

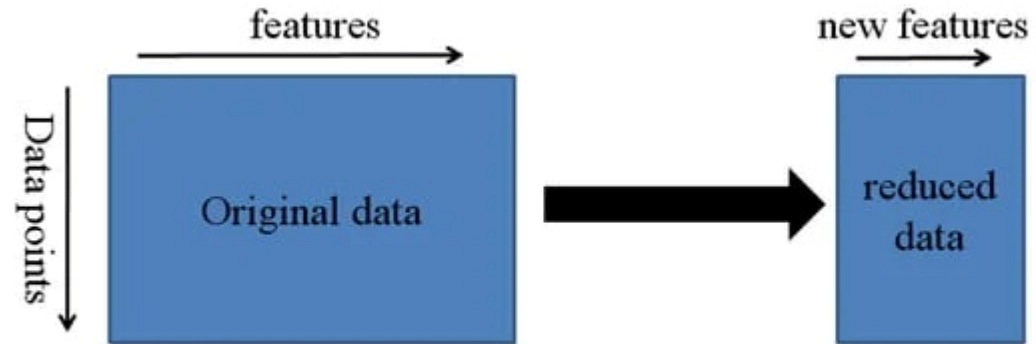
KONSEP *DIMENSIONALITY REDUCTION*

TEAM TEACHING

What is Dimensionality Reduction

- Dimensionality reduction merupakan proses pengurangan dimensi atau fitur dari dataset dengan mempertahankan informasi sebanyak mungkin dari dataset tersebut.
- Dimensionality reduction → mentransformasi data dari ruang dimensi tinggi ke ruang dimensi rendah sehingga representasi berdimensi rendah mempertahankan beberapa properti yang bermakna dari data asli, idealnya mendekati dimensi intrinsiknya.
- Dimensionality reduction prinsipnya sama dengan mengompres *file* yang berukuran besar menjadi *zip file*. Kompresi *file* tidak akan mengurangi informasi yang ada di dalam *file* tersebut, hanya membuatnya lebih sederhana sehingga mengurangi ukuran *file* sehingga mempercepat proses transfer *file*.

What is Dimensionality Reduction



Tujuan Dimensionality Reduction

- Mengurangi kompleksitas model.
- Meningkatkan performa learning algorithm.
- Mempermudah memvisualisasikan data.
- Mengurangi ruang penyimpanan dikarenakan dimensi lebih kecil sehingga waktu komputasi berkurang dan meningkatkan akurasi keseluruhan model machine learning.
- Menyederhanakan kumpulan data klasifikasi atau regresi agar lebih sesuai dengan model prediksi.
- Dapat menangani multikolinearitas dengan menghapus fitur yang berlebihan.

Teknik Dimensionality Reduction

- Feature selection
- Feature extraction

Feature Selection

- Teknik Feature selection mempertahankan fitur yang relevan (optimal) dan membuang fitur yang tidak relevan untuk memastikan akurasi model yang tinggi.
- Metode yang menggunakan Feature selection, antara lain:
 - a. Backward feature elimination
 - b. Forward feature selection
 - c. Random forest
 - d. dan sebagainya

Backward Feature Elimination

- Backward feature elimination biasanya digunakan untuk mengembangkan model regresi linier atau regresi logistik.
- Backward feature elimination dapat menentukan jumlah fitur yang penting untuk algoritma Machine Learning berdasarkan perkiraan kinerja model dan tingkat kesalahan yang dapat ditoleransi.
- Pada Scikit-learn dapat menggunakan **f_regression** (untuk regresi) dan **f_classif** (untuk klasifikasi)

Forward Feature Selection

- Forward feature selection kebalikan dari backward feature elimination.
- Forward feature selection mengandalkan penentuan karakteristik terbaik yang menghasilkan perolehan di atas rata-rata dalam performa model.
- Pada Scikit-learn dapat menggunakan `RFE()`

