ASSIGNMENT 1 - POLITECNICO DI TORINO

TECHNOLOGY REPLY

Dataset: https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification (altered)

Students will be given a dataset containing information regarding a market analysis in the mobile telephony sector. The dataset will contain information regarding the technical characteristics of the devices and price ranges.

The dataset is called: “A1_dataset.csv”.

The purpose of the analysis will be to study the various characteristics of the devices, in order to understand how these affect the price of the device.

Furthermore, three classification models will be implemented.

- `'battery_power': device battery power (mAh),`
- `'blue': Bluetooth availability,`
- `'clock_speed': micropocessor,`
- `'dual_sim': support for dual sim,`
- `'fc': front camera megapixels,`
- `'four_g': 4G,`
- `'int_memory': internal memory (Gigabyte),`
- `'m_dep': depth (cm),`
- `'mobile_wt': weigth,`
- `'n_cores' : processor cores,`
- `'pc': main camera megapixels,`
- `'px_height': pixels resolution,`
- `'px_width': pixels resolution,`
- `'ram': RAM memory (Megabyte),`
- `'sc_h': screen height(cm),`
- `'sc_w': screen width(cm),`
- `'talk_time': on-call battery life,`
- `'three_g': 3G,`
- `'touch_screen': touch screen,`
- `'wifi': wifi,`
- `'price_range': device price range (label)`

Tasks:

1. Load data: reading dataset information from the attached .csv file
2. Explorative analysis
    a. Visualize the first 5 entries of the dataset
    b. Check for null values
    c. Show info about the dataframe: columns, data types, non-null values and memory footprint
    d. Data summary. For each numerical column evaluate min, max, avg, …
    e. Using *pandas.plot* visualize all features distributions as boxplots

f. Check whether the dataset is balanced

3. Data manipulation
    a. Change null values for ‘dual_sim’ with zeros
    b. Change ‘wifi’ e ‘blue’ according to the following mapping:
        - Y : 1
        - N : 0
4. Correlation
5. Statistical analysis of features through graphic libraries
    a. Number of devices per ‘price_range’
    b. Column values distribution by target (boxplot)
    c. Relation between ‘px_width’ and ‘px_height’
    d. Relation between ‘fc’ e ‘pc’
    e. Visualize as barchart the number of devices for the various value of 'n_cores'
    f. Number of devices by ‘four_g’ and ‘three_g’

6. New features
    a. Add column ‘sc_dim’ defined as (‘sc_w’ * ’sc_h’)
    b. Add column ‘px_dim’ defined as ('px_width' * 'px_height')
    c. Add column ‘3g_4g’ which will be defined according to values ‘four_g’ and ‘three_g’, as follow:
        - 0-> 3G:n 4G:n
        - 1-> 3G:s 4G:n
        - 2-> 3G:n 4G:y
        - 3-> 3G:y 4G:y
7. Dropping features
    a. Drop the columns: ‘sc_w’, ‘sc_h’, 'px_width', 'px_height', ‘four_g’, and ‘three_g’
8. Analyse correlation on the modified dataset
9. Split target variable y (‘price_range’) from the other features (x)
10. Normalization
    a. Normalize the dataset (from sklearn.preprocessing import StandardScaler)
    b. Visualize using histograms all features distribution before and after normalization
11. Split X (normalized) e y in train and test set with a 70:30 ratio
12. Classification: using sklearn implement the following classification models
    a. Decision Tree
    b. Logisic Regression
    c. Support Vector Machines
13. Compare the end results by their accuracy

(Optional)

14. LDA: use LDA to achieve dimensionality reduction
15. Try classification again
    a. Decision Tree
    b. Logisic Regression
    c. Support Vector Machines
16. Compare the end results by their accuracy