Tarea 06

Nombres y código:

Ruiz Ortiz Cesar Adrián 208020817

Padilla Martin Juan Pablo 217294261

Álvarez Gutiérrez David Alejandro 217294016

Sección: D01

Materia: MINERIA DE DATOS

Profesor: ISRAEL ROMAN GODINEZ

CARRERA: INGENIERIA EN INFORMATICA

CICLO ESCOLAR: 2020 A

Descripción del problema

Desarrollar la fase de “Entendimiento de los datos” utilizando los conjuntos de atos “Iris Plant”, “Adult”. Para ello, elabore un documento con la siguiente

“iris”

a. Recolección de datos iniciales

i. Requerimientos de los datos

1. Orígenes de datos y su disponibilidad

Se accedio a los datos mediante los repositorios de UCI y la aplicación de orange

ii. Criterios de selección

1. Criterios de selección de los datos

El profesor de la materia eligio los repositorios a trabajr

iii. Inserción de datos

1. Si fueron diferentes orígenes de datos, indicar los mecanismos de adquisición

Para obtencion de la informacion se llevo acabo una recoleccion de datos que fueron almacenados y tratados para el uso educativo de conjuntos de datos y mineria de datos.

b. Descripción de los datos

i. Análisis volumétrico

1. Número de instancias

150

2. Número de atributos

4

3. Número de instancias por clase

200

ii. Definición del dominio del atributo (nombre, tipos de datos y formato)

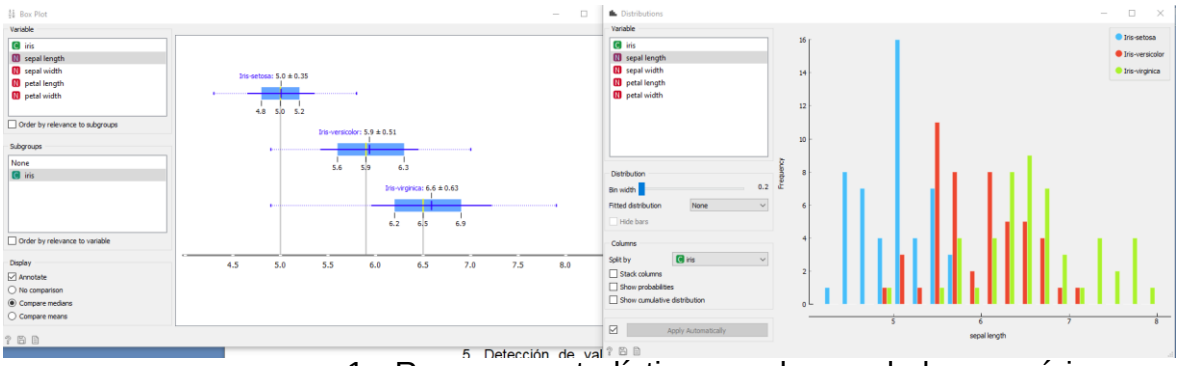
1. Diccionario de datos

Fuente	Bd/flat			
Nombre del atrubuto	Tipo	Dominio	Descripcion	caracteristicas
longitud del sépalo	Numerico	4 - 8	Valor que representa el largo del sepalo	Se mide en cm
ancho del sépalo	Numerico	2 – 4.5	Valor que representa el ancho del sepalo	Se mide en cm
longitud del pétalo	Numerico	1 – 7	Valor que representa el largo del petalo	Se mide en cm
ancho del pétalo	Numerico	0 – 2.5	Valor que representa el ancho del sepalo	Se mide en cm

c. Exploración de los datos

i. Análisis univariable (por atributo)

sepal length



1. Resumen estadístico, en el caso de los numéricos un Box plot con la descripción de los datos. En el caso de valores categóricos, histograma.

