JORAPI : Journal of Research and Publication Innovation
Vol. 3, No. 3, July 2025 **ISSN : 2985-4768**
Halaman : 274-283

ANALISIS DATA EKONOMI DAN KEUANGAN PUBLIK DI INDONESIA MENGGUNAKAN METODE DATA MINING

Nabila Putri Candra[1], Salsa Nabila[2], Intra Sepriansa[3], Serina Admi Yuliana[4], Okta Irawati[5]

[1]Department of Informatics Engineering, Universitas Pamulang, South Tangerang, Banten 15417, Indonesia
E-mail: [1]nabilaputricandraa543@gmail.com, [2]salsanabila2018@gmail.com, [3]intrasepriansa@gmail.com, [4]yulianaserinaadmi@gmail.com, [5]dosen02610@gmail.com

***Abstract*—** Penelitian ini menggabungkan empat studi kasus dalam analisis data publik Indonesia menggunakan pendekatan data mining. Tujuan dari penelitian ini adalah mengeksplorasi pola tersembunyi dari data nyata yang mencakup penyerapan anggaran kementerian tahun 2012, perkembangan UMKM periode 2015–2018, struktur penggajian di Direktorat Jenderal Pajak, serta saham sektor perbankan dalam indeks LQ45. Metode yang digunakan meliputi K-Means Clustering untuk pengelompokan berdasarkan kemiripan fitur, serta Decision Tree untuk klasifikasi efektivitas penyerapan anggaran. Dataset diperoleh dari sumber resmi seperti Seknas FITRA, Kementerian Keuangan, BPS, dan Bursa Efek Indonesia. Hasil analisis menunjukkan bahwa pendekatan data mining mampu mengelompokkan entitas berdasarkan karakteristik dominan, serta menemukan segmentasi yang relevan untuk mendukung pengambilan kebijakan berbasis data. Temuan ini menegaskan pentingnya pemanfaatan data mining dalam menganalisis data publik secara komprehensif dan efisien.

Kata Kunci: data publik, clustering, decision tree, UMKM, APBN, saham, tunjangan, data mining

1. PENDAHULUAN

Lembaga Keuangan Syariah (LKS) memegang peranan penting dalam menopang pertumbuhan ekonomi berbasis nilai-nilai Islam. Dalam kurun waktu 2005 hingga 2009, terjadi perkembangan signifikan terhadap jumlah jaringan kantor LKS, termasuk Bank Umum Syariah (BUS), Unit Usaha Syariah (UUS), Bank Pembiayaan Rakyat Syariah (BPRS), dan Office Channeling. Di sisi lain, Pegadaian Syariah sebagai lembaga pembiayaan mikro berbasis syariah juga mengalami transformasi dalam struktur produk dan rentang plafon pinjamannya.

Penelitian ini mencoba menggabungkan dua sudut pandang: perkembangan infrastruktur LKS serta pola distribusi pinjaman pada Pegadaian Syariah. Dengan menggunakan pendekatan analisis data kuantitatif melalui Python, diharapkan diperoleh wawasan yang komprehensif sebagai bahan pertimbangan dalam kebijakan keuangan syariah nasional.

2. PENELITIAN TERKAIT

Penelitian ini mengacupada beberapa teori dan pendekatan yang telah digunakan sebelumnya dalam analisis data publik menggunakan metode data mining:

- Han & Kamber (2011) menjelaskan bahwa data mining mampu mengekstrak pengetahuan dari data besar untuk keperluan kebijakan.
- Suyanto (2018) menekankan pentingnya penerapan Decision Tree dalam klasifikasi keputusan berbasis logika aturan.
- Fauzi & Nugroho (2022) membahas pengaruh pemanfaatan clustering dalam analisis efektivitas penyerapan anggaran APBN di tingkat kementerian.

Penelitian ini melengkapi studi-studi tersebut dengan pendekatan multikonteks yang mencakup sektor UMKM, pajak, saham, dan belanja pemerintah, yang belum banyak dijadikan satu dalam analisis terpadu.

3. METODE PENELITIAN

A. Penyerapan Anggaran Kementerian

Penelitian ini menggunakan data sekunder yang bersumber dari buku Manajemen Keuangan Sektor Publik yang disunting oleh Abdul Halim dan diterbitkan oleh Salemba Empat. Dataset yang digunakan merupakan data penyerapan anggaran kementerian di Indonesia tahun 2012, yang meliputi besaran APBNP, realisasi anggaran Semester I, dan persentase terhadap APBNP. Data ini merupakan data

nyata yang diperoleh dari Seknas FITRA dan Kementerian Keuangan.