Random Forest

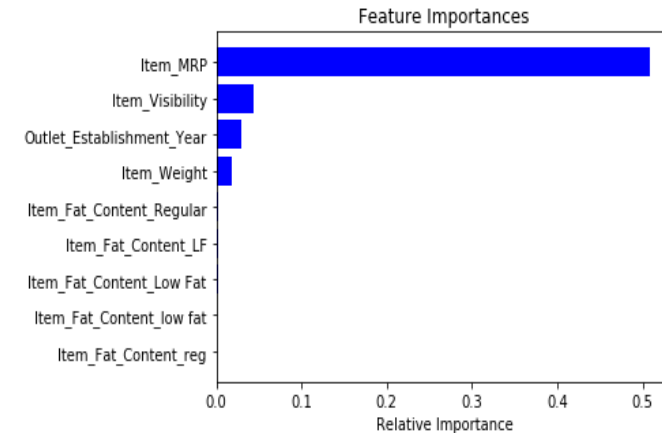
- Random forest merupakan pendekatan pemilihan fitur dengan paket signifikansi fitur bawaan yang mengidentifikasi kepentingan fitur yang mengakibatkan kebutuhan untuk memprogramnya secara individual dihilangkan.
- Dalam random forest, beberapa decision tree dibangun terhadap fitur target untuk mengidentifikasi subset variabel menggunakan statistik untuk setiap atribut.
- Random forest menerima input dalam bentuk data numerik, encoding process sangat penting untuk mengubah semua jenis data input menjadi data numerik.

Random Forest

- Contoh Program:

```
from sklearn.ensemble import RandomForestRegressor
df=df.drop(['Item_Identifier', 'Outlet_Identifier'], axis=1)
model = RandomForestRegressor(random_state=1, max_depth=10)
df=pd.get_dummies(df)
model.fit(df,train.Item_Outlet_Sales)

features = df.columns
importances = model.feature_importances_
indices = np.argsort(importances)[-9:] # top 10 features
plt.title('Feature Importances')
plt.barh(range(len(indices)), importances[indices], color='b', align='center')
plt.yticks(range(len(indices)), [features[i] for i in indices])
plt.xlabel('Relative Importance')
plt.show()
```



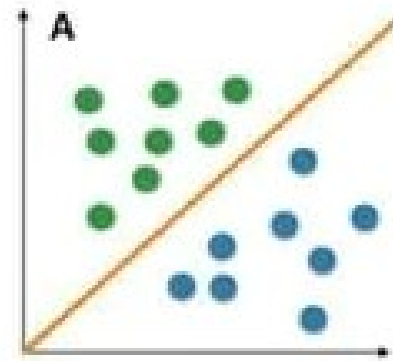
Dimensionality reduction

```
from sklearn.feature_selection import SelectFromModel
feature = SelectFromModel(model)
Fit = feature.fit_transform(df, train.Item_Outlet_Sales)
```

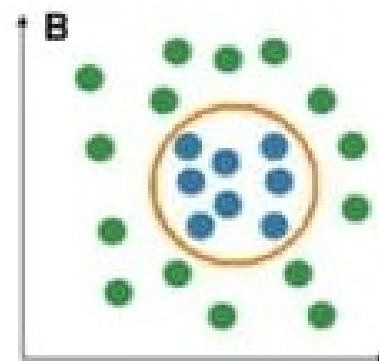
Feature Extraction

- Teknik Feature extraction disebut juga teknik feature projection.
- Teknik Feature extraction → ruang multidimensi diubah menjadi ruang dengan dimensi yang lebih rendah.
- Metode yang menggunakan Teknik Feature extraction, antara lain:
 - a. Metode Linear
 - b. Metode Non Linear (Manifold Learning)

Linear Vs Non Linear



Linear



Non Linear

Feature Extraction

- Metode Linear
 - a. Principal component analysis (PCA)
 - b. Linear discriminant analysis (LDA)
 - c. Factor Analysis (FA)
 - d. Truncated Singular Value Decomposition (SVD)
 - e. Independent component analysis (ICA)
 - f. dan sebagainya

Principal Component Analysis (PCA)

- PCA merupakan teknik Dimensionality reduction yang paling populer digunakan pada machine learning.
- PCA merupakan unsupervised machine learning algorithm yang digunakan untuk dimensionality reduction.
- PCA merupakan analisis data kunci dan teknik pemodelan prediktif.
- PCA menggunakan operasi matriks sederhana dari aljabar linier dan statistik untuk menghitung proyeksi dari data asli ke dalam dimensi dengan jumlah yang sama atau lebih sedikit.

