

## **Tarea 03**

### **Fuentes de conjuntos de datos**

EXAMEN DE DIAGNOSTICO

**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



1) ejemplifique (usando impresiones de Pantalla) las dinámicas que se piden en los puntos a, b, c..

a) Abra en su computadora la siguiente liga:



b. Navegue a través del sitio y conteste las siguientes preguntas:

i. ¿Qué es el UCI Machine Learning Repository?

Es un centro de aprendizaje de algoritmos de machine learning.

ii. ¿Qué universidad administra el repositorio?

Universidad de California

iii. ¿Cuál es el objetivo del repositorio?

una colección de bases de datos y generadores de datos que son utilizados para la comunidad que está aprendiendo sobre algoritmos de machine learning.

iv. ¿Cuál es la descripción de la página principal del sitio?

la pagina principal cuanta con una bienvenida y una descripción de que es la pagina,

apartado de noticias:

cuenta con un apartado de noticias por fecha.

Conjunto de datos mas nuevos:

Aparecen links de los conjuntos de datos separados por fechas.

Conjunto de datos mas populares:

Aparecen varios conjuntos de datos separados por fechas.

v. Lista las cinco bases de datos más populares.

1.-iris.

2.-adulto.

3.-vino.

4.-enfermedad del corazón.

5.-calidad del vino

vi. Lista las cinco bases de datos más nuevas

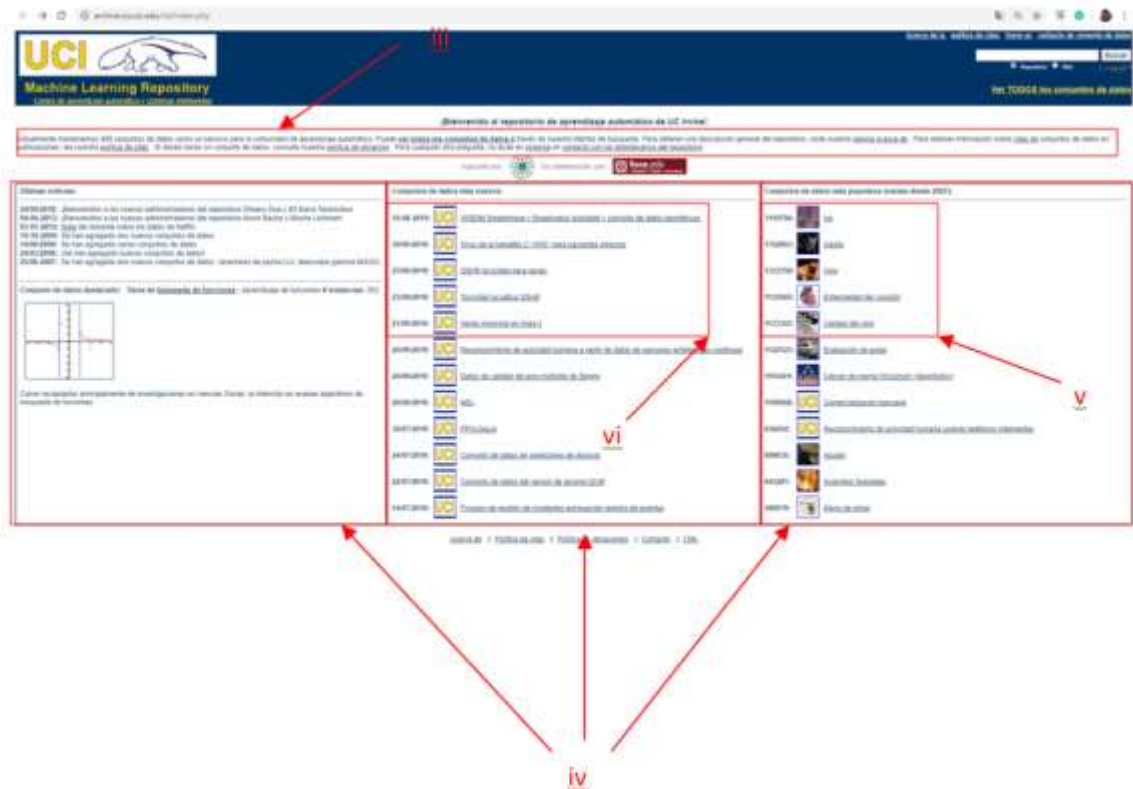
1.-WISDM Smartphone y Smartwatch Actividad y conjunto de datos biométricos

2.- Virus de la hepatitis C (VHC) para pacientes egipcios.