1) Sumber Data:
   - Data APBN: Seknas FITRA (2012)
   - Data UMKM: BPS (2015–2018)
   - Data struktur gaji: Direktorat Jenderal Pajak
   - Data saham sektor perbankan: Bursa Efek Indonesia, indeks LQ45

   Penelitian ini menggunakan data sekunder yang bersumber dari buku Manajemen Keuangan Sektor Publik yang disunting oleh Abdul Halim dan diterbitkan oleh Salemba Empat. Dataset yang digunakan merupakan data penyerapan anggaran kementerian di Indonesia tahun 2012, yang meliputi besaran APBNP, realisasi anggaran Semester I, dan persentase terhadap APBNP. Data ini merupakan data nyata yang diperoleh dari Seknas FITRA dan Kementerian Keuangan.

2) Variabel Penelitian:
   - Kementerian nama instansi pemerintah pusat
   - APBNP (miliar rupiah) jumlah anggaran yang dialokasikan
   - Realisasi Semester I (miliar rupiah): jumlah anggaran yang telah digunakan
   - Persentase Penyerapan (%): hasil perhitungan realisasi terhadap APBNP
   - Efektivitas (label klasifikasi): kategori penyerapan (rendah, sedang, tinggi)

3) Teknik Analisis Data
   Tahapan analisis data dilakukan sebagai berikut:

   a. Preprocessing Data
      - Data diubah ke dalam format digital (.csv)
      - Pembersihan data dilakukan untuk menghindari duplikasi dan kesalahan input
      - Normalisasi dilakukan pada variabel numerik jika diperlukan
      - Label klasifikasi dibuat berdasarkan kategori persentase penyerapan:

        < 10% = Rendah
        10–15% = Sedang
        15% = Tinggi

   b. Clustering (K-Means)
      - Digunakan untuk mengelompokkan kementerian ke dalam beberapa cluster berdasarkan pola kemiripan serapan anggaran.
      - Jumlah cluster (k) ditentukan secara manual (misal: 3 cluster).
      - Hasil divisualisasikan dalam bentuk scatter plot.

   c. Klasifikasi (Decision Tree)
      - Digunakan untuk membuat model klasifikasi efektivitas penyerapan.
      - Algoritma yang digunakan adalah Decision Tree Classifier dari Scikitlearn.
      - Output berupa pohon keputusan dan akurasi prediksi.

   d. Clustering
      K-Means untuk pengelompokan data yang memiliki karakteristik serupa (misalnya, kinerja kementerian atau sektor saham)

   e. Klasifikasi
      Decision Tree untuk menentukan efektivitas penyerapan anggaran

   f. Visualisasi
      Seaborn & matplotlib untuk grafik batang, pie, dan scatter plot

B. Perkembangan UMKM

Penelitian ini bertujuan untuk menganalisis perkembangan Usaha Mikro, Kecil, dan Menengah (UMKM) di Indonesia dalam rentang tahun 2015–2018 menggunakan pendekatan data mining. Metode yang digunakan meliputi regresi linier untuk memproyeksikan tren pertumbuhan, K-Means Clustering untuk segmentasi tahun berdasarkan indikator ekonomi, serta analisis anomali dan asosiasi antar fitur.

1) Jenis Data dan Sumber
   Penelitian ini menggunakan data sekunder bertipe deret waktu (time series) dari tahun 2015 hingga 2018 yang diperoleh dari:

- Badan Pusat Statistik (BPS)
- Kementerian Koperasi dan UKM Republik Indonesia

2) Variabel dan Fitur Data
Data terdiri dari beberapa fitur utama:

- Tahun
- Jumlah UMKM (unit)
- Pertumbuhan Jumlah UMKM (%)
- Jumlah Tenaga Kerja UMKM (orang)
- Pertumbuhan Tenaga Kerja UMKM (%)
- Kontribusi terhadap PDB UMKM (Rp Miliar)
- Pertumbuhan Kontribusi PDB UMKM (%)
- Nilai Ekspor UMKM (Rp Miliar)
- Pertumbuhan Nilai Ekspor UMKM (%)

3) Teknik Analisis
Beberapa teknik dalam data mining digunakan sebagai berikut:

- Regresi Linier: untuk memprediksi perkembangan indikator hingga tahun 2022.
- Clustering (K-Means): untuk mengelompokkan data tahun berdasarkan kesamaan
- karakteristik pertumbuhan.
- Deteksi Anomali: untuk menemukan nilai yang menyimpang dari tren umum.
- Asosiasi (Apriori): untuk mengeksplorasi hubungan antar fitur UMKM dan mencari
- aturan keterkaitan antar variabel.

