Studi Komparasi *Local Outlier Factor* (LOF) dan *Isolation Forest* (IF) pada Analisis Anomali Kinerja Dosen

Mutmainah[1], Wiyli Yustanti[2]

[1,2] Jurusan Teknik Informatika/Sistem Informasi, Universitas Negeri Surabaya
[1]mutma.20026@mhs.unesa.ac.id
[2]wiyliyustanti@unesa.ac.id

***Abstrak*—Pada setiap semester dalam universitas terdapat kuisioner berupa penilaian terhadap kinerja dosen. Evaluasi kinerja dosen yang terdapat di Universitas Negeri Surabaya merupakan proses penting untuk memastikan bahwa dosen telah memenuhi tugas dan tanggung jawabnya dalam menyampaikan pendidikan berkualitas terhadap mahasiswanya. Pada penelitian ini terdapat 22 instrumen pertanyaan menggunakan Skala Likert yang diisi oleh mahasiswa untuk menilai kinerja dosen. Terdapat 1055 dosen yang diolah untuk mendeteksi bagaimana kinerja dosen apakah sesuai dengan Rancangan Pembelajaran Semeste (RPS) atau terdapat dosen yang ketika mengajar tidak sesuai RPS. Oleh karena itu, metode deteksi anomali diterapkan untuk mengetahui kinerja dosen yang menyimpang atau tidak seperti biasanya. Dengan metode tersebut, maka dapat digunakan algoritma *Local Outlier Factor* (LOF) dan *Isolation Forest* karena lebih efisien dalam menangani data yang besar dan bekerja dengan cepat dalam ruang fitur. Data yang digunakan belum terdapat label untuk menghitung sehingga digunakan metode klastering kmeans untuk memperoleh label dari LOF dan IF. Kemudian pada *cluster* kmeans didapatkan 3 *cluster*, yaitu *cluster* 0 terdiri dari 279 data *points*, *cluster* 1 terdiri dari 597 data *points*, dan *cluster* 2 terdiri dari 179 data *points*. Dari hasil *cluster* tersebut akan digunakan untuk memperoleh nilai dari label LOF dan label IF dalam perhitungan evaluasi hasil komparasi. Pada anomali yang diterapakan dengan algoritma LOF yaitu terdapat 19 dosen terdeteksi anomali dan pada algoritma IF terdapat 22 dosen terdeteksi anomali. Pada evaluasi yang digunakan untuk memperoleh hasil komparasi yaitu menggunakan *rand index score* dan *silhouette score*. Didapatkan nilai dari *rand index* dari LOF sebesar 0.438 dan IF sebesar 0.441. Kemudian hasil dari *silhouette score* LOF sebesar 0.0019 dan IF sebesar 0.0377.**

Kata Kunci— **Kinerja dosen, LOF, IF, *rand index, silhouette score*.**

I. PENDAHULUAN

Pada setiap universitas terdapat aturan yang dibuat untuk mengatur segala aktivitas maupun perilaku yang telah ditetapkan dalam mencapai kondisi yang tertib dan kondusif. Menurut Kamus Besar Bahasa Indonesia (KBBI) aturan adalah tindakan atau perbuatan yang harus dijalankan. Hal ini juga diterapkan di universitas negeri maupun universitas swasta terutama peraturan yang diperuntukkan dosen selama mengajar mahasiswa baik di kelas maupun diluar kelas. Menurut Undang-undang No.4 Tahun 2005 tentang guru dan dosen yang menerangkan bahwa dosen merupakan pendidik profesional dan ilmuwan dengan tugas utama mengajar, mentransformasikan, teknologi dan seni melalui pendidikan, penelitian, dan pengabdian terhadap masyarakat[6]. Kinerja adalah kesediaan seseorang atau kelompok orang untuk melakukan sesuatu kegiatan dan menyempurnakannya sesuai dengan tanggung jawabnya dengan hasil seperti yang diharapkan. kinerja dosen merujuk sejauh mana dosen dapat bertanggung jawab dalam memenuhi tugasnya dari kemampuan, efektifitas, dan kontribusi dalam memberikan pendidikan yang berkualitas. Evaluasi kinerja dosen yang terdapat di universitas penting untuk dilakukan untuk mengetahui bahwa dosen telah memenuhi tugas dan tanggung jawabnya selama mengajar selama satu semester terhadap mahasiswanya. Selain itu, evaluasi kinerja dosen juga mendukung kualitas pendidikan tinggi yang dapat mengembangkan ilmu pengetahuan kepada peneliti dalam memberikan kontribusi. Dengan mengetahui hasil evaluasi kinerja dosen yang mengajar selama satu semester maka akan menghasilkan survei kepuasan terhadap kinerja dosen oleh mahasiswa.Terdapat jumlah data yang besar dari hasil survei kinerja dosen, maka digunakan data mining untuk mengolah data tersebut. Algoritma yang terdapat di dalam data mining yaitu klasifikasi (*classification*), asosiasi (*association*), klastering (*clustering*), dan deteksi anomali (*anomaly detection*). Pada saat ini diketahui bahwa belum ada standar yang jelas untuk mengetahui baik buruknya dari kinerja dosen di universitas dalam setiap semester. Pada Rencana Pembelajaran Semester (RPS) pasti terdapat rencana pembelajaran yang telah disusun rapi oleh setiap dosen supaya memberikan pembelajaran yang baik untuk mahasiswanya. Oleh sebab itu peneliti mengangkat permasalahan kinerja dosen selama mengajar satu semester untuk mencari kinerja dosen yang baik maupun kinerja dosen yang buruk. Maka sebagai alternatif pada penelitian ini digunakan metode *anomaly detection.*

Menurut KBBI anomali adalah penyimpangan atau kelainan. Jadi entitas dari peristiwa yang menunjukkan masalah yang terdapat penyimpangan tindakan dari kinerja dosen yang ada di Universitas Negeri Surabaya (UNESA) yaitu tujuan dari metode deteksi anomali. Namun sebelum data tersebut diolah kedalam deteksi anomali dibutuhkan algoritma klastering pada pengelompokan data kedalam jumlah *cluster*. Data kinerja dosen yang digunakan pada penelitian ini yaitu data pada tahun 2022 dengan jumlah dosen 2.397 sebelum melalui proses pengecekan data duplikat. Terdapat beberapa metode penelitian yang digunakan untuk mendeteksi anomali yaitu *Isolation Forest, Local Outlier Factor, Robust Covariance, One-class Support Vector Machine (SVM), dan One Class SVM with Stochastic Gradient Descent (SGD)*. Namun penelitian ini menggunakan algoritma Local Outlier Factor (LOF) dan

Isolation Forest (IF) karena skalabilitas algoritma LOF dan IF ini lebih efisien dalam menangani data yang besar dan bekerja dengan cepat dalam ruang fitur tinggi dibandingkan dengan algoritma yang lain.

