Análisis y aplicación de algoritmos de minería de datos

Data mining algorithms analysis and application

Salazar Torres, James Ir
Maestrante en Ingeniería Computacional –
https://orcid.org/0000-0003-1339-8794
Colegio integrado nacional oriente de caldas
CINOC

Edison Girón Cardenas
Maestrante en Ingeniería Computacional
https://orcid.org/0000-0003-4393-9854
Colegio integrado nacional oriente de caldas
CINOC

PERSPECTIVAS
https://revistas.uniminuto.edu/index.php/Pers/issue/view/195
ISSN 2145-6321
e-ISSN 2619-1687
71-88

Vol 1 - No. 21
ENERO - MARZO 2021

RECIBIDO : JULIO 12 -2020
ACEPTADO: DICIEMBRE 21 – 2020

RESUMEN

Introducción La Minería de Datos es utilizada en diferentes disciplinas para la búsqueda de patrones y modelos ocultos en las Bases de Datos. Esta generalmente es aplicada en las áreas de negocios y marketing. Sin embargo, su aplicación y uso quedan finalmente a disposición de quienes manejan este conocimiento, por lo que debe de ser trasformado en información útil para los niveles superiores. Materiales y métodos. Unos de los métodos más conocidos para describir atributos de una base de datos son tabla decisión, árbol de decisión, regresión lineal y M5. Conclusión. Se tomaron trece atributos de un cultivo de vinos, a los cuales se les hizo una discriminación y luego se agruparon en un conjunto denominado químicos. El óptimo dentro de este grupo resultaron ser los fenoles totales de acuerdo con los algoritmos aplicados. Por lo cual es el más recomendable de usar para el cultivo.

Palabras Clave: minería de datos, procesamiento de datos, algoritmo y interpretación de datos, árbol de un solo nivel.

ABSTRACT

Introduction Data Mining is used in different disciplines for searching for hidden patterns and models in databases. This is usually applied in the areas of business and marketing. However, its application and use are finally made available to those who handle this knowledge, so it must be transformed into useful information for the higher levels. Materials and methods. One of the best-known methods for describing attributes in a database is Decision Table, Decision Tree, Linear Regression, and M5. Conclusion. Thirteen attributes were taken from a wine crop, which was discriminated against and then grouped into a set called chemicals. The optimal one within this group turned out to be the total phenols according to the algorithms applied. Therefore, it is the most recommended to use for cultivation.

Key words. data mining, data processing, algorithm and data interpretation, single-level tree

Introducción

El almacenamiento de datos se ha convertido en una tarea rutinaria de los sistemas de información de las organizaciones. Esto es aún más evidente en las empresas de la nueva economía, el comercio, la telefonía, el marketing directo, etc. Los datos almacenados son un tesoro para las organizaciones, que es donde se guardan las interacciones pasadas con los clientes, la contabilidad de sus procesos internos, representan la memoria de la organización. Pero con tener memoria no es suficiente, hay que pasar a la acción inteligente sobre los datos para extraer la información que almacenan Bezerra (2009). Esto se puede realizar por medio del campo de la Minería de Datos, la cual es un área de las Tecnologías de Información que ha tomado gran relevancia para diversas industrias e instituciones académicas, dado que las metodologías y herramientas implementadas permiten un análisis objetivo de los procesos, basado en sus ejecuciones actuales C. M. Tomás, (2011). En BPM Chile. El interés por esta área ha llevado al desarrollo de diversos estudios en el tema, sin embargo, la mayoría de éstos se han enfocado en la modelación de comportamientos normales de un proceso Forina, (1991), dejando un amplio campo de estudio en la detección de anomalías y búsqueda de patrones en registros de procesos que presentan resultados no esperados, negativos o particulares. En tanto, la Minería de Datos se caracteriza por el uso de herramientas y algoritmos para analizar grandes cantidades de datos, con el objetivo de encontrar relaciones y patrones previamente desconocidos entre estos datos García M, (1997).

Los métodos tradicionales de Análisis de Datos incluyen el trabajo con variables estadísticas, varianza, desviación estándar, covarianza y correlación entre los atributos; análisis de componentes (determinación de combinaciones lineales ortogonales que maximizan una varianza determinada), análisis de factores (determinación de grupos correlacionados de atributos), análisis de clusters (determinación de grupos de conceptos que están cercanos según una función de distancia dada), análisis de regresión (búsqueda de los coeficientes de una ecuación de los puntos dados como datos), análisis multivariable de la varianza, y análisis de los discriminantes García, (2012). Todos estos métodos están orientados numéricamente. Son esencialmente cuantitativos Jeffrey w. (2010).