C. Penggajian DJP

Penelitian ini menggunakan pendekatan kuantitatif eksploratif dengan metode data mining. tujuan dari pendekatan ini adalah untuk menemukan pola tersembunyi dalam data tunjangan pegawai DJP dan mengelompokkan jabatan berdasarkan kesamaan karakteristik kompensasi. Proses analisis dilakukan melalui tahapantahapan utama: eksplorasi data, analisis statistik deskriptif, pemodelan klaster dengan algoritma K-Means, reduksi dimensi dengan PCA, dan pengujian hipotesis menggunakan uji t-test independen.

1) Dataset dan Atribut
Dataset diperoleh dari buku Pajak dan Pendanaan Peradaban Indonesia oleh Gatot Subroto Grade

- Kategori Jabatan: Pelaksana / Struktural
- Tunjangan A
- Tunjangan B
- Tunjangan C, yang berisi 24 entri (grade 1–24).

2) Statistik Deskriptif dan Korelasi
Tahap awal dilakukan analisis deskriptif terhadap seluruh fitur numerik. diperoleh rata-rata tunjangan sebagai berikut:

- Tunjangan A: ±3.849.276
- Tunjangan B: ±4.122.281
- Tunjangan C: ±8.260.352

Selain itu, dilakukan analisis korelasi antar variabel. Hasilnya menunjukkan korelasi yang sangat kuat ($r > 0,99$) antara nilai tunjangan dengan grade jabatan, menandakan pola kompensasi yang terstruktur dan linier.

3) Clustering dengan K-Means
Data dianalisis menggunakan algoritma K-Means Clustering. tahapan dimulai dengan normalisasi seluruh fitur numerik menggunakan StandardScaler untuk menyamakan skala antar fitur.

- Penentuan jumlah klaster (k) menggunakan metode Elbow Method, yang menunjukkan bahwa $k = 2$ merupakan jumlah optimal.
- Klaster pertama mencakup jabatan pelaksana, dan klaster kedua mencakup jabatan struktural.
- Model ini menghasilkan kolom baru yang menyimpan hasil label Cluster.

4) Visualisasi dengan PCA
Untuk membantu interpretasi hasil clustering, digunakan teknik Principal Component Analysis (PCA) untuk mereduksi dimensi dari 4 fitur numerik ke 2 komponen utama. Scatter plot hasil PCA menunjukkan pemisahan yang jelas antara dua kelompok jabatan.

5) Uji t-Test
Untuk menguji perbedaan rata-rata tunjangan antara kelompok pelaksana dan struktural, dilakukan Independent t-Test. Hasil pengujian

menunjukkan nilai p-value < 0.05 untuk semua fitur tunjangan, yang berarti terdapat perbedaan signifikan antara dua kelompok jabatan.

6) Tools dan Perangkat Lunak
Seluruh proses analisis dilakukan dengan menggunakan Python 3.x di lingkungan Jupyter Notebook, dengan library berikut:
- pandas, numpy: untuk manajemen dan analisis data
- scikit-learn: untuk implementasi K-Means dan PCA
- matplotlib, seaborn: untuk visualisasi
- scipy.stats: untuk pengujian t-Test

D. Klasterisasi Saham Sektor Perbankan
Penelitian ini bertujuan untuk mengelompokkan saham-saham perbankan dalam indeks LQ45 berdasarkan karakteristik rasio free float dan jumlah saham beredar, guna menemukan pola distribusi yang dapat mendukung pengambilan keputusan investasi. Pendekatan yang digunakan adalah unsupervised learning dengan algoritma K-Means Clustering.