Principal Component Analysis (PCA)

- PCA melakukan transformasi ortogonal untuk mengubah observasi karakteristik yang berkorelasi menjadi sekumpulan fitur yang berkorelasi linier.
- Pada scikit-learn menggunakan PCA().
- Contoh:

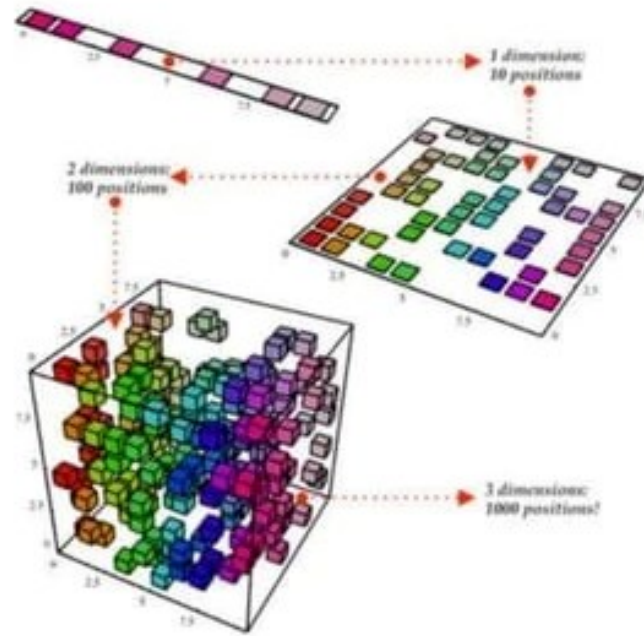
```
from sklearn.decomposition import PCA

pca = PCA(n_components=2)
fit_pca = pca.fit_transform(x)
pca_df = pd.DataFrame(data = fit_pca, columns = ['PCA_1', 'PCA_2'])
pca_df.tail()
```

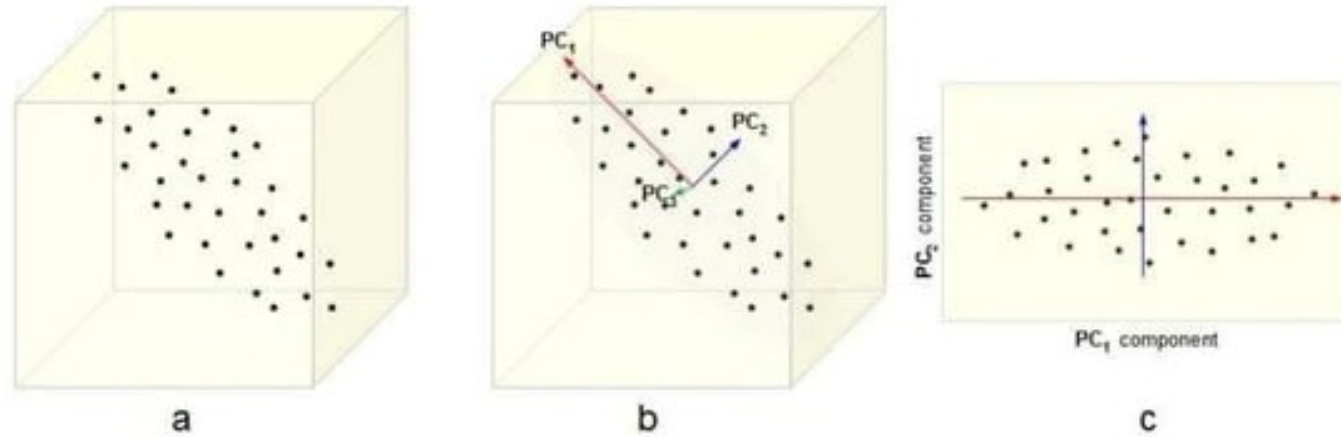


	PCA_1	PCA_2
95	-0.829311	-0.264008
96	0.517545	-0.270995
97	0.284249	0.614142
98	0.357980	-0.000400
99	-0.608530	-0.095601