3.- QSAR toxicidad para peces.

4.- Toxicidad acuática QSAR.

5.- Venta minorista en línea II.



2) Da click sobre la liga “view all data set” que se encuentra en la página principal del sitio.

i. Descripción de la página en el sitio.

En esta parte del sitio se puede apreciar un listado o una tabla con diferentes tipos de datos en cada celda, a su vez estas están acomodadas por tipo de dato, año, atributos e instancias. Al costado de la página podemos ver una columna la cual nos permite ver u organizar la información según el tipo de atributo, tipo de dato, área, atributos, instancias y tipo de formato.

ii. ¿Qué partes la conforma (en relación a la información que presenta)?

Se conforma de un conjunto de tablas cuyas columnas describen las características de la información, por ejemplo; se muestra el tipo de atributo, el número de instancias y de atributos que posee.

**UCI Machine Learning Repository**  
 Center for Machine Learning and Intelligent Systems

Búsqueda por atributos

Tabla con conjuntos de datos

489 conjuntos de datos

Navegar a través de:

- Tarea predeterminada:
  - Clasificación (386)
  - Regresión (147)
  - Asociación (36)
  - Clas. (34)
- Tipo de atributo:
  - Categoría (18)
  - Número (325)
  - Mixto (31)
- Tipo de dato:
  - Multivariante (374)
  - Univariante (14)
  - Secuencial (51)
  - Serie temporal (36)
  - Texto (55)
  - Texto del dominio (21)
  - OTR (11)
- Zona:
  - Centros de la vida (139)
  - Ciencias Físicas (52)
  - CI y Ingeniería (178)
  - Ciencias Sociales (26)
  - Biología (60)
  - Juegos (113)
  - OTR (14)
- # Atributos:
  - Menos de 15 (126)
  - 15 a 100 (279)
  - Más de 100 (84)
- # Instancias:
  - Menos de 100 (27)
  - 100 a 1000 (147)
  - Más de 1000 (295)
- Tipo de formato:
  - Matriz (146)
  - CSV (148)


Nombre	Tipo de dato	Tarea predeterminada	Tipo de atributo	# Instancias	# Atributos	Año
Abalone	Multivariante	Clasificación	Categoría, Entero, Real	4177	8	1995
Adulto	Multivariante	Clasificación	Categoría, Entero	48842	14	1996
UCI Recocido	Multivariante	Clasificación	Categoría, Entero, Real	798	38	
UCI Datos web similares de Microsoft		Sistema de recomendación	Categoría	37711	294	1996
Arctima	Multivariante	Clasificación	Categoría, Entero, Real	452	279	1990
Personajes Artificiales	Multivariante	Clasificación	Categoría, Entero, Real	6000	77	1992
Autobots (Gonemal)	Multivariante	Clasificación	Categoría	226		1997
Autobots (restanciada)	Multivariante	Clasificación	Categoría	226	49	1992
MPS automático	Multivariante	Regresión	Categoría, real	388	8	1993
Autobots	Multivariante	Regresión	Categoría, Entero, Real	265	26	1997
UCI Iris	Univariante, Texto	Clasificación		294	1	1994
Balance	Multivariante	Clasificación	Categoría	425	4.4	1994
Globo	Multivariante	Clasificación	Categoría	4.4		

iii. Descarga, en tu computadora, un conjunto de datos por cada una de las tareas que resuelven los algoritmos de aprendizaje automático (machine learning). (Se descargan dos carpetas “Data Folder” y “Data Set Description”)

**UCI Machine Learning Repository**  
 Center for Machine Learning and Intelligent Systems

**Iris Data Set**  
 Download: [Data Folder](#), [Data Set Description](#)

Abstract: Famous database; from Fisher, 1936



Data Set Characteristics:	Multivariate	Number of Instances:	150	Area:	Life
Attribute Characteristics:	Real	Number of Attributes:	4	Date Donated	1988-07-01
Associated Tasks:	Classification	Missing Values?	No	Number of Web Hits:	3121301

**Source:**

Creator:  
 R.A. Fisher

Donor:  
 Michael Marshall (MARSHALL%PLU '@' le.arc.nasa.gov)

**Data Set Information:**

This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to latter are NOT linearly separable from each other.

