Sebelumnya terimakasih atas kesempatan dan waktunya mas Egi. Selamat malam, perkenalkan nama saya Soni Adiyatma, saya merupakan mahasiswa S2 Statistika Institut Teknologi Sepuluh Nopember (ITS). Pada kesempatan kali ini saya akan sharing tentang data preprocessing.

Pengetahuan saya tentang data preprocessing masih belum seberapa ekspert namun disini saya akan sharing yang saya ketahui tentang data preprocessing, jika sekiranya saya kurang dalam penyampaiannya mohon dibenarkan. Baiklah akan saya mulai.

Data preprocessing merupakan suatu proses persiapan dalam melakukan analisis terhadap suatu data. Data seringkali tidak lengkap, tidak konsisten dan cenderung masih belum siap untuk dilakukan analisis.

Dalam melakukan data preprocessing, kita harus terlebih dahulu mengetahui tujuan menganalisis suatu data untuk apa sehingga dapat ditentukan sebaiknya menggunakan metode apa. Dalam data preprocessing, treatment yang dilakukan berbeda beda tergantung tujuan melakukan analisis data dan metodenya. Namun langkah awal secara umum menurut saya dalam melakukan preprocessing adalah Explanatory Data Analysis (EDA).

Tujuan EDA adalah untuk mengetahui karakteristik suatu data. Misal pada data timeseries akan dilakukan forecast. Sebelum melakukan forecast terlebih dahulu adalah melihat pola data apakah memiliki tren dan berpola tiap bulanan/tahunan, sehingga ketika mengunakan ARIMA maka harus melakukan pengujian stasioneritas. Preprocessing data untuk menggunakan ARIMA adalah transformasi data agar stasioner.

Misal ingin mengetahui hubungan sebab akibat menggunakan regresi, maka terlebih dahulu melakukan EDA dan melihat pola data. Misal variabel/feature merupakan kombinasi linier dengan variabel/feature yang lain dengan didapatnya dari hasil scatter plot maka preprocessing data salah satunya dengan menggunakan PCA.

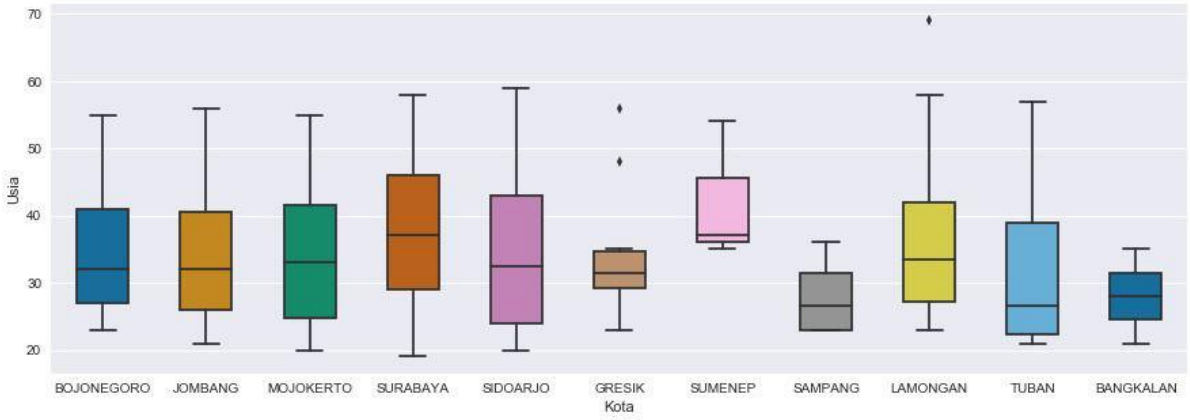
Langkah-langkah data preprocessing:
1. EDA
2. Data Cleaning
3. Data Transformation
4. Feature Reduction

langkah-langkah tersebut juga tergantung karakteristik suatu data dan tujuannya untuk apa;