En contraposición, los métodos basados en Aprendizaje Automático como los algoritmos, están orientados principalmente hacia el desarrollo de descripciones simbólicas de los datos, que puedan caracterizar uno o más grupos de conceptos Jeffrey w, diferenciar entre distintas clases, crear nuevas clases, crear una nueva clasificación conceptual, seleccionar los atributos más representativos, y ser capaces de predecir secuencias lógicas L. C. Peñuela. Son esencialmente cualitativos. Es decir que un algoritmo de minería de datos es un conjunto de cálculos y reglas heurísticas que permite crear un modelo de minería de datos a partir de los datos. Para crear un modelo, el algoritmo analiza primero los datos proporcionados, en busca de tipos específicos de patrones o tendencias. El algoritmo usa los resultados de este análisis para definir los parámetros óptimos para la creación del modelo de minería de datos. A continuación, estos parámetros se aplican en todo el conjunto de datos para extraer patrones procesables y estadísticas detalladas Magdalena, (2002).

Considerando lo anteriormente expuesto, se realizará un análisis y aplicación de cuatro diferentes algoritmos de minería de datos.

ALGORITMOS USADOS PARA CLASIFICACIÓN

La ordenación es el procedimiento de fraccionar un grupo de información en conjuntos con alto grado de exclusión, de tal forma que cada elemento de un conjunto se ubique lo próximo probable de los demás conjuntos y conjuntos opuestos estén lo más remoto probable de los demás, donde el alejamiento distancia se calcula con respecto a las variables especificadas, que se desean vaticinar

En 1979 Quinlan desarrolla el sistema ID3, que él denominaría simplemente herramienta porque la consideraba experimental. Conceptualmente es fiel a la metodología de CLS pero le aventaja en el método de expansión de los nodos, basado en una función que utiliza la medida de la información de Shannon. Quinlan tal como método de aprendizaje, es el sistema C4.5 que explica con cierta precisión en la obra C4.5: Programs for Machine Learning. El desarrollo -comercial- de ese método es otro llamado C5 del mismo autor, del que se puede conseguir un prototipo de manifestación tasada en cuanto a las virtudes; por ejemplo, el número máximo de modelos de ensayo.

Representación de un árbol de decisión

Un árbol de decisión puede interpretarse esencialmente como una serie de reglas compactadas para su representación en forma de árbol. en forma de árbol. Dado un grupo de modelos, ordenados como segmentos de pares organizados atributo-valor, de acuerdo con la configuración general en el aprendizaje inductivo a partir de modelos, el concepto que estos sistemas quieran durante el desarrollo de aprendizaje consiste en un árbol. Cada inflexión está marcada con un par atributo-valor y las hojas con una ralea, de forma que el camino que determinan desde la raíz los pares de un ejemplo de entrenamiento alcanza una hoja etiquetada -normalmente- con la clase del ejemplo. La clasificación de un ejemplo nuevo del que se desconoce su clase se hace con la misma técnica, solamente que en ese caso al atributo clase, cuyo valor se desconoce, se le asigna de acuerdo con la etiqueta de la hoja a la que se accede con ese ejemplo.

Problemas apropiados para este tipo de aprendizaje

Las características de los problemas apropiados para resolver mediante este aprendizaje dependen del sistema de aprendizaje específico utilizado, pero hay una serie de ellas generales y comunes a la mayoría y que se describen a continuación:

Que la representación de los ejemplos sea mediante vectores de pares atributo-valor, especialmente cuando los valores son disjuntos y en un número pequeño.

Los sistemas actuales están preparados para tratar atributos con valores continuos, valores desconocidos e incluso valores con una distribución de probabilidad.

Que el atributo que hace el papel de la clase sea de tipo discreto y con un número pequeño de valores, sin embargo existen sistemas que adquieren como concepto aprendido funciones con valores continuos.

Que las descripciones del concepto adquirido deban ser expresadas en forma normal disyuntiva.

Que posiblemente existan errores de clasificación en el conjunto de ejemplos de entrenamiento, así como valores desconocidos en

algunos de los atributos en algunos ejemplos Estos sistemas, por lo general, son robustos frente a los errores del tipo mencionado.