II. LANDASAN TEORI

Penelitian ini menggunakan beberapa landasan teori yang akan dijelaskan, sebagai berikut ;

A. Studi Literatur

Studi literatur dilakukan dengan mempelajari penelitian-penelitian terdahulu yaitu dengan mempelajari metode klastering, klasifikasi, dan deteksi anomali yang didapatkan dari jurnal-jurnal terdahulu. beberapa jurnal yang menjadi referensi tersebut diambil penerbitan 5 tahun terakhir. Jurnal yang didapatkan bermanfaat sebagai rujukan dalam menyelesaikan masalah pada penelitian yang diusulkan.

B. Data Mining

Data mining merupakan proses menemukan informasi berguna secara otomatis dalam repositori data yang besar untuk menentukan *pattern* (pola) dan *rule* (aturan) dari ekstraksi sejumlah data yang mendapatkan informasi yang penting didalamnya. Pernyataan ini dibuktian oleh peneliti Pang Ning [5] Ada bebrapa tahapan yang biasa digunakan dalam proses data mining berupa:

1. *Data selection*, proses penyeleksian data tertentu dari suatu kumpulan data
2. *Data transformation*, memanipulasi atau merubah data dari struktur tertentu kebentuk yang sesuai dengan sistem untuk dianalisis.
3. *Data mining*, pola tersembunyi dari *dataset* besar yang kompleks
4. *Pattern Evaluation*, mengevaluasi dan menginterpretasikan pola
5. *Knowledge presentation*, kumpulan data untuk pengembangan, pengujian, atau validasi model untuk mendapatkan wawasan baru

C. Kmeans

Kmeans adalah salah satu algoritma yang umum digunakan dalam menganalisis data dalam pembelajaran mesin klastering data. Tujuannya untuk mengelompokkan data menjadi klaster-klaster berdasarkan kemiripannya. Pusat awal yang dihasilkan secara acak untuk menunjukkan tahapan lebih rinci. *Background* ruang partisi hanya untuk ilustrsi dan tidak dihasilkan oleh algoritma Kmeans yang dibuktikan oleh peneliti Wakhida[]. . K-Means adalah algoritma yang populer dan sering digunakan dalam berbagai aplikasi seperti pengelompokan data, segmentasi gambar, dan kompresi data.

$$Distance(p,q) = \left(\sum_{k}^{n} \mu k\,|Pk - qk|r\right) 1/r$$

Keterangan:
N = Jumlah record data
K = Urutan field data
r = 2
μk = Bobot field yang diberikan user

D. Deteksi Anomali

Deteksi anomali merupakan proses memeriksa titik data tertentu dan mendeteksi kejadian langka yang tampak mencurigakan karena berbeda dari pola perilaku yang ditetapkan oleh Toni[7]. Tujuannya untuk menemukan entitas atau kejadian yang berbeda dari mayoritas data keamanan jaringan, pemantauan sistem, deteksi penipuan keuangan, atau bidang lainnya.

E. Local Outlier Factor (LOF)

Algoritma *Local Outlier Factor* adalah metode deteksi anomali yang berbasis pada konsep kerapatan yang mengukur seberapa anomali dengan membandingkan kerapatan sampel disekitarnya. LOF berguna untuk mendeteksi kecurangan, deteksi anomali dari jaringan, maupun kelompok dari kerapatan juga mempertimbangkan struktur kerapatan local dari *dataset* secara keseluruhan.

F. Isolation Forest (IF)

Algoritma *Isolation Forest* merupakan algoritma yang efektif untuk mendeteksi anomali dalam dataset berdimensi tinggi. Berdasarkan pemisahan (isolasi) nilai anomali dari nilai normal dalam *dataset*. Hanya memerlukan beberapa pemisahan untuk mencapai isolasi total dari anomali, waktu komputasi dapat lebih cepat dan keberhasilan tergantung pada sampel anomali. Dapat dikatakan anomali IF apabila sampel IF adalah -1 dan dapat dikatakan non anomali apabila nilai IF adalah 1.

G. Rand Index Score

Metrik yang mengukur seberapa mirip dua himpunan klastering yang diberikan sesuai dengan himpunan klastering yang sebenarnya atau yang dianggap sebagai referensi. *Rand score* berkisar antara 0 hingga 1, dimana nilai 1 menunjukkan kesamaan sempurna antara dua klaster. dan nilai 0 menunjukkan ketidaksamaan total. Ini digunakan karena memberi cara yang sederhana dan informatif untuk mengevaluasi kecocokan partisi yang dihasilkan. Rumus untuk menghitung rand index score yaitu:

$$R = (a+b) / ({}_{n}C_{2})$$

Keterangan :

a : Berapa kali sepasang elemen menjadi anggota *cluster* yang sama

b : Berapa kali sepasang elemen menjadi anggota *cluster* yang berbeda

${}_{n}C_{2}$: Banyaknya pasangan tak berurutan dalam himpunan yang terdiri dari *n* elemen.