1.Explanatory Data Analysis
- Statistika Deskriptif:
Tujuan:
- Mengetahui apakah ada missing value atau tidak
- Mengetahui apakah ada outlier atau tidak
- Mengetahui apakah ada format yang beda
- Visualisasi:
- mengetahui karakteristik/pola data
- Mengetahui apakah ada outlier atau tidak menggunakan plot
- Mengetahui persebaran data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 170 entries, 0 to 169
Data columns (total 12 columns):
Lokasi Dealer       170 non-null object
Tipe                170 non-null object
Sub Kategori        170 non-null object
Warna               170 non-null object
Jenis Kelamin       170 non-null object
Kota                170 non-null object
Propinsi            170 non-null object
Usia                170 non-null int64
Jenis Pembayaran    170 non-null object
Pekerjaan           170 non-null object
Pendidikan          170 non-null object
Harga Jual          170 non-null int64
dtypes: int64(2), object(10)
memory usage: 16.0+ KB
```

ini contoh melihat apakah data ada missing value atau tidak menggunakan python



berikut contoh juga melihat persebaran usia tiap daerah dan daerah mana yang terdapat outlier dengan menggunakan boxplot

Sebelum melakukan cleaning data, sebaiknya harus paham terlebih dahulu bedanya noise dan outlier. Outlier merupakan data yang beda/pencilan dan biasa disebut anomali dimana outlier memiliki arti dalam suatu data, sedangkan noise merupakan data yang beda dimana data tersebut tidak memiliki arti. Contohnya kesalahan format dalam menginputnya.

2. Data Cleaning
- Menyamakan format penulisan
- mengatasi missing value:
mengatasi missing value tergantung datanya seperti apa. Kalau datanya kategorik lebih baik inputasinya menggunakan modus, kalau datanya kontinyu lebih baik menggunakan mean atau median
- Mengatasi noise:
noise dalam data sebaiknya dihapus karena tidak dapat berpengaruh, namun bisa juga noise didefinisikan sebagai missing value, sehingga dapat diatasi menggunakan analisis missing value.

dalam missing value untuk menentukan kapan pake mean dan pake median adalah ketika data terdapat outlier lebih bagus menggunakan median.

TABLE 2 Comparison of Imputation Techniques for Missing Data

Imputation Method	*Advantages*	*Disadvantages*	*Best Used When:*
		Imputation Using Only Valid Data	
Complete Data	• Simplest to implement • Default for many statistical programs	• Most affected by nonrandom processes • Greatest reduction in sample size • Lowers statistical power	• Large sample size • Strong relationships among variables • Low levels of missing data
All Available Data	• Maximizes use of valid data • Results in largest sample size possible without replacing values	• Varying sample sizes for every imputation • Can generate "out of range" values for correlations and eigenvalues	• Relatively low levels of missing data • Moderate relationships among variables
		Imputation Using Known Replacement Values	
Case Substitution	• Provides realistic replacement values (i.e., another actual observation) rather than calculated values	• Must have additional cases not in the original sample • Must define similarity measure to identify replacement case	• Additional cases are available • Able to identify appropriate replacement cases
Hot and Cold Deck Imputation	• Replaces missing data with actual values from the most similar case or best known value	• Must define suitably similar cases or appropriate external values	• Established replacement values are known, or • Missing data process indicates variables upon which to base similarity
		Imputation by Calculating Replacement Values	
Mean Substitution	• Easily implemented • Provides all cases with complete information	• Reduces variance of the distribution • Distorts distribution of the data • Depresses observed correlations	• Relatively low levels of missing data • Relatively strong relationships among variables
Regression Imputation	• Employs actual relationships among the variables • Replacement values calculated based on an observation's own values on other variables • Unique set of predictors can be used for each variable with missing data	• Reinforces existing relationships and reduces generalizability • Must have sufficient relationships among variables to generate valid predicted values • Understates variance unless error term added to replacement value • Replacement values may be "out of range"	• Moderate to high levels of missing data • Relationships sufficiently established so as to not impact generalizability • Software availability
