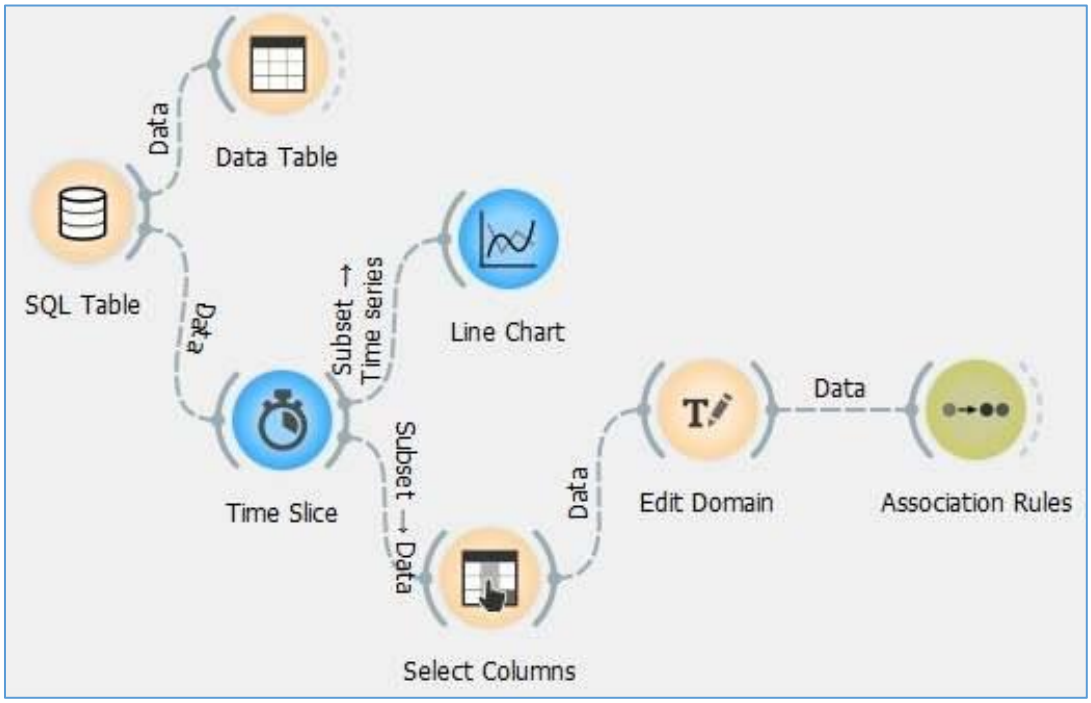


Figura 33. Diseño para evaluar Asociaciones

d) Evaluar el modelo

Se procesa los datos para ser evaluados por las técnicas de minería de datos

Datos evaluados por Kmeans en un día.

Como se puede observar la herramienta Orange nos provee de cálculos estadísticos indicándonos cuál sería el mejor clúster para asignar a K y observar resultados relevantes, como se puede ver la siguiente figura, estos puntajes varían de -1 a 1 donde un valor alto indica que la configuración del agrupamiento es apropiada, la

silueta (Silhouette Scorres) contrasta la distancia promedio a los elementos del mismo con la distancia promedio a los elementos de otros grupos.

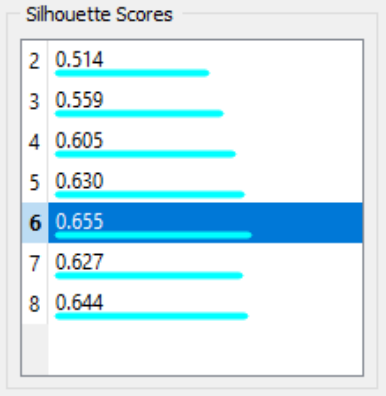


Figura 34. Kmeans por día puntajes para K

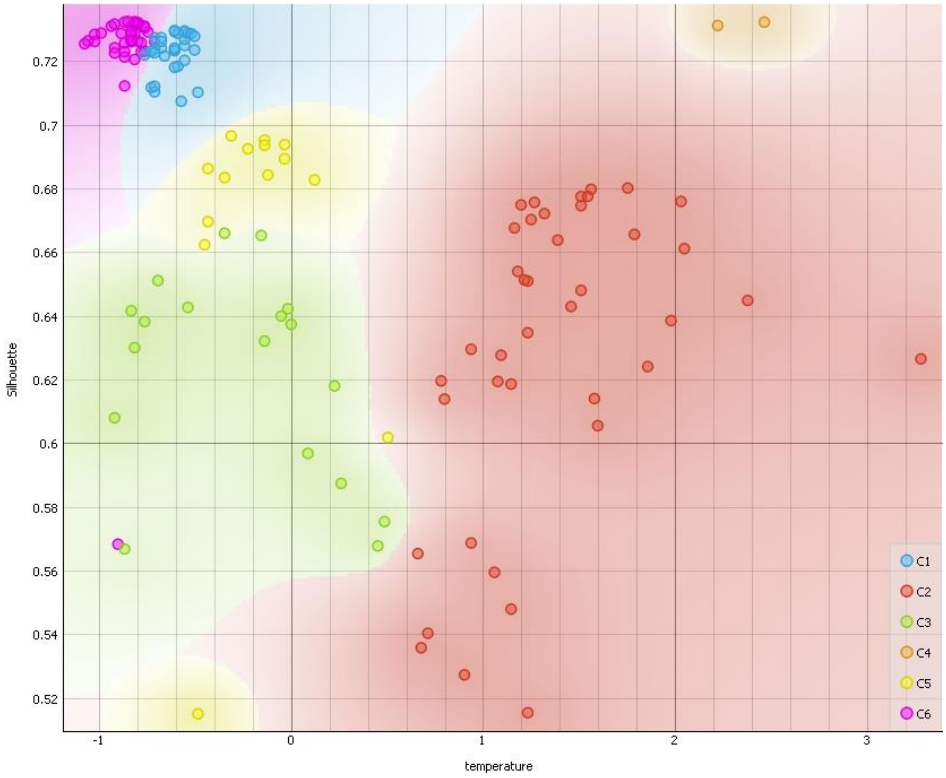


Figura 35. Keas con K=6

Datos evaluados por Kmeans en una semana

Se puede observar que el mejor valor de agrupamiento para la semana es 3 clúster con un puntaje 0.618.

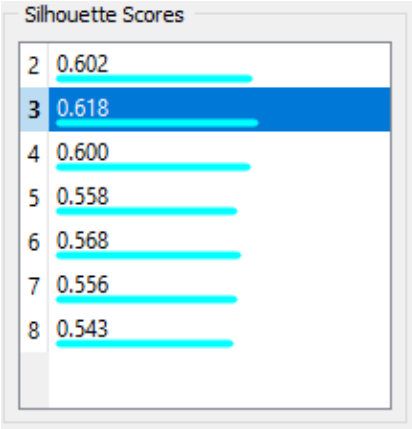


Figura 36. *Kmeans por Semana*

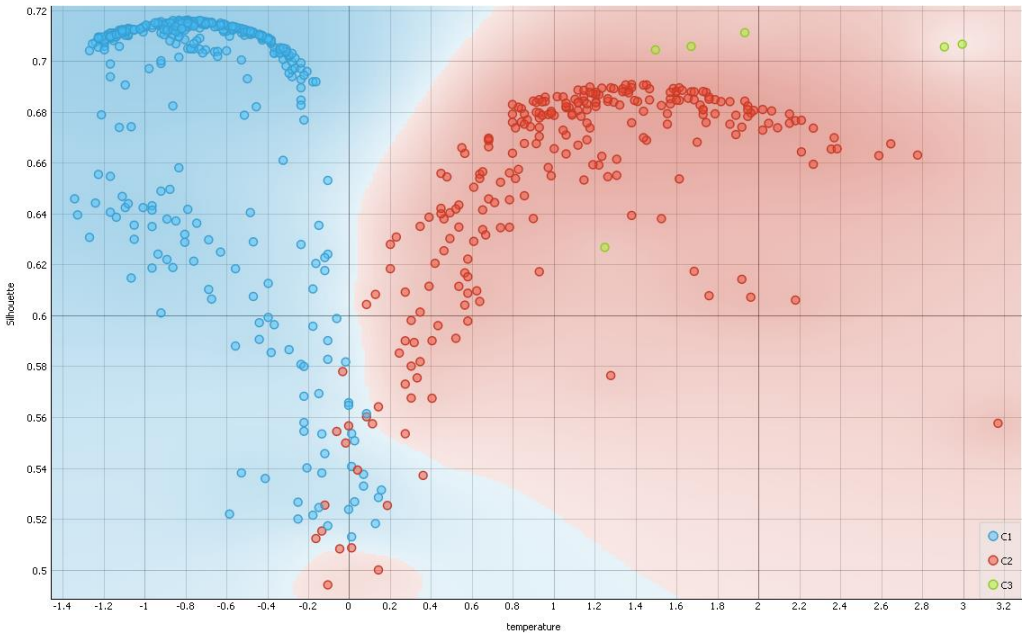


Figura 37. Kmeans por Semana K=3

Datos evaluados por Kmeans en un mes

Se puede observar que el mejor valor de agrupamiento para el mes es 3 clúster con un puntaje 0.576.