   Se presenta una distribucion de menor a mayor en todas la mediana es menor que la media por lo cual contamos con un sesgo positivo

2. Determinación de sesgo.

Iris setosa: sesgo positivo

Iris versicolor: sesgo positivo

Iris virginica:sesgo positivo

2. Identificación de valores faltantes

Uenta con valores faltantes

3. Identificación de valores fuera de dominio

Todos los datos se encuentran dentro del dominio

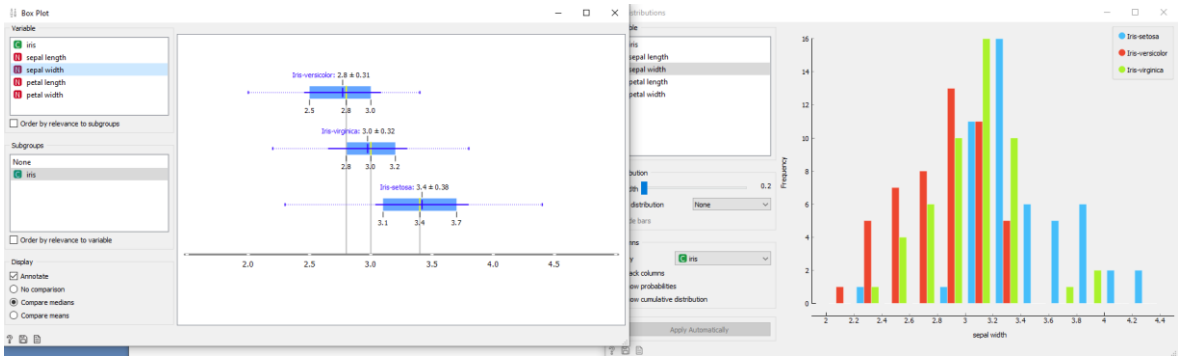
4. Detección de valores erróneos (Outliers). Considere un utlier a los valores que estén por encima de 1.5 el rango intercuartil.

Iris setosa: Q1= 4.8 -1.5 = outlier Q3=5.2 + 1.5 outlier, outlaiers= no hay outliers

Iris versicolor: Q1= 5.6 - 1.5 = outlier Q3=6.3 + 1.5 outlier, outlaiers=no hay

Iris virginica: Q1= 6.2 - 1.5 = outlier Q3=6.9 + 1.5 outlier, outlaiers= no hay

sepal width



1. Resumen estadístico, en el caso de los numéricos un Box plot con la descripción de los datos. En el caso de valores categóricos, histograma.

La variacion en los sesgos se nota en este atributo ya que hay sesgos negativos y positivos

2. Determinación de sesgo.

Iris setosa: sesgo negativo

Iris versicolor: sesgo negativo

Iris virginica:sesgo positivo

2. Identificación de valores faltantes

No cuenta con valores faltantes

3. Identificación de valores fuera de dominio

Todos los datos se encuentran dentro del dominio

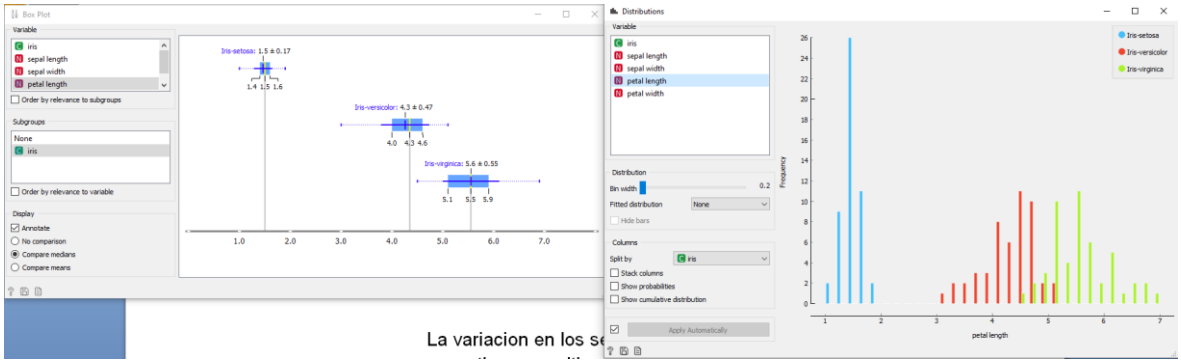
4. Detección de valores erróneos (Outliers). Considere un utlier a los valores que estén por encima de 1.5 el rango intercuartil.

Iris setosa: Q1= 2.5 -1.5 = outlier Q3=3.0 + 1.5 outlier, outlaiers= no hay outliers

Iris versicolor: Q1= 2.8 - 1.5 = outlier Q3=3.2 + 1.5 outlier, outlaiers=id= no hay

Iris virginica: Q1= 3.1 - 1.5 = outlier Q3=3.7 + 1.5 outlier, outlaiers= no hay

petal length



1. Resumen estadístico, en el caso de los numéricos un Box plot con la descripción de los datos. En el caso de valores categóricos, histograma.