Principal Component Analysis (PCA)



Principal Component Analysis (PCA)



Linear Discriminant Analysis (LDA)

- LDA merupakan salah satu algoritma machine learning untuk klasifikasi yang juga dapat digunakan untuk reduksi dimensi.
- Cara kerja LDA → dengan menghitung statistik ringkasan untuk fitur-fitur input menurut label, seperti *mean* dan standar deviasi.
- Contoh:

```
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis

lda = LinearDiscriminantAnalysis(n_components=2)
fit_lda = lda.fit(x,y).transform(x)
lda_df = pd.DataFrame(fit_lda, columns = ['LDA_1', 'LDA_2'])
lda_df.tail()
```



	LDA_1	LDA_2
95	-0.598649	1.172701
96	1.307304	-0.589163
97	0.939720	-0.418005
98	-0.965420	0.484157
99	0.386594	-0.712115

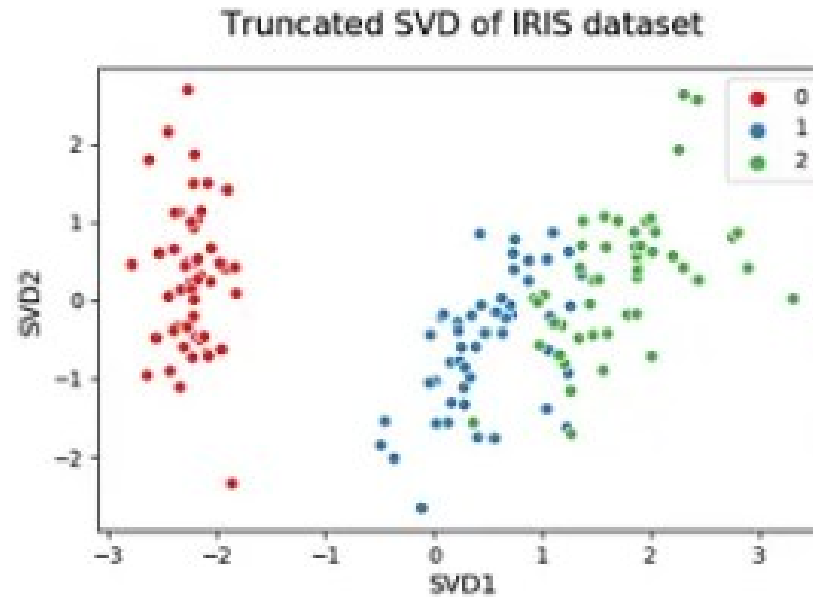
Factor Analysis (FA)

- FA → menentukan hubungan antara sekelompok variabel dan kemudian memutuskan untuk mempertahankan variabel tertentu berdasarkan korelasi variabel yang kuat.
- Dengan demikian, variabel dalam suatu kelompok mungkin berkorelasi kuat tetapi mungkin menunjukkan korelasi yang lemah dengan kelompok lain.
- Setiap variabel dipertahankan atau dihilangkan berdasarkan faktor korelasi tersebut.

Truncated Singular Value Decomposition (SVD)

- SVD bekerja dengan baik dengan sparse data dengan banyak nilai barisnya nol. Sebaliknya, PCA bekerja dengan baik dengan dense data.
- Truncated SVD juga dapat digunakan dengan dense data.
- Perbedaan utama lainnya antara truncated SVD dan PCA adalah faktorisasi untuk SVD dilakukan pada matriks data sedangkan faktorisasi untuk PCA dilakukan pada matriks kovarians.

