Estrella 

Facultad de Ingeniería Informática

Tema: Modelos para el estudio de las variables que más influyen en la estimación de la legitimidad en el modelo probabilístico de autenticación propuesto por Legón.

Autor:

Estrella Cárdenas Blanco

Tutor:

DrC. Raisa Socorro Llanes

La Habana 2021

Resumen

En la modernidad los usuarios necesitan acceder a muchos servicios digitales imprescindibles para su vida cotidiana por lo que la autenticación de un usuario para concederle acceso a un sistema es un aspecto esencial para la seguridad de la información. Las contraseñas alfanuméricas son las más usadas pero tienen una contradicción entre su seguridad y su usabilidad, pues para ser seguras deben ser aleatorias, largas y no predecibles, mientras que para ser usables deben ser memorizables. Una de las alternativas de solución a este problema es la autenticación gráfica. Los sistemas de autenticación gráfica basan su funcionamiento en que los usuarios reconozcan patrones de imágenes o partes de ellas, en lugar de recordar largas y complejas secuencias de caracteres. Entre las técnicas de autenticación gráfica, *el PassPoint* ha recibido atención especial, es una de las técnicas basadas en la memorización de puntos como contraseña. Los modelos probabilísticos en autenticación gráfica se aplican para estimar la clave más probable a seleccionar, en cada imagen, por el usuario que se va a registrar. Existe un sistema que gestiona los intentos de autenticación de los usuarios incorporando al esquema PassPoint en el modelo probabilístico implementado en la tesis del Dr. Legón. Estos datos serán utilizados para aplicarles Técnicas de Minería de Datos con el fin de identificar patrones entre las variables: el sexo, la edad, el nivel educativo de las personas, el tiempo, el radio de tolerancia y la imagen que influyen en la estimación de la legitimidad del modelo probabilístico de autenticación gráfica propuesto en la tesis del Dr. Legón.

Palabras claves: Autenticación Gráfica, Modelo Probabilístico, *PassPoint*.

Abstract

In modernity, users need access to many essential digital services for their daily lives, so the authentication of a user to grant access to a system is an essential aspect for information security. Alphanumeric passwords are the most used but they have a contradiction between their security and usability, because to be secure they must be random, long and not predictable, while to be usable they must be memorized. One of the alternative solutions to this problem is graphical authentication. Graphical authentication systems base their operation on users recognizing image patterns or parts of them, instead of remembering long and complex sequences of characters. Among the graphical authentication techniques, the PassPoint has received special attention; it is one of the techniques based on memorizing points as a password. Probabilistic models in graphical authentication are applied to estimate the most likely key to be selected, in each image, by the user to be registered. There is a system that manages user authentication attempts incorporating the PassPoint scheme in the probabilistic model implemented in Dr. Legón's thesis. These data will be used to apply Data Mining Techniques in order to identify patterns between the variables: sex, age, educational level of people, time, tolerance radius and image that influence the estimation of the legitimacy of the graphic authentication probabilistic model proposed in Dr. Legón's thesis.

Keywords: Graphical Authentication, Probabilistic Model, PassPoint.

Índice General

[Introducción 1](#_Toc103199411)

[Capítulo 1 Fundamentos Teóricos 8](#_Toc103199412)

[1.1 Introducción 8](#_Toc103199413)

[1.2 Contraseñas Gráficas 8](#_Toc103199414)

[1.2 Evaluación experimental de las técnicas Cued-Recall Based 12](#_Toc103199415)

[1.3 Minería de Datos 16](#_Toc103199416)

[1.3.1 Las fases del proceso de extracción del conocimiento 17](#_Toc103199417)

[1.3.2 Modelos de minería de datos 22](#_Toc103199418)

[1.3.3 Técnicas y algoritmos de minería de datos (aquí voy a poner las técnicas que voy a explicar) 23](#_Toc103199419)

[1.3.4 Metodologías utilizadas para la minería de datos 23](#_Toc103199420)

[1.3.5 Evaluación de arquitecturas candidatas. 28](#_Toc103199421)

[1.3.6 Konstanz Information Miner 29](#_Toc103199422)

[1.4 Conclusiones Parciales 32](#_Toc103199423)

[Capítulo 2 Desarrollo de la aplicación 32](#_Toc103199424)

[2.1 Introducción 32](#_Toc103199425)

[2.2 Diseño Experimental 32](#_Toc103199426)

[2. Conclusiones 33](#_Toc103199427)

[Capítulo 3 Validación de la herramienta 33](#_Toc103199428)

[3.1 Introducción 33](#_Toc103199429)

[3. Conclusiones 33](#_Toc103199430)

[Conclusiones Generales 33](#_Toc103199431)

[Recomendaciones 33](#_Toc103199432)

[Referencias 34](#_Toc103199433)

Índice de Figuras

[Figura 1 Clasificación de los métodos de autenticación 8](#_Toc103199527)

[Figura 2 Evolución de las técnicas basadas en la memorización de puntos contraseñas 10](#_Toc103199528)

[Figura 3 Fases del proceso de descubrimiento de conocimiento. 18](#_Toc103199529)

[Figura 4 Encuesta realizada por la KDnuggets en el año 2007 23](#_Toc103199530)

[Figura 5 Fases de la metodologíaSEMMA 24](#_Toc103199531)

[Figura 6 Iteración de las fases de SEMMA 25](#_Toc103199532)

[Figura 7 Metodología CRISP-DM 26](#_Toc103199533)

[Figura 8 Flujo de Trabajo. 31](#_Toc103199534)

[Figura 9 Estado de los nodos. 31](#_Toc103199535)

Índice de las Tablas

# Introducción

La autenticación de un usuario para acceder a un sistema es un tema clave para la seguridad de la información ya que determina si el usuario puede tener acceso [[1](#_ENREF_1)]. En la modernidad los usuarios necesitan acceder a muchos servicios digitales imprescindibles para su vida cotidiana tales como: (correo, cuentas bancarias, redes sociales, dispositivos móviles, servicios de comunicaciones, etc. Estos servicios deben poseer seguridad y confiabilidad, siendo necesario algún tipo de autenticación que identifique al usuario de forma única. Los métodos de autenticación se clasifican según [[2](#_ENREF_2)] en: sistemas basados en conocimiento, *Tokens* o información Biométrica. Los basados en conocimiento pueden ser alfanuméricos o gráficos

En [[3](#_ENREF_3)] se plantea que las contraseñas alfanuméricas son las más usadas pero tienen una contradicción entre su seguridad y su usabilidad, pues para ser seguras deben ser aleatorias, largas y no predecibles, mientras que para ser usables deben ser memorizables, esta contradicción suele denotarse como *The Password Problem*. Estas debilidades reducen el espacio de búsqueda y las hacen predecibles y vulnerables a diversos ataques basados en modelos probabilístico[[4](#_ENREF_4)]. La autenticación gráfica surge como alternativa a las contraseñas alfanuméricas. En algunos trabajos como [[5](#_ENREF_5), [6](#_ENREF_6)] se justifica su uso a partir del criterio de que las personas son más hábiles para recordar imágenes que texto.

Los sistemas de autenticación gráfica basan su funcionamiento en que los usuarios reconozcan patrones de imágenes o partes de ellas, en lugar de recordar largas y complejas secuencias de caracteres. Las contraseñas gráficas según varios actores [[7](#_ENREF_7), [8](#_ENREF_8)] se clasifican en tres tipos: *Recognition Based Technique*, *Recall Based Technique* y *Cued-Recall Based Technique*. En [[9](#_ENREF_9)] se presenta un resumen muy completo hasta 2011 de los sistemas de autenticación basados en contraseñas gráficas, Shendage Swapnil en [[10](#_ENREF_10)] se enfoca en los sistemas del tipo *Cued click points*. Uno de los primeros exponentes de los esquemas *Cued click points* es el *PassPoint*, quien ha recibido una atención especial [[4](#_ENREF_4)], es una de las técnicas basadas en la memorización de puntos como contraseña. La técnica de *PassPoint* requiere que el usuario seleccione como contraseña gráfica cincos puntos de una imagen escogida durante la fase de registro, y luego en la fase de autenticación tiene que seleccionar 5 puntos dentro de la radio de tolerancia definida para cada punto en el registro. Si todos los puntos se encuentran dentro de la región de tolerancia correspondiente entonces el usuario se autentica satisfactoriamente. Para lograr que cada punto de la contraseña pertenezca a una región de tolerancia surge la discretización de la imagen.

Estudios realizados por Wiedenbeck en [[11](#_ENREF_11)] plantean la influencia del tamaño de región de tolerancia en la eficacia del modelo original, además de lo difícil que resulta seleccionar con precisión los posibles puntos de la contraseña, y además se conoce que algunas personas mayores o poco acostumbradas a utilizar la computación presentan dificultades para manipular con precisión dispositivos con el mouse.

Wiedenbeck en [[12](#_ENREF_12)] propone la técnica PassPoints, un sistema de contraseñas gráficas nuevo y más seguro, donde realizaron un diseño experimental donde los participantes crearon y practicaron una contraseña gráfica o alfanumérica, realizaron 3 pruebas longitudinales para ingresar su contraseña en el transcurso de 6 semanas, las secciones ocurrieron en la semana 1, 2 y 6. Los resultados muestran que los usuarios de contraseñas gráficas crearon una contraseña válida con menos dificultades que los usuarios alfanuméricos, mientras practicaban sus contraseñas. Los participantes fueron 40 miembros de una comunidad universitaria, incluidos personal, estudiantes y profesores. El rango de edad fue de 20 a 55 años.

Wiedenbeck en [[11](#_ENREF_11)] realizó un estudio sobre el efecto de la tolerancia en las contraseñas gráficas usando la técnica PassPoints, donde participaron 32 estudiantes de pregrado, 10 mujeres y 22 hombres. La edad media de los participantes fue de 22,7 y usaban computadoras con frecuencia. Además realizó otro experimento sobre el efecto de las imágenes en las contraseñas gráficas de 83 estudiantes de pregrado, 42 hombres y 8 mujeres. La edad media fue 22.7 y usaban computadoras con frecuencia.

Chiasson en [[1](#_ENREF_1)] propone la técnica *Cued Click Points(CCP),* un sistema de contraseñas gráficas de recuperación de claves. Los usuarios hacen clic en un punto por una secuencia de imágenes. Las siguientes imágenes se basan en el punto anterior. Además realizaron un estudio con 24 participantes con un rango de edad de 17 a 26 años, eran estudiantes universitarios de diversa procedencia, ninguno estudiaba específicamente seguridad informática, pero todos eran usuarios regulares en la web. Los participantes completaron una sesión de una hora en el laboratorio y en total se realizaron 257 ensayos reales.

Chiasson en [[13](#_ENREF_13)] propone la técnica *Persuasive Cued Click-Points* (PCCP) como mejora al CCP y busca persuadir al usuario en la selección de puntos más aleatorios. Además realizaron un estudio de 39 participantes con un rango de edad entre de 17 a 37 años, la mayoría eran estudiantes universitarios de diversos campos, todos eran usuarios regulares con computadoras que se sentían cómodos con las contraseñas y el mouse. Los participantes completaron una sesión de una hora en el laboratorio y en total se recopilaron 307 ensayos.

Dirik en [[14](#_ENREF_14)] proponen un modelo que predice probabilidades de posibles puntos de clic, esto permite predecir la entropía de un punto de una contraseña gráfica para una imagen dada. El modelo permite evaluar automáticamente si una imagen determinada es adecuada para el sistema PassPoints y analizar posibles ataques de diccionario contras el sistema. Se comparó las predicciones proporcionada por el modelo con los resultados de experimentos que involucran a usuarios humanos. Los experimentos realizados fueron pequeños y limitados. Los participantes eran en su mayoría estudiantes de pregrado y posgrado a quienes se les pidió que crearan una contraseña gráfica en una de dos imágenes ya definidas.