Predicted attribute: class of iris plant.





This is an exceedingly simple domain.






This data differs from the data presented in Fishers article (identified by Steve Chadwick, [gschadwick '@' ecomadaz.net](mailto:gschadwick '@' ecomadaz.net)). The 35th sample should be: 4.9,3.1.

**Attribute Information:**

- sepal length in cm
- sepal width in cm
- petal length in cm
- petal width in cm
- class:

- Iris Setosa
- Iris Versicolour
- Iris Virginica

Nombre	Estado	Fecha de modificación	Tipo	Tamaño
 agrupamiento	✓	08/02/2020 09:59 p. m.	Carpeta de archivos	
 asosiacion	✓	08/02/2020 09:57 p. m.	Carpeta de archivos	
 clasificacion yo	✓	08/02/2020 09:58 p. m.	Carpeta de archivos	
 regresion yo	✓	08/02/2020 09:58 p. m.	Carpeta de archivos	

 Index	✓	08/02/2020 09:25 p. m.	Archivo	1 KB
 link	✓	08/02/2020 09:58 p. m.	Documento de te...	1 KB
 wine (1)	✓	08/02/2020 09:26 p. m.	Archivo NAMES	3 KB
 wine	✓	08/02/2020 09:25 p. m.	Archivo DATA	11 KB
 wine	✓	08/02/2020 09:24 p. m.	Archivo NAMES	3 KB

b) Para cada uno de los conjuntos de datos identifique:

i. Concepto

ii. Autores

iii. Número de instancias

iv. Atributos de las instancias

v. Tipos de datos para cada una de las instancias

## Agrupamiento (esponjas marinas)

### **Concepto:**

Estas son esponjas marinas atlántico-mediterráneas que pertenecen a O.Hadromerida (Demospongiae.Porifera).

### **Autores:**

Iosune Uriz

Marta Domingo

### **Número de instancias:**

76

### **Tipos de dato para cada una de las instancias:**

27 atributos son no numéricos y nominales.

15 atributos son booleanos y toman los valores (NO SI).

3 atributos son numéricos y toman números naturales.

## **Asociación (familias)**

### ***Concepto:***

Esta base de datos relacional consta de 24 nombres únicos en dos familias (tienen estructuras equivalentes). Hinton utilizó una unidad de salida única para cada persona y estaba interesado en predecir las siguientes relaciones: esposa, esposo, madre, padre, hija, hijo, hermana, hermano, tía, tío, sobrina y sobrino. Hinton usó 104 pares de vectores de entrada-salida (desde un espacio de  $12 \times 24 = 288$  pares posibles). La tarea de predicción es la siguiente: dado un nombre y una relación, las salidas deben estar activadas solo para aquellos individuos (entre los 24) que satisfacen la relación. Los resultados para todos los demás individuos deberían estar apagados.

Resultados de Hinton: Utilizando 100 vectores como entrada y 4 para la prueba, sus resultados en dos pases arrojaron 7 respuestas correctas de 8. Su red de 36 unidades de entrada, 3 capas de unidades ocultas y 24 unidades de salida utilizaron 500 barridos del conjunto de entrenamiento. durante el entrenamiento.

Resultados de Quinlan: Utilizando FOIL, repitió el experimento 20 veces (en lugar de las 2 veces de Hinton). FOIL fue correcto 78 de 80 veces en los casos de prueba.