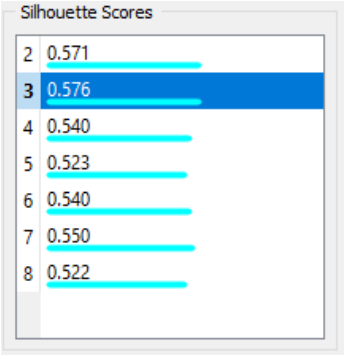


Figura 38. *Kmeans por Mes*

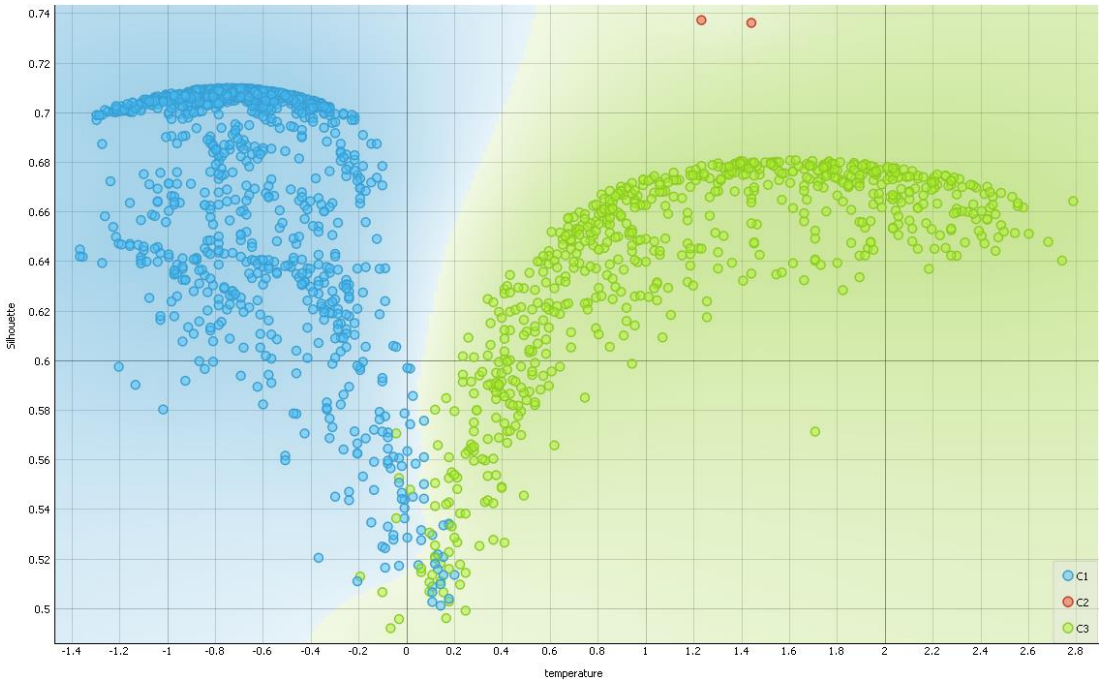


Figura 39. Kmeans por Mes K=3

Técnica de Correlación.

Valores con correlación.

A continuación se puede observar los coeficientes de correlación entre las variables continuas humedad, temperatura, luminosidad, Factor UV, calidad de aire, de los cuales se consideró los coeficientes más altos (> + o - 0.7) en relación en base a lo explicado en la sección anterior.

Correlaciones de un día.

Con respecto al análisis correlacional efectuado para un día se puede observar la existencia de una relación negativa fuerte entre Humedad y Temperatura con un coeficiente Pearson r=-0.973, para Luminosidad y Humedad la relación es positiva fuerte con un coeficiente Pearson r=+0.809, para Temperatura y Factor UV la relación es positiva fuerte con un coeficiente Pearson r=+0.798, para Luminosidad y Temperatura la relación es negativa fuerte con un coeficiente Pearson r=-0.756 y del mismo modo para las variables Humedad y Factor UV la relación es negativa fuerte con un coeficiente Pearson r=-0.744.

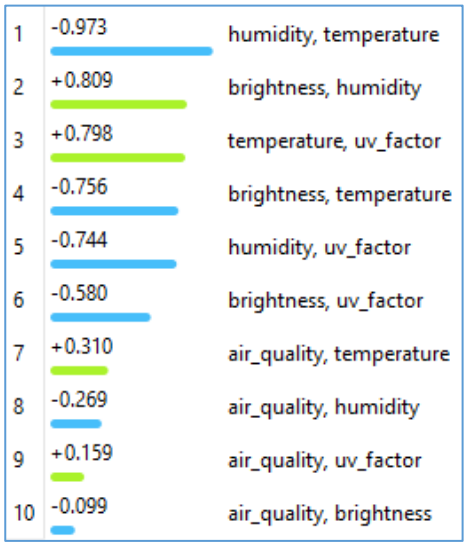
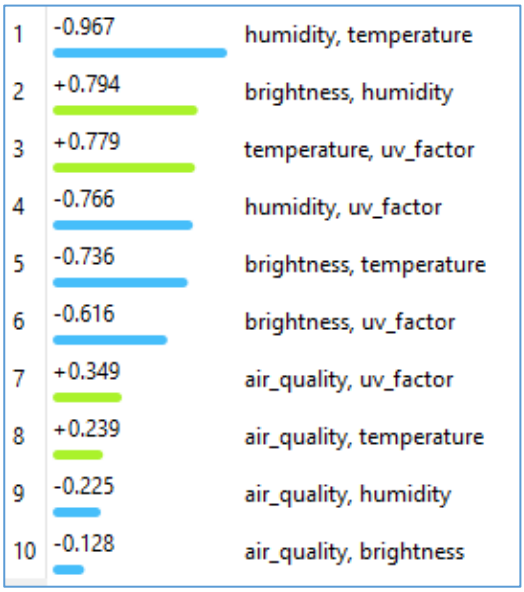


Figura 40. Correlación de un día

Correlaciones de un semana.

Con respecto al análisis correlacional efectuado para una semana se puede observar la existencia de una relación negativa fuerte entre Humedad y Temperatura con un coeficiente Pearson r=-0.967, para Luminosidad y Humedad la relación es positiva fuerte con un coeficiente Pearson r=+0.794, para Temperatura y Factor UV la relación es positiva fuerte con un coeficiente Pearson r=+0.779, para Luminosidad y Temperatura la relación es negativa fuerte con un coeficiente Pearson r=-0.766 y del mismo modo para las variables Humedad y Factor UV la relación es negativa fuerte con un coeficiente Pearson r=-0.736.



1	-0.967	humidity, temperature
2	+0.794	brightness, humidity
3	+0.779	temperature, uv_factor
4	-0.766	humidity, uv_factor
5	-0.736	brightness, temperature
6	-0.616	brightness, uv_factor
7	+0.349	air_quality, uv_factor
8	+0.239	air_quality, temperature
9	-0.225	air_quality, humidity
10	-0.128	air_quality, brightness

Figura 41. Correlaciones de una semana

Correlaciones de un mes.

Con respecto al análisis correlacional efectuado para un mes se puede observar la existencia de una relación negativa fuerte entre Humedad y Temperatura con un coeficiente Pearson r=-0.942, para Luminosidad y Humedad la relación es positiva fuerte con un coeficiente Pearson r=+0.881, para Temperatura y Factor UV la relación es positiva fuerte con un coeficiente Pearson r=+0.825, para Luminosidad y Temperatura la relación es negativa fuerte con un coeficiente Pearson r=-0.756 y del mismo modo para las variables Humedad y Factor UV la relación es negativa fuerte con un coeficiente Pearson r=-0.712.