Como plantea Legón, en su tesis de maestría [[4](#_ENREF_4)], en la actualidad existen pocos modelos probabilísticos para el uso en las contraseñas gráficas. La mayoría de estos modelos están dirigidos al tratamiento digital de imágenes. Estos modelos son usados para pronosticar la contraseña más probable en cada imagen, además son usados para escoger las imágenes más adecuadas o para desarrollar ataques de diccionarios. En [[4](#_ENREF_4)] se propone un modelo probabilístico para el uso en el esquema de autenticación gráfica *PassPoint*, el cual tiene como objetivo cuantificar la autenticidad de los usuarios, para ello asigna una probabilidad de que este sea un usuario legítimo. Dicho modelo no puede ser empleado en sistemas que emplea la discretización robusta ya que este necesita conocer las coordenadas originales de las contraseñas en el momento del registro, esto hace imposible su implementación en el esquema *PassPoint*, el cual basa su funcionamiento en el uso de la discretización robusta. Al usar esta discretización no es posible durante la autenticación recuperar las coordenadas exactas de los puntos de la contraseña registrada. Este modelo puede implementarse satisfactoriamente si se utiliza la discretización centrada, la cual si permite durante la autenticación del usuario determinar las coordenadas de los puntos originales de la contraseña. Las pruebas realizadas a este modelo se ejecutaron utilizando un solo usuario, generando de forma aleatoria diferentes intentos de autenticación del mismo. Identificando en cada caso el nivel de legitimidad de cada intento utilizando el modelo probabilístico. Además los límites del estadígrafo para establecer las clasificaciones de los usuarios se obtuvieron de forma experimental, y no utilizando una distribución probabilística del estadígrafo propuesto. Esto se debe a que no se contaba con una implementación del modelo que pudiera registrar los intentos de autenticación de los usuarios y poder estimar la distribución de probabilidad del estadígrafo a partir de estos datos.

En la mayoría de los experimentos se tuvieron en cuenta las variables: el sexo, la edad, el nivel educativo de las personas, el tiempo, el radio de tolerancia y la imagen. Wiedenbeck en [[11](#_ENREF_11)] realizó un análisis con técnicas estadísticas para comprender el efecto de diferentes tamaños de tolerancia alrededor de los puntos de la contraseña y los efectos que tienen las imágenes en el rendimiento del usuario. En los demás experimentos no hicieron un análisis sobre la influencia de las variables en el resultado de la autenticación. Es importante saber la influencia de estas variables, una de las técnicas que se utiliza es las técnicas de minería de datos.

Se define entonces como **problema** de la investigación: ¿Cómo obtener modelos que permitan analizar el comportamiento de las variables: el sexo, la edad, el nivel educativo de las personas, el tiempo, el radio de tolerancia y la imagen en la estimación de la legitimidad en el modelo probabilístico de autenticación propuesto en [[4](#_ENREF_4)]?

El **objeto de estudio** de este trabajo son los sistemas de autenticación y la minería de datos.

El **campo de acción** es la detección de patrones en los intentos de autenticación realizados por los usuarios.

**Objetivo general**: Identificar patrones entre las variables: el sexo, la edad, el nivel educativo de las personas, el tiempo, el radio de tolerancia y la imagen que influyen en la estimación de la legitimidad del modelo probabilístico de autenticación gráfica propuesto en [[4](#_ENREF_4)].

**Objetivos específicos**:

1- Analizar el estado del arte de los métodos de autenticación gráficos y los experimentos realizados en este tipo de autenticación.

2- Diseñar los experimentos a realizar al modelo probabilístico *Passpoint* con vistas a identificar las características de los usuarios y las imágenes utilizadas por este en la efectividad del modelo para determinar legitimidad del usuario

3- Ejecutar los experimentos y con estos poblar la base de datos de intentos de autenticación utilizando el modelo probabilístico de *PassPoint* teniendo en cuenta los criterios que se evaluarán en los experimentos.

4- Análisis exploratorio de los datos de autenticación obtenidos durante los experimentos.

5- Aplicar diferentes técnicas de Minería de Datos a los experimentos para identificar las características de los usuarios y las imágenes empleadas por estos que influyen en los niveles de autenticidad de un usuario empleando el modelo probabilístico *Passpoint*.

6- Interpretar y evaluar los resultados obtenidos en los experimentos

**Preguntas:**

¿Cómo realizo los experimentos?

¿Qué técnicas de minería de datos le aplico a los experimentos?

¿Cómo influyen la edad, la profesión y el radio de tolerancia en los tipos de autenticación?

El **valor práctico** de este proyecto está dado en el marco práctico. Se aplican algoritmos de minería de datos a la base de datos de los intentos de autenticación con el objetivo de obtener las características de los usuarios que más influyen en la precisión del modelo, siendo así **el artefacto de salida** un flujo de trabajo KNIME que permite obtener los patrones de los intentos de autenticación.

Este trabajo está divido en tres capítulos. El capítulo 1 se describe el marco teórico que contiene los conceptos fundamentales relaciones con la autenticación gráfica y la minería de datos, en el capítulo 2 se explica el flujo de trabajo KNIME con el que se aplican los algoritmos pertinentes a las base de datos de los intentos de autenticación para obtener reglas y clasificadores. Finalmente el capítulo 3 se describe los resultados obtenidos de los algoritmos.

# Capítulo 1 Fundamentos Teóricos

## Introducción

En el presente capítulo se tratan los aspectos teóricos fundamentales para desarrollar la propuesta del trabajo. Se exponen las principales técnicas de autenticación gráfica, los experimentos realizados en los modelos y los conceptos de minería de datos.

## 1.2 Contraseñas Gráficas

Los métodos de autenticación se clasifican en tres tipos: Los basados en token, los basados en información biométrica y los basados en conocimiento [[15](#_ENREF_15)] como se muestra en la figura 1.

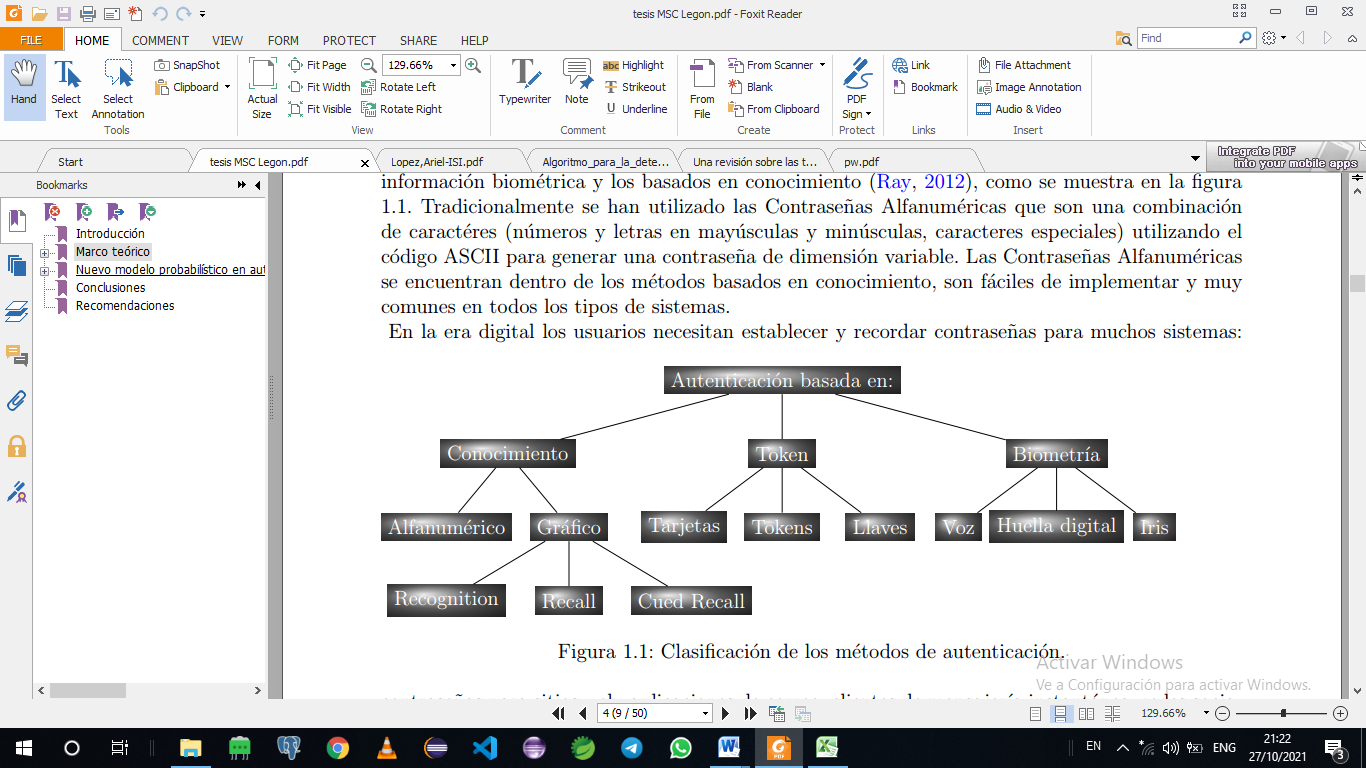


Figura 1 Clasificación de los métodos de autenticación

En la era digital los usuarios necesitan establecer y recordar contraseñas para muchos sistemas, las contraseñas alfanuméricas son las más usadas y para ser seguras deben ser aleatorias, largas y no predecibles, lo cual trae como consecuencia que sea difícil de recordar. Los usuarios casi siempre ignoran las recomendaciones para elaborar una contraseña segura y tienden a ser cortas y pocas aleatorias, y se convierten en contraseñas débiles siendo vulnerables a ataques de diccionarios.

Por otra parte, los métodos de autenticación basados en información biométrica son utilizados en escenarios muy específicos dado que requieren de la utilización de dispositivos de hardware dedicado para capturar los datos y el proceso de captura y autenticación puede ser largo y molesto para los usuarios.

Las Contraseñas Gráficas se clasifican dentro de los Sistemas de Autenticación Basados en Conocimiento. La idea original fue planteada por Blonder [[16](#_ENREF_16)], para los humanos es más fácil reconocer patrones que un carácter alfanumérico complejo, por tanto hace que esta técnica tenga ventajas sobre las contraseñas alfanuméricas y a su vez sea resistente a los ataques clásicos a los que estas son vulnerables [[10](#_ENREF_10)].

Las contraseñas gráficas pueden ser clasificadas en tres grupos: *Recognition Based Technique, Recall Based Technique y Cued-Recall Based Technique* [[7](#_ENREF_7), [8](#_ENREF_8)].

Recognition Based Technique (Técnicas basadas en el reconocimiento de patrones en imágenes). Este tipo de técnicas se basa en el principio del reconocimiento de patrones de imágenes siendo una de las más sencillas para la memoria humana. Se le presenta al usuario un grupo de imágenes en orden aleatorio y de composición también aleatoria. El usuario debe seleccionar las que considere para formar su clave, luego para autenticarse debe recordar y repetir su elección dentro de un grupo mayor de imágenes aleatorias.

Recall Based Technique (Técnicas basadas en la memorización de patrones). También conocidos como Pure Recall Based Techniques. En este tipo de sistemas los usuarios escriben su contraseña en un lienzo o sobre una imagen. El usuario reproduce en forma de dibujo lo que ha establecido en el proceso de registro.

Cued-Recall Based Technique (Técnicas basadas en la memorización de puntos contraseñas). En estos sistemas el usuario tiene que memorizar un conjunto de puntos en áreas de la imagen, en la idea original de Blonder [[16](#_ENREF_16)], el usuario debe marcar con un mouse determinados puntos de la imagen, si lo hacía de forma correcta seria aceptado por el sistema de lo contrario es rechazado, pero con el tiempo se han hechos evoluciones de estas técnicas que se muestra en la figura 2.