Truncated Singular Value Decomposition (SVD)



Independent Component Analysis (ICA)

- ICA mencakup **blind source separation** dan **cocktail party problem**.
- ICA mengacu pada metode reduksi dimensi linier yang bertujuan untuk mengidentifikasi komponen independen dalam kumpulan data yang disediakan.

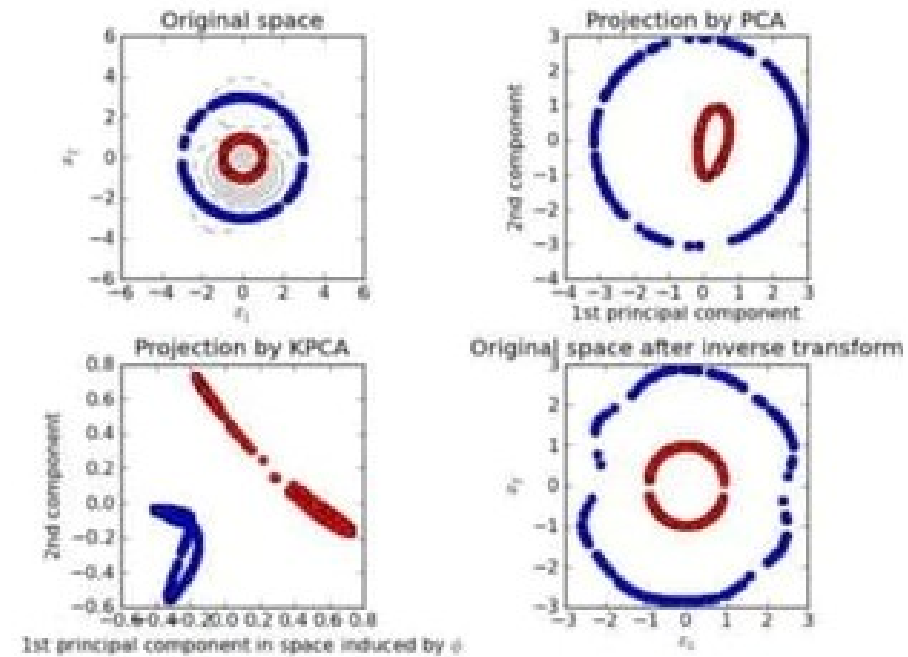
Feature Extraction

- Metode Non-Linear
 - a. Kernel PCA (K-PCA)
 - b. t-distributed Stochastic Neighbor Embedding (t-SNE)
 - c. Multidimensional Scaling (MDS)
 - d. Isometric mapping (Isomap)
 - e. Non linear component analysis (NLCA)
 - f. dan sebagainya

Kernel PCA

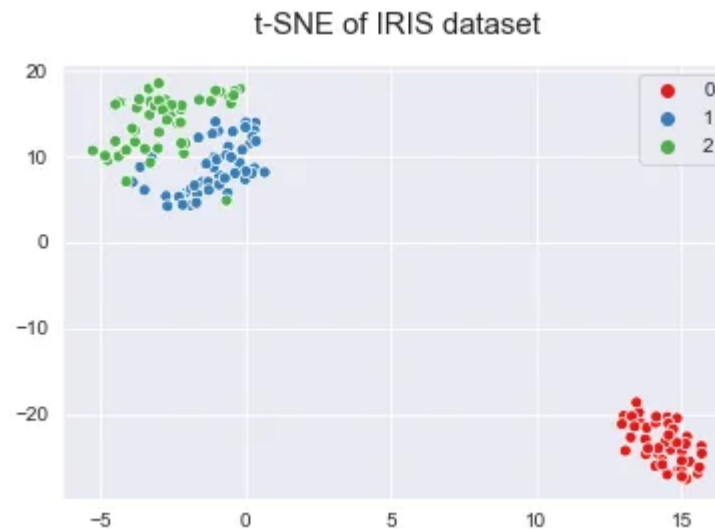
- Kernel PCA merupakan teknik dimensionality reduction non-linear yang menggunakan kernel.
- Kernel PCA merupakan bentuk non-linear dari PCA normal.
- Kernel PCA bekerja dengan baik dengan kumpulan data non-linier di mana PCA normal tidak dapat digunakan secara efisien.