1) Sumber Data dan Atribut
Dataset diperoleh dari buku The Dividend Investor karya Jefferly Helianthusonfri, khususnya bagian lampiran LQ45 (Oktober 2021). Dataset berisi:
- Kode Saham
- Nama Perusahaan
- Rasio Free Float (% saham publik)
- Jumlah Saham Beredar

Data dikonversi ke format .csv untuk kebutuhan analisis menggunakan Python.

2) Preprocessing dan Normalisasi
Tahap preprocessing meliputi:
- Pengubahan format penulisan angka pada kolom jumlah saham
- Penyesuaian tipe data menjadi numerik
- Normalisasi menggunakan StandardScaler untuk menyetarakan skala antar fitur

3) Preprocessing dan Normalisasi
Tahap preprocessing meliputi:
- Pengubahan format penulisan angka pada kolom jumlah saham
- Penyesuaian tipe data menjadi numerik
- Normalisasi menggunakan StandardScaler untuk menyetarakan skala antar fitur

4) Klasterisasi dengan K-Means
Algoritma K-Means Clustering digunakan dengan nilai k = 3, yang ditentukan secara eksploratif berdasarkan visualisasi data.
- Fitur input: free float, jumlah saham
- Output: Kolom cluster baru yang menandai hasil pengelompokan saham
- Visualisasi: Scatter plot dua dimensi dengan pewarnaan sesuai klister

5) Interpretasi Klaster
Berdasarkan hasil klasterisasi:
- Cluster 0 (Merah): Saham dengan free float rendah dan jumlah saham kecil cenderung kurang likuid
- Cluster 1 (Biru): Free float sedang dan jumlah saham menengah saham stabil
- Cluster 2 (Hijau): Jumlah saham sangat besar saham bluechip / BUMNil klasterisasi:

E. ALAT

1. Python
2. Pustaka: pandas, matplotlib, seaborn, scikit-learn
3. Preprocessing: missing values, normalisasi, dan encoding data kategorik

4. PEMBAHASAN DAN HASIL

A. Penyerapan Anggaran Kementerian

1) Hasil K-Means Clustering
Tahap pertama dalam analisis adalah pengelompokan data kementerian berdasarkan variabel-variabel numerik seperti APBNP, realisasi Semester I, dan persentase penyerapan. Algoritma K-Means digunakan untuk membagi data menjadi tiga cluster yang merepresentasikan kelompok kementerian dengan pola penyerapan yang mirip. Setelah menjalankan algoritma K-Means dengan nilai k = 3, diperoleh tiga cluster yang menggambarkan:

- Cluster 0: Kementerian dengan anggaran besar namun tingkat serapan sangat rendah.
- Cluster 1: Kementerian dengan anggaran menengah dan serapan sedang.
- Cluster 2: Kementerian dengan serapan tinggi, baik dalam nilai nominal maupun persentase.

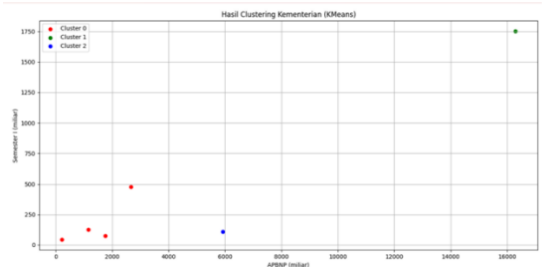
Gambar di bawah menunjukkan hasil pengelompokan 3 cluster berdasarkan persentase serapan

Interpretasi Hasil

```
313/python.exe c:/Users/Nabila/OneDrive/Desktop/Project_Anggaran_Kementerian/decision_tree.py
|--- Semester I (miliar) <= 117.95
|   |--- Semester I (miliar) <= 58.75
|   |   |--- class: Tinggi
|   |--- Semester I (miliar) >  58.75
|   |   |--- class: Rendah
|--- Semester I (miliar) >  117.95
|   |--- APBNP (miliar) <= 1912.90
|   |   |--- class: Sedang
|   |--- APBNP (miliar) >  1912.90
|   |   |--- APBNP (miliar) <= 9479.15
|   |   |   |--- class: Tinggi
|   |   |--- APBNP (miliar) >  9479.15
|   |   |   |--- class: Sedang

PS C:\Users\Nabila\OneDrive\Desktop\Project_Anggaran_Kementerian>
```

dengan sangat baik. Hal ini menunjukkan bahwa efektivitas anggaran tidak hanya ditentukan oleh jumlah dana yang diberikan, tetapi juga oleh kapasitas kelembagaan, perencanaan program, dan efisiensi birokrasi di masing-masing kementerian. Clustering memberikan gambaran umum kelompok kinerja, sedangkan decision tree memberi aturan logis yang bisa digunakan sebagai dasar evaluasi atau monitoring di masa mendatang



