

ELENCO DELLE CATEGORIE DI ESERCIZI PER PROVE TRATTATE

07_07_23 → https://github.com/simonebugo/Big_Data/blob/main/7_7_23.ipynb
https://drive.google.com/drive/u/0/folders/1dFmf8jmfl_eKKJq2ggqfPQpAnyUG0vS)

Analisi: map, groupby, pivot_table, istogramma

Transformazione e analisi: decision tree + KNC + dummy. Grid search su dectree.

Demographic parity -> la probabilità di predire una classe con un determinato attributo è maggiore di un'altra

Equalized dott -> accuratezza con attributi diversi è diversa

Simple imputer

Linear regressor può sostituire il dec tree? (2 classi -> si applica soglia)

25_07_23 → https://github.com/simonebugo/Big_Data/blob/main/25_07_23.ipynb
<https://drive.google.com/drive/u/0/folders/1FssncbTd9tdX1hJCAm2WMSN5p811KEbd>)

Analisi: map, groupby, pivot. Isto e boxplot

Transformazione e analisi: ordinal encoder. decision tree + KNC + dummy. Grid search su dectree. Simle imputer, kbin disc, feature union (versione feature union con replicazione delle feature trasformate)

Linear regressor può sostituire il dec tree? (+ classi -> si round + clip)

13_06_24 → https://github.com/simonebugo/Big_Data/blob/main/13_06_24.ipynb
https://drive.google.com/drive/u/0/folders/1_sk836OXdBEd_EfEEjEIUDBAB5F7AEa9)
(mobile price class)

Analisi : Istogramma + Istogramma su vari gruppi. Pivot. Scatter plot (con condizioni)

Transformazione e analisi: decision tree + KNC + dummy. K-fol cross validation. Grid search.

Correlazione tra feature nel dataset (5 feature + correlate)

Select k best in pipeline -> grid search per migliore k

Truncated svd in Feature Union con k fold

26_07_24 → https://github.com/simonebugo/Big_Data/blob/main/26_07_24.ipynb
<https://drive.google.com/drive/u/0/folders/19z8kd6nmvxtvFULUhBbaLrtg73cA57RT>)

Analisi: group by, pivot table. Istogramma e istogramma su vari gruppi, utilizzo di cut per creare nuova feature descr. Scatter plot

Transformazione e analisi: linear regresso + sgd. Parametri MSLE e R2. Grid search con 2 misure. Select k best

Caratteristiche particolari: esercizio pipeline richiede che si aggiunga una colonna contenente i valori di una feature discretizzati (oltre a quelle già presenti)

10_09_24 → https://github.com/simonebugo/Big_Data/blob/main/10_09_24.ipynb
<https://drive.google.com/drive/u/0/folders/1CR3O9CfGUoUsU0EkVGvIZZNwNDAXcof>)

Analisi. Cut per discretizzare+ group by. Pivot table. Istogramma con 2 gruppi

Transformazione e analisi: decision tree + KNC + dummy. K fold cross val.

Probabilità su gender e accuracy (Demographic parity e equalized dot)

Aggiunta nuova colonna come eq delle altre

Grid search

MinMax (attributo scalato 0-1)

Trunc svd in normal pipeline + grid search

10_01_25 → https://github.com/simonebugo/Big_Data/blob/main/10_01_25.ipynb
(<https://drive.google.com/drive/u/0/folders/1CohSaUvVkJZlZ4jwO7FpGcnhFbCt0ICx>)

Analisi: Scatter su nuova colonna. Cut per nuova colonna. Pivot table

Box Plot di seaborn

Transformazione e analisi:

eliminare attributi con + del 50% di valori nulli

decision tree + KNC + dummy

k score cross val

permutation feature importante (PFI) per analizzare feature importance dataset

pipeline differenziata per num_cols e categorical_cols

trunc svd

18_02_25 → https://github.com/simonebugo/Big_Data/blob/main/18_02_25.ipynb
(<https://drive.google.com/drive/u/0/folders/1-wiWMgHKwQFie7FyWjc-X6kF7viCYRU>)

Analisi: nuova colonna + group by. Cut per discretizzare. Pivot table. Scatter plot

Transformazione e analisi: Extre tree classifier, k fold cross val

Probabilità su gender e accuracy (Demographic parity e equalized dot) per extra tree e per knc

Correlazione tra feature (2 più pos e 2 più negativ)

Gird search

16_04_25 → https://github.com/simonebugo/Big_Data/blob/main/16_04_25.ipynb
(https://drive.google.com/drive/u/0/folders/1fnKgOLBJDCVBMTjb5A_2LLMllSPPLb8G)

Analisi: pivot, cut per disc, heatmap , e istogramma

Transformazione e analisi: decision tree + KNC + dummy. K fold cross.

LOCO -> leave one covariate out (feature importance)

Grid search. Select k best

Feature union

18_06_25 → https://github.com/simonebugo/Big_Data/blob/main/18_06_25.ipynb
(https://drive.google.com/drive/u/0/folders/1FLLe_xJhTVEBiWn9MjVxn9fnVtqAlIV-)
(simulazione)

Analisi: verifica media attributi, cut per discretizzare, map per cambiare i valori, pivot table.

Adozione servizio extra riduce tasso abbandono

Nuova feature che conti numero di servizi + isto

Transformazione e analisi:

log regression + random forest + dummy

Probabilità su gender e accuracy (Demographic parity e equalized dot) per extra tree e per knc
Grid search

Correlazione feature (positive e negative (abs))

Pipeline su num e cat cols

One hot encoder

Select k best

Fairness →

https://github.com/simonebugo/Big_Data/blob/main/Esercizio_in_aula_Fairness.ipynb

(<https://drive.google.com/drive/u/0/folders/1mVZNCuyZGfCjzTT9eLLWO8FTkcoTGi2X>)

Col transf su dataset

SGD + model_coef per vedere i pesi assegnati ad ogni feature

demographic_ratio della libreria fairlearn

PFI + boxplot

LOCO + scatter

8_01_26 → https://github.com/simonebugo/Big_Data/blob/main/8_1_26.ipynb
(<https://drive.google.com/drive/u/0/folders/1PIDaGYhLAGJZuEAR98JwMVEWH-DDH1cm>)

Analisi:

confronto tra prob di categorie

cut con fasce precise

pivot table

group by + mean e median

boxplot

Transformazione e analisi:

dummy extra tree e knc

Probabilità su gender e accuracy (Demographic parity e equalized dot) per extra tree

Minmax e std scaler

Trasformazione su colonne categoriche e numeriche (ricordati di gestire prima il target)

K best