### ***Autores:***

Geoff Hinton

J. Ross Quinlan

### ***Número de instancias:***

104

***Atributos de instancias:***

Esposa, esposo, madre, padre, hija, hijo, hermana, hermano, tía, tío, sobrina, sobrino.

Tipos de dato para cada una de las instancias:

104 atributos son no numéricos y nominales.

## **Clasificación (iris)**

***Concepto:***

Clasificación de tipos de flor de la especie iris la cual se clasifica por diferentes instancias para aplicarle clustering.

***Autores:***

Creador:

RAFisher

Donante:

Michael Marshall

***Números de instancias:***

3 clases de 50 instancias

***Atributos de las instancias:***

1.- longitud del sépalos en cm

2.- ancho del sépalos en cm

3.- longitud del pétalo en cm

4.- ancho del pétalo en cm

5.- clase:

***Tipos de datos para cada una de las instancias:***

Atributos del 1 al 4 son de tipo numérico ordinal de razón

El 5to atributo es nominal.



## Regresión (automóvil):

### **Concepto:**

Consta del estudio de 3 puntos importantes a tomar en cuenta en los automóviles, especificación, riesgo, perdidas comparadas con otro carro.

### **Autores:**

Creador      Donante:

Jeffrey C. Schlimmer

### **Números de instancias:**

26 instancias

### **Atributos de las instancias:**

1. símbolo: -3, -2, -1, 0, 1, 2, 3.
2. pérdidas normalizadas: continua de 65 a 256.
3. marca:  
alfa-romero, audi, bmw, chevrolet , esquivar, honda,  
isuzu, jaguar, mazda, mercedes-benz, mercurio,  
mitsubishi, nissan, peugot, plymouth, porsche,  
renault, saab, subaru, toyota, volkswagen, volvo
4. tipo de combustible: diesel, gas.
5. aspiración: estándar, turbo.
6. número de puertas: cuatro, dos.
7. estilo de carrocería: techo rígido, vagón, sedán, hatchback, descapotable.
8. ruedas motrices: 4wd, fwd, rwd.
9. ubicación del motor: delantero, trasero.
10. distancia entre ejes: continua desde 86.6 120.9.
11. longitud: continua de 141.1 a 208.1.
12. ancho: continuo de 60.3 a 72.3.
13. altura: continua de 47.8 a 59.8.
14. peso en vacío: continuo de 1488 a 4066.
15. tipo de motor: dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
16. número de cilindros: ocho, cinco, cuatro, seis, tres, doce, dos.
17. tamaño del motor: continuo de 61 a 326.
18. sistema de combustible: 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
19. diámetro interior: continuo de 2.54 a 3.94.
20. carrera: continua de 2.07 a 4.17.
21. relación de compresión: continua de 7 a 23.
22. potencia: continua de 48 a 288.

- 23. rpm máximas: continua de 4150 a 6600.
- 24. mpg en ciudad: continua de 13 a 49.
- 25. carretera- mpg: continuo de 16 a 54.
- 26. precio: continuo de 5118 a 45400.

***Tipos de datos para cada una de las instancias:***

Los tipos de dato son ordinal de intervalo, también cuenta con tipos de dato nominales definidos.

**2.-** Busque dos sitios webs o repositorios donde sea posible encontrar diferentes conjuntos de datos, ya sea en formatos como los de UCI-ML o en algún otro formato que podamos leer, es decir, que podamos acomodar en forma de atributos e instancias.

1.- <https://data.world/>

2.- <https://www.kaggle.com/>

**3.-Para cada uno de estos sitios web, elabore un documento de texto donde se indique:**

**<https://data.world/>**

**a. ¿Cuál es el tipo de información que gestiona?**

Conjuntos de datos para el análisis de sistemas expertos de bases de datos

**b. ¿Qué información se puede encontrar en el sitio?**

Diferentes conjuntos de datos creados por una comunidad para que la misma pueda utilizarla para aprender y administrar sistemas expertos de información de datos de datos.

