ESCUELA POLITÉCNICA NACIONAL
FACULTAD DE INGENIERÍA DE SISTEMAS
INGENIERÍA DE SISTEMAS INFORMÁTICOS Y DE COMPUTACIÓN

Laboratorio de: Inteligencia de negocios
Práctica No.: 3 de Datamining
Tema: Regresión lineal

Nombre: Díaz Padilla Danny Sebastián **Fecha:** 22/01/2020

1. Objetivos:

1.1. Objetivo General

Generar un modelo de regresión lineal con WEKA

1.2. Objetivos Específicos

- Crear un archivo ARFF propio.
- Determinar las variables más importantes para el modelo.

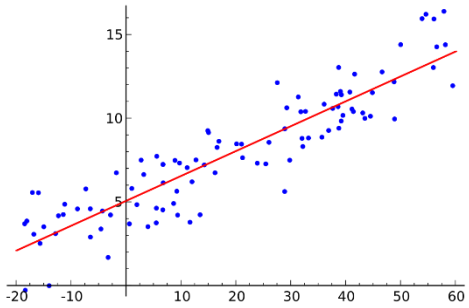
2. Marco teórico:

WEKA

Es un software programado en Java que está orientado a la extracción de conocimientos desde bases de datos con grandes cantidades de información.[1]

Regresión lineal

Es un conjunto de procesos estadísticos para estimar las relaciones entre una variable dependiente (a menudo llamada 'variable de resultado') y una o más variables independientes (a menudo llamadas 'predictores', 'covariables' o 'características'). [2]



Coeficiente de correlación

Es una medida de regresión que pretende cuantificar el grado de variación conjunta entre dos variables. [3]

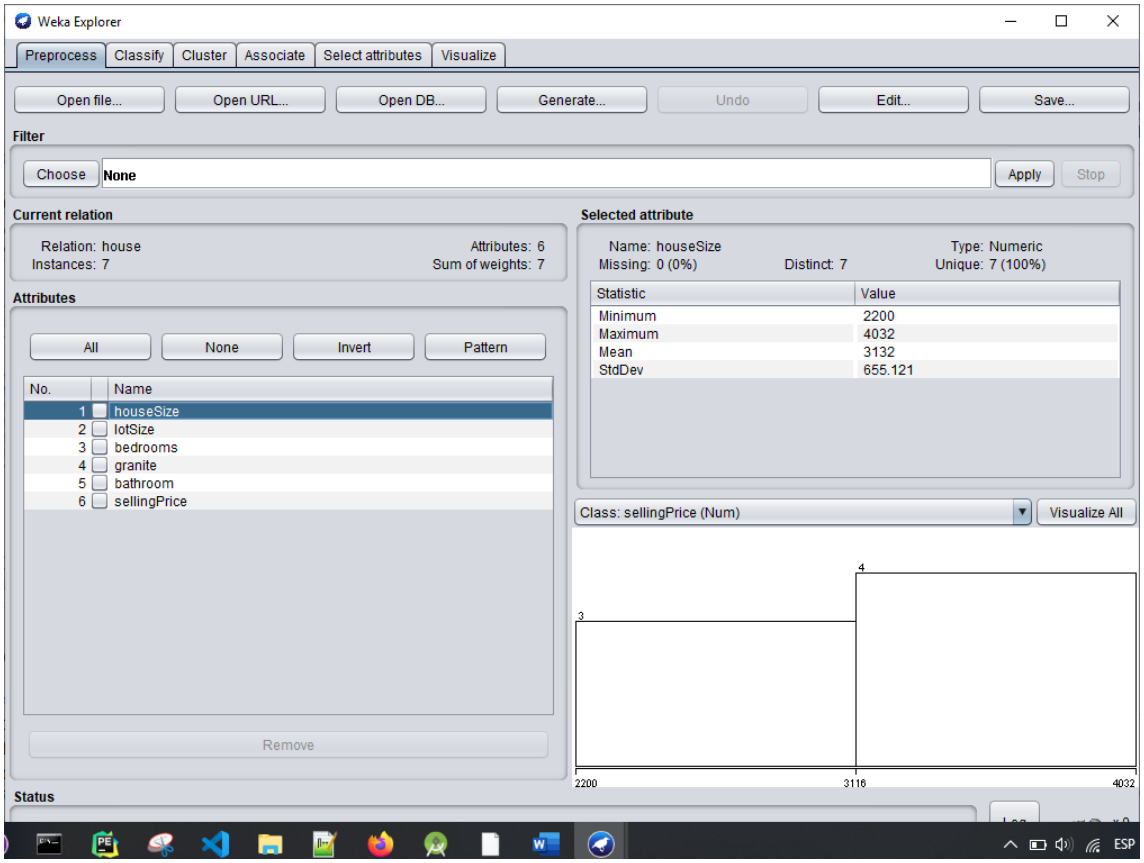
3. Desarrollo de la práctica:

Primero se trasladan los datos de la tabla de casas a un archivo arff.