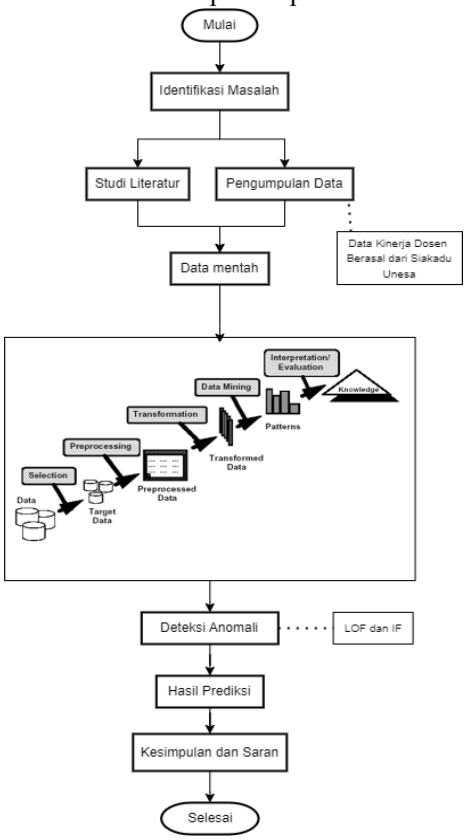
H. Silhouette Score

Meruupakan metrik untuk mengukur kualitas dari klastering pada analisis data. Nilai berkisar antara -1 sampai dengan 1, dimana skor yang mendekati 1 menunjukkan bahwa klaster tersebut terpisah baik dan padat. Jika klaster 0 menunjukkan bahwa klaster terdapat tumpang tindih antara klaster. Untuk score negatif memungkinkan bahwa klaster data salah dalam pengelompokan.

III. METODE PENELITIAN

Pada penelitian ini peneliti menggunakan metode *Knowledge Discovery in Databases* (KDD) dalam proses yang

mencakup identifikasi pola berharga dan pengetahuan yang bermanfaat dari suatu data. Tahapan dari metode KDD ini termasuk pemilihan data, *preprocessing, transformation,* inteepretasi dan *data mining.* Dengan proses KDD ini digunakan untuk pengambilan keputusan atau pemahaman lebih lanjut tentang suatu fenomena. Diketahui metode ini digunakan dalam industri data selama lebih dari satu dekada, tepatnya sejak tahun 1989 oleh Piatetsky Shapiro. proses KDD berfokus pada pemetaan *low-level* data dan mengubahnya menjadi bentuk yang lebih padat, jelas, dan bermakna. Metode Berikut adalah tahapan ari proses KDD.



Gbr. 1 Proses KDD

A. *Dataset*

Dataset memiliki nama lain record, point ,vector ,pattern, event, observasi, case, atau data. Yang mengumpulkan dataset adalah orang yang profesional dibidang data seperti data analis. Terdapat beberapa tahapan pengolahan data yang bisa digunakan untuk dataset seperti data cleaning dan kategorisasi. Sehingga kategori dari masing-masing dataset terkumpul dan dapat digunakan dengan profesional yang didalamnya terdapat variabel yang saling berhubungan.

B. *Selection*

Pada tahap selection ini berasal dari data yang dilakukan oleh peneliti yang bersifat sekunder sebab data tersebut diperoleh melalui observasi melalui website Single Sign On (SSO). Populasi yang digunakan dalam penelitian ini adalah dosen aktif UNESA tahun 2022. Terdapat 22 variabel dan setiap variabel mewakili nilai dari 1 sampai 4. Dengan tercapainya semua kinerja dosen yang sesuai dengan kriteria akan menciptakan kinerja dosen yang sangat baik dan bermanfaat bagi Universitas dan mahasiswa. Kemudian untuk mengisi jawaban dari responden/mahasiswa dari data survei yang digunakan dalam penelitian ini terdapat instrumen yang digunakan dalam skala likert.

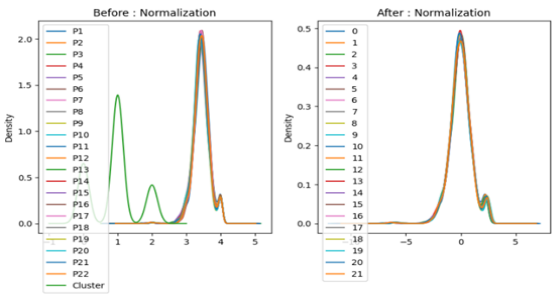
C. *Preprocessing*

Pada proses preprocessing yaitu untuk meningkatkan akurasi pada saat melakukan klasterisasi data untuk menghilangkan data yang tidak dibutuhkan atau noise sehingga algoritma lebih mudah dalam mengenali objek yang dituju. Proses preprocessing mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsistensi, dan memperbaiki kesalahan pada data. Karena pada dataset terdapat data duplikat maka dilakukan penghapusan data duplikat yang awalnya jumlah data sebanyak 2.397 menjadi 1055 dataset. Setelah itu yaitu terdapat proses pemerikssaan data hilang.

```
<class 'pandas.core.frame.DataFrame'>
Index: 1055 entries, 0 to 2396
Data columns (total 24 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   nidn    1055 non-null   int64
 1   nama    1055 non-null   object
 2   P1      1055 non-null   float64
 3   P2      1055 non-null   float64
 4   P3      1055 non-null   float64
 5   P4      1055 non-null   float64
 6   P5      1055 non-null   float64
 7   P6      1055 non-null   float64
 8   P7      1055 non-null   float64
 9   P8      1055 non-null   float64
 10  P9      1055 non-null   float64
 11  P10     1055 non-null   float64
 12  P11     1055 non-null   float64
 13  P12     1055 non-null   float64
 14  P13     1055 non-null   float64
 15  P14     1055 non-null   float64
 16  P15     1055 non-null   float64
 17  P16     1055 non-null   float64
 18  P17     1055 non-null   float64
 19  P18     1055 non-null   float64
 20  P19     1055 non-null   float64
 21  P20     1055 non-null   float64
 22  P21     1055 non-null   float64
 23  P22     1055 non-null   float64
dtypes: float64(22), int64(1), object(1)
memory usage: 206.1+ KB
```

Gbr. 2 Pemeriksaan Data Hilang

Membuat berbagai jenis visualisasi data sehingga memudahkan dalam interpretasi hasil normalisasi yang umum digunakan dalam python dalam membuat fungsi plotting yaitu dengan import matplotlib.pyplot. Manfaat yang diperoleh dari hasil proses normalisasi ini adalah basis data menjadi mudah diakses, data mudah dikelola, dan meminimalkan tempat penyimpanannya[4]. Perbandingan dilakukan sebelum dan sesudah proses dari normalisasi dalam memahami bagaimana distribusi data berubah setelah proses normalisasi.



Gbr. 3 Visualisasi Normalisasi Data

Metode yang digunakan dalam penelitian ini menggunakan klasterisasi kmeans dan anomali LOF dan IF dimana algoritma klasterisasi tergolong dalam algoritma *unsupervised learning.* Proses preprocessing ini dilakukan karena dataset yang digunakan dalam penelitian ini tidak terdapat atribut *class* atau tidak memiliki variabel.