Contamos con un conjunto simetrico en el caso de la iris virginica

2. Determinación de sesgo.

Iris setosa: sesgo negativo

Iris versicolor: sesgo negativo

Iris virginica:distribucion simetrica

2. Identificación de valores faltantes

No cuenta con valores faltantes

3. Identificación de valores fuera de dominio

Todos los datos se encuentran dentro del dominio

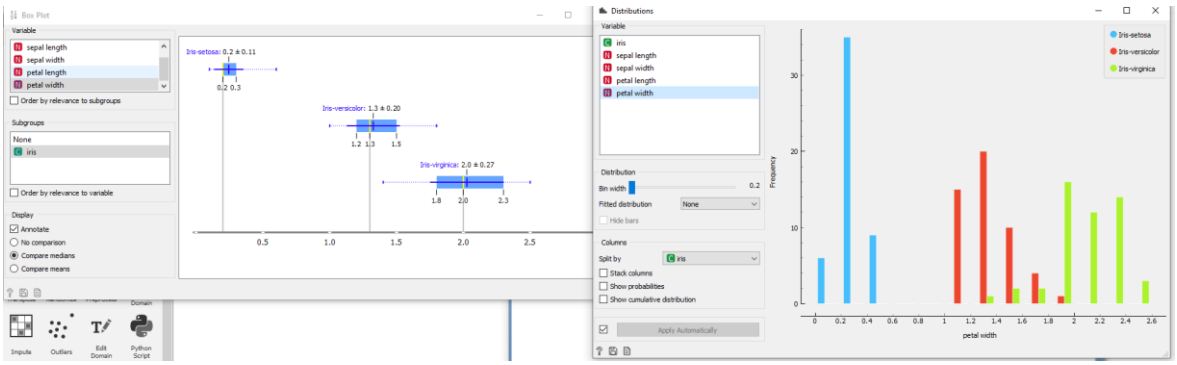
4. Detección de valores erróneos (Outliers). Considere un utlier a los valores que estén por encima de 1.5 el rango intercuartil.

Iris setosa: Q1= 1.4 -1.5 = outlier Q3=1.6 + 1.5 outlier, outlaiers= no hay outliers

Iris versicolor: Q1= 4.0 - 1.5 = outlier Q3=4.6 + 1.5 outlier, outlaiers=id= no hay

Iris virginica: Q1= 5.1 - 1.5 = outlier Q3=5.9 + 1.5 outlier, outlaiers= no hay

petal width



1. Resumen estadístico, en el caso de los numéricos un Box plot con la descripción de los datos. En el caso de valores categóricos, histograma.

Contamos con un conjunto simetrico en el caso de la iris virginica

2. Determinación de sesgo.

Iris setosa: sesgo positivo

Iris versicolor: sesgo positivo

Iris virginica: distribucion positivo

2. Identificación de valores faltantes

No cuenta con valores faltantes

3. Identificación de valores fuera de dominio

Todos los datos se encuentran dentro del dominio

4. Detección de valores erróneos (Outliers). Considere un outlier a los valores que estén por encima de 1.5 el rango intercuartil.

Iris setosa: Q1= 0.2 -1.5 = outlier Q3=0.3 + 1.5 outlier, outlaiers= no hay outliers

Iris versicolor: Q1= 1.2 - 1.5 = outlier Q3=1.5 + 1.5 outlier, outlaiers=id= no hay

Iris virginica: Q1= 1.8 - 1.5 = outlier Q3=2.3 + 1.5 outlier, outlaiers= no hay

“adult”

a. Recolección de datos iniciales

i. Requerimientos de los datos

1. Orígenes de datos y su disponibilidad

La extracción fue realizada por Barry Becker de la base de datos del Censo de 1994. Se extrajo un conjunto de registros razonablemente limpio

ii. Criterios de selección

1. Criterios de selección de los datos

siguientes condiciones: ((AAGE> 16) && (AGI> 100) && (AFNLWGT> 1) && (HRSWK> 0)) La tarea de predicción es determinar si una persona gana más de 50K a año.