1	-0.942	humidity, temperature
2	+0.881	temperature, uv_factor
3	-0.825	humidity, uv_factor
4	+0.756	brightness, humidity
5	-0.712	brightness, temperature
6	-0.613	brightness, uv_factor
7	-0.083	air_quality, humidity
8	+0.079	air_quality, temperature
9	-0.078	air_quality, brightness
10	+0.070	air_quality, uv_factor

Figura 42. Correlaciones de un mes

Humedad - Temperatura

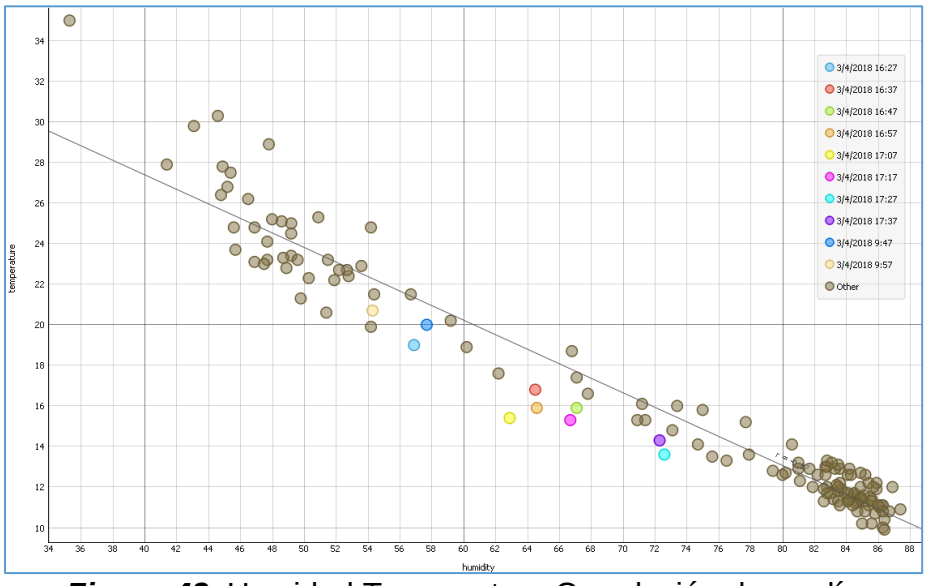


Figura 43. Humidad Temperatura Correlación de un día

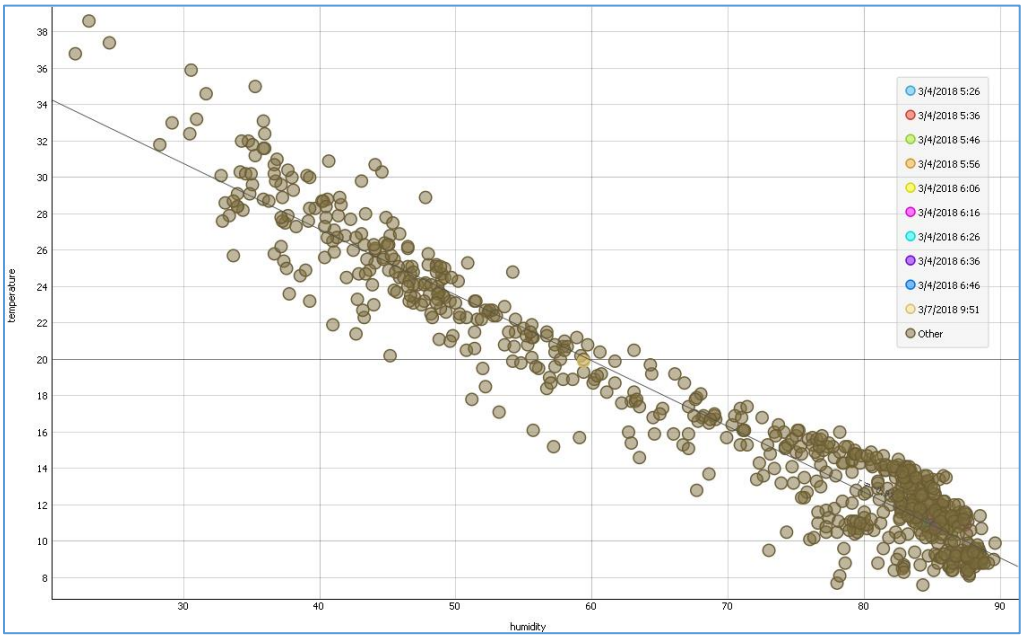


Figura 44. Humedad Temperatura Correlación de una semana

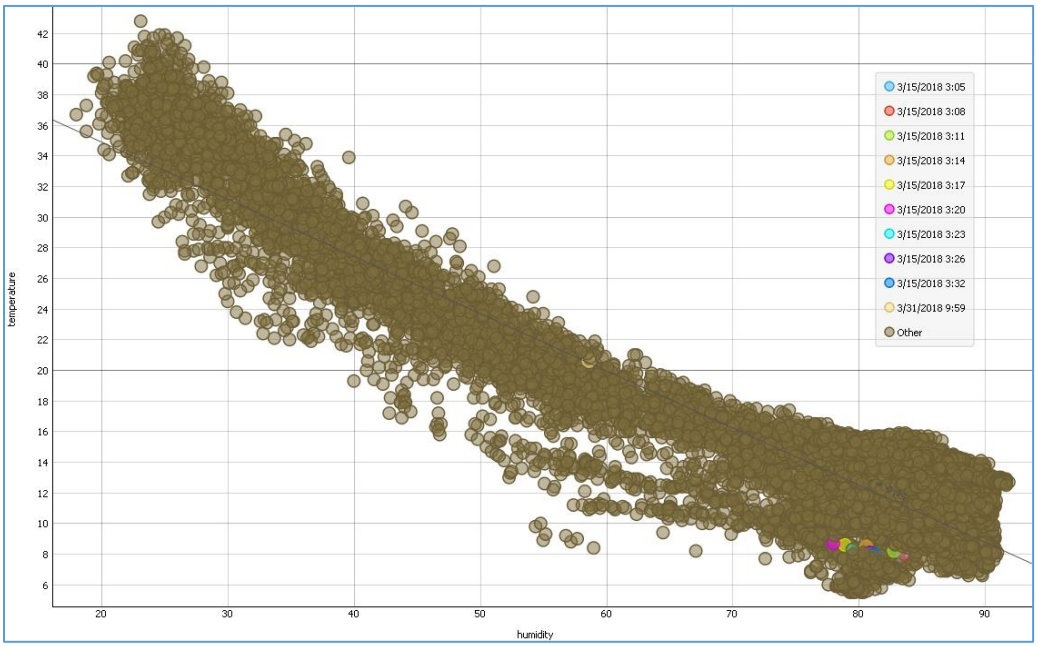


Figura 45. Humedad Temperatura Correlación de un mes

En las figuras 43, 44 y 45 se puede observar la correlación negatica fuerte que existe entre temperatura y humedad relativa (mide la cantidad de agua en el aire en forma de vapor), lo que indica que estas variables son inversamente proporcionales es decir a medida que aumenta la temperatura la humedad disminuye, por ende las condiciones óptimas de temperatura que se dieron dentro del invernadero se ubican en este horario: mañana (9:30 a 12:00), tarde (16:40 a 18:30), noche y madrugada (18:30 a 8:30), con respecto a la Humedad Relativa las condiciones óptimas se encuentran por la: mañana (8:00 a 8.30) y tarde (17:00 a 18:30).

Con el análisis realizado y en relación a otras zonas florícolas se determinó y corroboró que estos 2 factores no siempre van a coincidir o van a ser los más adecuados en una determinada hora, lo cual influye en las decisiones que debe tomar el técnico encargado del manejo del invernadero.

No obstante, también es importante mencionar que no existieron saltos térmicos sobre los 10 ºC, temperaturas mayores a 42 °C o menores a 4ºC ya que derivarían en problemas de producción o calidad de las rosas, con respecto a la humedad relativa se evidenció que en horas de la noche y madrugada esta sobrepasaba el 80% lo cual repercute en la presencia de plagas y enfermedades como: botrytis, cinerea, etc.