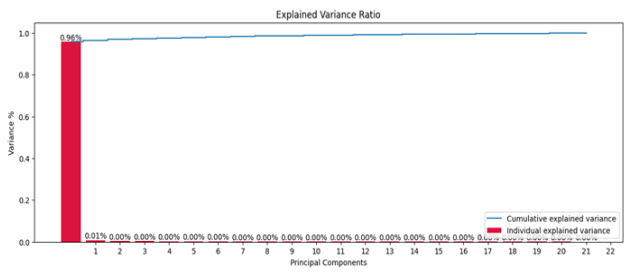
D. Transformation

Pada tahap *transformation* yaitu proses transformasi pada data yang telah dipilih sehingga data tersebut siap untuk *proses data mining* karena sebelum diaplikasikan metode *data mining* membutuhkan format data yang khusus. PCA dapat digunakan untuk mereduksi dimensi suatu data tanpa mengurangi karakteristik data tersebut secara signifikan[3]. Dalam meningkatkan keakuratan transformasi data ini juga terdapat normalisasi yang dilakukan untuk mengubah nilai-nilai dalam dataset akan memiliki skala atau rentang yang seragam. Untuk memastikan konsistensi data dan mencegah bias yang mungkin timbul dari penanganan juga dengan melibatkan pengisian nilai yang hilang atau menghapus baris maupun kolom yang hilang.

```
import numpy as np
from sklearn.decomposition import PCA
#menentukan komponen utama
pca = PCA(n_components= 22).fit(X_scaled)
principal_components = range(1, pca.n_components_+1)
#membuat subplot
fig, ax =  plt.subplots(figsize = (15, 5))
#menyiapkan data untuk diplotting
x_pca = principal_components
y_pca = pca.explained_variance_ratio_
cum_sum_eigenvalues = np.cumsum(y_pca)
```

Gbr. 4 Source code Explained Variance

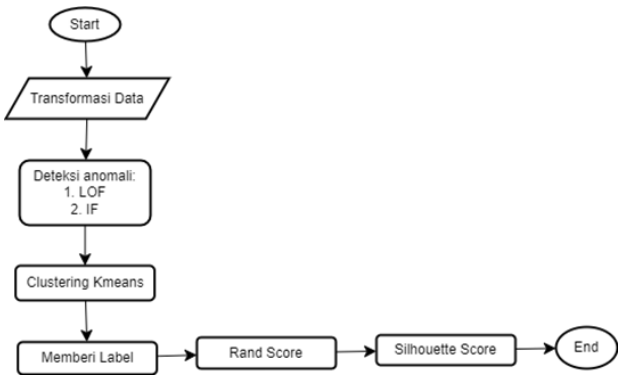
Dataset pada penelitian ini terdapat 22 komponen utama dimana setiap komponen utama telah ditunjukkan oleh Gambar (diatas). Dalam dataset menjelaskan 96% varians dari komponen pertama yang ditunjukkan. Untuk mengukur komponen terbesar utama biasanya terdapat pada varians komponen utama yang pertama dan dianggap sebagai komponen yang paling penting. Oleh sebab itu, hanya 2 komponen utama yang digunakan dalam penelitian ini karena komponen tersebut telah menunjukkan >90% varians dalam dataset.



Gbr. 5 Visualisasi Individual Expalined Variances

D. Data Mining

Data mining yang akan diterapkan oleh peneliti pada penelitian ini dengan *dataset* sesuai dengan proses KDD. Karena jumlah dimensi di dalam data meningkat, pendekatan baru untuk penemuan pola sangat diperlukan. tersebut dapat ditarik kesimpulan bahwa *Knowledge Discovery in Database* (KDD) adalah proses yang bertujuan untuk menggali dan menganalisis sejumlah besar himpunan data dan mengekstrak informasi[1]. Tahapan dari penelitian ini dalam mencari data yang termasuk dari anomali, peneliti mengumpulkan *dataset* untuk menerapkan algoritma *clustering* ke dalam sejumlah klaster. Kemudian setelah melakukan *clustering* maka data tersebut akan dibandingkan dengan hasil nilai dari label LOF dan label IF.



Gbr. 6 Diagram data mining

Pertama, deteksi anomali diimplementasikan terlebih dahulu sehingga akan menghasilkan anomali baik dan anomali buruk dengan nilai -1. Deteksi anomali yang digunakan yaitu algoritma LOF dan algoritma IF, dari setiap algoritma memiliki model dengan parameter yang telah diatur sehingga terbentuk sedemikian rupa.

D. *Interpretation/Evaluation*

Interpretation/evaluation adalah tahap untuk melakukan evaluasi hasil setelah melakukan semua proses *data mining*. Untuk menerjemahkan dari pola-pola yang dihasilkan oleh metode *clustering*, dimana hasil klaster yang terbentuk dideskripsikan sesuai dengan karakteristiknya kedalam jumlah data yang masuk pada klaster yang sama. Untuk memastikan bahwa hasil yang diperoleh dari model dan analisis data dapat diandalkan, berguna, dan mudah dipahami oleh pengambil keputusan. Visualisasi data dilakukan terhadap model setiap algoritma yaitu klastering dan anomali yang akan memudahkan bagi pembaca

1) Klasterisasi Kmeans

Pada visualisasi hasil klasterisasi Kmeans dengan menampilkan data dalam dua dimensi berdasarkan dua komponen utama yang berasal dari PCA. Yaitu membuat *figure* baru terlebih dahulu yang kemudian akan disimpan pada ‘visualizationz Kmeans’. Setiap klaster diberi warna yang berbeda untuk membedakan warna klaster satu sama lain. Titik-titik yang dihasilkan oleh klaster menunjukkan data dalam dua dimensi berdasarkan dua komponen utama yang dihasilkan oleh PCA. atribut cluster_centers_ digunakan untuk menambahkan centroid pada tiap *cluster*. Setiap klaster diberi

warna yang berbeda untuk membedakan warna klaster satu sama lain.

```
#membuat figure dan scatter plot
plt.figure(figsize=(10,5))
#scatter plot cluster 1, 2 & 3
plt.scatter(X_reduced[kmeans_labels==0,0],
X_reduced[kmeans_labels==0,1], s=60, c='pink',
        edgecolor='crimson', label = 'Cluster 1')
plt.scatter(X_reduced[kmeans_labels==1,0],
X_reduced[kmeans_labels==1,1], s=60, c='plum',
        edgecolor='indigo', label = 'Cluster 2')
plt.scatter(X_reduced[kmeans_labels==2,0],
X_reduced[kmeans_labels==2,1], s=60, c='lightskyblue',
        edgecolor='navy', label = 'Cluster 3')

#mengatur judul dan label sumbu
plt.title('Visualization of K-Means')
plt.xlabel('1st principal component')
plt.ylabel('2nd principal component')
plt.legend(loc="best")
plt.show()
```