Kernel PCA



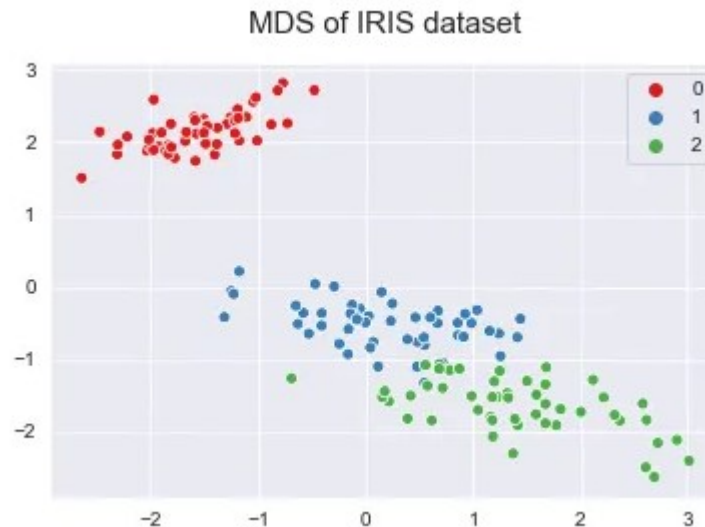
t-distributed Stochastic Neighbor Embedding (t-SNE)

- t-SNE merupakan teknik dimensionality reduction non-linear yang banyak digunakan untuk visualisasi data.
- T-SNE juga banyak digunakan dalam pemrosesan gambar dan NLP.
- Dalam Dokumentasi Scikit-learn merekomendasikan untuk menggunakan PCA atau truncated SVD sebelum t-SNE jika jumlah fitur dalam dataset lebih dari 50.
- Pada scikit-learn menggunakan TSNE().



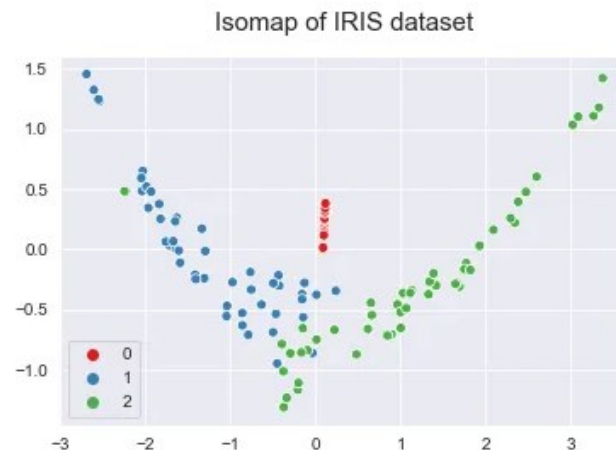
Multidimensional Scaling (MDS)

- MDS merupakan teknik dimensionality reduction non-linear yang pemetaan isometrik.
- MDS mencoba mempertahankan jarak antar instance sambil mengurangi dimensi data non-linier.
- Ada dua jenis algoritma MDS: Metrik dan Non-metrik.
- Kelas MDS() di Scikit-learn mengimplementasikan keduanya dengan menyetel hyperparameter metrik ke True (untuk tipe Metrik) atau False (untuk tipe Non-metrik).
- Pada scikit-learn menggunakan MDS().



Isometric mapping (Isomap)

- Isomap merupakan teknik dimensionality reduction non-linear yang pemetaan isometrik.
- Isomap merupakan perpanjangan dari MDS atau Kernel PCA.
- Isomap menghubungkan setiap instance dengan menghitung jarak melengkung atau geodesik ke tetangga terdekatnya dan mengurangi dimensi.
- Pada scikit-learn menggunakan Isomap().



Aplikasi Dimensionality Reduction

- Neuroscience
- Image retrieval
- Text Categorization
- Biology – gene expressions
- Image processing – face recognition
- Social Network
- Recommender system
- Dan sebagainya

TERIMA KASIH