Màster en Ciència de Dades	Juan Antonio Vera
Visualització de dades	19-10-2024

PART I: Selecció Conjunt De Dades

Sumari

Índex de figures

Part I: selecció del conjunt de dades

Aquesta activitat, primera part de la pràctica final, consisteix en la selecció per part de l'estudiant d'un conjunt de dades del vostre interès que serà usat en el projecte de creació de la visualització de dades, d'acord amb uns criteris establerts. Bàsicament, la temàtica és lliure, però es valoraran els aspectes següents:

1. [10%] Justifiqueu breument la vostra selecció

Per aquesta pràctica farem servir dos datasets:

PRIMER DATASET

S'ha decidit fer servir el dataset «Avaluació de quart d'Educació Secundària Obligatòria» disponible a https://datos.gob.es/gl/catalogo/a09002970-evaluacion-de-cuarto-de-educacion-secundaria-obligatoria per els següents motius:

1. Seguint les indicacions de l'enunciat, i tot i que son uns molt bons datasets per realizar estudis fent servir ciència de dades, s'ha optat per no fer servir els datasets disponibles a webs com kaggle i d'altres tal i com s'ha indicat a l'enunciat per dos motius principals:
    1. La majoria d'ells ja estan resolts per alguna altre persona i es tindria el perill de que el treball realitzat a la pràctica no es considerés original.
    2. Trobar un altre dataset que no estigués tan treballat.

2. Conèixer si hi han diferències a la comunitat d'alumnes que justifiquin aconseguir resultats diferents en les proves.

SEGON DATASET

El segon dataset serà «Alumnes matriculats per ensenyament i unitats dels centres docents» i el trobarem a https://analisi.transparenciacatalunya.cat/Educaci-/Alumnes-matriculats-per-ensenyament-i-unitats-dels/xvme-26kg/about_data

Ens servirà per intentar obtenir més informació per a l'estudi a realitzar que si només es tingués en compte el primer.

2.[10%] La rellevància del conjunt de dades en el context.

Són dades actuals? Tracten un tema important per algun col·lectiu concret? S'ha tingut en compte la perspectiva de gènere?

El resultat de les competències dels alumnes de 4rt de la ESO diferenciat per matèria tipus de centre ens ha de marcar la qualitat del nostre sistema educatiu i, en bona part, el futur d'una societat.

S'ha volgut ampliar amb el segon dataset per realitzar l'estudi amb algunes característiques dels tipus de centres que tenim que no es troben en el primer dataset per tal de poder veure si trobem alguna diferència.

Al primer dataset trobarem una variable de gènere que ens ajudarà a avaluar si hi han diferències significatives entre les noies i els nois en els resultats acadèmics.

3. [25%] La complexitat (mesura, variables disponibles, tipus de dades, etc.).

Heu de tenir de l'ordre de milers de registres mínim. I ha de tenir un mínim de l'ordre de desenes de variables. Combina dades categòriques i quantitatives? Inclou altres tipus de dades? Evita els conjunts excessivament simples.

Màster en Ciència de Dades	Juan Antonio Vera
Visualització de dades	19-10-2024

PRIMER DATASET

Darrera actualització	10/10/2024 22:00 (UTC)
URL	https://datos.gob.es/gl/catalogo/a09002970-evaluacion-de-cuarto-de-educacion-secundaria-obligatoria
Registres	96300
Columnes	36

El dataset té més de 96300 registres i 36 columnes/variables. La majoria de les variables son numèriques i es corresponen al resultat de les preguntes en les diferents matèries, així que també afegirem alguna variable més per poder agrupar el resultat.

També ens trobem variables que descriuen el gènere, així com el tipus de centre i població, distingint per número d'habitants (categòriques).