Gbr. 7 Source code Visualisasi Kmeans

2) *Anomaly Detection*

Pada visualisai hasil dari *anomaly detection* DataFrame 'df_nilai' yang memvisualisasikan nilai *mean* (rata-rata) dari penilaian kinerja dosen. plot.scatter() yang digunakan untuk membuat *scatter plot*. Untuk lamda function menerjemahkan kode ini menjadi warna yang sesuai. Misalnya, 'r' diterjemahkan menjadi 'red' dan 'b' diterjemahkan menjadi 'blue.

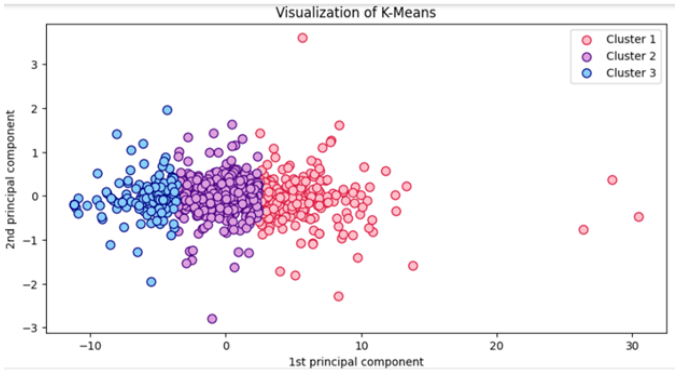
```
#Membuat kolom baru dengan nama colors untuk menyimpan warna setiap
kondisi diatas
df_nilai['colors'] =
np.where(greater_than_95,'r',np.where(smaller_than_5,'r','b'))

#membuat Scatter Plot Menggunakan Pandas DataFrame
import matplotlib.pyplot as plt
df_nilai.plot.scatter(x='nama_dosen',y='mean',c = df_nilai['colors'].
            apply(lambda x: dict(r='red', b='blue')[x]),
        figsize=(10, 5),title='Mean dari Penilaian Kinerja Dosen UNESA',
            xlabel='Nama Dosen',ylabel='Mean')
#Rotasi Label pada Sumbu X
plt.xticks(rotation=65);
```

Gbr. 8 Source code Visualisasi Anomali

IV. ANALISIS HASIL PENELITIAN

A. Hasil Analisis Klastering K-Means

Pada analisis klastering Kmeans, implementasi algoritma Kmeans terhadap *transformed* data telah berhasil dilakukan. Ada banyak pendekatan untuk membuat *cluster*, diantaranya adalah membuat aturan yang mendikte keanggotaan dalam kelompok yang sama berdasarkan tingkat persamaan di antara anggota-anggotanya[10]. Berdasarkan hasil klasterisasi algoritma Kmeans terlihat bahwa klaster terpisah menjadi 3 klaster.



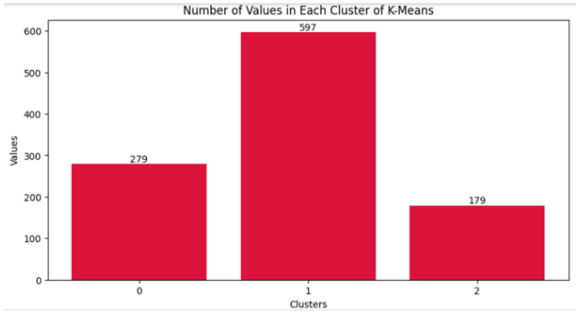
Gbr. 9 Visualisasi algoritma K-Means

Berikut adalah keterangan dari setiap titik yang ada pada *scatter plot*, sebagai berikut:

a. *Cluster* 1, data *points* yang ditandai dengan marker '*point*' yang berwarna merah muda (pink)
b. *Cluster* 2, data *points* yang ditandai dengan marker '*point*' yang berwarna ungu (plum)
c. *Cluster* 3, data *points* yang ditandai dengan marker '*point*' yang berwarna biru (lightskyblue)

Berdasarkan hasil visualisasi dari perhitungan jumlah *data points* yang telah di *mapping* sehingga menghasilkan *cluster* yang awal mulanya 1 menjadi 0, *cluster* 2 menjadi *cluster* 1, dan *cluster* 3 menjadi *cluster* 2. Dari klasterisasi dengan algoritma Kmeans yang berjumlah 1055 dataset dan terpecah menjadi 3 *cluster*

a. *Cluster* 0, terdiri dari 279 *data points*
b. *Cluster* 1, terdiri dari 597 *data points*
c. *Cluster* 2, terdiri dari 179 *data points*



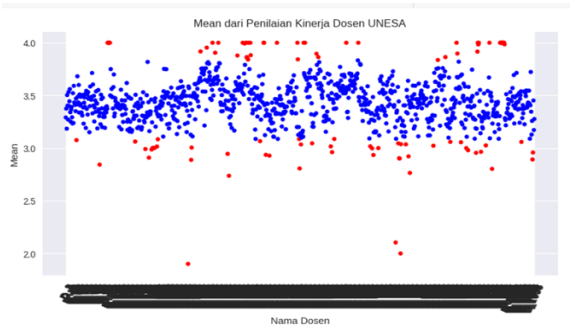
Gbr. 10 Visualisasi data Point Kmeans

Pada data klaster yang telah ditemukan, kemudian hasil klaster Kmeans di unduh dan mencari nilai statistik deskriptif untuk mengetahui karakteristik utama dari ketiga klaster. Analisis statistik deskriptif dilakukan dengan menghitung nilai rata-rata dari setiap variabel penelitian yang berjumlah 22 tersebut. Kemudian nilai dari *mean* setiap variabel di hitung rata-rata ulang sehingga menghasilkan nilai *grand mean* untuk setiap klaster yang terbentuk. *Grand mean* yang didapatkan yaitu *cluster* 0 sebesar **3.169,** *cluster* 1 sebesar **3.458,** *cluster* 2 sebesar **3.778**.

a. *Cluster* 0 diberi label nama "Buruk" sebab klaster tersebut mempunyai nilai *grand mean* yang sangat rendah sehingga kinerja dosen UNESA yang dihasilkan **Kurang Baik**
b. *Cluster* 1 diberi label nama "Normal" sebab klaster tersebut mempunyai nilai *grand mean* yang normal sehingga kinerja dosen UNESA yang dihasilkan **Normal** sebagaimana semestinya
c. *Cluster* 2 diberi label nama "Baik" sebab klaster tersebut mempunyai nilai *grand mean* yang baik, sehingga kinerja dosen UNESA yang dihasilkan **Baik** atau lebih dari segabaimana semestinya.