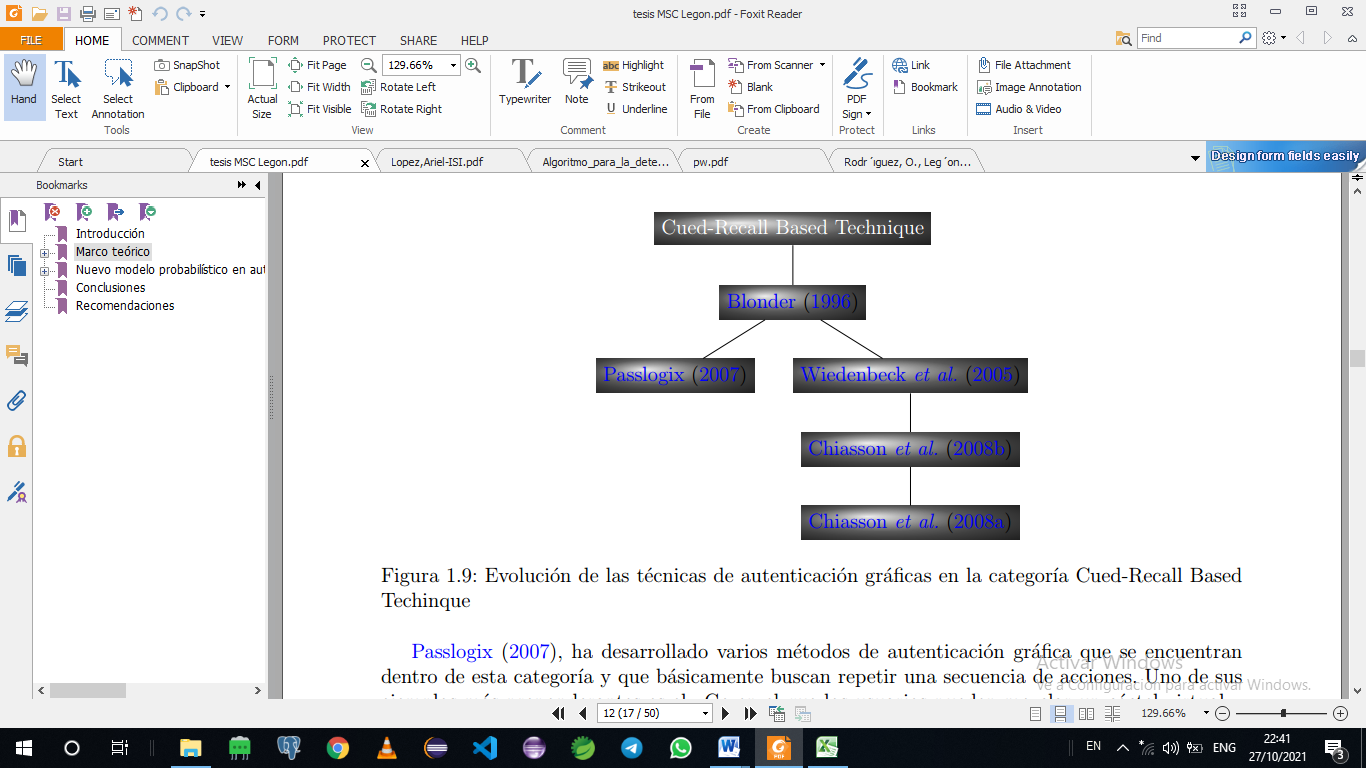


Figura 2 Evolución de las técnicas basadas en la memorización de puntos contraseñas

Passlogix [[17](#_ENREF_17)], ha desarrollado varios métodos de autenticación gráfica que se encuentran dentro de esta categoría y que básicamente buscan repetir una secuencia de acciones. Sus propuestas poseen un espacio de contraseñas pequeñas y no existe forma de prevenir que el usuario escoja contraseñas fáciles de adivinar.

Entre las técnicas de autenticación gráfica, *el PassPoint* ha recibido atención especial [[4](#_ENREF_4)]. Passpoint Wiedenbeck [[12](#_ENREF_12)] extiende el funcionamiento del trabajo inicial de Blonder. En este método puede ser utilizada cualquier imagen (pinturas, fotos naturales, fotos familiares, fotos de arquitectura, etc.) lo que hace al sistema más flexible. Las imágenes podían ser seleccionadas por el usuario o proveídas por el sistema. Wiedenbeck [[12](#_ENREF_12)] en su investigación asegura que las imágenes más deseadas para el proceso debían tener contenido que tuviera significado para el usuario por lo que estas debían contener escenas concretas. También era un requisito adicional que las imágenes seleccionadas fueran intrincadas en el contenido y que tuvieran cientos de puntos memorables diseminados de forma homogénea. El usuario podía seleccionar en estas imágenes cualquier conjunto de puntos para crear su contraseña. Luego en el proceso de autenticación este debe seleccionarlos en el orden preciso, con un margen de error (tolerancia) alrededor de cada punto. Como una imagen puede contener cientos de miles de puntos, el espacio teórico de este sistema es suficientemente grande. Según Wiedenbeck [[12](#_ENREF_12)] no se necesitarían muchos puntos para hacer la contraseña segura, con 5 o 6 se pueden lograr más contraseñas que con 8 caracteres dentro de un alfabeto estándar de 64 letras. Para que el proceso de autenticación fuera efectivo y rápido para el usuario debía existir una tolerancia asociada a cada punto (aproximadamente 0, 25cm). Además por razones de seguridad se requiere que el sistema no almacene la contraseña de forma explícita. Passpoint almacena un hash con la concatenación de los puntos de la contraseña.

Cued Click Points (CCP) Chiasson [[1](#_ENREF_1)] se propone por los autores como una alternativa más eficiente que el Passpoint, basa su funcionamiento en la idea original de Blonder pero sugiere la utilización de una secuencia de imágenes (total de 5) y de un solo punto en cada una de ellas. La secuencia de imágenes variaría dependiendo de los puntos que el usuario seleccione. Cuando el usuario establece su contraseña, selecciona un punto en cada una de un grupo de 5 imágenes que se le muestran de forma consecutiva. Para el proceso de autenticación este debe seleccionar correctamente los puntos que estableció con anterioridad en el proceso de registro de la contraseña. Si selecciona en cada imagen el punto correcto se le muestran las imágenes que el escogió en el proceso de registro (algo que solo él conoce). De lo contrario se le guiará por un grupo de imágenes distinto hasta que al finalizar se le informará que el proceso de autenticación ha sido incorrecto. Al aumentar el número de imágenes y solo un punto en cada una de ellas aumenta considerablemente el espacio de contraseña.

Persuasive Cued Click-Points (PCCP) Chiasson [[13](#_ENREF_13)] se proponen como mejora al CCP y busca persuadir al usuario en la selección de puntos más aleatorios. En principio es muy parecido al CCP, solo que en el proceso de creación de la contraseña en la imagen se muestra un cuadro que resalta un ´área aleatoria limitando el espacio donde el usuario puede seleccionar el punto. Existirá un botón que variará a selección del usuario la localización del cuadro. Para el proceso de autenticación se mantiene la idea original.

## 1.2 Evaluación experimental de las técnicas Cued-Recall Based

Wiedenbeck en [[12](#_ENREF_12)] realizaron un diseño experimental donde los participantes crearon y practicaron una contraseña gráfica o alfanumérica, realizaron 3 pruebas longitudinales para ingresar su contraseña en el transcurso de 6 semanas, las secciones ocurrieron en la semana 1, 2 y 6. Los participantes fueron 40 miembros de una comunidad universitaria, incluidos personal, estudiantes y profesores. El rango de edad fue de 20 a 55 años, 23 fueron hombres y 17 mujeres. En la semana 1 los participantes primero crearon una contraseña, luego la aprendieron al ingresarla varias veces. En la fase de creación de contraseñas se midió el número de intentos y la cantidad de tiempo para crear una contraseña válida. En la fase de aprendizaje se midió el número de intentos y la cantidad de tiempo para cumplir con un criterio de aprendizaje fijo. Después los participantes completaron un cuestionario sobre sus percepciones del sistema de contraseña que utilizaron. Finalmente al final de la primera sesión, los participantes realizaron su primera prueba de retención. La segunda sesión tuvo lugar en la semana 2, esta fue la segunda prueba de retención de los participantes. La tercera sesión en la semana 6 donde fue la última prueba de retención. En cada prueba de retención se midió con técnicas estadísticas la cantidad de intentos y la cantidad de tiempo para que los participantes ingresen una contraseña válida. En este experimento se utilizó una sola imagen con muchos elementos que podían servir como puntos de clic memorable, el cuadrados de la tolerancia que se estableció fue de 20x20 píxeles. Los resultados muestran que los usuarios de contraseñas gráficas crearon una contraseña válida con menos dificultades que los usuarios alfanuméricos, mientras practicaban sus contraseñas. Sin embargo los usuarios tardaron más e ingresaron más contraseñas no válidas que los usuarios alfanuméricos mientras practicaban sus contraseñas.

Chiasson en [[1](#_ENREF_1)] realizaron un estudio con 24 participantes, 12 eran mujeres y 12 hombres, tenían un rango de edad de 17 a 26 años. Eran estudiantes universitarios de diversa procedencia, ninguno estudiaba específicamente seguridad informática, pero todos eran usuarios regulares en la web. Los participantes completaron una sesión de una hora en el laboratorio, primero leyeron y firmaron el formulario de consentimiento y se les dio una introducción a las tareas que completarían durante la sesión. Esto incluyo mostrarles una imagen de ejemplo con cuadros superpuestos, demostrando cual preciso debían ser al volver a ingresar sus puntos. Se utilizó un conjunto de 330 imágenes y el cuadro de tolerancia utilizado en este estudio fue de 19x19 píxeles. También explicaron que la secuencia de la imagen dependía de donde hicieran clic en la imagen actual, se les dijo que si de repente veían una imagen que no reconocieran durante la fase de confirmación o inicio de sesión, entonces probablemente estaban en el camino equivocado. Los participantes completaron dos ensayos de práctica seguidos, en total se completaron 257 ensayos reales. Un ensayo consta de 5 pasos: fase de creación, fase de confirmación, dos preguntas, completar un rompecabezas de prueba de rotaciones mentales, fase de inicio de sesión. Las fases 1, 2 y 5 representan las fases de contraseñas utilizadas en análisis posteriores. Fase de creación (2): cree una contraseña haciendo clic en un punto de cada una de las cinco imágenes seleccionada por el sistema presentadas en secuencia. Fase de confirmación (1): confirme esta contraseña volviendo a introducirla correctamente, los usuarios que confirmaban correctamente su contraseña podían volver a intentar la confirmación o regresar al paso 1, una nueva contraseña comenzaba con la misma imagen inicial, pero generalmente incluía diferentes imágenes dependiendo de los puntos del clic. Fase de inicio de sesión (5): Inicie sesión con su contraseña actual, si los usuarios notaron un error durante el inicio de sesión, podrían cancelar su intento de inicio y volver a intentarlo. Alternativamente, si no sabían su contraseña, podían crear una nueva contraseña, regresando efectivamente al paso 1 de la prueba con la misma imagen inicial como punto de partida. Si los usuarios se sentían demasiados frustrados con las imágenes en particular como para volver a intentarlo, podían omitir esta prueba y pasar a la siguiente. En total 201 de 257 (79%) fueron completados sin reinicio en ninguna fase. Las tasas de éxitos fueron altas para todas las fases. El rendimiento fue muy bueno en términos de velocidad, precisión y números de errores, los usuarios prefirieron CCP a PassPoint, diciendo que seleccionar y recordar un punto por imágenes era más fácil, y que ver cada imagen activaba su memoria de donde estaba ubicado el punto correspondiente. También CCP proporciona mayor seguridad que PassPoint porque la cantidad de imágenes aumenta el espacio de búsqueda.

Chiasson en [[13](#_ENREF_13)] realizaron un estudio de 39 participantes con un rango de edad entre de 17 a 37 años, la mayoría eran estudiantes universitarios de diversos campos, todos eran usuarios regulares con computadoras que se sentían cómodos con las contraseñas y el mouse. Los participantes completaron una sesión de una hora en el laboratorio y en total se recopilaron 307 ensayos. Un ensayo consta de 5 pasos: fase de creación, fase de confirmación, dos preguntas, completar un rompecabezas de prueba de rotaciones mentales, fase de inicio de sesión. Se utilizó un conjunto de 347 imágenes, con una ventana gráfica era un cuadro de 75x75 píxeles. Los registros del sistema registraron las coordenadas del punto en cada imagen, la ubicación de la ventana gráfica para cada reproducción aleatoria y las marcas de tiempo para cada acción del usuario. A los participantes se les presento primero el sistema y se les dijo que crearían contraseñas gráficas. Además, se les indicó que simulen que estas contraseñas protegen su información bancaria y, por lo tanto, deben seleccionar contraseñas que sean memorables pero difíciles de adivinar para otros. Se les dijo que la ventana gráfica era una herramienta para ayudarlos a seleccionar contraseñas más seguras, pero que podían barajar tantas veces como quisieran encontrar un punto adecuado. Las tasas de éxitos fueron altas para todas las fases. Las tasas de éxito de inicio de sesión de CCP y PCCP no son significativamente diferentes. Este esquema reduce significativamente los puntos de acceso mientras mantiene su usabilidad.