Luminosidad – Humedad

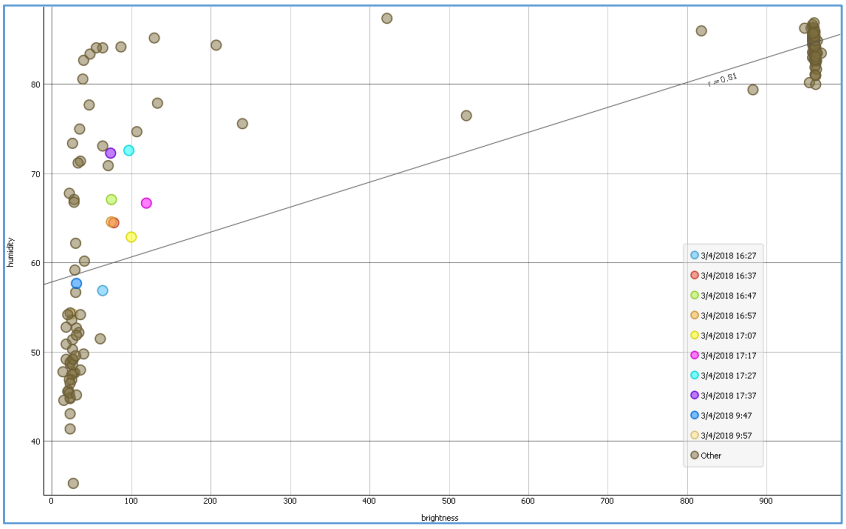


Figura 46. Luminosidad Humedad Correlación de un día

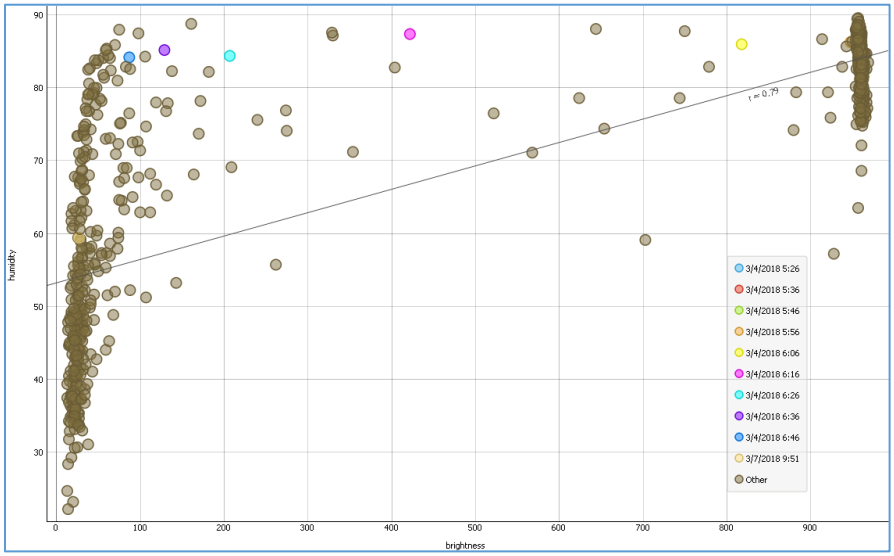


Figura 47. Luminosidad Humedad Correlación de una semana

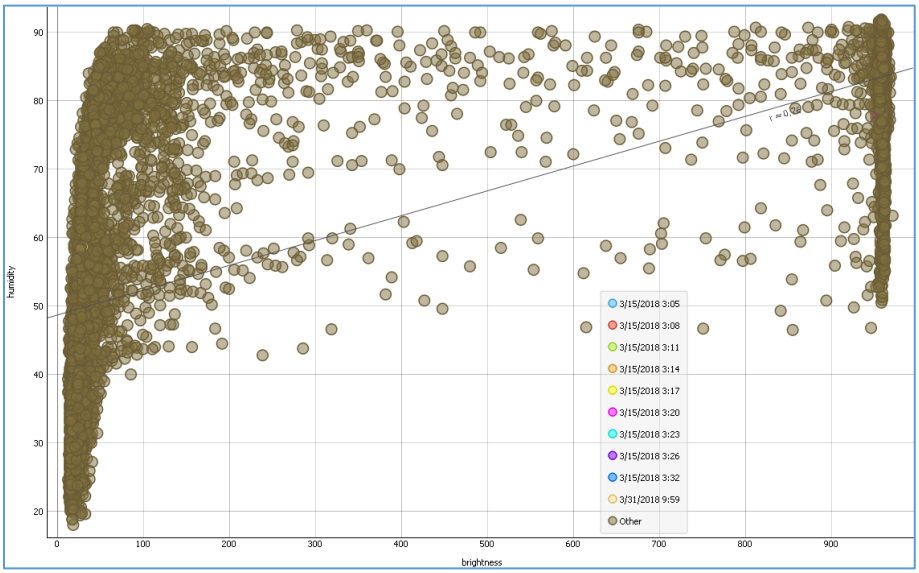


Figura 48. Luminosidad Humedad Correlación de un mes

En las figuras 46, 47 y 48 se puede observar una correlación positiva fuerte que existe entre Luminosidad y Humedad relativa, dándonos a entender que a mayor luminosidad en el invernadero la humedad relativa también aumenta, esto resulta beneficioso ya que la fotosíntesis puede alcanzar su punto máximo; caso contrario si hay poca luminosidad pueden descender las necesidades de las otras variables.

Cabe mencionar que el rango de luminosidad con el que trabaja este sensor varia de 0 (muy claro) en el día a 1024 (muy oscuro) en la noche, los mismos que no pueden alcanzar los puntos máximos ya que por ejemplo una baja intensidad lumínica puede producir abortos en su flor y un mayor número de brotes ciegos.

Temperatura – Factor UV

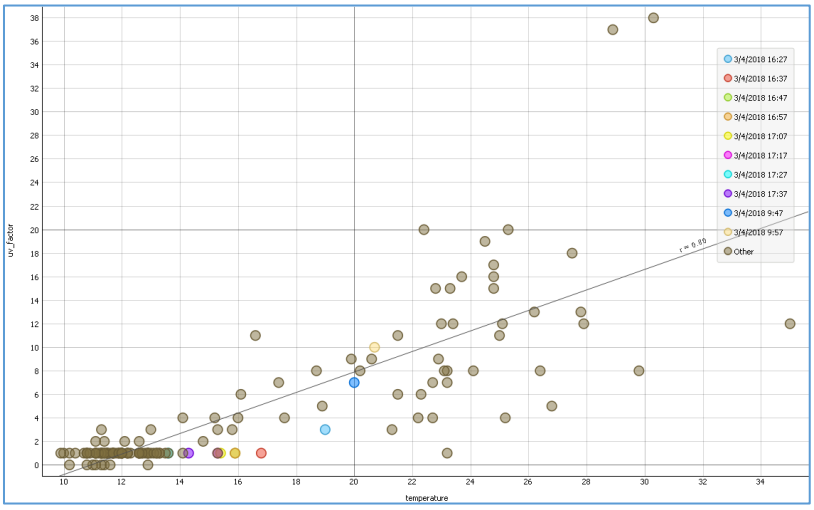


Figura 49. Temperatura Factor UV de un día

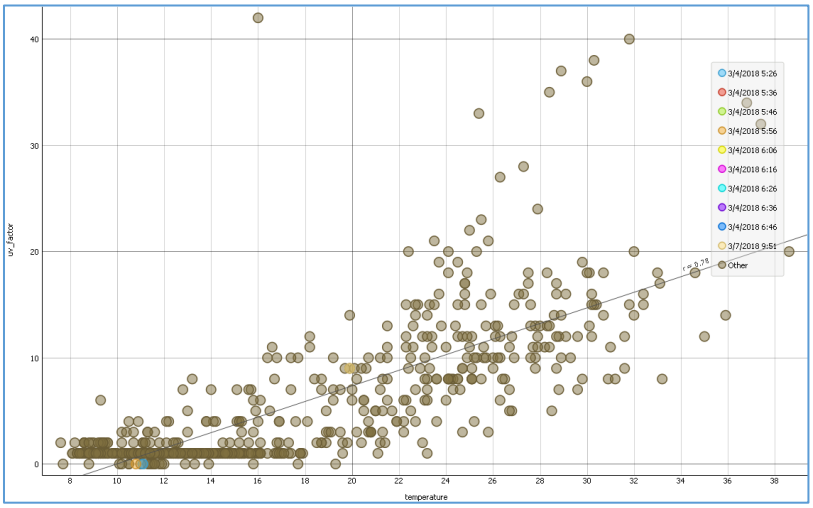


Figura 50. Temperatura Factor UV de una semana

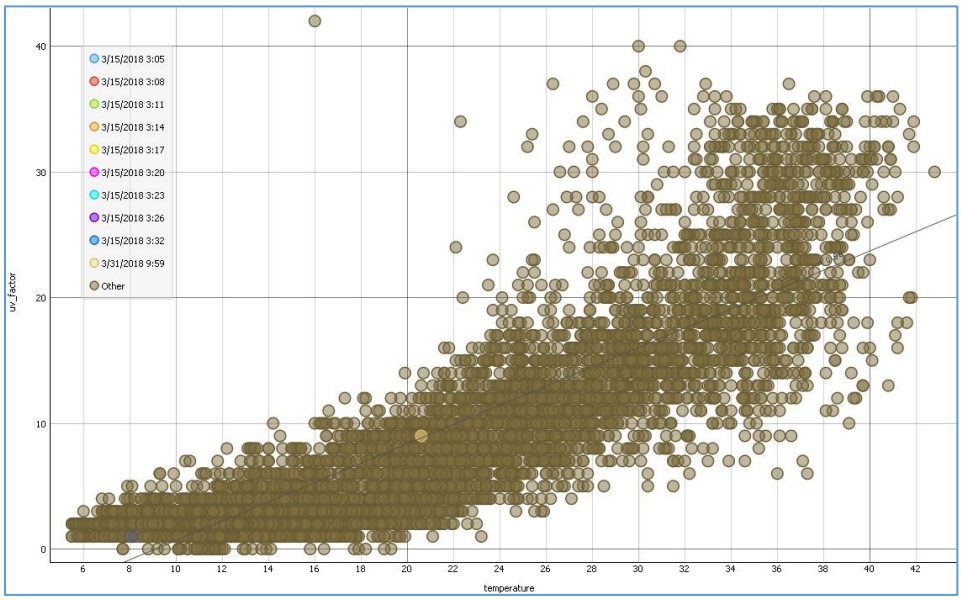


Figura 51. Temperatura Factor UV de un mes

En las figuras 49, 50 y 51 se puede evidenciar que existe una correlación positiva fuerte entre Temperatura y Factor UV, dándonos a entender que ha mayor radiación UV en el exterior la temperatura bajo invernadero aumenta, lo cual genera problemas de producción, incremento de botones ciegos y la aparición de flores más pequeñas de lo normal con colores más cálidos y escasos pétalos.