**i. Objetivo**

En data.world, queremos derribar las barreras entre las personas y los datos mediante la construcción del recurso de datos más significativo, colaborativo y abundante del mundo. Estamos construyendo una plataforma donde las personas pueden:

- Descubra, prepare y comparta datos relevantes de una amplia gama de fuentes.
- Reduzca la sobrecarga de administrar numerosos formatos de archivo desde muchos lugares diferentes

- Explore y cree una comprensión compartida de sus datos
- Contribuya y discuta datos en un solo lugar para ayudar a una colaboración rápida, eficiente y fructífera.

## **ii. Tipos de usuarios objetivo**

personas que están aprendiendo sobre sistemas expertos de información creando una comunidad que cree y utilice los conjuntos de información

## **iii. Como se accede a los conjuntos de datos (Crawler, API, descarga directa, entre otros)**

se puede acceder a la información via web en un visualizador de archivos .txt y se puede realizar la descarga directa de los archivos.

## **iv. Formato (ej. Relación de base de datos, archivo de texto plano, entre otros)**

el formato de los archivos son .txt en su mayoría

**<https://www.kaggle.com/>**

### **a. ¿Cuál es el tipo de información que gestiona?**

Gestiona bases de datos y códigos en formato jupyter para el procesamiento de GPU

### **b. ¿Qué información se puede encontrar en el sitio?**

Diferentes conjuntos de datos creados por la comunidad en formatos de tablas de datos

### **i. Objetivo**

Kaggle ofrece un entorno Jupyter Notebooks sin configuración, personalizable. Acceda a GPU gratuitas y a un gran depósito de datos y códigos publicados por la comunidad.

## **ii. Tipos de usuarios objetivo**

Personas que utilizan Jupyter para aprender sistemas de aprendizaje automatico

**iii. Como se accede a los conjuntos de datos (Crawler, API, descarga Directa, entre otros)**

Mediante Júpiter y archivos .csv

**iv. Formato (ej. Relación de base de datos, archivo de texto plano, entre Otros)**

.CSV

4.- Descargue en su computadora un conjunto de datos por cada sitio encontrado en el punto 3 e identifique la siguiente información:

**<https://data.world/>**

a. Autores

Marko Nohanec, Blaz Zupan.

b. Concepto

Base de datos de evaluación de automóviles se derivó de un modelo de decisión jerárquico simple desarrollado originalmente para la demostración de DEX, M. Bohanec, V. Rajkovic: sistema experto para la toma de decisiones. Sistemica 1 (1), pp. 145-157, 1990).

d. Los atributos y su correspondiente tipo de datos

buying: vhigh, high, med, low.

maint: vhigh, high, med, low.

doors: 2, 3, 4, 5more.

persons: 2, 4, more.

lug\_boot: small, med, big.

safety: low, med, high

e. Número de instancias

1728

**<https://www.kaggle.com/>**

a. Autores

Khashayar Baghizadeh Hosseini

b. Concepto

Este es un conjunto de datos del repositorio de aprendizaje automático UCI que se refiere a los valores de la vivienda en los suburbios de Boston.

c. Cantidad de atributos

14

d. Los atributos y su correspondiente tipo de datos

CRIM: tasa de criminalidad per cápita por ciudad

ZN: proporción de tierra residencial dividida en zonas para lotes de más de 25,000 pies cuadrados.

INDUS: proporción de acres de negocios no minoristas por ciudad

CHAS: variable ficticia Charles River (= 1 si el tramo limita con el río; 0 en caso contrario)

NOX: concentración de óxidos nítricos (partes por 10 millones)

RM: número medio de habitaciones por vivienda

EDAD: proporción de unidades ocupadas por el propietario construidas antes de 1940

DIS: distancias ponderadas a cinco centros de empleo de Boston

RAD: índice de accesibilidad a carreteras radiales

IMPUESTO: tasa de impuesto a la propiedad de valor total por \$ 10,000

PTRATIO: relación alumno-profesor por localidad

B:  $1000 (B_k - 0.63)^2$  donde  $B_k$  es la proporción de negros por ciudad

LSTAT:% menor estado de la población

MEDV: valor medio de viviendas ocupadas por sus propietarios en \$ 1000

Valores de atributos faltantes: ninguno

e. Número de instancias

506