Dins de l'estudi que es realitzarà, categoritzarem en noves variables de treball els tipus de preguntes per matèries (científiques, de llengües, etc)

Amb una mica més de detall, es mostrà una taula amb el significat de les columnes del dataset

ANY	Any que va tenir lloc l'avaluació	any	Text
CODI_ALUMNE	Codi que identifica unívocament els alumnes que van participar en l'avaluació (anonimitzat)	codi_alumne	Text
PCAT	Puntuació global ponderada de la competència lingüística en llengua catalana	pcat	Nombre
PCAT_CL	Puntuació global de comprensió lectora en llengua catalana	pcat_cl	Nombre
PCAT_EE	Puntuació global d'expressió escrita en llengua catalana	pcat_ee	Nombre
PCAST	Puntuació global ponderada de la competència lingüística en llengua castellana	pcast	Nombre
PCAST_CL	Puntuació global de comprensió lectora en llengua castellana	pcast_cl	Nombre
PCAST_EE	Puntuació global d'expressió escrita en llengua castellana	pcast_ee	Nombre
PANG	Puntuació global ponderada de la competència lingüística en llengua anglesa	pang	Nombre
PANG_CO	Puntuació global de comprensió oral en	pang_co	Nombre

	llengua anglesa		
PANG_CL	Puntuació global de comprensió lectora en llengua anglesa	pang_cl	Nombre
PANG_EE	Puntuació global d'expressió escrita en llengua anglesa	pang_ee	Nombre
PFRAN	Puntuació global ponderada de la competència lingüística en llengua francesa	pfran	Nombre
PFRAN_CO	Puntuació global de comprensió oral en llengua francesa	pfran_co	Nombre
PFRAN_CL	Puntuació global de comprensió lectora en llengua francesa	pfran_cl	Nombre
PFRAN_EE	Puntuació global d'expressió escrita en llengua francesa	pfran_ee	Nombre
PMAT	Puntuació global ponderada de competència matemàtica	pmat	Nombre
PMAT_EFM	Puntuació global d'espai, forma i mesura en competència matemàtica	pmat_efm	Nombre
PMAT_CR	Puntuació global de canvi i relacions en competència matemàtica	pmat_cr	Nombre
PMAT_EST	Puntuació global d’estadística en competència matemàtica	pmat_est	Nombre
PMAT_NC	Puntuació global de numeració i càlcul en competència matemàtica	pmat_nc	Nombre
PALE	Puntuació global ponderada de la competència lingüística en llengua alemanya	pale	Nombre
PALE_CO	Puntuació global de comprensió oral en llengua alemanya	pale_co	Nombre
PALE_CL	Puntuació global de comprensió lectora en llengua alemanya	pale_cl	Nombre
PALE_EE	Puntuació global d'expressió escrita en llengua alemanya	pale_ee	Nombre
PCIEN	Puntuació global de competència cientificotecnològica	pcien	Nombre
PCIEN_COMP1	Puntuació global de la competència d'explicar fenòmens naturals i aplicacions tecnològiques utilitzant coneixements científic i tecnològics de competència cientificotecnològica	pcien_comp1	Nombre
PCIEN_COMP2	Puntuació global de la competència de reconeixement dels aspectes principals de la investigació científica de competència cientificotecnològica	pcien_comp2	Nombre
PCIEN_COMP3	Puntuació global de la interpretació	pcien_comp3	Nombre

	d'informació de caràcter cientificotecnològic proporcionada en forma de dades i proves de competència cientificotecnològica		
PCIEN_COMP4	Puntuació global de la competència d'anàlisi i avaluació d'explicacions tecnològiques d'especial rellevància de la competència cientificotecnològica.	pcien_comp4	Nombre
GENERE	Gènere de l'alumne/a que es presenta a l'avaluació	genere	Text
MES_NAIXEMENT	Mes de naixement de l'alumne/a que es presenta a l'avaluació	mes_naixement	Nombre
ANY_NAIXEMENT	Any de naixement de l'alumne/a que es presenta a l'avaluació	any_naixement	Text
NATURALESA	Determina si el centre de l'alumne/a és públic, privat o concertat	naturalesa	Text
ÀREA TERRITORIAL	Nom de l'Àrea Territorial del centre on està matriculat l'alumne	area_territorial	Text
HÀBITAT	Municipis per trams de població	h_bitat	Text

SEGON DATASET

Darrera actualització	15-10-2024
URL	https://194.69.254.91/es/catalogo/a09002970-alumnos-matriculados-por-ensenanza-y-unidades-de-los-centros-docentes
Registres	389K
Columnes	36