Con relación a la radiación solar si esta es excesiva, puede causar estrés a la planta, afectar a su proceso fotosintético, generar mutaciones y presencia de plagas y enfermedades. En rosas los daños se pueden apreciar a simple vista, ennegrecimiento de los botones en variedades rojas, pérdida de pigmentación y quemaduras en los márgenes de los pétalos. (Clusterflor, 2017).

Una medida correctiva para radiaciones mayores a 60 nm es la utilización de materiales de cubierta fotoselectivos u otro tipo que ayude a evitar el paso directo de la radiación al interior del invernadero.

Luminosidad – Temperatura

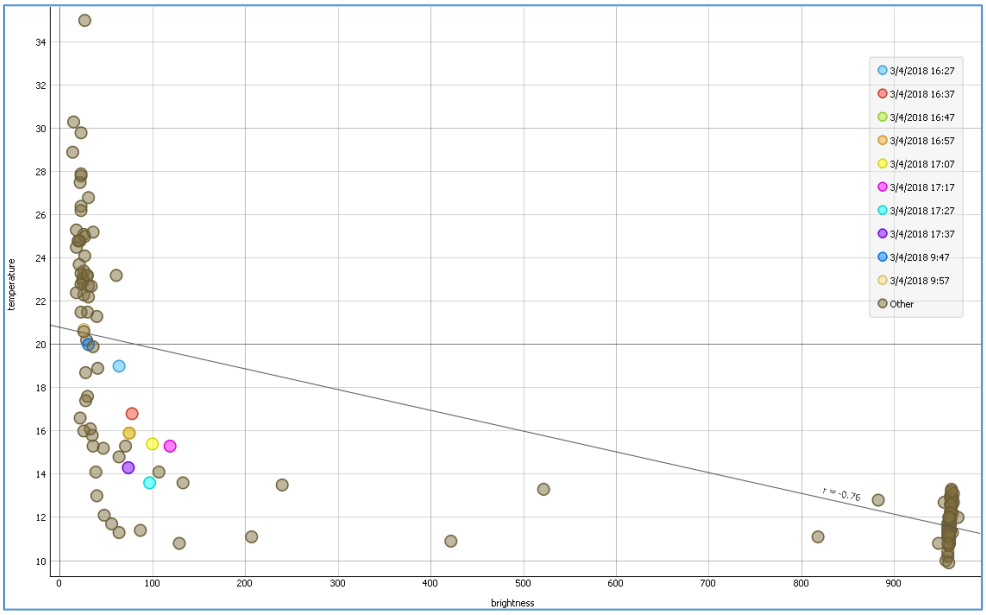


Figura 52. Luminosidad Temperatura Correlación de un día

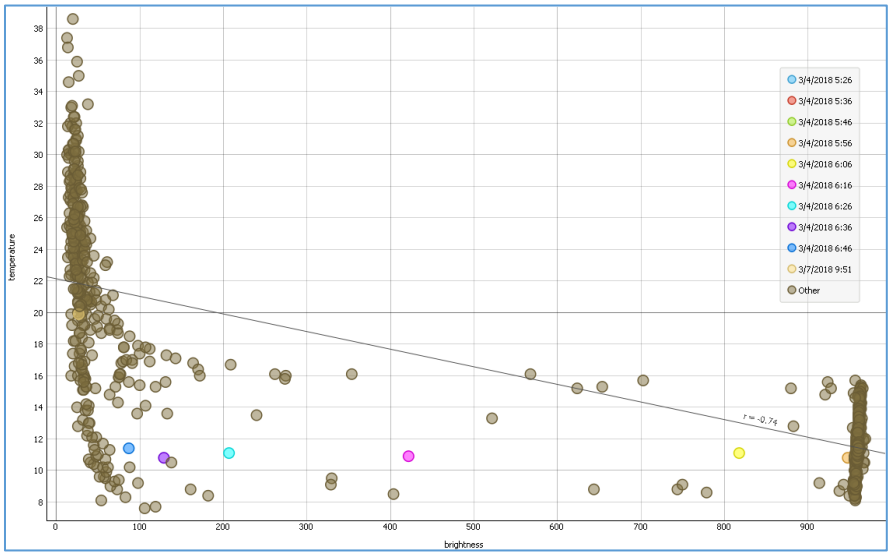


Figura 53. Luminosidad Temperatura Correlación de una semana

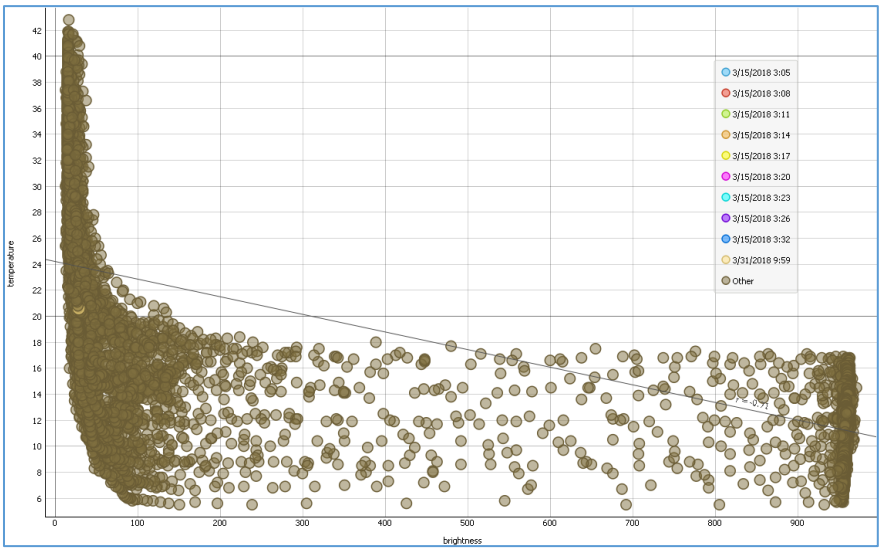


Figura 54. Luminosidad Temperatura Correlación de un mes

En primer lugar se debe tener encuenta que el rango de luminosidad varia de 0 (muy claro) a 1024 (muy oscuro), entonces en las figuras 52, 53 y 54 se

puede observar la correlación negativa fuerte que existe entre Luminosidad y Temperatura, dándonos a entender que a mayor luz en el exterior la temperatura bajo invernadero aumenta y esto a su vez va conforme al paso del dia, ocasionando los problemas de temperatura mencionados en los apartados anteriores.