Sistemas que se pueden usar para un Árbol de Decisión

Sistema ID3

.El procedimiento ID3 es un numero sencillo y, sin embargo, potente, cuya función es la realización de un árbol de decisión. El procedimiento para producir un árbol de decisión reside en, como se estableció precedentemente en escoger una cualidad o atributo como raíz del árbol y crear un vástago con cada uno de los posibles valores de dicha cualidad. Con cada vástago resultante (nuevo nodo del arbusto), se realiza el mismo procedimiento, esto es, se escoge otra cualidad y se crea un nuevo vástago para cada factible coste del atributo. Este proceso sigue hasta que los modelos se clasifiquen a través de uno de los caminos del arbusto. El nodo concluyente de cada camino será un nodo hoja, al que se le otorgará la ordene correspondiente. Así, el meta objetivo de los arbustos de decisión es obtener regulaciones o vinculaciones que favorezcan ordenar desde de las cualidades

En cada nodo del arbusto de resolución se debe escoger una cualidad para seguir fraccionando, y el principio que se toma para escogerlo: se determina la cualidad que mejor divida (ordene) los modelos de acuerdo con los tipos. Para ello se emplea la entropía, que es una medida de cómo está organizado el cosmos. La hipótesis de la información (basada en la entropía) determinada cifra de bits (información, preguntas sobre las cualidades) que hace falta dar para determinar el tipo a la que pertenece un modelo. Cuanto menor sea la cuantía de la entropía, menor será la inseguridad y más útil será el atributo para la ordenación de los elementos.

Sistema C4.5

El ID3 es capaz de tratar con atributos cuyos valores sean discretos o continuos. En el primer caso, el árbol de decisión generado tendrá tantas ramas como valores posibles tome el atributo. Si los valores del atributo son continuos, el ID3 no clasifica correctamente los ejemplos dados. Por ello, Quinlan propuso el C4.5, como extensión del ID3, que permite:

Empleo del concepto razón de ganancia

Construir árboles de decisión cuando algunos de los ejemplos presentan valores desconocidos para algunos de los atributos.

Trabajar con atributos que presenten valores continuos.

La poda de los árboles de decisión

Obtención de Reglas de Clasificación

Decisión Stump (Árbol de un solo nivel)

Aun hay un numero o algoritmo más simple que proporciona un arbusto de decisión de un solo nivel. Radica en un algoritmo, que utiliza un único atributo para realizar un arbusto de resolucion. La dterminacion de un solo atributo que creara parte del árbol se realizará teniendo en cuenta en el dividendo de la información, y a pesar de su sencillez, en algunos problemas puede llegar a conseguir resultados significativos. No tiene opciones de conformación, pero la implementación es muy completa, dado que admite tanto atributos aritméticos como simbólicos y clases. En árbol de decisión habrá tres arbustos: una de ellas será para el caso de que el atributo sea no conocido, y las otras dos serán para el caso de que el valor del atributo del test sea igual a un valor concreto del atributo o distinto a dicho valor, en caso de los atributos simbólicos, o que el valor del ejemplo de test sea mayor o menor a un determinado valor en el caso de atributos aritméticos. En los atributos simbólicos cada valor posible del mismo y se calcula la rentabilidad de los datos con el atributo igual a la cuantía, opuesto a la cuantía y valores no conocidos del atributo.

Considerarse cuatros casos al determinar la rentabilidad de los datos: que sea un atributo figurado y la clase sea figurada o que la clase sea aritmética, o que sea un atributo aritmético y la clase sea figurada o que la clase sea aritmética.

ALGORITMO USADO PARA PREDICCIÓN

Es el proceso que intenta determinar los valores de una o varias variables, a partir de un conjunto de datos. La predicción de valores continuos puede planificarse por las técnicas estadísticas de regresión.

Regresión lineal

Los modelos lineales generalizados representan el fundamento teórico en que la regresión lineal puede aplicarse para modelar las categorías de las variables dependientes. En los modelos lineales generalizados, la variación de la variable y es una función del valor medio de y, distinto a la regresión lineal dónde la variación de y es constante. Los tipos comunes de modelos lineales generalizados incluyen regresión logística y regresión del Poisson. La regresión logística modela la probabilidad de algún evento que ocurre como una función lineal de un conjunto de variables independientes. Frecuentemente los datos exhiben una distribución de Poisson y se modelan normalmente usando la regresión del Poisson Peñuela, (2013).