B. Hasil Analisi Anomali

Pada analisis anomali yang terdapat pada penelitian ini terdapat 2 algoritma yang dilakukan, yaitu anomali *Local Outlier Factor* (LOF) dan anomali *Isolation Forest* (IF). Data yang bersifat anomali dapat dideteksi keberadaannya dengan melihat perbedaan karakteristiknya dengan kondisi data yang dianggap normal[9]. Berdasarkan hasil dari anomali yang didapatkan dari *dataset* sebelum dilanjut pada anomali LOF dan anomali IF terdapat anomali dosen berjumlah 106 nama dosen yang terdeteksi anomali. Sedangkan yang terdeteksi normal terdapat 949 dosen.



Gbr. 11 Data Anomali Dosen Unesa 2022

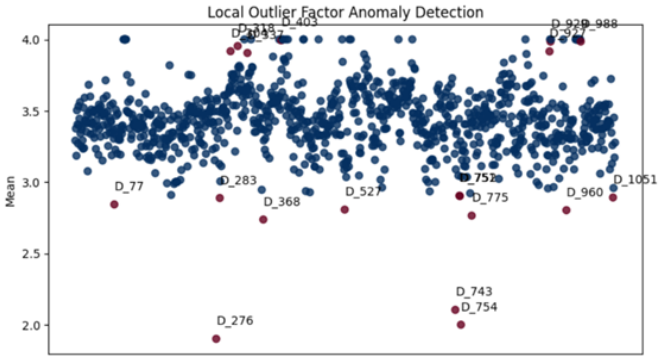
1. Anomali *Local Outlier Factor* (LOF)

Hasil prediksi yang akan muncul adalah nilai 1 dan -1. Untuk model model_LOF_scores, atribut ini menyimpan menyimpan nilai *negative outlier factor* untuk setiap data *point* dalam *dataset.* Dengan cara ini, kita dapat mendeteksi dan menganalisis outlier dalam dataset, memberikan wawasan tambahan untuk analisis data.

```
#menonaktifkan warning
import warnings
warnings.filterwarnings('ignore')
#membuat figur dan sub plot
fig, ax = plt.subplots(1, figsize=(9, 5), sharex=True, sharey=False)
#Scatter Plot untuk Deteksi Anomali
ax.scatter(df_nilai['nama_dosen'], df_nilai['mean'],
        c=df_nilai['LOF_anomaly'],cmap='RdBu',alpha = 0.8)
#Mengatur Judul dan Label Sumbu
ax.set_title("Local Outlier Factor Anomaly Detection")
ax.set_xlabel("Nama Dosen")
ax.set_ylabel("Mean")
```

Gbr. 12 Source code Prediksi Visualisasi Anomali LOF

Menambahkan anotasi *pada scatter plot* yang telah dibuat dengan matplotlib. Anotasi ini bertujuan untuk menandai data yang diidentifikasi sebagai *outlier.* Anotasi ini membantu untuk dengan jelas memvisualisasikan mana data yang dianggap sebagai *outlier.*



Gbr. 13 Visualisasi Anomali LOF

Dari yang ditunjukkan tersebut, berikut adalah keterangan mengenai titik-titik pada scatter plot. Sebagai berikut:

a. Titik 'biru' pada data *points* dianggap sebagai data normal yang bernilai 1
b. Titik 'merah' pada data *points* dianggap sebagai data anomali yang bernilai -1

Hasil yang didapatkan dari anomali LOF yaitu terdapat 19 nama dosen yang ditetapkan sebagai anomali baik anomali baik maupun anomali buruk.

TABEL 1
DATA DOSEN ANOMALI LOF

NO	nama_dosen	Mean	LOF_anomaly
1	D_77	2,843182	-1
2	D_276	1,900909	-1
3	D_283	2,888182	-1
4	D_304	3,917273	-1
5	D_318	3,953636	-1
6	D_337	3,905	-1
7	D_368	2,737273	-1
8	D_403	3,993636	-1
9	D_527	2,806818	-1
10	D_743	2,104091	-1
11	D_751	2,904091	-1
12	D_752	2,901818	-1
13	D_754	2	-1
14	D_775	2,764545	-1
15	D_927	3,915909	-1
16	D_929	3,984545	-1
17	D_960	2,802727	-1
18	D_988	3,985	-1
19	D_1051	2,891818	-1

Dari data tersebut dapat diketahui jumlah dosen yang terdeteksi anomali LOF dengan nilai -1. Dari nilai tersebut

dapat diketahui dosen yang terdeteksi anomali baik dan anomali buruk. Berikut rinciannya dari datapoints tersebut:

a. Dosen dengan nilai mean 'rendah' yaitu dibawah nilai 3,9 maka dianggap sebagai anomali **buruk**

b. Dosen dengan nilai 'tinggi' yaitu nilai diatas 2,9 maka dianggap sebagai anomali **baik**

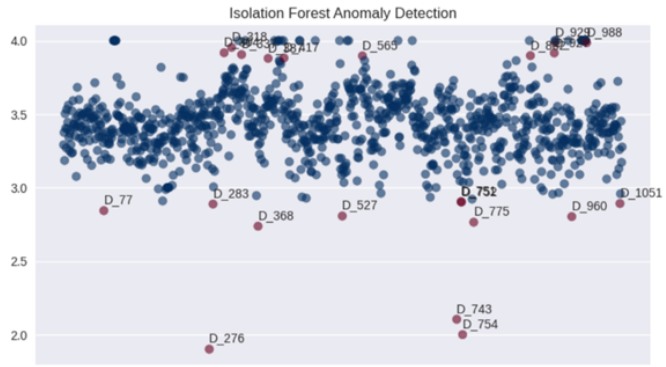
2. Anomali *Isolation Forest* (IF)