Wiedenbeck en [[11](#_ENREF_11)] realizó un estudio para comprender el efecto de diferentes tamaños de tolerancia alrededor de los puntos de la contraseña con 32 estudiantes de pregrado, 10 mujeres y 22 hombres. La edad media de los participantes fue de 22,7 y usaban computadoras con frecuencia. La tolerancia puede variar en este sistema, la pregunta es como la tolerancia afecta el éxito en las contraseñas gráficas. Se utilizó una sola imagen con dos tolerancia una de 10x10 y otra de 14x14 pixeles. Los participantes fueron asignados aleatoriamente a las dos condiciones de tolerancia. La sesión comenzó con una presentación explicando el experimento y los conceptos de contraseñas gráficas. Pasando al experimento, las instrucciones en la pantalla guiaron a los participantes a crear una contraseña válida de 5 puntos, se les dijo que ellos tendrían que recordar los puntos y en el orden que fueron de entrada. Luego viene la fase de aprendizaje, para reforzar la contraseña el participante ingreso la contraseña repetidamente hasta que logró 10 entradas correctas. La retención de la contraseña se midió al final de la primera sesión y una semana después. El participante tuvo que ingresar la contraseña correctamente una vez, si ingresaba una contraseña incorrecta se le indicaba que la volviera a ingresar sino termino la sesión. Se utilizaron pruebas de ANOVA para el análisis con la tolerancia. Los resultados muestran que la memoria precisa para la contraseña se reduce considerablemente cuando se usa una tolerancia pequeña alrededor de los puntos de la contraseña del usuario. Esto puede ocurrir porque los usuarios no pueden codificar los puntos de la contraseña en la memoria de la manera precisa que es necesaria para recordar la contraseña durante un lapso de tiempo.

Además realizó otro experimento para comprender los efectos que tienen las imágenes en el rendimiento del usuario con 83 estudiantes de pregrado, 42 hombres y 8 mujeres. La edad media fue 22.7 y usaban computadoras con frecuencia. La pregunta es como la variación de la imagen afecta el éxito en el uso de contraseñas gráficas. La configuración del sistema fue la misma que en el estudio de la tolerancia pero solo se utilizó una tolerancia de 20x20 píxeles y 4 imágenes. Los participantes fueron asignados aleatoriamente a las 4 condiciones de la imagen. Se utilizaron pruebas de ANOVA los análisis con la imagen. Los resultados indican que hubo pocas diferencias significativas en el rendimiento de la imagen, esto sugiere que muchas imágenes pueden admitir la memorización en sistemas de contraseñas gráficas.

Dirik en [[14](#_ENREF_14)] proponen un modelo para identificar las regiones más probables deben hacer click para crear contraseñas gráficas. Los experimentos realizados fueron pequeños y limitados. Los participantes eran en su mayoría estudiantes de pregrado y posgrado a quienes se les pidió que crearan una contraseña gráfica en una de dos imágenes ya definidas. A cada usuario se le presento una de las dos imágenes y se le pidió que seleccionaran una contraseña que conste de 5 clics, que no fuera fácil de adivinar, pero que deberían recordar. Para asegurarse de que los usuarios ingresaran contraseñas realistas, se les pidió que volvieran a ingresar la contraseña. En este experimento solo se utilizaron contraseñas reingresadas correctamente. También se usó el modelo para predecir tantos puntos como se obtuvieron en el experimento. Para probar el rendimiento del modelo de predicción de las distribuciones de probabilidad de las posiciones de los puntos de los clics, se aplicó la prueba de hipótesis de bondad de ajuste de Kolmogorov-Smirnov. El modelo permite evaluar automáticamente si una imagen determinada es adecuada para el sistema PassPoints y analizar posibles ataques de diccionario contra el sistema.

En la mayoría de los experimentos se tuvieron en cuenta las variables: el sexo, la edad, el nivel educativo de las personas, el tiempo, el radio de tolerancia y la imagen. Wiedenbeck en [[11](#_ENREF_11)] realizó un análisis con técnicas estadísticas para comprender el efecto de diferentes tamaños de tolerancia alrededor de los puntos de la contraseña y los efectos que tienen las imágenes en el rendimiento del usuario. En los demás experimentos no hicieron un análisis de las variables en la calidad del resultado. Es importante saber la influencia de estas variables, una de las técnicas que se utiliza es las técnicas de minería de datos.

## 1.3 Minería de Datos

El proceso de hurgar en los datos para descubrir conexiones ocultas y predecir tendencias futuras tiene una larga historia. Conocido algunas veces como “descubrimiento de conocimientos en base de datos”, el término “minería de datos” no se acuño hasta la década de 1990. Pero su base comprende tres disciplinas científicas entrelazadas: estadística (el estudio numérico de relaciones de datos), inteligencia artificial (inteligencia similar a la humana exhibida por software) y machine learning (algoritmos que pueden aprender de datos para hacer predicciones). Lo que era antiguo es nuevo otra vez, ya que la minería de datos continua evolucionando para igualar el ritmo del potencial sin límites del big data y poder de cómputo asequible.

En la última década, los avances en el poder y la velocidad de procesamiento nos han permitido llegar más allá de las prácticas manuales, tediosas y que toman mucho tiempo al análisis rápido, fácil y automatizado. Cuanto más complejo son los conjuntos de datos recopilados, mayor es el potencial que hay para descubrir insights relevantes.

En [[18](#_ENREF_18)] definen la minería de datos como “el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos”. Empleando una amplia variedad de técnicas, puede utilizar esta información para incrementar sus ingresos, recortar costos, mejorar sus relaciones con clientes, reducir riesgos y más. Esta ciencia es un paso dentro de un proceso más amplio cuyo objetivo es el Descubrimiento de Conocimiento en Bases de Datos, (*Knowledge Discovery in Database,* KDD) [[18](#_ENREF_18)].

Según [[19](#_ENREF_19)] KDD es “el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en una última instancia, comprensibles a partir de los datos”. El KDD es un proceso complejo que incluye no solo la obtención de los modelos o patrones (el objetivo de la minería), sino también la evaluación y posible interpretación de los mismos.

### 1.3.1 Las fases del proceso de extracción del conocimiento

El proceso se KDD se organiza entorno a cinco fases como se muestra en la figura 3. En la fase **de integración y recopilación de los datos** se determinan las fuentes de información que son necesarias para el análisis. A continuación, se eliminan o corrigen los datos incorrectos y se decide la estrategia a seguir con los datos incompletos; además, se seleccionan los atributos a los que se les realizará el análisis. A los datos resultantes de esta fase se les conoce como **fase de selección, limpieza y transformación**. A las dos primeras fases se suele englobar bajo el nombre “preparación de los datos”. En la **fase de minería de datos** se decide cuál es la tarea que se va a realizar (agrupar, clasificar, etc.), y se elige el método que se va a utilizar, además se obtiene un conjunto de patrones que describen el comportamiento de los datos. En la **fase de evaluación e interpretación** se evalúa y analiza el modelo obtenido, en caso de no mostrar los resultados esperados se vuelve a una de las fases anteriores para una nueva iteración. En la última **fase difusión** se hace uso del nuevo conocimiento a través de analistas de datos y se hace partícipe de él a todos los posibles usuarios.

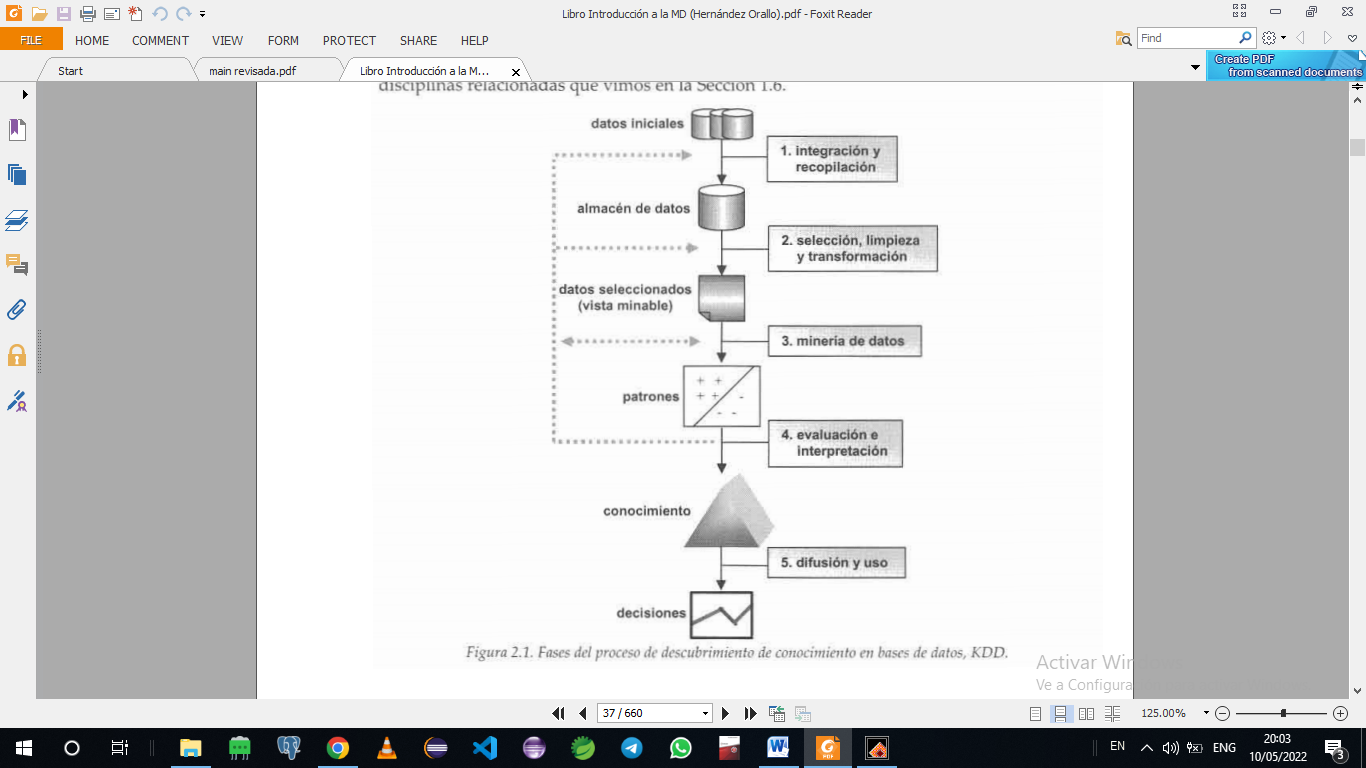


Figura 3 Fases del proceso de descubrimiento de conocimiento.

A partir de [[18](#_ENREF_18)] se muestra a continuación la explicación de las fases.

**Fase de integración y recopilación**

En esta fase se reconocen y reúnen los datos con los que se trabajará en el futuro. Estos datos pueden proveer de 2 fuentes diferenciadas: fuentes internas y fuentes externas.

Las fuentes internas son las más fáciles de obtener y representa las bases de datos transaccionales propias de la organización utilizadas tradicionalmente para el procesamiento transaccional en línea u OLTP (*On-Line Transaction Processing*), aunque pueden no ser utilizadas para tal propósito.

Las fuentes externas representan la información que no pertenece al ámbito de las operaciones de la empresa pero que son importantes para el análisis. En muchos casos habrá que adquirir datos desde bases de datos públicas como puede ser el censo, los datos demográficos o climatológicos o desde otras bases de datos privadas a la que la organización tenga acceso.

A este respecto es importante señalar la utilización de data *warehouses* o almacenes de datos que proporciona metodologías y la tecnología para efectuar a la vez procesamiento transaccional en tiempo real (OLTP) y procesamiento analítico en tiempo real (OLAP). Su propósito es la recuperación de información, realizar análisis e informes y suele ser utilizado por usuarios con capacidad de decisión, como pueden ser ejecutivos o analistas.