Los modelos lineales logarítmicos aproximan las distribuciones de probabilidad multidimensionales discretas, y pueden usarse para estimar el valor de probabilidad asociado con los datos de las células cúbicas. Por ejemplo, suponiendo que se tienen los datos para los atributos ciudad, artículo, año y ventas. En el método logarítmico lineal, todos los atributos deben ser categorías; por lo que los atributos estimados continuos deben ser previamente discretizados.

Algunas de las propiedades de la regresión lineal para tener en cuenta al momento de la implementación son:

Admite atributos numéricos y nominales. Los nominales con k valores se convierten en k-1 atributos binarios.

La clase debe ser numérica.

Se permite pesar cada ejemplo

En la técnica de regresión lineal la filosofía de funcionamiento es diferente. En este caso, se trata de predecir el valor numérico de cada uno de los atributos de los datos de entrada. El algoritmo de regresión lineal implementado por WEKA es muy sencillo; las reglas consisten en funciones lineales de los atributos. Así, en nuestro caso, para predecir el valor de puntos por minuto de un determinado dato de entrada, el algoritmo establece una función lineal del resto de atributos (número de asistencias por minuto, altura, tiempo jugado,

edad). De esta forma, al aplicar un dato a la función, se toman los valores de estos atributos, se aplican a la función lineal y se obtiene el número de puntos por minuto estimado Molina y Gracia (2012).

M5

En cuanto a la implementación concreta que se lleva a cabo en esta herramienta, cabe destacar lo siguiente:

Admite atributos simbólicos y numéricos; la clase debe ser, por supuesto, numérica.

Para la generación de las regresiones lineales se emplea la clase que implementa la regresión lineal múltiple en WEKA.

El número mínimo de ejemplos que deben clasificarse a través de un nodo para seguir dividiendo dicho nodo, definido en la constante SPLIT_NUM es 3.5, mientras la otra condición de parada, que es la desviación típica de las clases en el nodo respecto a la desviación típica de todas las clases del conjunto de entrenamiento, está fijada en 0.05.

No puede manejar instancias ponderadas por pesos.

No puede ser actualizado de forma incremental (soportar añadir nuevos datos sin reclasificar a los anteriores).

Cuando se encuentra con un valor de atributo no determinado, M5' reemplaza dicho hueco por la media global o la moda del conjunto de datos de entrenamiento antes de que se construyera el árbol. Permite diferentes tipos de salida: árbol modelo, árbol de decisión sin modelos lineales en las hojas y regresión lineal. Presenta un proceso automático de suavizado que puede ser deshabilitado y también se puede controlar la profundidad del podado, así como la cantidad de información.

2. Materiales y métodos

Se presentará a continuación de forma concisa las técnicas metodológicas que se empleará para el análisis de datos por medio de la herramienta de minería de datos WEKA; como software de gratis reparto hecho en Java. Está hecho por una serie de paquetes

de código abierto con diferentes metodologías de pre- procesado, ordenación, agrupamiento, asociación, y representación, así como posibilidades para su desarrollo y análisis cuando son aplicadas a la información de entrada seleccionados. Estos paquetes pueden ser integrados en cualquier proyecto de análisis de datos, e incluso pueden prorrogarse con distribuciones y contribuciones de los usuarios que desarrollen nuevos algoritmos. Con objeto de ayudar su uso por un mayor número de usuarios, WEKA además incluye una interfaz gráfica de usuario para acceder y configurar las diferentes herramientas integradas.

Los datos de entrada a la herramienta, sobre los que operarán las técnicas implementadas, deben estar codificados en un formato específico, denominado Attribute-Relation File Format (extensión "arff"). La herramienta permite cargar los datos en tres soportes: fichero de texto, acceso a una base de datos y acceso a través de internet sobre una dirección URL de un servidor web. Michalski, a. B. Baskin, and k. A. Spackman, (1982).

Los atributos pueden ser principalmente de dos tipos: numéricos de tipo real o entero (indicado con la palabra real o integer tras el nombre del atributo), y simbólicos, en cuyo caso se especifican los valores posibles.

Se usará una base de datos, que fue tomada de Machine Learning Repository L. C. Peñuela, (2013), la cual consistió en un análisis químico sobre vinos. Se tomaron tres diferentes cultivos a los cuales se les llevo un control de trece atributos:

- Alcohol
- Ácido Málico
- Cenizas
- Alcalinidad de cenizas de Magnesio
- Fenoles totales
- Flavonoides
- Fenoles no flavonoides
- Proantocianinas
- Intensidad de color
- Matiz
- OD280/OD315 de vinos diluidos
- Prolina

Esta investigación se llevó a cabo en una región de Italia.