Model ini akan belajar pola dari data untuk mendeteksi anomali. Menghitung skor anomali untuk setiap data menggunakan fungsi decision_function dari model yang telah dilatih. Prediksi ini menghasilkan nilai 1 untuk data yang dianggap normal dan -1 untuk data yang dianggap sebagai anomali IF. Untuk memvisualisasikan data dengan *highlight* pada anomali yang terdeteksi oleh model *Isolation Forest* kita dapatkan dengan jelas melalui warna 'RdBu' ang memberikan gradasi warna dari merah ke biru. Hasil *scatter plot* dari setiap anomali ditandai dengan teks anotasi yang menunjukkan nama dosen. Ini akan mempermudah interpretasi visual dari hasil deteksi anomali.

```
#menonaktifkan warning
import warnings
warnings.filterwarnings('ignore')

#Membuat figur dan Subplot
fig, ax4 = plt.subplots(1, figsize=(9, 5), sharex=True, sharey=False)
#Membuat Scatter Plot
ax4.scatter(df_nilai['nama_dosen'],
df_nilai['mean'],c=df_nilai['IF_anomaly'],cmap='RdBu',alpha=0.6 )
#Mengatur Judul Plot
ax4.set_title("Isolation Forest Anomaly Detection")
```

Gbr. 14 Source code Prediksi Visualisasi Anomali IF

Berdasarkan hasil dari deteksi anomali IF yang telah didapatkan terlihat bahwa terdapat titik-titik anomali yang berjumlah 22 nama dosen.



Gbr. 15 Visualisasi Anomali IF

Dari yang ditunjukkan tersebut, berikut adalah keterangan mengenai titik-titik pada scatter plot. Sebagai berikut.

a. Titik 'biru' pada data *points* dianggap sebagai data normal yang bernilai 1

b. Titik 'merah' pada data *points* dianggap sebagai data anomali yang bernilai -1

Untuk mengetahui jumlah data normal dan data anomali IF, yaitu dengan mendownload hasil tersebut kedalam bentuk file excel. Berikut adalah hasil yang didapatkan dari anomali IF yaitu terdapat 22 nama dosen yang ditetapkan sebagai anomali baik anomali baik maupun anomali buruk.

TABEL 2
DATA DOSEN ANOMALI IF

No	nama_dosen	mean	IF_anomaly
1	D_77	2,843182	-1
2	D_276	1,900909	-1
3	D_283	2,888182	-1
4	D_304	3,917273	-1
5	D_318	3,953636	-1
6	D_337	3,905	-1
7	D_368	2,737273	-1
8	D_403	3,993636	-1
9	D_417	3,881818	-1
10	D_527	2,806818	-1
11	D_565	3,895909	-1
12	D_743	2,104091	-1
13	D_751	2,904091	-1
14	D_752	2,901818	-1
15	D_754	2	-1
16	D_775	2,764545	-1
17	D_882	3,897727	-1
18	D_927	3,915909	-1
19	D_929	3,984545	-1
20	D_960	2,802727	-1
21	D_988	3,985	-1
22	D_1051	2,891818	-1

3. Analisis Hasil *Rand Index Score*

Pada tahap analisis hasil tersebut digunakan untuk membandingkan hasil dari setiap metode, dilakukan evaluasi dengan menggunakan Rand Index[2]. Yaitu nilai antara label Kmeans *cluster* dengan label anomali LOF dan label Kmeans *cluster* dengan label anomali IF. Sebelum melakukan pengecekan pada rand index hal pertama yang dilakukan adalah mencari nilai dari label anomali LOF dan nilai label anomali IF apakah terdapat jumlah yang sama *dataset* yang sama dengan label Kmeans *cluster*. Setelah didapatkan nilai dari label Kmeans, label anomali LOF dan label anomali IF maka langkah selanjutnya untuk menghitung nilai dari *rand index*.Kemudian menghitung *rand index* antara label LOF dan label kmeans, yaitu pertama import fungsi rand_score() dari *library* scikit-learn. Hasil dari *rand index* yang didapatkan yaitu bernilai 1(100%) dan dapat diartikan bahwa pada label LOF dan label Kmeans menunjukkan kedua pengelompokan tersebut identik.

```
from sklearn.metrics.cluster import rand_score
 lof_labels= df_kmeansLOF['LOF Label']
 kmeans_labels = df_kmeansLOF['Kmeans Label']
 #Rand score LOF dan Kmeans rand_index_LOF =
rand_score(kmeans_labels, lof_labels)
 print(f"Rand Index LOF: {rand_index_LOF}")
```

Gbr. 16 Source code Menghitung Rand index LOF

Kemudian menghitung *rand index* antara label IF dan label kmeans, Selanjutnya yaitu menghitung nilai *rand index* dari label anomali IF dan label Kmeans yaitu baca *dataset* 'data anomaly IF'. Hasil dari *rand index* yang didapatkan yaitu bernilai 1(100%) dan dapat diartikan bahwa pada label IF dan label Kmeans menunjukkan kedua pengelompokan tersebut identik.

```
from sklearn.metrics.cluster import rand_score
 if_labels= df_kmeansIF['IF Label']
 kmeans_labels = df_kmeansIF['Kmeans Label']
 #Rand score IF dan Kmeans
 rand_index_IF = rand_score(kmeans_labels, if_labels)
 print(f"Rand Index IF: {rand_index_IF}")
```

Gbr. 17 Source code Menghitung Rand Index IF

Namun, jika nilai label LOF, label IF dan label Kmeans digunakan secara keseluruhan dalam arti semua dataset dengan jumlah 1055 baik yang normal atau anomali dihitung untuk mendapatkan nilai *rand index* maka terdapat perbedaan nilai. Nilai *rand index* LOF yang diperoleh dari *dataset* tersebut yaitu sebesar **0,438** dan nilai *rand index* dari IF dari *dataset* tersebut sebesar **0,441**.