El objetivo de esta fase es crear un almacén de datos que contenga la información que será utilizada.

**Fase de selección, limpieza y transformación**

Esta fase tiene como objetivo la mejora en la calidad del almacén de datos y trata de mostrar el conjunto de datos de la manera más apropiada para la siguiente fase: la minería de datos.

Extraer la información, tal y como se ha explicado en la fase anterior, de diversas fuentes da lugar a diversos tipos de imperfecciones que será necesario resolver. A un mayor número de imperfecciones en el almacén este verá disminuida su calidad. La tarea de eliminar estas imperfecciones se denomina limpieza.

Un tipo de datos sobre los que hay que efectuar tareas de limpieza son los datos irrelevantes o innecesarios.

También existen valores que no se adecuan al comportamiento general de los datos, los denominados valores *outliers* o valores atípicos.

Otro tipo de datos que pueden llegar a presentar algún tipo de problema son los valores faltantes o missing values.

Además de tratar con estos valores que representan un problema se requiere realizar una transformación de algunos valores de manera que estos se registren siguiendo la misma estructura y las mismas variables. Esta transformación estructuración es debida a que las distintas fuentes de donde se han extraído los valores han podido almacenarlos utilizando diferentes formatos.

La información elaborada de esta fase recibe el nombre de vista minable y será utilizado por la siguiente fase, la fase de minería de datos.

**Fase de minería de datos**

Esta es la fase más característica del proceso KDD. Tanto que, en muchas ocasiones, todo el proceso recibe su nombre. Es la más característica en cuanto a que su aplicación ofrece como resultado conocimiento que puede ser utilizado, pero no por ello es la más importante, ya que todas y cada una de las fases cumple una función esencial para la correcta obtención del conocimiento.

En esta fase se construye un modelo basándose en la vista minable que ha sido obtenida de la fase anterior. Este modelo puede ser utilizado para diferentes fines, por ello es necesario tomar algunas decisiones antes de comenzar con el proceso de minería: Determinar el tipo de tarea apropiado, elegir el tipo de modelo y elegir el algoritmo que más convenga.

Es una fase muy amplia que daría para un estudio individualizado.

**Fase de evaluación e interpretación**

Una vez obtenido el conocimiento y el patrón resultante de la fase de minería, el siguiente paso es evaluar como ese conocimiento se aproxima a la realidad en cuanto a las metas definidas en la primera fase.

La evaluación se realiza con dos conjuntos de datos pertenecientes a la vista minable. Uno de ellos se denomina training-set o conjunto de entrenamiento y el otro se denomina test-set o conjunto de prueba. El *training-set* será el utilizado para construir el modelo y el *test-set* se utilizará para aplicar el modelo obtenido con el *training-set* y comparar tal resultado con la realidad. Esta separación en dos conjuntos es necesaria ya que si se creara el modelo con el mismo conjunto con el que se realiza la prueba, el modelo resultante utilizará la misma información que ha sido utilizada para su construcción en la prueba, obteniendo estimaciones muy optimistas.

Existen diversas técnicas de evaluar el modelo en cuanto a la forma de separar el conjunto de entrenamiento del conjunto de prueba. La más sencilla, denominada validación simple, utiliza un porcentaje de la base de datos como conjunto de prueba y el resto como conjunto de entrenamiento.

Dependiendo de la tarea que realice la minería de datos, la evaluación se hará de forma diferente. Por ejemplo:

* En una clasificación se medirá el número de entradas clasificadas correctamente entre el número de entradas de prueba.
* En una regresión se medirá la distancia (generalmente al cuadrado, que tendrá más en cuenta las distancias más grandes) entre el valor que se ha predicho y el valor real.
* En un agrupamiento se medirá la distancia al punto medio del grupo y la distancia entre grupos.
* En una tarea de reglas de asociación se evaluará de forma separada cada una de las reglas.
* También pueden utilizarse medidas subjetivas como pueden ser interés, la novedad, la simplicidad o la comprensibilidad.

**Fase de difusión y uso**

Esta fase tiene como fin la aplicación correcta del conocimiento adquirido, en forma de modelo, en el dominio para el cual fue creado. Esto es usar el modelo de forma correcta por parte de los usuarios para los que se comenzó el proceso KDD.

La capacidad de un modelo para ser entendida por un usuario se denomina comprensibilidad. Una alta comprensibilidad hará que el usuario conozca el por qué el modelo ha tomado una decisión u otra, es decir, que conozca la técnica que se está utilizando y en qué elementos se basa.

Con la finalidad de que la aplicación sea correcta, en esta fase también se realizan pruebas periódicas sobre el modelo ya que el dominio del problema puede variar en el tiempo haciendo que este quede desactualizado y, en ocasiones, completamente invalidado.

### 1.3.2 Modelos de minería de datos

**Modelo predictivo:** El objetivo es estimar valores futuros o desconocidos de algunas variables de interés a partir de otras variables independientes. Las tareas de minería de datos de modelos predictivos comprenden regresión y clasificación[[18](#_ENREF_18)].

Clasificación: La clasificación asume que hay un conjunto de objetos caracterizados por algún atributo o rasgo que pertenece a distintas clases. La etiqueta de clase es un valor discreto y es conocido para cada objeto. El objetivo de esta tarea es asignar la etiqueta de clase correcta objetos nuevos y sin etiqueta dados los valores de sus atributos. La clasificación es una las tareas más comunes en la minería de datos[[18](#_ENREF_18)].

Regresión: Esta tarea es muy parecida a la clasificación ya que a cada elemento se le asigna únicamente un valor de salida, con la diferencia de que este valor de salida es un valor numérico, es decir, puede ser un valor entero o real [[18](#_ENREF_18)].

**Modelo de descripción:** El objetivo es identificar patrones en los datos que los explican o resumen. Las tareas de minería de datos de modelos descriptivos son la agrupación en clústeres, las reglas de asociación y la correlación [[18](#_ENREF_18)].

Agrupación: El objetivo de esta tarea es agrupar los elementos de un conjunto de objetos, de tal manera que los elementos que formen parte de un mismo grupo tengan características similares. La diferencia con respecto a la clasificación es que en el agrupamiento son los grupos y la pertenencia a dichos grupos lo que se quiere determinar, y además a priori no se sabe cuántos grupos hay ni cómo son[[18](#_ENREF_18)].

Reglas asociativas: Esta tarea es parecida a la de las correlaciones y factorizaciones pero en este caso se utilizan atributos nominales en lugar de numéricos. Se utilizan para descubrir hechos que ocurren en común dentro de un determinado conjunto de datos[[18](#_ENREF_18)].

Correlación: Se usa para examinar el grado de similitud de los valores de dos variables numéricas. Esta tarea resulta muy útil para establecer reglas de elementos correlacionados[[18](#_ENREF_18)].

### 1.3.3 Técnicas y algoritmos de minería de datos (aquí voy a poner las técnicas que voy a explicar)

### 1.3.4 Metodologías utilizadas para la minería de datos

Las metodologías nos permiten llevar a cabo el proceso de minería de datos en forma sistemática y no trivial. Ayudan a las organizaciones a entender el proceso de descubrimiento de conocimiento y proveen una guía para la planificación y ejecución de los proyectos. Una metodología no sólo define las fases de un proceso sino también las tareas que deberían realizarse y cómo llevarlas a cabo. Entre las metodologías dominantes para el proceso de la minería de datos se encuentra CRISP-DM y SEMMA [[20](#_ENREF_20)].

Como se puede observar en la Figura 4, CRISP-DM se ha convertido en la metodología más utilizada, según un estudio publicado en el año 2007 por la comunidad KDnuggets (Data Mining Community's Top Resource).

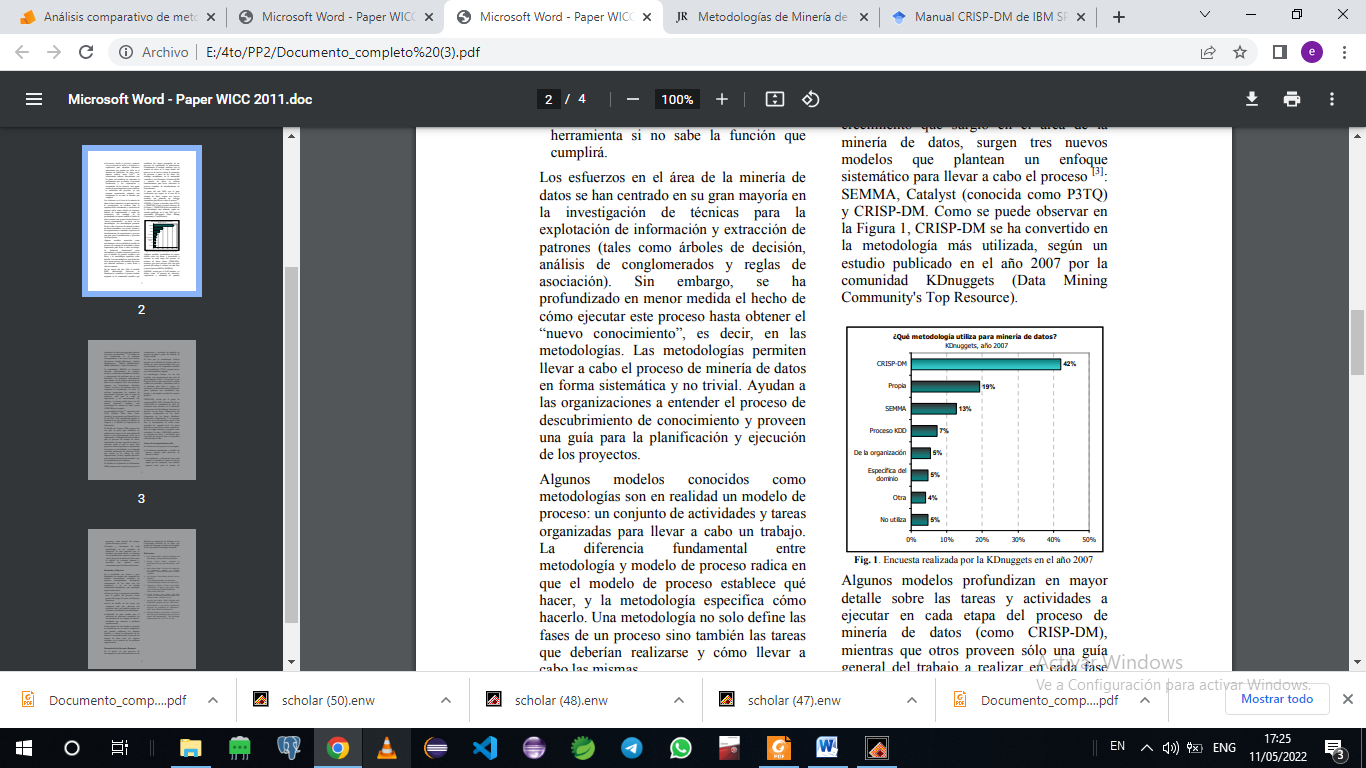


Figura 4 Encuesta realizada por la KDnuggets en el año 2007

**SEMMA**

La metodología llamada SEMMA, desarrollada por el instituto SAS que es el acrónimo de SAMPLE, EXPLORE, MODIFY, MODEL, ASSESS, que podemos interpretar como muestrea, explora, modifica, modela y evalúa, que se refiere al proceso básico para realizar Minería de Datos[[21](#_ENREF_21)] como se muestra en la figura 5. Esto es a partir de una muestra representativa de los datos, se aplican técnicas estadísticas de exploración y visualización, se seleccionan y transforman variables, se modela con las variables para predecir los y se evalúa la exactitud del modelo.