Nom columna	Descripció	Nom curt	Tipus de dades
Curs	Curs de referència	curs	Text
Any	Any inferior del curs	any	Nombre
Codi centre	Codi de centre educatiu	codi_centre	Text
Denominació completa	Nom complert del centre educatiu	denominaci_completa	Text
Codi naturalesa	Codi naturalesa Públic/Privat	codi_naturalesa	Text
Nom naturalesa	Públic/Privat	nom_naturalesa	Text
Codi	Codi de la titularitat de la propietat	codi_titularitat	Text

titularitat	del centre educatiu		
Nom titularitat	Titularitat de la propietat del centre educatiu	nom_titularitat	Text
Codi àrea territorial	Codi de l'Àrea Territorial del centre educatiu	codi_delegaci	Text
Nom àrea territorial	Àrea Territorial	nom_delegaci	Text
Codi comarca	Codi comarca del centre educatiu	codi_comarca	Text
Nom comarca	Nom comarca del centre educatiu	nom_comarca	Text
Codi municipi_5	Codificació a 5 dígits	codi_municipi_5	Text
Codi municipi_6	Codificació a 6 dígits	codi_municip_6	Text
Nom municipi	Nom del municipi del centre educatiu	nom_municipi	Text
Codi districte municipal	Codi del districte del municipi del centre educatiu	codi_districte_municipal	Text
Nom districte municipal	Nom del districte del municipi del centre educatiu	nom_dm	Text
Coordenades UTM X	ETRS89	coordenades_utm_x	Nombre
Coordenades UTM Y	ETRS89	coordenades_utm_y	Nombre
Coordenades GEO X	WGS84	coordenades_geo_x	Nombre
Coordenades GEO Y	WGS84	coordenades_geo_y	Nombre
Codi estudis	Codi de l’estudi	codi_estudis	Text
Nom estudis	Nom de l’estudi	nom_estudis	Text
Temàtica	Família professional dels estudis	tem_tica	Text
Grau	Grau de l’ensenyament	grau	Text
Codi ensenyament	Codi de l’ensenyament	codi_ensenyament	Text
Nom ensenyament	Nom de l’ensenyament	nom_ensenyament	Text
Nivell	Nivell de l’ensenyament	nivell	Text

Matrícula concertada	Informació sobre si l’ensenyament al centre es concertat o rep fons públics	matr_cula_concertada	Text
Modalitat	Modalitat en que es cursa l’ensenyament (presencial, semipresencial, distància o lliure)	modalitat	Text
Període matrícula	Període de matriculació	per_ode_matr_cula	Text
Matrícules. Total	Nombre de matrícules. Total	matr_cules_total	Nombre
Matrícules. Dones	Nombre de matrícules. Dones	matr_cules_dones	Nombre
Matrícules. Homes	Nombre de matrícules. Homes	matr_cules_homes	Nombre
Unitats	Nombre d'unitats o grups	unitats	Nombre
Georeferència		georefer_ncia	Text

4. [25%] L'originalitat.

Es valora no repetir els conjunts de dades clàssiques o molt treballades Links to an external site. Ni temes ja molt tractats (p. ex. Covid-19, trànsit, criminalitat...) Podeu combinar o millorar el conjunt de dades. En el primer cas, enriquir el conjunt de dades amb altres de diferents per donar un enfocament nou. En el segon cas, generant noves mètriques o indicadors amb les variables existents mitjançant transformacions. Hi ha altres visualitzacions basades en aquest conjunt de dades? És una evolució o una actualització d'un conjunt anterior? Heu enriquit un conjunt de dades ja existent?

S’ha intentat evitar els datasets com COVID19, etc .. perquè es troben molts de treballats i seria susceptible a que el resultat de la PAC no fos original.

Com ja s’ha comentat, generarem noves variables dins de l’estudi que ens agrupin el tipus d’estudis. A més, el fet de fer servir una segon font de dades li donarà més context a l’estudi realitzat.

5. [30%] Les qüestions que respondreu amb la visualització de dades, tenen en compte els punts anteriors?

No hem trobat estudis publicats tret de les notícies a TV a on es comenten els resultats dels estudis realitzats (de ben segur que hi existiran, però no hem trobat notebooks al respecte a plataformes com kaggle).

Pretenem conèixer:

1. Evolució dels resultats per grups de coneixements
2. Distinció dels resultats entre tipus de centres (concertat vs públic)
3. Avaluar si els resultats sobre algunes matèries es traslladen a d'altres.