iii. Inserción de datos

1. Si fueron diferentes orígenes de datos, indicar los mecanismos de adquisición

El origen de los datos es por parte del UCI repositorio de conjunto de datos

b. Descripción de los datos

i. Análisis volumétrico

1. Número de instancias

48842

2. Número de atributos

14

3. Número de instancias por clase

14

ii. Definición del dominio del atributo (nombre, tipos de datos y formato)

1. Diccionario de datos

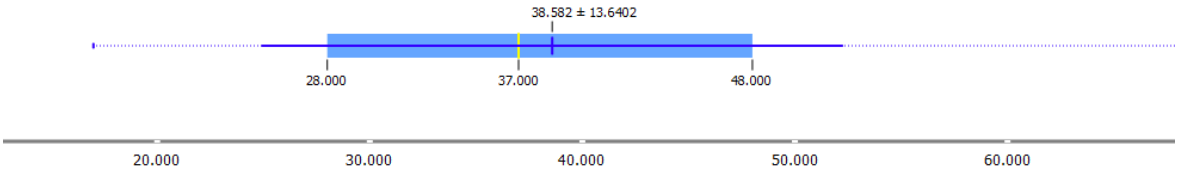
fuente	Bd/flat			
Nombre del atrubuto	Tipo	Dominio	Descripcion	caracteristicas
edad	numerico	0-100	Edad del individuo	De tipo numerico reales
clase de trabajo	nominal	Privado, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Sin pago, Nunca trabajado.	Las funciones que desempeña en el trabajo	Se predetermina un dominio
fnlwgt	numerico	0 - 1	na	0 o 1
educación	nominal	Bachiller, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool	Escolaridad del individuo	
educación-num	numerico	0 - 20	grado	
estado civi	nominal	casado-civil-cónyuge, divorciado, nunca casado, separado, viudo, casado-cónyuge ausente, casado-AF-cónyuge.	Estado civil del individuo	
ocupación	nominal	soporte técnico, reparación de artesanía, otro servicio, ventas, ejecutivo de dirección, especialidad profesional, manipuladores-limpiadores, máquina-op-inspct, administrativo-adm, agricultura-pesca, transporte-mudanza, Priv-house- serv, serv de protección, fuerzas armadas.	Puesto de trabajo	
relación	nominal	esposa, hijo propio, esposo, no familiar, otro pariente, soltero.	Relacion en el hogar	
raza	nominal	blanco, asiático-isleño-pac, amer-indio-esquimal, otro, negro.	Raza del individuo o color de piel	
sexo	nominal	femenino, masculino.	Sexo del individuo	
ganancia de capital	numerico	0 – 100,000	Cantidad de dinero que obtiene	

pérdida de capital	numerico	0 – 50,000	Cantidad de dinero que se gasta	
horas por semana	numerico	0 - 100	Horas trabajadas	
país de origen	nominal	Estados Unidos, Camboya, Inglaterra, Puerto Rico, Canadá, Alemania, Estados Unidos periféricos (Guam-USVI-etc.), India, Japón, Grecia, Sur, China, Cuba, Irán, Honduras, Filipinas, Italia , Polonia, Jamaica, Vietnam, México, Portugal, Irlanda, Francia, República Dominicana, Laos, Ecuador, Taiwán, Haití, Columbia, Hungría, Guatemala, Nicaragua, Escocia, Tailandia, Yugoslavia, El-Salvador, Trinadad y Tobago, Perú, Hong , Holanda-Países Bajos.	Pais en el que nacio el individuo	

c. Exploración de los datos

i. Análisis univariable (por atributo)

“edad”



1. Resumen estadístico, en el caso de los numéricos un Box plot con la descripción de los datos. En el caso de valores categóricos, histograma.

Cuenta con un sesgo pozitivo cuenta con un Q1=28 una media de 38.5 y q3 de 48

2. Determinación de sesgo.

positivo

2. Identificación de valores faltantes

ninguno

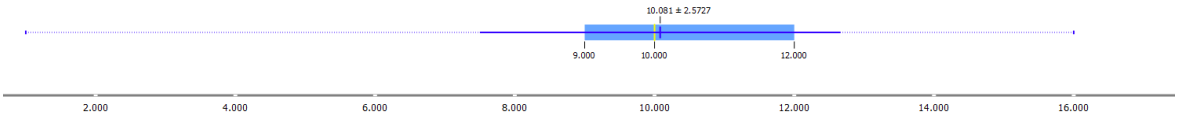
3. Identificación de valores fuera de dominio

ninguno

5. Detección de valores erróneos (Outliers). Considere un utlier a los valores que estén por encima de 1.5 el rango intercuartil.