Humedad – Factor UV

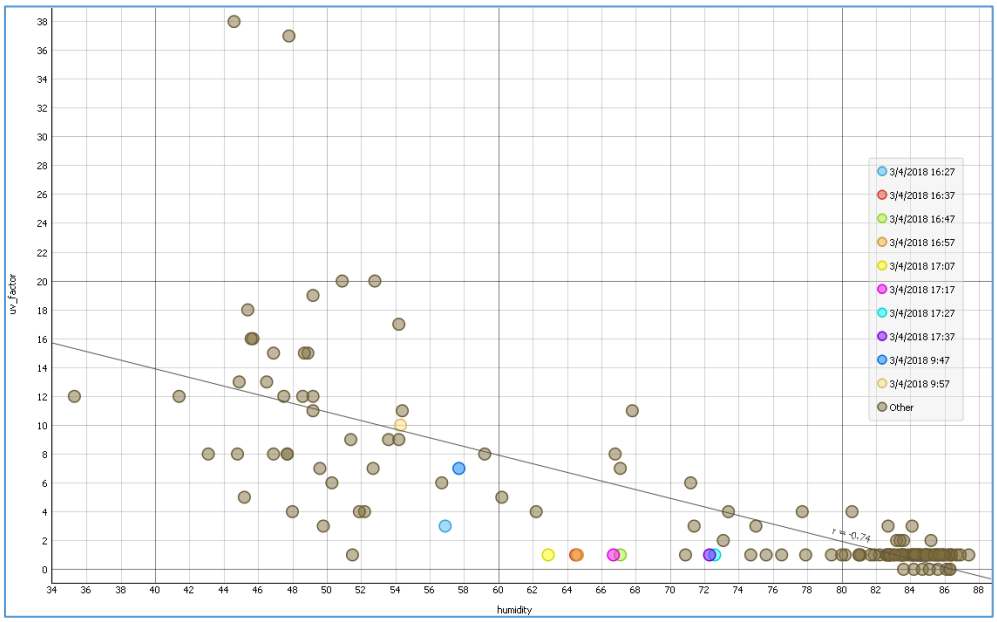


Figura 55. Humedad Factor UV Correlación de un día

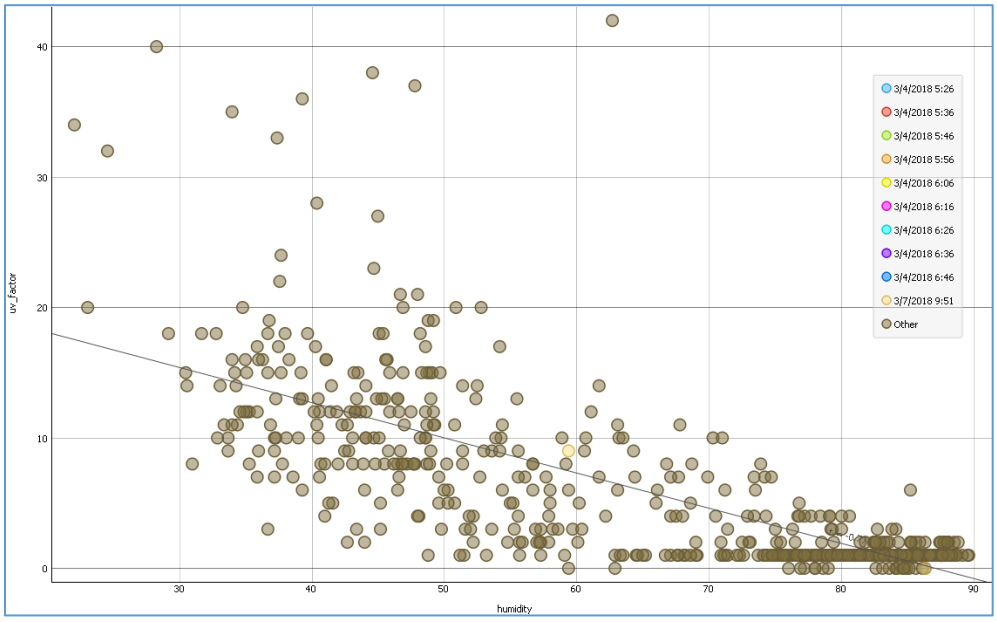


Figura 56. Humedad Factor UV Correlación de una semana

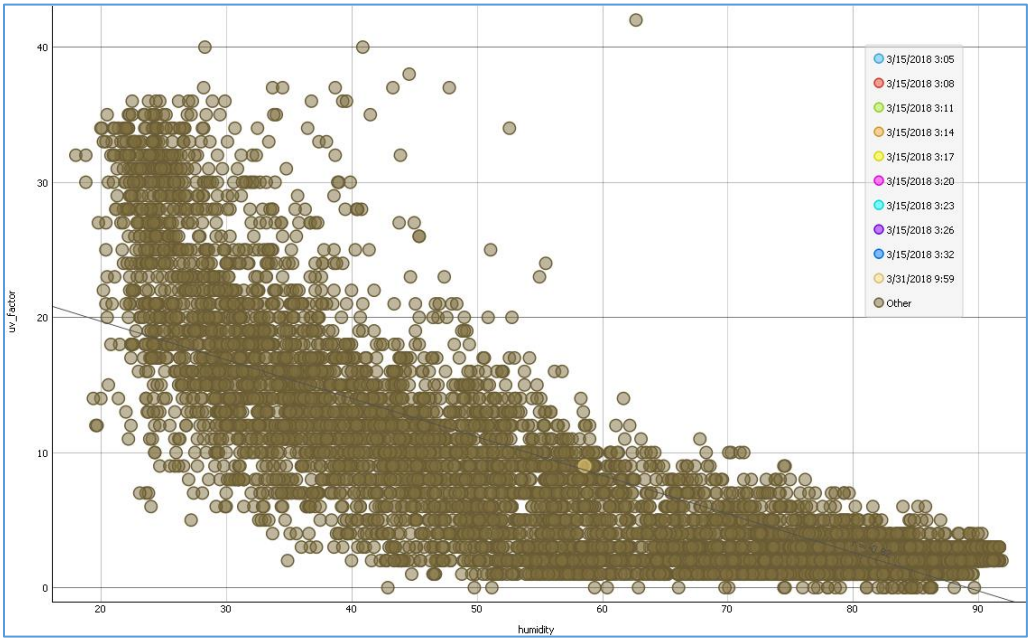


Figura 57. Humedad Factor UV Correlación de un mes

En las figuras 55, 56 y 57 se puede observar la correlación negativa fuerte que existe entre Humedad relativa y Factor UV, lo cual indica que ha mayor radiación UV en el exterior la temperatura bajo invernadero aumenta y por consecuente la humedad relativa disminuye, desencadenando los problemas de HR mencionados en los apartados anteriores.

Se debe tener en cuenta que la radiación solar es un factor muy influyente en las otras variables que no puede ser controlada.

Valores sin correlación.

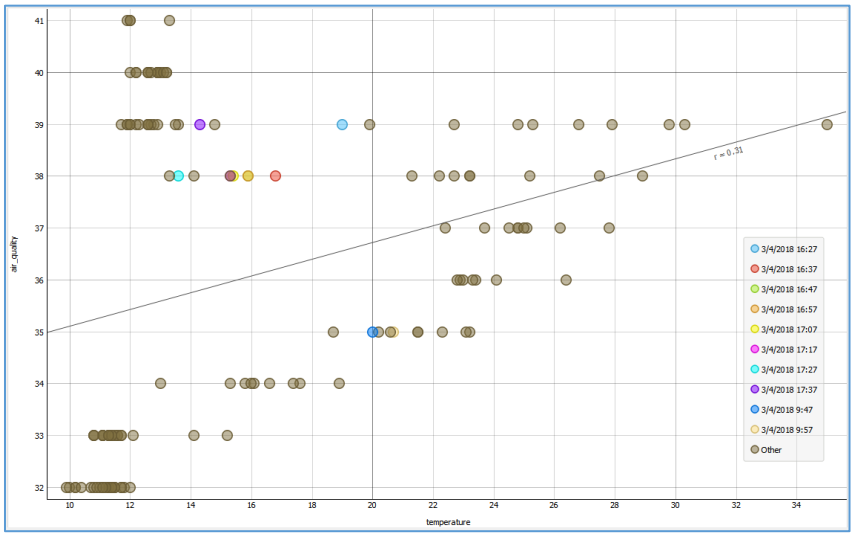


Figura 58. Temperatura Calidad de aire Correlación por un día

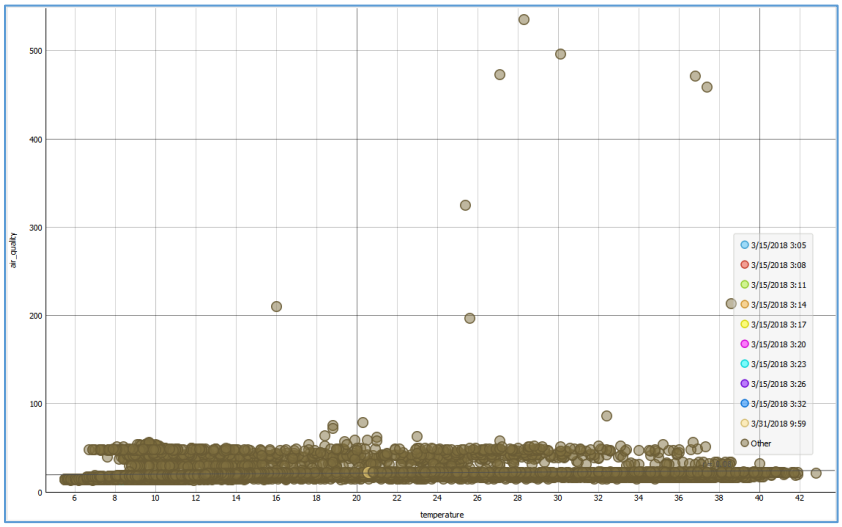


Figura 59. Temperatura Calidad de aire Correlación por una semana

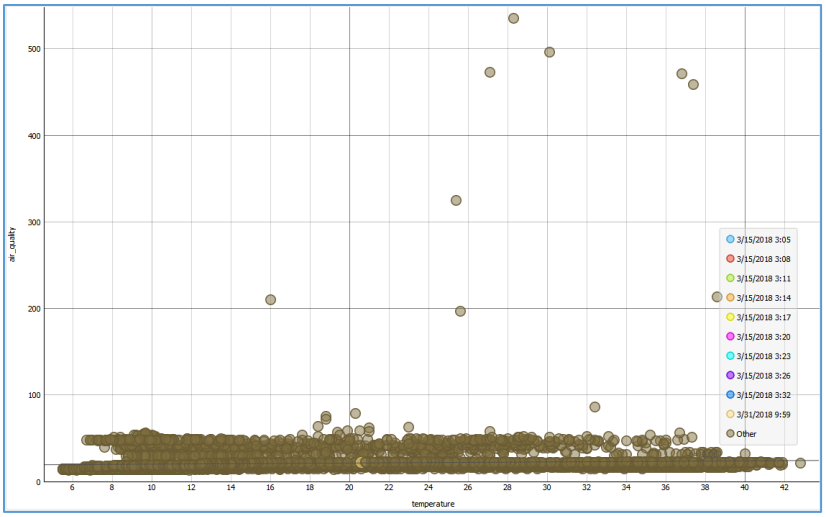


Figura 60. Temperatura Calidad de aire Correlación por un mes

En las figuras 58 y 59 se puede observar que existe una correlación moderada entre temperatura y factor V, pero en la figura 60 también se puede apreciar que dicha correlación ya no existe porque el coeficiente de Pearson r=0.079 tiende a cero, de igual forma esto sucede con las variables de humedad, luminosidad y factor UV

dándonos a entender que no se pudo establecer un patrón con este factor abiótico, mismos que pueden ser corroborados con los resultados obtenidos en las figuras 40, 41 y 42.