		Model-Based Methods for MAR Missing Data Processes	
Model-Based Methods	• Accommodates both nonrandom and random missing data processes • Best representation of original distribution of values with least bias	• Complex model specification by researcher • Requires specialized software • Typically not available directly in software programs (except EM method in SPSS)	• Only method that can accommodate nonrandom missing data processes • High levels of missing data require least biased method to ensure generalizability

ini beberapa metode dalam imputasi missing value

3. Data Transformation

- Label encoder

mengubah format data teks menjadi suatu kode. misal : Jenjang pendidikan=[SD,SMP,SMA] menjadi Jenjang pendidikan=[1,2,3]

- Diskritisasi/binning

mengubah bilangan kontinyu menjadi diskrit atau code juga. misal Usia 1-5 diubah menjadi code balita, 6-15 menjadi anak-anak dan usia 16-22 menjadi remaja.

- scalling

mengubah suatu data yang memiliki varia

si besar sehingga memiliki skala yang tidak terlalu besar

- normalization
untuk mengatasi outlier dan data mengikusi distribusi normal

4. Data Reduction
Data reduction memiliki tujuan untuk mempercepat running suatu metode tertentu dan efisiensi dalam model suatu metode
-Feature Selection
Dalam feature selection salah satu contohnya RFE
-Feature Extraction
Tujuannya adalah menggabungkan/mengelompokkan feature menjadi beberapa feature saja. salah satu contoh metodenya yaitu PCA

Sesi Tanya jawab;

Pertanyaan pertama dari @Saipul :

RFE singkatan dari apa ya ?

Jawaban :

Recursive Feature Elimination

Pertanyaan kedua dari @Andreas Chandra :

Adakah hal khusus ketika suatu variable memiliki ribuan kategori? contohnya kota/kab tmpt tinggal. dan ga bisa dibikin faktorial. kalo pun one hot. bakalan ada ratusan column?

Jawaban :

Metode balik lagi ke tujuannya. tujuan analisisnya seperti apa Pak kalo boleh tau? Dikelompokkan lagi berdasarkan provinsi, misal menggunakan algorima naive bayes atau decision tree bagus untuk yang memiliki feature paling banyak kategorik.

Pertanyaan ketiga dari @Rian Apriansyah

Perkenalkan saya Rian. saya masih baru dalam dunia data. Saya ingin menanyakan tentang bagian feature extraction. Ketika kita melakukan preprocess data text misal data tweet, kita akan mengubah data text menjadi angka dalam bentuk vektor melalui proses feature extraction apakah itu memakai TF-IDF atau metode2 yang lain. Pertanyaan saya, bisa kah kita menggabungkan 2 atau lebih feature yang sudah berbentuk vektor menjadi 1 feature baru?

Jawaban :

Saya belum seberapa expert didalam NLP tapi kalo misal ekstraktion dan ukurannya gak sama featurenya maka tidak bisa digabung Pak

Pertanyaan ke-empat dari @Wimi Sartika

Mas soni @sadiyatma .. nanya donk.. konsep RFE itu apakah sama dengan pemodelan? Hasil outputnya sprt apa? Ada bbrp script jg mengeluarkan masing2 acc, senstivity. Kalau sudah melakukan RFE dengan rekomendasi features2 yg diberikan, apakah perlu melakukan pemodelan lain?

Jawaban :

konsepnya mungkin seperti ini Bu, jadi diawal kan kita udah nentuin pake metodenya lalu RFE itu bergantung pada suatu metode yang sudah kita tentukan berdasarkan tujuannya. Dan konsepnya itu

melihat feature mana yang memiliki pengaruh paling kecil dalam model menggunakan metode tersebut akan dihilangkan

kalo misal pake python ini ada beberapa documentasinya https://www.scikit-yb.org/en/latest/api/features/rfecv.html

pertanyaan kelima dari @Ahmad Ilham

Mau nanya, sy awam juga di bagian ini, sy nanya di scope umum di penggunana feature extraction (FE) untuk kasus supervised. Kapan FE digunakan mas? Dan apa pentingnya FE untuk pemodelan prediksi misalnya!?

Jawaban :

kalo saya pribadi balik lagi ke tujuannya. feature extraction kan tujuannya menggabungkan beberapa feature menjadi lebih sedikit dan tidak mengurangi informasi pada data

kalo menggunakan feature selection kan kita mengurangi feature yang otomatis informasinya ada yang berkura