AG1 – Actividad Guiada 1

Tema 1: Introducción al proceso KDD

Guía de resolución de actividades

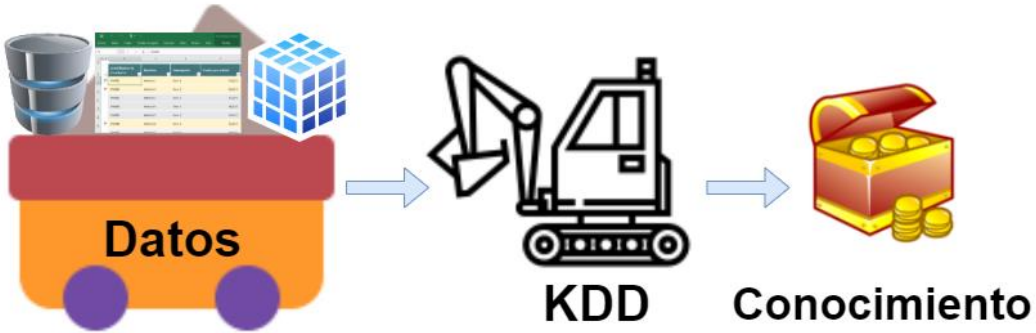
Minería de Datos

viu Universidad Internacional de Valencia

De:
Planeta Formación y Universidades

¿Qué es el proceso KDD?

- El acrónimo KDD hace referencia a un proceso, compuesto de múltiples etapas, cuyo objetivo principal es **la extracción de conocimiento**, que ha de **resultar útil y no ser trivial,** a partir de los datos a los que se tiene acceso.



¿Qué necesitas saber?

- Esta tarea implica generalmente preprocesar los datos, hacer minería de datos (data mining) y presentar resultados
- KDD se puede aplicar en diferentes dominios:
  - determinar perfiles de clientes fraudulentos (evasión de impuestos)
  - descubrir relaciones implícitas existentes entre síntomas y enfermedades, entre características técnicas y diagnóstico del estado de equipos y máquinas
  - determinar perfiles de estudiantes “académicamente exitosos” en términos de sus características socioeconómicas
  - determinar patrones de compra de los clientes en sus canastas de mercado

¿Qué necesitas saber?

- ¿Qué problema quieres resolver?
- ¿Cómo puedes ayudar a tu empresa a generar conocimiento?
- ¿Qué proceso manual se está realizando que se podría realizar de manera automática a través de los datos?

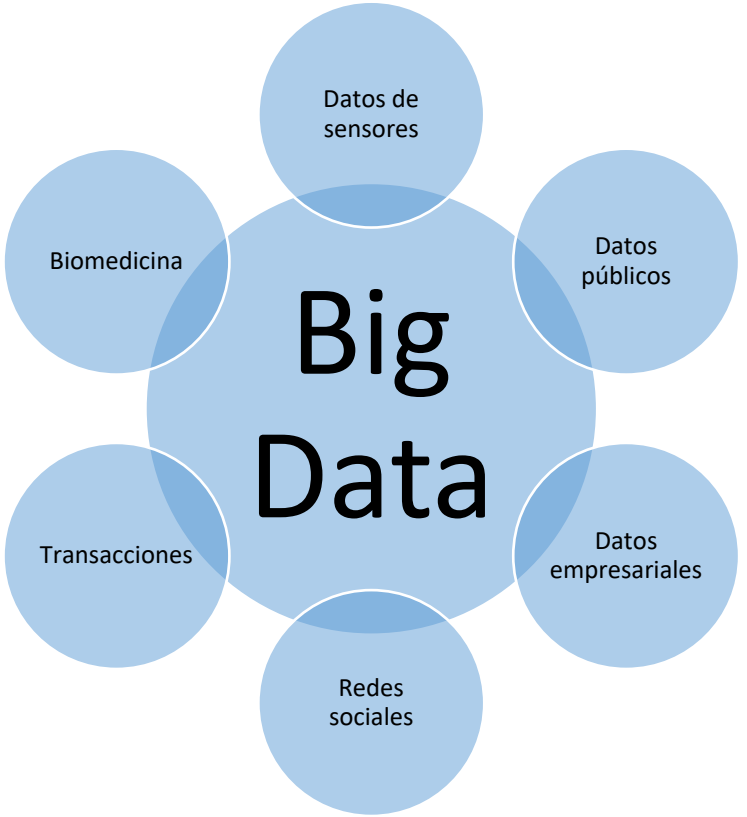
Una vez definido el objetivo

- ¿Dónde saco los datos?
- ¿Es suficiente con una fuente de datos?
- Analizar si los datos iniciales se encuentran procesados de alguna manera