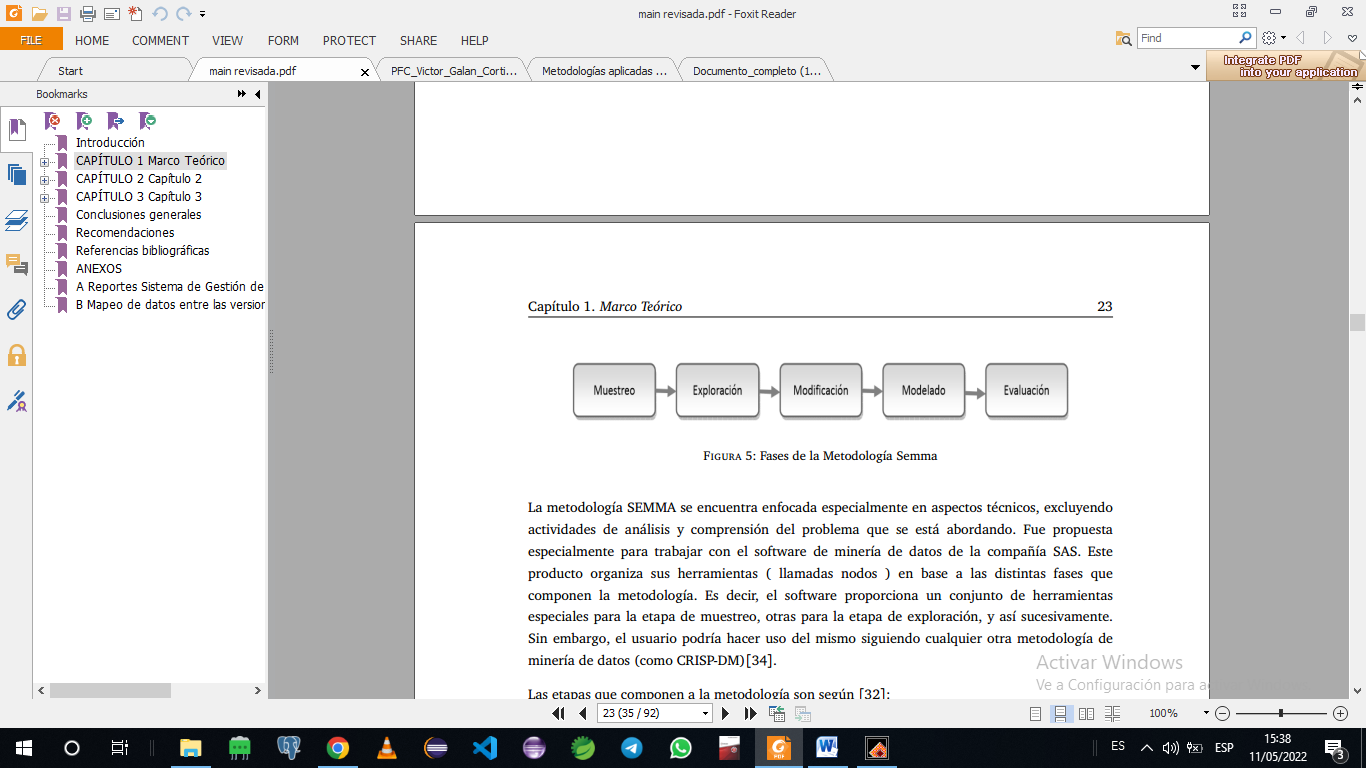


Figura 5 Fases de la metodologíaSEMMA

Las etapas que componen a la metodología son según [[20](#_ENREF_20)] :

1. *Sample* (Muestreo): En esta etapa se toma una muestra del conjunto de datos disponible, que debe ser lo suficientemente grande para contener la información relevante, y lo suficientemente pequeña como para correr el proceso rápidamente. La etapa de muestreo es opcional, aconsejable cuando el tamaño del conjunto de datos es demasiado extenso.

2. *Explore* (Exploración): Consiste en explorar los datos en búsqueda de relaciones y tendencias desconocidas. Es una etapa especial para familiarizarse con los datos, y formular nuevas hipótesis a partir de su análisis.

3. *Modify* (Modificación): Consiste en una etapa de preparación de los datos, donde se limpian los valores anómalos, se realiza un tratamiento de los datos faltantes, y se seleccionan, crean y modifican las variables con las que se trabajarán.

4. *Model* (Modelado): Consiste en la creación del modelo que permitirá predecir las variables de respuesta a partir de las variables explicativas, utilizando algunas de las técnicas predictivas como árboles de decisión, redes neuronales, análisis discriminante o análisis de regresión.

5. *Assess* (Evaluación): En esta fase se evalúa la utilidad y la exactitud de los modelos obtenidos en el proceso de minería de datos, por ejemplo analizando la capacidad predictiva de los mismos.

SEMMA propone que luego de la fase de evaluación, se generan nuevas hipótesis que llevan a repetir el proceso iterativamente como muestra la Figura 6

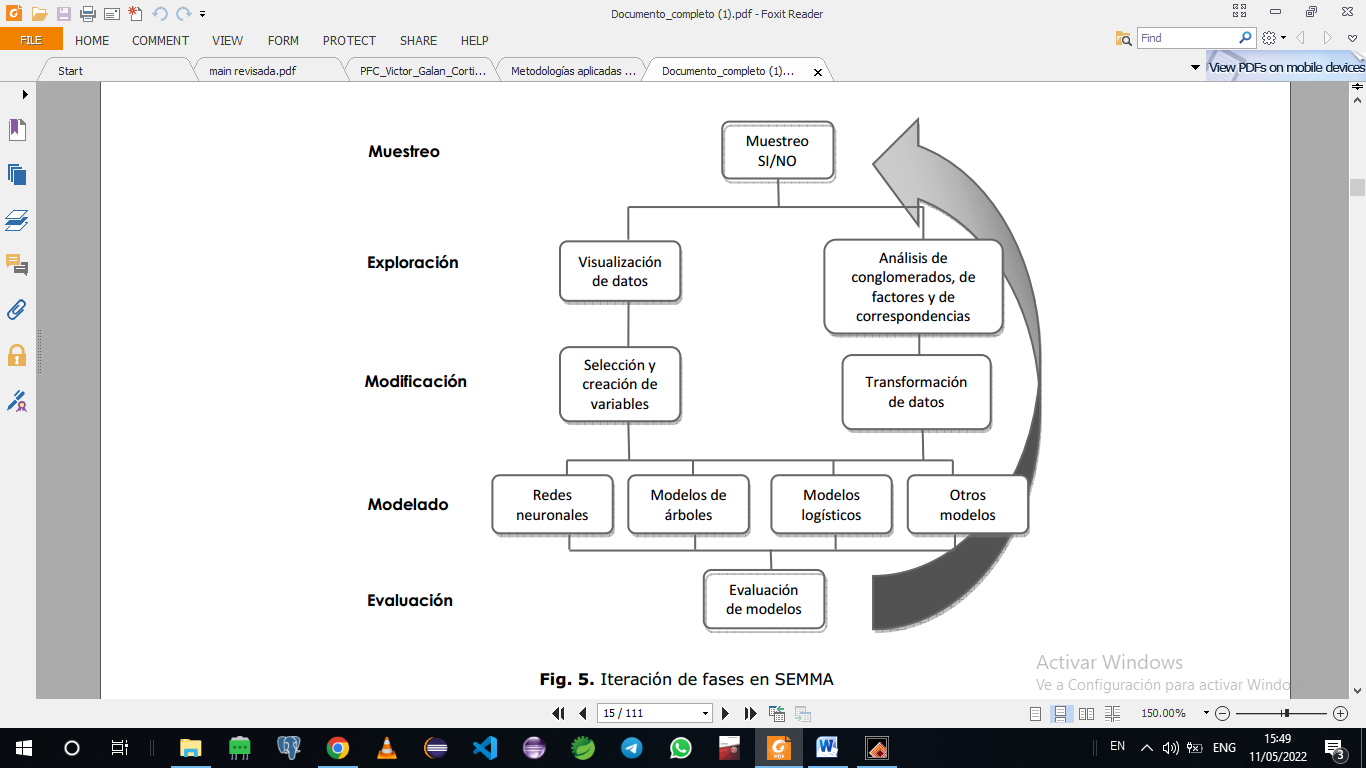


Figura 6 Iteración de las fases de SEMMA

**CRISP-DM**

CRISP-DM (*CRoss-Industry Standard Process for Data Mining*) es una metodología de Minería de datos para desarrollo de proyectos analíticos. CRISP-DM se explica como un proceso jerárquico, que tiene cuatro niveles de abstracción: Fase, tareas generales, tareas específicas e instancias de proceso[[22](#_ENREF_22)].

CRISP-DM, propone en el nivel más alto seis fases para el proceso de minería de datos como se muestra en la figura 7: entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación e implementación. La sucesión de fases, no es necesariamente rígida.

Cada fase se descompone en un conjunto de tareas genéricas (o generales) de segundo nivel. Estas tareas son genéricas ya que tratan de abarcar la mayoría de las situaciones posibles en minería de datos. A partir del tercer nivel de abstracción, se realiza un “mapeo” de las tareas genéricas definidas en el modelo a situaciones específicas. De esta forma, las tareas genéricas se traducen en tareas específicas para casos y proyectos concretos. En el cuarto nivel, encontramos las instancias de proceso, donde se describen las acciones, decisiones y resultados de un proyecto particular de minería de datos.

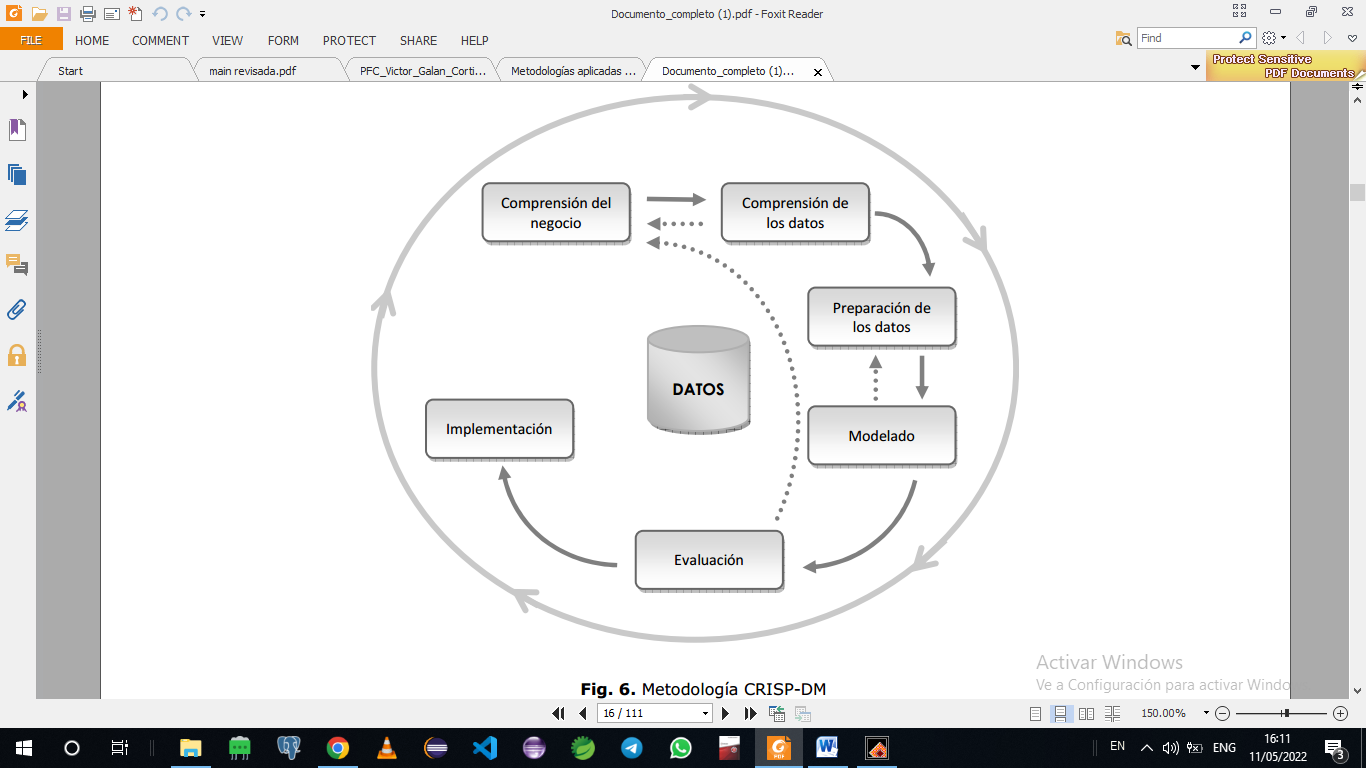


Figura 7 Metodología CRISP-DM

La metodología proporciona un modelo de referencia y una guía de usuario. El modelo de referencia presenta un resumen de las fases y tareas a llevar a cabo en cada una (junto con sus salidas). Es decir, describe “que” debería hacerse en un proyecto de minería de datos. La guía de usuario proporciona sugerencias para la ejecución de cada tarea del modelo de referencia [[22](#_ENREF_22)].