Ejecución en WEKA

WEKA se reparte como un archivo ejecutable comprimido de java (fichero "jar"), que se convoca sobre la máquina virtual JVM. En las primeras impresiones de WEKA se necesitaría la máquina virtual Java 1.2 para convocar la interfaz gráfica, hecho con el paquete gráfico de Java Swing. En el caso de la última versión, WEKA 3-6, es la que se ha usado para desarrollar esas notas, se necesita Java 1.3 o mejorada. El instrumento se invoca desde el intérprete de Java, cuando se utiliza un contexto windows, con una ventana de comandos para invocar al intérprete Java sería necesario. Una vez convocada, se muestra la ventana de ingreso a la interfaz gráfica (GUIChooser), la nos da cuatro posibilidades posibles de trabajo: Simple CLI, Explorer, Experimenter, KnowledgeFlow L. C. Peñuela, (2013).

Es de anotar que en este artículo se utilizara opción, Explorer. Una vez escogida, se abre una ventana con 6 pestañas en la parte superior que contienen diferentes clases de actuaciones, en etapas independientes, que se pueden hacer sobre la informacion (Preprocess, Clasify, Cluster, Associate, Select Attributes, Visualize). Además de estas pestañas de escogencia, en la parte inferior de la ventana salen dos componentes comunes. Uno es el botón de "Log", que al usarlo muestra una ventana textual donde se indica la dinámica de todas las operaciones que se han conllevado dentro del "Explorer", sus tiempos de starting y end, así como los avisos equívocos más usuales L. C. Peñuela, (2013).

Algoritmos para usar en WEKA

WEKA para Tabla de decisiones

El algoritmo de tabla de decisión implementado en el instrumento WEKA se encuentra en la clase weka.classifiers.DecisionTable.java. Las posibildades de creacion de que disponen son:

DisplayRules: Por defecto no se muestran las reglas del clasificador, concretamente la tabla de decisión construida.

MaxStale: Indica el número máximo de conjuntos que intenta mejorar el algoritmo para encontrar una tabla mejor sin haberla encontrado en los últimos n-1 subconjuntos.

CrossVal: Por defecto se evalúa el sistema mediante el proceso leave-one-out. Si se aumenta el valor 1 se realiza validación cruzada con n carpetas.

WEKA para Árbol de decisiones (Decisión Stump)

La clase en la que se implementa el algoritmo tocón de decisión en la herramienta WEKA es weka.classifers.DecisionStump.java. Así, en WEKA se llama a este algoritmo tocón de decisión. No tiene opciones de configuración, pero la implementación es muy completa, dado que admite tanto atributos numéricos como simbólicos y clases de ambos tipos también. El árbol de decisión tendrá tres ramas: una de ellas será para el caso de que el atributo sea desconocido, y las otras dos serán para el caso de que el valor del atributo del ejemplo de test sea igual a un valor concreto del atributo o distinto a dicho valor, en caso de los atributos simbólicos, o que el valor del ejemplo de test sea mayor o menor a un determinado valor en el caso de atributos numéricos.

WEKA para Regresión Lineal

Es en la clase weka.classifers.LinearRegression.java L. C. Peñuela, (2013) en la que se hace una la regresión lineal múltiple. Las posibilidades que permite este algoritmo son:

AttributeSeleccionMethod (M5 method): Método de escogencia del atributo a borrar de la regresión. Las opciones son M5 Method, Greedy y None.

Debug (False): Muestra el proceso de construcción del ordenador

WEKA para M5

La clase en la que se implementa el algoritmo M5 en la herramienta WEKA es weka.classifers.m5.M5Prime.java. Molina and j. García, (2012 Las opciones que permite este algoritmo son:

ModelType: ayuda a escoger como ejemplo a desarrollar entre un árbol de modelos, un árbol de regresión o una regresión lineal.