viu Universidad Internacional de Valencia

Identificando fuente de datos potenciales

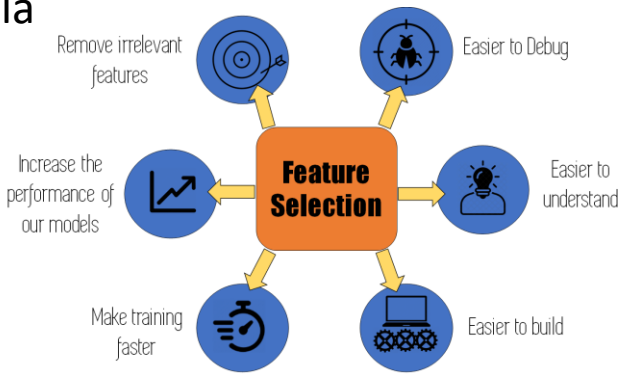
- https://data.nasdaq.com/
- https://opendata.socrata.com/
- https://cloud.google.com/bigquery/public-data/
- https://github.com/datagovsg/datagovsg-datasets
- https://www.ncdc.noaa.gov/cdo-web/datasets
- https://www.who.int/data/gho/gho-search
- https://www.ncbi.nlm.nih.gov/datasets/



viu Universidad Internacional de Valencia

Revisar el / los datasets encontrados me ayudan a resolver el objetivo planteado

- ¿Cuántos registros hay?
  - ¿Están todas las filas completas o tenemos campos con valores nulos?
  - ¿Son demasiado pocos?
- ¿Qué tipos de datos tenemos?
- ¿Vas a trabajar un problema supervisado? -> Identifica la clase.
- Una vez definido el objetivo
- ¿Qué características son más importantes?
- ¿Es suficiente con una fuente de datos?



Revisar el / los datasets encontrados me ayudan a resolver el objetivo planteado

- ¿Hay correlación entre algunas de las variables?
- ¿Tengo conocimiento experto que puedo agregar al dataset?
- ¿Están los datos en la misma unidad?
- ¿la clase de salida o los datos de entrada siguen alguna distribución?

viu Universidad Internacional de Valencia

La importancia de los datos

- Datos numéricos (mínimo 3-5) :
  - Enteros
  - Reales
- Fecha:
  - Hora y minutos
  - Días, meses y años
- String:
  - Cadenas de texto sin patrón común
  - Enumerado: texto que se repite como una categoría (mínimo 2-3)
- Blob:
  - Imagines
  - Audio
  - Video

viu Universidad Internacional de Valencia

Ejemplo de un conjunto de datos (dataset)