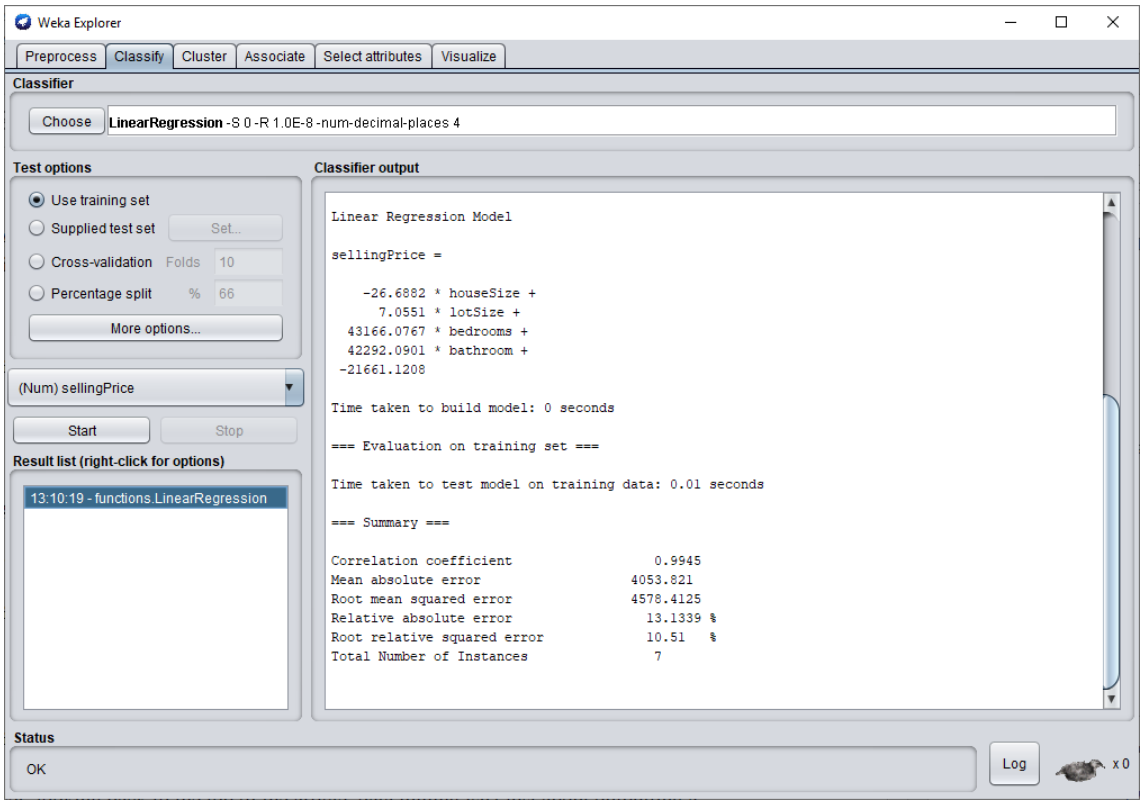
houses.arff | new 20 | activity_main.xml

```
@RELATION house
@ATTRIBUTE houseSize NUMERIC
@ATTRIBUTE lotSize NUMERIC
@ATTRIBUTE bedrooms NUMERIC
@ATTRIBUTE granite NUMERIC
@ATTRIBUTE bathroom NUMERIC
@ATTRIBUTE sellingPrice NUMERIC
@DATA
3529,9191,6,0,0,205000
3247,10061,5,1,1,224900
4032,10150,5,0,1,197900
2397,14156,4,1,0,189900
2200,9600,4,0,1,195000
3536,19994,6,1,1,325000
2983,9365,5,0,1,230000
```

Se utiliza “Open file” para acceder al archivo creado.



En la pestaña de clasificación se selecciona la función: LinearRegression y se dejan los parámetros por defecto.



La fórmula de regresión lineal tiene la siguiente estructura matemática.

```
Linear Regression Model

sellingPrice =

    -26.6882 * houseSize +
      7.0551 * lotSize +
  43166.0767 * bedrooms +
  42292.0901 * bathroom +
 -21661.1208
```

Y los datos finales de errores y correlación se muestran a continuación:

```
Correlation coefficient                  0.9945
Mean absolute error                   4053.821
Root mean squared error               4578.4125
Relative absolute error                 13.1339 %
Root relative squared error             10.51   %
Total Number of Instances                7
```

4. Análisis de resultados:

El granito no importa: WEKA solo usará columnas que estadísticamente contribuyan a la precisión del modelo, este modelo de regresión nos dice que el granito en la cocina no afecta el valor de la casa.

Los baños son importantes, tener entre 0 y 1 marca una diferencia muy importante.

Las casas más grandes reducen el valor. Esto es por la mala calidad de los datos ya que por sentido común realmente debería ser al contrario.

El modelo no es perfecto y requiere muchos más datos.

5. Conclusiones y recomendaciones:

- Se generó un modelo de regresión lineal, sin embargo, debido a la calidad de los datos el modelo no es perfecto.
- Se utilizó un archivo ARFF creado manualmente, lo que implica que se aprendió sobre este método de transporte de datos.
- Las variables considerablemente más importantes son el número de baños y cuartos. Mientras que la variable menos importante es el granito en la cocina.

- Es recomendable hacer una limpieza de datos para mejorar el modelo.

- Al trabajar con 7 instancias el modelo no comprendió el problema en su totalidad, se recomienda agregar muchas más instancias reales para direccionar el modelo a un hecho de la vida real.

6. Bibliografía:

[1]"Introducción a la minería de datos con Weka", *Locualo.net*, 2007. [Online]. Available: http://www.locualo.net/programacion/introduccion-mineria-datos-weka/00000018.aspx. [Accessed: 21- Jan- 2020].

[2]"Regression analysis", *En.wikipedia.org*. [Online]. Available: https://en.wikipedia.org/wiki/Regression_analysis. [Accessed: 21- Jan- 2020].

[3]A. Ucha, "Coeficiente de correlación lineal - Definición, qué es y concepto | Economipedia", *Economipedia*. [Online]. Available: https://economipedia.com/definiciones/coeficiente-de-correlacion-lineal.html. [Accessed: 21- Jan- 2020].