Analizando el nivel más alto de abstracción del modelo, las seis fases que  
componen el proceso de minería de datos son según [[20](#_ENREF_20)]:

* Comprensión del negocio: en esta fase se determinan los objetivos y requerimientos del proyecto desde una perspectiva del negocio, definiendo el problema de minería y el plan de trabajo.
* Comprensión de los datos: fase que consiste en la recolección de datos que se utilizarán en el proyecto y la familiarización con los mismos. En esta etapa es posible el surgimiento de las primeras hipótesis acerca de la información que podría estar oculta.
* Preparación de los datos: comprende aquellas actividades de  
  tratamiento de los datos para construir la vista minable o conjunto de  
  datos final sobre el cual se aplicarán las técnicas de minería.
* Modelado: en esta etapa se aplican las diversas técnicas y algoritmos de minería sobre el conjunto de datos para obtener la información oculta y los patrones implícitos en ellos.
* Evaluación: fase en la que se analizan los patrones obtenidos en función de los objetivos organizacionales. En esta etapa se debería determinar si se ha omitido algún objetivo importante del negocio y si el nuevo conocimiento será implementado, es decir, si se pasará a la próxima etapa.
* Implementación: consiste en la comunicación e implementación del  
  nuevo conocimiento, el cual debe ser representado de forma entendible para el usuario.

Cada una de estas fases generales se compone de un conjunto de tareas en las que se definen las salidas o entregables que se generan.

Si bien la metodología no especifica detalladamente cómo llevar a cabo cada tarea, los consejos de la sección “Guía de Usuario” resultan de mucha utilidad y orientación al momento de ejecutarlas.

**Estudio comparativo**

Tomando como base las fases generales que componen a un proceso de minería de datos (análisis del negocio, selección y preparación de los datos, modelado, evaluación, implementación) se ha llegado a las siguientes conclusiones:

CRISP–DM, creada por el grupo de empresas SPSS, NCR y Daimer Chrysler en el año 2000, es actualmente la guía de referencia más utilizada en el desarrollo de proyectos de Data Mining [[23](#_ENREF_23)].

CRISP-DM contempla el análisis y comprensión del problema antes de comenzar el proceso de minería, mientras que SEMMA excluye esta actividad del modelo.

En todos los modelos se contempla la selección y preparación de los datos. SEMMA propone trabajar con un muestreo de los datos originales (en caso de tener un gran volumen de datos).

La fase de modelado constituye el núcleo del proceso y está presente en todas las metodologías.

SEMMA se encuentra enfocada especialmente en aspectos técnicos, excluyendo actividades de análisis y comprensión del problema que se está abordando. Fue propuesta especialmente para trabajar con el software de minería de datos de la compañía SAS [[21](#_ENREF_21)].

### 1.3.5 Evaluación de arquitecturas candidatas.

Para comenzar el desarrollo de una herramienta se hace necesario seleccionar con qué arquitectura trabajar. Ello implica realizar estudios de todas las arquitecturas que puedan servirle a la solución que se propone y analizar sus ventajas y desventajas.

Se deben evaluar los atributos de calidad considerando cada atributo de forma aislada y analizar las arquitecturas candidatas utilizando el análisis de sensibilidad [[24](#_ENREF_24)] .

Entre estas herramientas se encuentran:

**RapidMiner:** (anteriormente, YALE, *Yet Another Learning Environment*) es un programa informático para el análisis y minería de datos. Permite el desarrollo de procesos de análisis de datos mediante el encadenamiento de operadores a través de un entorno gráfico. Se usa en investigación, educación, capacitación, creación rápida de prototipos y en aplicaciones empresariales [[25](#_ENREF_25)].

**Weka:** *Waikato Environment for Knowledge Analysis*, es un software programado en Java que está orientado a la extracción de conocimientos desde bases de datos con grandes cantidades de información. Weka contiene una colección de herramientas de visualización y algoritmos para el análisis de datos y el modelado predictivo, junto con interfaces gráficas de usuario para facilitar el acceso a estas funciones[[26](#_ENREF_26)].

**R:** R es un entorno y lenguaje de programación con un enfoque al análisis estadístico. Se trata de uno de los lenguajes de programación más utilizados en investigación científica, siendo además muy popular en los campos de aprendizaje automático (machine learning), minería de datos, investigación biomédica, bioinformática y matemáticas financieras. A esto contribuye la posibilidad de cargar diferentes bibliotecas o paquetes con funcionalidades de cálculo y graficación. Como deficiencia se puede señalar que no siempre se utilizan de la misma forma, ya que al provenir de desarrolladores diferentes, dificulta la realización de tareas propias de la herramienta[[27](#_ENREF_27)].

**SAS:** *Statistical Analysis System* es un paquete de software estadístico desarrollado por el Instituto SAS para la gestión de datos, análisis avanzado, análisis multivariado, inteligencia empresarial, investigación criminal y análisis predictivo. Se caracteriza por su gran escalabilidad pues permite aumentar progresivamente su eficiencia aumentando los recursos de hardware o de cualquier otro tipo. No obstante, solo puede utilizarse de forma gratuita si una institución pública proporciona una licencia, es decir que en la mayoría de los casos esta herramienta es de pago obligado[[28](#_ENREF_28)].

**KNIME:** *Konstanz Information Miner* es una plataforma de minería de datos que permite el desarrollo de modelos en un entorno visual. Está construido bajo la plataforma Eclipse. Está concebido como una herramienta gráfica y dispone de una serie de nodos (que encapsulan distintos tipos de algoritmos) y flechas (que representan flujos de datos) que se despliegan y combinan de manera gráfica e interactiva. Esta herramienta orienta de manera fácil el proceso de descubrimiento de conocimiento en los datos a través del diseño de un flujo de trabajo [[29](#_ENREF_29)].

Entre todas las herramientas para realizar minería y análisis de datos KNIME muestra la ventaja de ser gratuito, funciona stand-alone (en el equipo y no en la red), es altamente visual lo que evita la gran cantidad de código de programación y permite una mejor trazabilidad del cálculo. El carácter abierto de la herramienta hace posible su extensión mediante la creación de nuevos nodos que implementen algoritmos a la medida del usuario.

### 1.3.6 Konstanz Information Miner

*Konstanz Information Miner* (KNIME, disponible en http://www.knime.org) es una plataforma de MD libre que permite el desarrollo de modelos de Minería de Datos (MD).

KNIME fue creada en 2006, desarrollado originalmente en el departamento de bioinformática y minería de datos de la Universidad de Constanza, Alemania, bajo la supervisión del profesor Michael Berthold. Está desarrollado sobre la plataforma Eclipse y programado, esencialmente en Java; es un software libre[[29](#_ENREF_29)].

KNIME integra diversos componentes para aprendizaje automático y minería de datos a través de su concepto de fraccionamiento de datos (data pipelining) modular. La interfaz gráfica de usuario permite el montaje fácil y rápido de nodos para preprocesamiento de datos (ETL: extracción, transformación, carga), para el análisis de datos y modelado y visualización. Esta herramienta guía el proceso de KDD en todas sus fases a través de un ambiente de trabajo visual e interactivo, lo que le permite al usuario ir conformando paso a paso el flujo de KDD de una manera fácil e intuitiva. Una de las características principales de KNIME es su posibilidad de instalar diferentes extensiones que se pueden agregar en función de las necesidades del usuario.

Los autores de [[30](#_ENREF_30)] explican que esta herramienta fue diseñada bajo tres principios fundamentales:

1. Ambiente de trabajo visual e interactivo: Los flujos de trabajo se forman arrastrando los elementos al área de trabajo, de forma que sea fácil e intuitivo para el usuario.

2. Modularidad: Las unidades de procesamiento o contenedoras de datos no deben depender una de las otras. De esta forma se puede distribuir la computación de cada unidad de la forma que se desee. Además permite la implementación de algoritmos de forma independiente. No hay tipos de datos predefinidos, por lo que se pueden definir nuevos tipos con sus especificaciones propias. Los nuevos tipos de datos pueden declararse compatibles con otros existentes.

3. Extensibilidad de forma sencilla: La adición de nuevas unidades de procesamiento, visualización y tratamiento de datos, debe ser una tarea fácil de realizar, evitando complicar el proceso con procedimientos engorrosos de instalación /desinstalación

**Diseño de un flujo de trabajo**

Knime (Konstanz Information Miner) es una plataforma de exploración de datos modular que guía el proceso de descubrimiento de conocimiento en bases de datos Knowledge Discovery in Databases (KDD). Se basa en el diseño de flujos de trabajo (workflows) de forma visual, y permite además ejecutar selectivamente algunos o todos los pasos de análisis, para luego investigar los resultados a través de vistas interactivas de los datos y modelos [[30](#_ENREF_30)].

Un flujo de trabajo es una secuencia de pasos para alcanzar un resultado. Cada flujo de trabajo cuenta con diferentes nodos y flechas donde se combinan de manera gráfica e interactiva. Para crear un flujo de trabajo se debe tener en cuenta que las salidas de los nodos se usan como entradas de otros. En la figura 8 se muestra un ejemplo de flujo de trabajo.

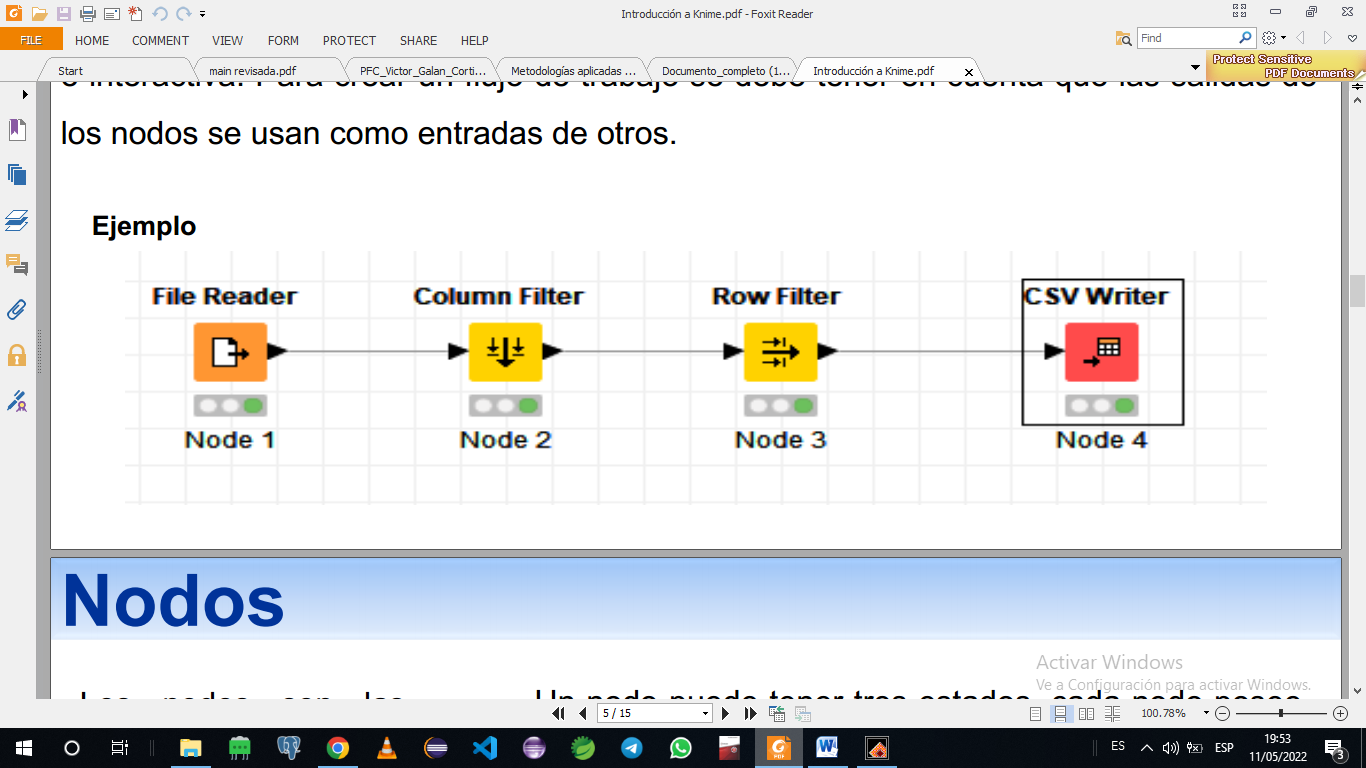


Figura 8 Flujo de Trabajo.