Selanjutnya yaitu menghitung nilai dari *silhouette score* antara nilai PCA, label IF dan label LOF untuk membandingkan nilai tertinggi dari *silhouette score* label LOF dan IF. Hasil nilai dari *silhouette score* LOF yang didapatkan adalah **0.0019**. Kemudian menghitung nilai dari *silhouette score* dari PCA dan label IF yaitu pertama *import* silhouette_score digunakan untuk menghitung *silhouette score* dari sebuah pengelompokan PCA dan label IF. . Hasil nilai dari *silhouette score* IF yang didapatkan adalah **0.0377**.

V. KESIMPULAN

Berdasarkan hasil implementasi antara algoritma Kmeans *cluster* dan algoritma anomali LOF dan algoritma anomali IF terhadap *dataset* kinerja dosen UNESA pada tahu 2022 dapat disimpulkan bahwa:

1. Pada klasterisasi kmeans telah didapatkan nilai dari statistik deskriptif berupa *grand mean* yaitu *cluster* 0 sebesar 3.169, *cluster* 1 sebesar 3.458, *cluster* 2 sebesar 3.778. Dari hasil analisis statistik deskriptif ini digunakan untuk mendapatkan nilai dari label anomali LOF dan label anomali IF. Dari *dataset* kinerja dosen didapatkan nama-nama dosen yang terdeteksi sebagai anomali yang menggunakan algoritma *Local Outlier Factor* (LOF) dan algoritma *Isolation Forest* (IF). Dari algoritma LOF didapatkan sebanyak 19 dosen anomali yaitu terdiri dari 12 dosen terdeteksi anomali buruk dan 7 dosen terdeteksi anomali baik. Pada algoritma IF didapatkan sebanyak 22 dosen yaitu terdiri dari 12 dosen terdeteksi anomali buruk dan 10 dosen terdeteksi anomali baik.
2. Untuk mendapatkan nilai yang lebih valid dari hasil anomali maka dilakukan analisis dari *Rand index score* dan *silhoutte score*. Dari hasil tersebut didapatkan nilai dari *rand index* LOF sebesar 0.438 dan *rand index* IF sebesar 0.441. Dapat disimpulkan bahwa nilai *rand index* IF lebih besar dari *rand index* LOF. Kemudian pada perhitungan dari *silhoutte score* antara LOF dan IF didapatkan nilai dari LOF sebesar 0.0019 dan nilai dari IF sebesar 0.0377. Dapat disimpulkan bahwa nilai *silhouette score* jika mendekati angka 1 maka nilai *silhouette score* lebih akurat. Jadi dari perolehan nilai *rand index* dan *silhouette score* didapatkan bahwa pada penelitian ini algoritma IF lebih valid daripada algoritma LOF.

VI. SARAN

Berdasarkan hasil penelitian yang diperoleh dari dataset kinerja dosen UNESA pada tahun 2022 peneliti ingin mengemukakan beberapa saran yang mungkin bermanfaat bagi penelitian selanjutnya. Adapun saran yang peneliti ajukan adalah sebagai berikut:

1. Untuk melakukan komparasi hasil penelitian kinerja dosen dengan LOF dan IF mungkin bisa menggunakan variasi seperti *Adjusted Rand Index* (ARI) yang menyesuaikan untuk kemungkinan pengelompokan elemen secara acak.
2. Hendaknya peneliti selanjutnya menggunakan algoritma lain untuk mendeteksi anomali, mengingat penelitian ini belum sepenuhnya bisa menggambarkan dengan baik hasil dari anomali tersebut.
3. Menggunakan layanan SynapseML dan Azure AI di Apache Spark untuk deteksi anomali multivariat. Deteksi anomali yang efektif seringkali melibatkan kombinasi berbagai teknik dan pemahaman mendalam tentang data yang dianalisis.

UCAPAN TERIMA KASIH

Rasa terima kasih penulis ucapkan kepada Allah SWT. yang telah memberikah karunia dan rahmatnya kepada penulis sehingga dapat menyelesaikan jurnal ilmiah ini dengan baik. Penulis berterimakasih juga kepada pihak yang telah membimbing, memotivasi, dan memberikan bantuannya yaitu oleh orang tua, dosen pembimbing,, para saudara, sahabat, dan teman seperjuangan. Semoga kebaikan dan keikhlasannya dibalas oleh Allah dengan keberkahan yang lebih-lebih.

REFERENSI

[1] Arta, I. K. J., Indrawan, G., & Rasben Dantes, G. (2019). Data Mining Rekomemdasi Calon Mahasiswa Berprestasi di STMIK Denpasar Menggunakan Metode Technique For Other Reference By Similarity to Ideal Solution.

[2] Ashari, I. F., Dwi Nugroho, E., Baraku, R., Novri Yanda, I., & Liwardana, R. (2023). Analysis of Elbow, Silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-Index Evaluation on K-Means Algorithm for Classifying Flood-Affected Areas in Jakarta.

[3] Firliana, R. (2014). Implementasi Pengenalan Wajah Manusia Menggunakan Principal Component Analysis (Pca). 2(April), 135–138.
[4] Mulyati, S., Sujatmoko, B. A., Wira, T. I. M., Afif, R., & Pratama, R. A. (2018). Normalisasi Database Dan Migrasi Database Untuk Memudahkan Manajemen Data. Sebatik, 22(2), 124–129.
[5] Pang Ning, T. (2020). Introduction to Data Mining (Issue July) (
[6] Undang Undang No.4 Tahun 2005 tentang Guru dan Dosen, 1 1 (2005).
[7] *Toni, A. (2008). Pengertian Deteksi Anomali. 282.*.
[8] Wakhidah, N. (2014). Clustering Menggunakan K-Means Algorithm (K-Means Algorithm Clustering). Fakultas Teknologi Informasi, 21(1), 70–80.
[9] Zulfikar, A., Rahmani, F. A., Azizah, N., Perbendaharaan, D. J., Keuangan, K., &Pinang, P. (2023). Deteksi Anomali Menggunakan Isolation Forest Belanja Barang Persediaan Konsumsi Pada Satuan Kerja Kepolisian Republik Indonesia. Jurnal Manajemen Perbendaharaan, 4(1), 1–15.
[10] Talakua, M. W., Leleury, Z. A., & Taluta, A. W. (2017). Analisis Cluster Dengan Menggunakan Metode K-Means Untuk Pengelompokkan Kabupaten/Kota Di Provinsi Maluku Berdasarkan Indikator Indeks Pembangunan Manusia Tahun 2014.