	A	B	W	X	Y	Z	AA	AB	AE	AF	AS	BP	BQ	BR	BS	BT	BU	BV	BW
1	domain	requiremen	numberVari	highLevelVa	totalEdgesC	totalWeight	veRatio	weRatio	inputEdgeC	inputEdgeC	inputEdgeH	inputWeigh	outputWeig	outputWeig	outputWeig	num_releva	num_action	h_ff_ratio	rp_fact_bala
2	grounded-st	2	717	30	48129	778846	0.014898	16.1825	716	67.1255	39.4703	27.0938	54425	544.444	3.33778	757	12125	5.352941	-23
3	spider	4	704	44	73979	1967154	0.009516	26.5907	703	104.947	80.4137	322.615	68207	1281.14	49.9313	1500	44229	9.470589	-49
4	termes	2	13	12	73	1606	0.178082	22	12	5.61538	2.29999	3.84023	120	21.3462	1.28008	58	468	2	-5
5	snake	2	208	21	39157	486000	0.005312	12.4116	207	188.255	60.93	15.4224	54291	986.372	3.55371	369	21405	2.222222	-15
6	grounded-st	2	965	4	21901	142440	0.044062	6.50381	200	22.1244	6.48053	2.78165	821	41.735	0.514524	?	?	?	?
7	grounded-st	2	1117	3	30586	186376	0.03652	6.09351	363	26.6634	10.857	1.2406	1547	58.6106	0.403569	?	?	?	?
8	grounded-st	2	80	3	1069	3022	0.074836	2.82694	58	11.85	6.44834	0.276015	207	-9000	0.055203	85	101	1.6	-2
9	data-networ	5	49	3	241	6966	0.20332	28.9046	12	4.91837	0.479483	0.139942	1533	35.132	0.069971	105	1530	2.333333	-8
10	grounded-st	2	1755	6	49537	383406	0.035428	7.73979	394	27.7385	9.41081	2.66004	1525	53.9484	0.322068	?	?	?	?
11	grounded-st	2	345	4	12936	106593	0.02667	8.24003	278	34.7449	16.2396	4.18721	1761	47.9391	0.348934	528	1273	1.588235	-12
12	termes	2	13	12	73	1606	0.178082	22	12	5.61538	2.29999	7.16843	120	19.9615	2.30414	58	468	4	-10
13	grounded-st	2	435	6	8831	52703	0.049258	5.96795	433	17.4506	21.0545	17.6994	1555	31.6333	0.574033	365	1299	1.764706	-13
14	snake	2	150	19	19923	244905	0.007529	12.2926	149	132.82	47.9821	16.6671	28783	686.393	3.60919	271	10549	1.833333	-6
15	organic-synt	3	440	5	10911	43364	0.040326	3.97434	73	24.0409	1.15206	0.118912	1889	62.0477	0.023782	473	5633	2.2	-3
16	data-networ	5	74	1	455	7984	0.162637	17.5473	16	6.14865	0.230919	0.114677	1258	29.9532	0.057338	173	1939	1.2	-4
17	data-networ	5	49	2	247	4823	0.198381	19.5263	11	5.04082	0.395729	0.139942	1158	27.5407	0.069971	113	1148	2.2	-4
18	grounded-st	2	310	4	5948	19823	0.052118	3.33272	256	17.1516	14.5168	0.141532	1200	37.8366	0.028307	186	345	2	-4
19	grounded-st	2	709	7	19469	277870	0.036417	14.2724	685	24.3315	26.3897	21.3766	9472	90.586	0.806311	714	7459	2.470588	-20
20	nurikabe	2	1241	14	20868	1188097	0.059469	56.9339	1233	16.8155	0	0.099273	147807	-9000	0.014182	4120	21354	5.857143	-11
21	nurikabe	2	893	15	14910	591375	0.059893	39.663	879	16.6965	0	0.116992	88089	-9000	0.016713	3561	12777	5.15	-7
22	grounded-st	2	1202	2	33057	198414	0.036361	6.00218	394	26.8103	11.363	2.29115	1575	57.555	0.5908	?	?	?	?
23	data-networ	5	57	1	299	6333	0.190635	21.1806	12	5.24561	0.262572	0.130129	1361	30.1863	0.065065	125	1480	1.4	-4
24	data-networ	5	131	3	867	15763	0.151096	18.1811	19	6.61832	0.299174	0.086704	2839	34.0026	0.043352	258	3803	3.2	-5
25	organic-synt	3	4207	32	318309	1396736	0.013217	4.38799	1030	73.8659	11.928	3.02882	15525	-9000	3.16754	?	?	?	?
26	caldera	4	277	5	28391	2076813	0.009757	73.1504	275	92.7617	27.9582	0.359204	14055	1136.18	0.179602	?	?	?	?
27	data-networ	5	41	1	199	4197	0.20603	21.0905	10	4.85366	0.308515	0.152365	836	26.5504	0.076182	101	996	1.2	-4
28	snake	2	214	27	40830	516727	0.005241	12.6556	213	190.794	69.1257	20.7179	60355	1007.32	4.49058	387	22234	2.571429	-13
29	caldera	4	230	4	19687	1171503	0.011683	59.5064	228	75.6913	21.831	0.689339	6025	442.904	0.328257	?	?	?	?
30	termes	2	13	12	73	3118	0.178082	42.7123	12	5.61538	2.29999	7.29644	228	32.6209	2.43215	91	888	3.857143	-20
31	snake	2	220	33	42653	559999	0.005158	13.1292	219	193.877	76.4675	18.3556	69004	1050.51	3.85903	405	23441	1.714286	-7
32	termes	2	13	12	73	2614	0.178082	35.8082	12	5.61538	2.29999	6.14437	192	29	2.04812	80	748	3.875	-22
33	grounded-st	2	224	4	4287	20195	0.052251	4.71075	217	16.6161	15.1162	22.383	808	31.7746	4.12406	288	697	1.588235	-13
34	snake	2	159	28	21913	285593	0.007256	13.033	158	137.818	58.3468	11.3875	36887	749.603	2.32479	298	11788	3	-7
35	grounded-st	2	410	21	20815	239639	0.019697	11.5128	409	50.7146	27.2571	45.4479	15723	294.027	5.32073	433	4405	5.25	-19
36	organic-synt	3	2797	19	192577	779728	0.014524	4.04892	747	67.3826	10.6782	21.3399	9108	203.1	6.23753	?	?	?	?
37	snake	2	193	34	32413	435151	0.005954	12.4252	192	167.943	71.2959	11.8512	56543	935.939	2.39889	362	17850	2.875	18

viu Universidad Internacional de Valencia

¿Qué se necesita?

- Datos adecuados
- Poder de computación
- Software de **minería de datos**
- Operador cualificado que conoce tanto la naturaleza de los datos como las herramientas de software
- Razón, teoría o corazonada

viu Universidad Internacional de Valencia

Gracias

viu Universidad Internacional de Valencia

De:
Planeta Formación y Universidades