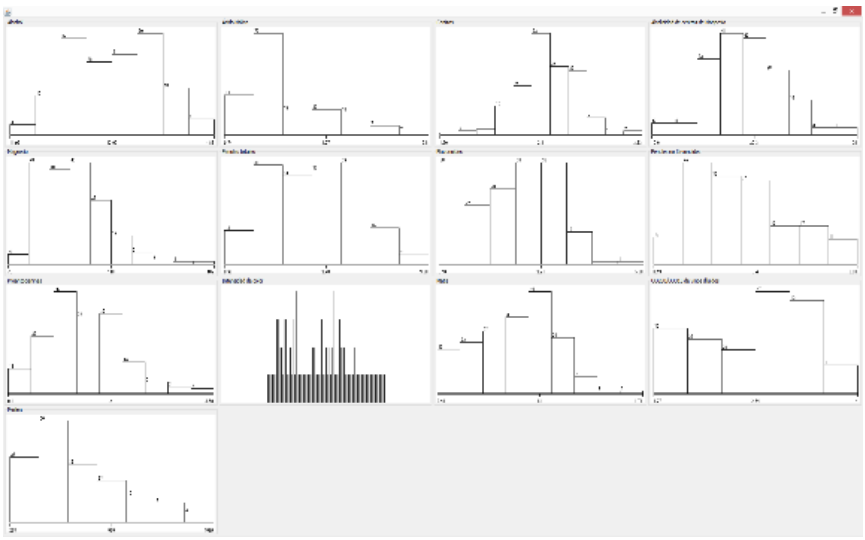
UseUnsmoothed: muestra la realización proceso de suavizado (False) o si no se realizará (True).

PruningFactor: determina el factor de poda.

Verbosity: Sus posibles valores son 0, 1 y 2, y permite definir las estadísticas que se mostrarán con el ejemplo.

3. Resultados

Tabla 1. *Resultado de todos los datos y atributos ingresados a WEKA*



La Tabla 1 nos muestra una similaridad entre los 13 atributos, por lo cual se seleccionaron 4 de ellos que pertenecen a la rama de los químicos; se les tomo a cada uno el error cuadrático medio (ver Tabla 2.). Mostrando que el atributo de Fenoles totales aplicando el algoritmo de Regresión Lineal frente a los otros algoritmos, es el que presenta un menor error cuadrático medio con un valor de 0.1491

Tabla 2. *Error cuadrático medio para 4 algoritmos y 4 atributos*

	Error cuadrático medio			
Atributo	Regresión lineal	M5	Arboles de primer nivel	Tablas de decisión
Alcohol	0.2612	0.3007	0.6481	0.6203
Ácido Málico	0.5244	0.3809	0.9003	0.8733
Magnesio	80.153	65.649	12.4013	12.2049
Fenoles totales	0.1491	0.2139	0.3575	0.2976

Tabla 3. Coeficiente de correlación para 4 algoritmos y 4 atributos

	Coeficiente de correlación			
Atributo	Regresión lineal	M5	Arboles de primer nivel	Tablas de decisión
Alcohol	0.9465	0.9284	0.8196	0.879
Ácido Málico	0.8823	0.9404	0.5992	0.6425
Magnesio	0.8266	0.8886	0.589	0.6209
Fenoles totales	0.971	0.9394	0.4918	0.5154

Teniendo en cuenta que los Fenoles totales son los que presentan un menor error cuadrático medio y a su vez el coeficiente de correlación más cercano a 1, se presenta a continuación con detalle la aplicación de cada algoritmo para este atributo, los cuales fueron de Clasificación mediante la opción Use training set.

Algoritmo "decision table" fenoles totales

=== Evaluation on training set ===

=== Summary ===

Correlation coefficient	0.879
Mean absolute error	0.2218
Root mean squared error	0.2976
Relative absolute error	41.3645 %
Root relative squared error	47.6824 %
Total Number of Instances	178

Algoritmo "árbol de primer nivel" fenoles totales:

=== Evaluation on training set ===

=== Summary ===

Correlation coefficient	0.8196
Mean absolute error	0.2869
Root mean squared error	0.3575
Relative absolute error	53.4929 %
Root relative squared error	57.2875 %
Total Number of Instances	178

Algoritmo "regresión lineal" fenoles totales:

=== Evaluation on training set ===

=== Summary ===

Correlation coefficient	0.971
Mean absolute error	0.1034
Root mean squared error	0.1491
Relative absolute error	19.29 %
Root relative squared error	23.8913 %
Total Number of Instances	178

Algoritmo "M5":

=== Evaluation on training set ===

=== Summary ===

Correlation coefficient	0.9394
Mean absolute error	0.1597
Root mean squared error	0.2139
Relative absolute error	29.7775 %
Root relative squared error	34.2687 %
Total Number of Instances	178