Cuenta con bastantes todos los numero por debajo de 26.5 y por encima de 49.5

“educacion-num”



Resumen estadístico, en el caso de los numéricos un Box plot con la descripción de los datos. En el caso de valores categóricos, histograma.

Cuenta con un sesgo pozitivo cuenta con un Q1=9 una media de 10 y q3 de 12

2. Determinación de sesgo.

positivo

4. Identificación de valores faltantes

ninguno

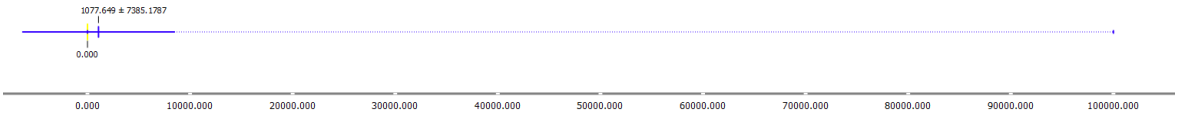
5. Identificación de valores fuera de dominio

ninguno

5. Detección de valores erróneos (Outliers). Considere un utlier a los valores que estén por encima de 1.5 el rango intercuartil.

Cuenta con bastantes todos los numero por debajo de 7.5 y por encima de 13.5

“ganancia de capital”



Resumen estadístico, en el caso de los numéricos un Box plot con la descripción de los datos. En el caso de valores categóricos, histograma.

Cuenta con un sesgo pozitivo cuenta con un Q1=0 una media de 1077 y q3 de 10000

2. Determinación de sesgo.

positivo

6. Identificación de valores faltantes

Cuenta con bastantes todos los que están determinados en 0 son faltantes

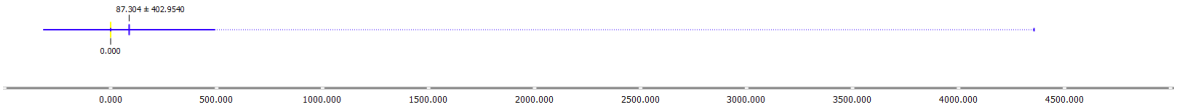
7. Identificación de valores fuera de dominio

ninguno

5. Detección de valores erróneos (Outliers). Considere un utlier a los valores que estén por encima de 1.5 el rango intercuartil.

No hay por debajo y por encima de no esta determinado

“perdida de capital”



Resumen estadístico, en el caso de los numéricos un Box plot con la descripción de los datos. En el caso de valores categóricos, histograma.

Cuenta con bastantes valores faltantes solo podemos visualizar la media

2. Determinación de sesgo.

positivo

8. Identificación de valores faltantes

Cuenta con bastantes todos los que están determinados en 0 son faltantes

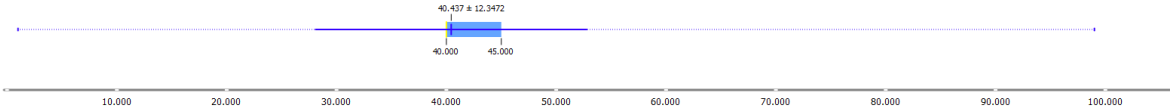
9. Identificación de valores fuera de dominio

ninguno

5. Detección de valores erróneos (Outliers). Considere un utlier a los valores que estén por encima de 1.5 el rango intercuartil.

La tabla no esta determinada debido a que hay demaciados números faltantes

“horas por semana”



Resumen estadístico, en el caso de los numéricos un Box plot con la descripción de los datos. En el caso de valores categóricos, histograma.

Cuenta con bastantes valores faltantes solo podemos visualizar la media

2. Determinación de sesgo.

Positivo

10. Identificación de valores faltantes

Ninguno

11. Identificación de valores fuera de dominio

Ninguno

5. Detección de valores erróneos (Outliers). Considere un utlier a los valores que estén por encima de 1.5 el rango intercuartil.

Por debajo de 38500 es un outlier

Por ensima de 46500