Esta anomalía se detectó al momento de verificar los datos emitidos por el sensor ya que la información que este generaba era en mili voltios (mV), muy diferente a la unidad de medida dióxido de carbono (CO2) requerida por el especialista para llevar a cabo este tipo de análisis.

Técnica Reglas de asociación.

Valores analizados por un día.

En la figura 58 se identificaron las siguientes reglas de asociación:

- Cuando hay obscuridad la temperatura fue óptima pero la humedad relativa no, en el 49% de los casos y con un nivel de confianza del 98%;
- Cuando hay obscuridad la temperatura es óptima en el 49% de los casos y con un nivel de confianza de 98%;
- Cuando hay humedad relativa inadecuda con obscuridad la temperatura será óptima en el 49% de los casos y con un nivel de confianza de 98%.

Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		Consequent
0.493	0.986	0.500	1.569	1.257	0.101	brightness=0	→	temperature=1, humidity=0
0.493	0.628	0.785	0.637	1.257	0.101	temperature=1, humidity=0	→	brightness=0
0.493	0.986	0.500	1.681	1.174	0.073	brightness=0	→	temperature=1
0.493	0.986	0.500	1.681	1.174	0.073	humidity=0, brightness=0	→	temperature=1
0.500	1.000	0.500	1.875	1.067	0.031	brightness=0	→	humidity=0
0.493	1.000	0.493	1.901	1.067	0.031	temperature=1, brightness=0	→	humidity=0
0.785	0.837	0.938	0.896	0.996	-0.003	humidity=0	→	temperature=1
0.785	0.934	0.840	1.116	0.996	-0.003	temperature=1	→	humidity=0
0.438	0.875	0.500	1.875	0.933	-0.031	brightness=1	→	humidity=0
0.292	0.840	0.347	2.700	0.896	-0.034	temperature=1, brightness=1	→	humidity=0
0.347	0.694	0.500	1.681	0.826	-0.073	brightness=1	→	temperature=1
0.292	0.667	0.438	1.921	0.793	-0.076	humidity=0, brightness=1	→	temperature=1

Figura 61. Reglas de asociación por Día

Valores analizados por una semana

En la figura 59 se identificaron las siguientes reglas de asociación:

- Cuando la temperatura y humedad relativa no son óptimas la luminosidad fue adecuada en el 25% de los casos y con un nivel de confianza del 71%;
- Cuando hay obscuridad la temperatura fue óptima pero la humedad relativa inadecuada en el 36% de los casos y con un nivel de confianza del 76%;
- Cuando hay obscuridad la temperatura es óptima en el 37% de los casos y con un nivel de confianza de 77%;
- Cuando hay humedad relativa no óptima con obscuridad la temperatura es óptima en el 36% de los casos y con un nivel de confianza del 77%.

Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		Consequent
0.259	0.681	0.381	1.262	1.418	0.076	temperature=0	→	humidity=0, brightness=1
0.274	0.719	0.381	1.362	1.388	0.076	temperature=0	→	brightness=1
0.259	0.711	0.365	1.422	1.371	0.070	temperature=0, humidity=0	→	brightness=1
0.367	0.763	0.482	1.222	1.296	0.084	brightness=0	→	temperature=1, humidity=0
0.367	0.624	0.589	0.818	1.296	0.084	temperature=1, humidity=0	→	brightness=0
0.375	0.778	0.482	1.286	1.256	0.076	brightness=0	→	temperature=1
0.375	0.605	0.619	0.778	1.256	0.076	temperature=1	→	brightness=0
0.367	0.777	0.473	1.310	1.255	0.075	humidity=0, brightness=0	→	temperature=1
0.473	0.982	0.482	1.979	1.030	0.014	brightness=0	→	humidity=0
0.367	0.980	0.375	2.543	1.029	0.010	temperature=1, brightness=0	→	humidity=0
0.365	0.958	0.381	2.504	1.005	0.002	temperature=0	→	humidity=0
0.589	0.618	0.953	0.650	0.997	-0.002	humidity=0	→	temperature=1
0.589	0.950	0.619	1.539	0.997	-0.002	temperature=1	→	humidity=0
0.259	0.947	0.274	3.481	0.993	-0.002	temperature=0, brightness=1	→	humidity=0
0.480	0.927	0.518	1.839	0.972	-0.014	brightness=1	→	humidity=0
0.221	0.904	0.245	3.898	0.949	-0.012	temperature=1, brightness=1	→	humidity=0

Figura 62. Reglas de asociación por Semana

4.6. Implementación

Elaboración y validación de la propuesta con la Dra. Elizabeth Urbano

Tabla 6
Propuesta de mejora invernadero

MATRIZ DE APLICACIÓN DE LOS DATOS RECOLECTADOS A TRAVES DE UNA WSN Y MINERÍA DE DATOS

	Condición ideal		**Condición actual**						**Propuesta de mejoramiento del invernadero**
	Día	**Noche**	**Madrugada**	**Mañana**	**Tarde**	**Noche**	**Promedio**		En horas de 12h00 a 06h30 y 18h30 a 00h00
			00:00-06:30	06:30-12:00	12:00-18:30	18:30-00:00	**Día**	**Noche**	1. Incorporar sistemas de calefacción 2. Evaluar la diferencia en el ciclo de cultivo cuando existe sistemas de calefacción que mantienen la T° ideal vs a los que no disponen.
Temperatura	20 – 25 ºC	10 – 16 ºC	9 - 12 ºC	10 - 24 ºC	16 – 35 ºC	11 – 13 ºC	20.41 °C	11.86 °C	
Humedad Relativa	70-75 %		82 – 87 %	45 – 85 %	35 - 78 %	76 – 87 %	58.93%	83.90%	En horas de 12h00-18h30 1. Regar caminos 2. Incorporar sistema de nebulización 3. Evaluar diferencias en ciclo de cultivo si se dispone de estos equipos

CONTINÚA

Factor UV	40 A 60 nm	0 – 20 nm	10 – 60 nm	10 – 100 nm	0 - 20 nm	-	En horas de 12h00-18h30 1. Existe exceso de radiación UV lo cual perjudica variedades de color rojo por efecto de blackenig, por tanto en esos casos adquirir materiales de cubierta que filtren el 100 % de radiación UV.
Calidad de aire	CO2: 1000 - 2000 ppm.	No se pudo definir por el tipo de sensor implementado				-	Implementar sensores que nos permitan medir este parámetro.
Luminosidad	60.000-80.000 Lux	Rango de medida del sensor 0 “muy claro” a 1024 “muy obscuro”				-	Implementar sensores que nos permitan medir este parámetro en Lux para relacionar con cantidad de fotosíntesis realizada por la planta.

CAPÍTULO 5

4.1. Conclusiones

Al realizar esta investigación se pudo constatar que ese tipo proyectos pueden ser aplicados en otros campos como la piscicultura, avicultura, cunicultura etc. para el descubrimiento de patrones de comportamiento o identificación de interacción de factores abióticos y bióticos que inciden en la producción o crecimiento animal.

A pesar de lo novedoso del tema, puede afirmarse que el uso de técnicas de minería sobre streams de datos provenientes de este tipo de redes WSN beneficia el entorno florícola en gran medida, debido a que es una importante herramienta para obtener conocimiento. Por ende su uso hizo posible, entre otros objetivos, identificar fortalezas y/o debilidades mediante el análisis de los datos para de esta manera tomar acciones correctivas y contribuir a que se lleve a cabo un mejor manejo de los procesos internos del invernadero.

Del análisis realizado podemos concluir que los factores más significativos son: humedad y temperatura, en los cuales se determinó que nunca llegan a mantenerse en las condiciones óptimas requeridas dentro del invernadero en el día, caso contrario sucede en la noche y madrugada donde estos mismos factores fluctúan en los rangos óptimos (10 °C -16 °C y 70 – 75% respectivamente), por tal motivo en el día el técnico del invernadero debe tomar las medidas correctivas ya que la temperatura y humedad

alcanzan puntos extremadamente altos o bajos que repercuten de alguna manera en la producción y la calidad de las rosas.