Si comparamos todos los errores cuadráticos (los que se encuentran resaltados) se puede notar que el más apropiado para tomar como referente y que tendrá menor incidencia en el cultivo es el algoritmo de regresión lineal, debido a que tiene un menor valor, 0.149. Además, cabe notar que el coeficiente de correlación (ver Tabla 3.) en relación a los otros atributos es el más apropiado, debido a que su valor 0.971 es el más cercano a 1, indicando que es óptimo y tiene una correlación positiva perfecta la cual conlleva a una dependencia total entre los atributos, es decir cuando una de ellas aumenta, los otros también lo hacen en proporción constante. (Imagen 1).

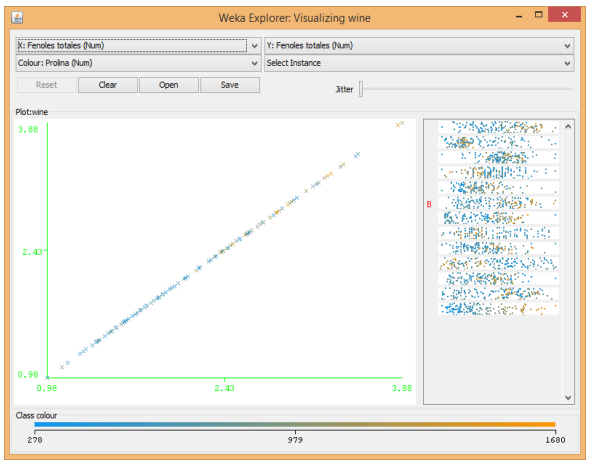


Imagen 1. Regresión lineal para Fenoles

4. Conclusiones

Al momento de tener una gran cantidad de datos es recomendable hacer una agrupación de los atributos semejantes, como en el caso de esto documento que se agruparon los químicos estando comprendido por alcohol, ácido málico, magnesio y fenoles totales. De esta forma se pueda garantizar un análisis de datos preciso generando resultados más confiables al momento de tomar una decisión. Si se debe elegir entre alcohol, ácido málico, magnesio y fenoles totales; para que sea el óptimo para la cosecha y que a su vez afecte positivamente a los otros compuestos deben ser los Fenoles totales.

Referencias

Bezerra (2009). f, wainer j, and v. D. Aaalst, "anomaly detection using process mining. Lecture notes in business information processing," vol. 29, p. 12, 2009.

C. M. Tomás, (2011) "desarrollo y análisis de la utilización de algoritmos de minería de datos para la búsqueda de anomalías y patrones secuenciales en minería de procesos," *pontificia universidad catolica de chile escuela de ingenieria* p. 167, 2011

Forina, (1991) parvus, "using chemical analysis determine the origin of wines," *machine learning repository,* 1991.

García martínez, (1997). "sistemas autónomos: aprendizaje automático," *nueva librería, buenos aires, argentina,* 1997.

García, (2012)"tecnicas de minería de datos basadas en aprendizaje automatico."

José m. Molina and j. "técnicas de minería de datos basadas en aprendizaje automático," 2012.

Jeffrey w. (2010) , "data mining: an overview," *congressional research service ˜ the library of congress,* vol. 19

L. C. Peñuela, (2013)."algoritmos para mineria de datos con redes neuraonales " *universidad politécnica de madrid facultad de informática* p. 170, 2013.

Magdalena, (2002) "algoritmos tdidt aplicados a la mineria de datos inteligente," p. 358, 2002

Molina and j. García, (2012). "técnicas de minería de datos basadas en aprendizaje automático," 2012.

Peñuela, (2013) "algoritmos para mineria de datos con redes neuraonales " *universidad politécnica de madrid facultad de informática* p. 170, 2013.

S. Michalski, a. B. Baskin, and k. A. Spackman, (1982). a logic-based approach to conceptual database analysis, sixth annual symposium on computer applications on medical care," *george washington university, medical center, washington, dc, ee.uu.,* 1982.

S. Michalski and g. E. Tecuci, (2012). "machine learning: a multistrategy approach," *morgan kauffinan, ee.uu,* vol. Iv, 1994.microsoft, "data mining algorithms (analysis services - data mining)," *microsoft,* vol. 4, 2012. "tecnicas de minería de datos basadas en aprendizaje automatico.