Los nodos son las unidades básicas de procesamiento del flujo de trabajo. Cada nodo tiene un número de entrada y/o salida. Un nodo puede tener tres estados, cada nodo posee en la parte inferior una luz indicadora (semáforo) que muestra el estado que presenta ese nodo, puede tener tres posibles colores: rojo, amarillo y verde. En la figura 9 se muestra un ejemplo de estado de los nodos.

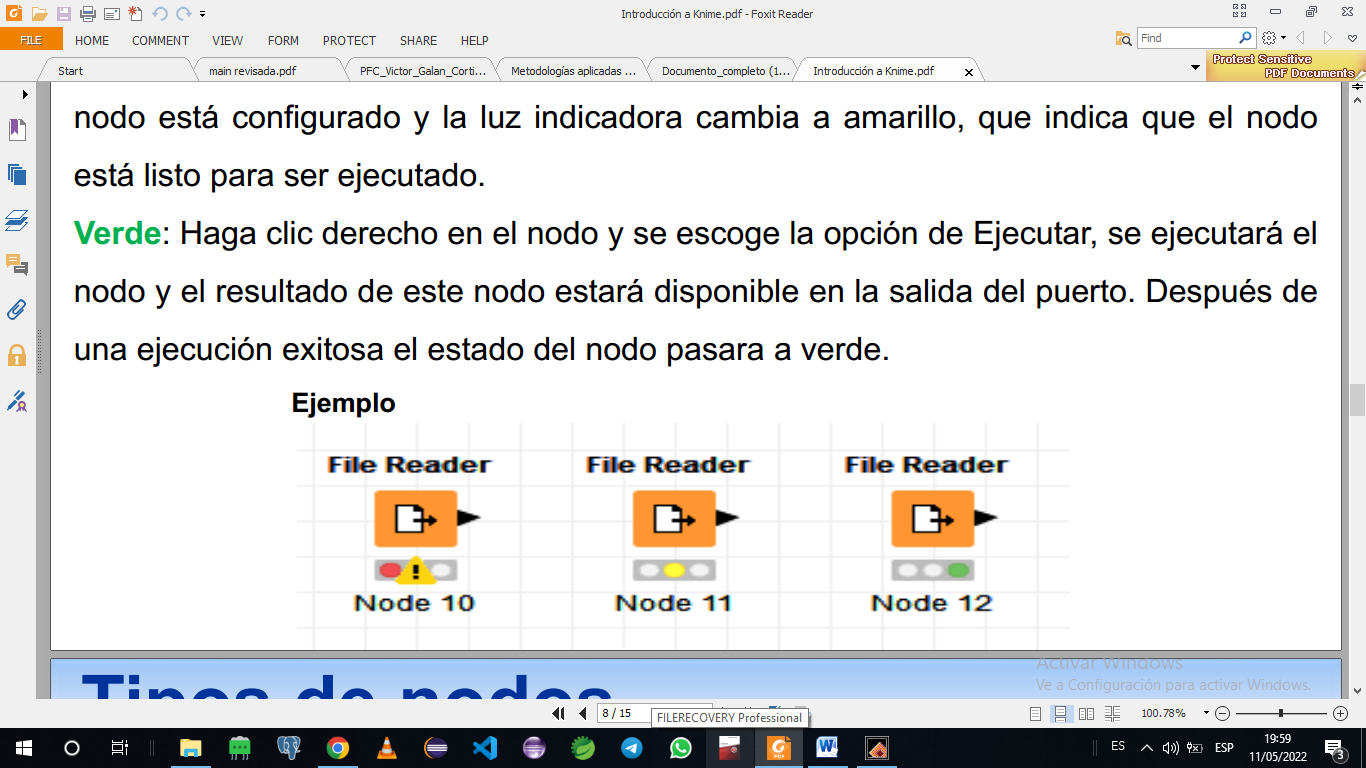


Figura 9 Estado de los nodos.

Rojo: Cuando un nodo se arrastra en el editor de flujo de trabajo la luz de estado está en rojo, lo que significa que el nodo tiene que ser configurado para posteriormente ser ejecutado. En el caso de que exista algún error en el momento de configurar el nodo, la herramienta lo indica con un símbolo de exclamación de color amarillo, mostrando la información del error, para que el usuario pueda brindar una solución.

Amarillo: Cuando el diálogo de configuración se cierra pulsando el botón Aceptar el nodo está configurado y la luz indicadora cambia a amarillo, que indica que el nodo está listo para ser ejecutado.

Verde: Haga clic derecho en el nodo y se escoge la opción de Ejecutar, se ejecutará el nodo y el resultado de este nodo estará disponible en la salida del puerto. Después de una ejecución exitosa el estado del nodo pasara a verde.

Knime brinda diferentes tipos de nodos, los cuales están agrupados por colores. Cada grupo está compuesto por varios nodos que realizan diferentes acciones.

* Amarillo: Preprocesamiento.
* Rojo: Exportar información en diferentes formatos.
* Carmelita: Cargar Datos.
* Verde: Algoritmos de Minería de Datos.
* Azul: Acciones que permiten conocer estadísticas sobre la base de datos, pero que no la modifican.

## 1.4 Conclusiones Parciales

# Capítulo 2 Desarrollo de la aplicación

## Introducción

## Diseño Experimental

El diseño de experimentos que se va a utilizar para ver cómo estas variables influyen en el éxito de los intentos de autenticación se describe a continuación.

Los participantes son 50 personas con un rango de edad de 15-20 años, 21-35 años, 36-50 años, 51-60 años y más de 60 años. Van a tener distintas profesiones y algunas personas pueden que usen computadoras con frecuencia y otras no. El tolerancia puede variar y las imágenes también. Se realizaran 5 sesiones con una diferencia de al menos 1 día por cada sesión. La primera sesión comenzara explicando el experimento y los conceptos básicos. Pasando al experimento, el usuario tiene que crear una contraseña de 5 puntos en la imagen que escoja de las establecidas por el sistema y ingresar la tolerancia que va a tener su contraseña, luego va a introducir la pregunta y la respuesta de seguridad. Una vez creada la contraseña el usuario va a autenticarse hasta 5 veces en las 5 secciones establecidas. Se utilizaran pruebas de técnicas de minería de datos para ver como las variables influyen en los intentos de autenticación.

## 2. Conclusiones

# Capítulo 3 Validación de la herramienta

## Introducción

## 3. Conclusiones

# Conclusiones Generales

# Recomendaciones

# Referencias

1. Chiasson, S., P.C. Van Oorschot, and R. Biddle. *Graphical password authentication using cued click points*. in *European Symposium on Research in Computer Security*. 2007. Springer.

2. Rodriguez Valdés, O., C.M. Legón, and R. Socorro Llanes, *Seguridad y usabilidad de los esquemas y técnicas de autenticación gráfica.* Revista Cubana de Ciencias Informáticas, 2018. **12**: p. 13-27.

3. Walkup, E., *The password problem*, 2016, Sandia National Lab.(SNL-NM), Albuquerque, NM (United States).

4. Legón, C.M., et al., *Nuevo modelo probabilístico en autenticación gráfica.* Ingeniería Electrónica, Automática y Comunicaciones, 2019. **40**(3): p. 92-104.

5. Itti, L. and C. Koch, *Computational modelling of visual attention.* Nature reviews neuroscience, 2001. **2**(3): p. 194-203.

6. Valdés, O.R., et al., *Patrones en el orden de los clics y su influencia en la debilidad de las claves de la Técnica de Autenticación Gráfica Passpoints.* Rev. Cuba. Cienc. Inform, 2019. **12**: p. 37-47.

7. Lashkari, A.H., et al., *A new algorithm on Graphical User Authentication (GUA) based on multi-line grids.* Scientific Research and Essays, 2010. **5**(24): p. 3865-3875.

8. Nikhil, T. and D. Arati, *Graphical passwords authentication: A survey.* International Journal of, 2015.

9. Biddle, R., S. Chiasson, and P.C. Van Oorschot, *Graphical passwords: Learning from the first twelve years.* ACM Computing Surveys (CSUR), 2012. **44**(4): p. 1-41.

10. Sunil, S.S., D. Prakash, and Y.R. Shivaji, *Cued click points: Graphical password authentication technique for security.* IJCSIT) International Journal of ComputerScience and Information Technologies, 2014. **5**(2).

11. Wiedenbeck, S., et al. *Authentication using graphical passwords: Effects of tolerance and image choice*. in *Proceedings of the 2005 symposium on Usable privacy and security*. 2005.

12. Wiedenbeck, S., et al., *PassPoints: Design and longitudinal evaluation of a graphical password system.* International journal of human-computer studies, 2005. **63**(1-2): p. 102-127.

13. Chiasson, S., et al., *Influencing users towards better passwords: persuasive cued click-points.* People and Computers XXII Culture, Creativity, Interaction 22, 2008: p. 121-130.

14. Dirik, A.E., N. Memon, and J.-C. Birget. *Modeling user choice in the PassPoints graphical password scheme*. in *Proceedings of the 3rd symposium on Usable privacy and security*. 2007.

15. Ray, P.P., *Ray’s scheme: Graphical password based hybrid authentication system for smart hand held devices.* Journal of Information engineering and Applications, 2012. **2**(2): p. 1-11.

16. Blonder, G.E., *Graphical password, us patent 5559961*, 1996, September.

17. Sarohi, H.K. and F.U. Khan, *Graphical password authentication schemes: current status and key issues.* International Journal of Computer Science Issues (IJCSI), 2013. **10**(2 Part 1): p. 437.

18. Hernández, J., C. Ferri, and M. Ramírez, *Introducción a la Minería de Datos.* 2004.

19. Fayyad, U.M., G. Piatetsky-Shapiro, and P. Smyth. *Knowledge Discovery and Data Mining: Towards a Unifying Framework*. in *KDD*. 1996.

20. Moine, J.M., *Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo*, 2013, Universidad Nacional de La Plata.

21. Moine, J.M., A.S. Haedo, and S.E. Gordillo. *Estudio comparativo de metodologías para minería de datos*. in *XIII Workshop de Investigadores en Ciencias de la Computación*. 2011.

22. Chapman, P., et al. *The CRISP-DM user guide*. in *4th CRISP-DM SIG Workshop in Brussels in March*. 1999. sn.

23. Chapman, P., et al., *CRISP-DM 1.0: Step-by-step data mining guide.* SPSS inc, 2000. **9**: p. 13.

24. Roger, S.P., *Ingeniería de Software: Un enfoque práctico*2002: McGraw Hill New York.

25. Anandarajan, M., C. Hill, and T. Nolan, *Learning-based sentiment analysis using rapidminer*, in *Practical Text Analytics*2019, Springer. p. 243-261.

26. Hall, M., et al., *The WEKA data mining software: an update.* ACM SIGKDD explorations newsletter, 2009. **11**(1): p. 10-18.

27. Aria, M. and C. Cuccurullo, *bibliometrix: An R-tool for comprehensive science mapping analysis.* Journal of informetrics, 2017. **11**(4): p. 959-975.

28. Duling, D., W. Thompson, and S. Schubert, *SAS Institute, Inc.* 2003.

29. Dietz, C. and M.R. Berthold, *KNIME for open-source bioimage analysis: a tutorial.* Focus on Bio-Image Informatics, 2016: p. 179-197.

30. Berthold, M.R., et al., *KNIME-the Konstanz information miner: version 2.0 and beyond.* ACM SIGKDD explorations newsletter, 2009. **11**(1): p. 26-31.