Cuando se desea aplicar minería de datos se debe tener claro el objetivo a alcanzar con respecto de los datos, ya que puede tratarse de un problema de clasificación como detección de correo basura o de un problema de clustering como recomendarle un libro a un cliente, siempre hay que hacerse la pregunta ¿qué es lo que quiero hacer?, aunque no se tengan datos con un objetivo definido.

4.2. Recomendaciones

En minería de datos stream se recomienda primero evaluar los objetivos del proyecto para determinar si existe la necesidad de realizar un análisis en tiempo real ya que este tipo de proyectos se enfocan en recibir y analizar un gran cantidad de datos de diversos orígenes y a diferentes tasas de arribo.

En relación a los sensores se recomienda remplazar el sensor de Luminosidad por un TSL2591 el cual permite cálculos exactos en Lux, con respecto al sensor Factor UV se debe adquirir un VEML6070 el cual posee un sensor real de luz en el espectro UV y con relación al sensor de Calidad de Aire se sugiere realizar una configuración adecuada para obtener los datos requeridos ya que este sensor detecta varios gases tal como alcohol, benceno, humo, CO2, etc.

Cuando se realiza un proyecto de minería de datos se debe aplicar la minería descriptiva y predictiva a la vez, ya que estas se relacionan mutuamente y

complementan entre sí, ya que otorgan modelos confiables y efectos inesperados que se transforma en un valor añadido a la empresa.

Se recomienda que para establecer un patrón de comportamiento a nivel correlacional, los resultados obtenidos por este análisis deben ser similares por más de tres veces y con diferentes grupos de datos, ya que de lo contrario puede tomarse como una coincidencia.

4.3. Líneas de trabajo futuro

Como trabajos futuros se propone la implementación de un invernadero automatizado y una arquitectura de minería de datos en streaming similar a la usada en este proyecto con la finalidad de incorporar alertas y predicciones.

Para una futura investigación más profunda se debería incorporar sensores que realicen el análisis de factores bióticos y combinarlos con los usados en esta investigación para descubrir nuevo conocimiento que ayude a mejorar el cultivo.

REFERENCIAS BIBLIOGRAFICAS

Astellia. (14 de 005 de 2015). *Orange big data offer for businesses*. Obtenido de astellia.com: https://www.astellia.com/resources/flux-vision-orange-big-data-offer-for-businesses-and-public-authorities/

Biolab. (2 de 05 de 2016). *Form survery to orange*. Obtenido de blog.biolab.si: https://blog.biolab.si/tag/data/

Camejo, I. M. (17 de 05 de 2012). *Desempeño de algoritmos de mineria en indicadores académicos*. Obtenido de scielo.sld.cu: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2227-18992015000400008

Clusterflor. (21 de 05 de 2017). *Clima bajo invernadero*. Obtenido de flor.ebizor.com: http://flor.ebizor.com/clima-bajo-invernadero/?fbclid=IwAR1rWWxFBqBPNp2ztYQUu0sbZe-Xd7NyBgqw-964XjIbx-9B0WodMVEiqFg

Conexionesan. (25 de 4 de 2014). *El valor de la segmentación predictiva*. Obtenido de esan.edu.pe: https://www.esan.edu.pe/apuntes-empresariales/2018/06/el-valor-de-la-segmentacion-predictiva-en-la-mineria-de-datos/

Confluent. (5 de 05 de 2015). *Introducing kafka streams*. Obtenido de confluent.io: https://www.confluent.io/blog/introducing-kafka-streams-stream-processing-made-simple/

Hortonworks. (5 de 05 de 2015). *Introduction to spark streaming*. Obtenido de es.hortonworks.com: https://es.hortonworks.com/tutorial/introduction-to-spark-streaming/

InfoAgro. (12 de 02 de 2012). *Rosas para corte*. Obtenido de abcAgro.com: http://www.abcagro.com/flores/flores/rosas.asp

Infoagro. (5 de 05 de 2015). *Cultivo de rosas para corte*. Obtenido de infoagro.com: http://www.infoagro.com/flores/flores/rosas.htm

León Guzmán, E. (22 de 02 de 2015). *Metricas para validación de clustering*. Obtenido de disi.unal.edu.co: http://www.disi.unal.edu.co/profesores/eleonguz/cursos/md/presentaciones/Sesion13_validacion_Clustering.pdf

Lizarraga, R. E. (14 de 05 de 2015). *Comparativa entre metodos de regresion lineal y identificación de variables asociadas*. Obtenido de researchgate.net: https://www.researchgate.net/publication/236855935_Comparativa_entre_los_Metodos_de_Regresion_Lineal_y_Mineria_de_Datos_para_la_Identificacion_de_Variables_Asociadas_al_Exito_Academico_en_Estudiantes_de_Educacion_Superior

Luca. (16 de 10 de 2017). *Como solucionar tu problema de negocio con big data*. Obtenido de empresas.blogthinkbig.com: https://empresas.blogthinkbig.com/como-solucionar-tu-problema-de-negocio-con-big-data/

Martinez de Pinson Ascacibar, F. (14 de 3 de 2015). *Mineria de datos series temporales*. Obtenido de lsi.us.es: http://www.lsi.us.es/redmidas/CEDI/papers/143.pdf

Minitab. (22 de 03 de 2012). *Interpretación de los resultados clave*. Obtenido de support.minitab.com: https://support.minitab.com/es-mx/minitab/18/help-and-how-to/statistics/basic-statistics/how-to/correlation/interpret-the-results/key-results/?fbclid=IwAR1Q_lxBQFXO2-Ui9b8oX2pSSCn0udZQQadSLegqDJkYk1BtzJl6T8v9BLE

Minitab. (15 de 5 de 2014). *Interpretación de resultados para correlación*. Obtenido de support.minitab.com: https://support.minitab.com/es-mx/minitab/18/help-and-how-to/statistics/basic-statistics/how-to/correlation/interpret-the-results/key-results/?fbclid=IwAR0lpfDKRaI5p-M21Sx-TdBU28aEUA3JTQlDhzCgz47msSBoleV-IIriHkM

Oldemar, R. (12 de 02 de 2015). *Metodos descriptivos en mineria de datos*. Obtenido de oldemarrodriguez.com: http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037

Orange. (12 de 05 de 2015). *Fluz vision real time statistics on mobility patterns*. Obtenido de orange-business.com: https://www.orange-business.com/en/products/flux-vision

Orange. (12 de 05 de 2016). *Data mining fruitful and fun open source machine learning*. Obtenido de orange.biolab.si: https://orange.biolab.si/

Rodriguez Montequín, M., & Alvarez Cabal, J. (12 de 05 de 2014). *Metodologias para la realización de proyectos de mineria de datos.* Obtenido de Aeipro.com: https://www.aeipro.com/files/congresos/2003pamplona/ciip03_0257_0265.2134.pdf

Rodriguez, O. (12 de 5 de 2014). *Metodos predictivos en mineria de datos*. Obtenido de oldemarrodriguez.com: http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037

Santos, P. R. (2 de 05 de 2015). *Tipos de aprendizaje machine learning*. Obtenido de empresas.blogthinkbig.com: https://data-speaks.luca-d3.com/2017/11/que-algoritmo-elegir-en-ml-aprendizaje.html

Solsoft. (14 de 05 de 2014). *Descubrimiento de patrones secuenciales utilizando logica temporal*. Obtenido de solsoft.biz: http://solsoft.biz/prensa/articulo/article/descubrimiento-de-patrones-secuenciales-utilizando-razonamiento-logico-temporal.html

Villena Román, J. (8 de 08 de 2016). *Metodologia para poner orden en data science*. Obtenido de data.sngular.com: https://data.sngular.com/es/art/25/crisp-dm-la-metodologia-para-poner-orden-en-los-proyectos-de-data-science

Webmining. (12 de 6 de 2014). *Procesos para obtener conocimiento sobre mineria de datos*. Obtenido de webmining.cl: http://www.webmining.cl/2011/01/proceso-de-extraccion-de-conocimiento/

Weka. (12 de 06 de 2015). *Practica machine learning tools and techniques*. Obtenido de cs.waikato.ac.nz: https://www.cs.waikato.ac.nz/ml/weka/book.html

Weka. (5 de 12 de 2017). *Weka 3 machine learning software in java*. Obtenido de cs.waikato.ac.nz: https://www.cs.waikato.ac.nz/ml/weka/

Wikipedia. (20 de 03 de 2015). *Metodologia para mineria de datos semma*. Obtenido de en.wikipedia.org: https://en.wikipedia.org/wiki/SEMMA

Xataka. (12 de 02 de 2014). *Las redes neuronales: qué son y por qué están volviendo*. Obtenido de Xataka.com: https://www.xataka.com/robotica-e-ia/las-redes-neuronales-que-son-y-por-que-estan-volviendo