Gambar 1. Scatter Plot Clustering Kementerian

Visualisasi hasil clustering dilakukan menggunakan scatter plot, dengan masingmasing cluster diberi warna berbeda. dari hasil plot terlihat bahwa kementerian tidak selalu berkinerja tinggi meskipun memiliki anggaran besar. Beberapa kementerian dengan anggaran kecil justru berada dalam cluster dengan efektivitas serapan tinggi

2) Hasil Decision Tree Classification

Setelah dilakukan pengelompokan, tahap berikutnya adalah klasifikasi menggunakan Decision Tree untuk memahami aturan atau pola yang membedakan efektivitas penyerapan anggaran. Label efektivitas dibuat berdasarkan kategori berikut:

- Rendah: Persentase serapan < 10%
- Sedang: 10% – 15%
- Tinggi: > 15%

Struktur pohon keputusan berikut dihasilkan dari algoritma Decision Tree. Pohon ini menunjukkan alur klasifikasi efektivitas serapan anggaran.

Gambar 2. Struktur Decision Tree

Algoritma Decision Tree digunakan untuk mempelajari hubungan antara variabel input (APBNP, Realisasi Semester I) terhadap label efektivitas tersebut. Hasil dari training model menunjukkan bahwa persentase penyerapan adalah variabel paling dominan dalam menentukan kelas efektivitas. Struktur pohon keputusan yang dihasilkan sederhana namun informatif. Misalnya:

- Jika persentase serapan < 10%, maka kementerian diklasifikasikan sebagai "rendah".
- Jika persentase serapan > 15%, maka diklasifikasikan sebagai "tinggi".
- Sisanya masuk ke kategori "sedang".

B. Perkembangan UMKM di Indonesia

1) Analisis Regresi

Analisis regresi linier digunakan untuk memprediksi arah perkembangan indikator utama UMKM dari tahun 2015 hingga 2018, serta memperkirakan tren hingga tahun 2022. Dengan memetakan data historis terhadap waktu (tahun), model regresi linier memberikan gambaran tren pertumbuhan secara kuantitatif. Berdasarkan hasil regresi, seluruh indikator menunjukkan tren positif. Jumlah unit UMKM mengalami peningkatan setiap tahun, yang berarti semakin banyak individu dan kelompok masyarakat yang terlibat dalam sektor usaha kecil. Peningkatan ini secara tidak langsung mendorong pertumbuhan tenaga kerja, karena setiap usaha baru umumnya menyerap setidaknya satu tenaga kerja tambahan.

Kontribusi UMKM terhadap Produk Domestik Bruto (PDB) juga menunjukkan kecenderungan meningkat. Hal ini mencerminkan efisiensi dan produktivitas UMKM dalam kegiatan ekonomi riil. Dengan demikian, model regresi dapat dijadikan dasar dalam penyusunan proyeksi kebijakan ekonomi, terutama untuk mendukung pertumbuhan sektor UMKM secara nasional. Selain itu, nilai ekspor UMKM pun meningkat, walaupun secara fluktuatif. Tren ini menunjukkan bahwa UMKM Indonesia mulai mampu menjangkau pasar internasional. Proyeksi model regresi linier memperkirakan bahwa apabila tren saat ini dipertahankan, kontribusi UMKM terhadap perekonomian akan semakin signifikan dalam lima tahun ke depan

2) Clustering

Clustering atau pengelompokan digunakan untuk membagi data ke dalam beberapa kelompok berdasarkan kemiripan karakteristik. Dalam penelitian ini, algoritma K Means digunakan untuk membagi data UMKM per tahun ke dalam dua kelompok berdasarkan pertumbuhan jumlah UMKM, tenaga kerja, kontribusi PDB, dan ekspor. Hasil analisis menunjukkan bahwa:

- Cluster 1 mencakup tahun-tahun dengan pertumbuhan sedang atau rendah, yang mungkin disebabkan oleh tantangan ekonomi makro atau keterbatasan kebijakan dukungan UMKM.

- Cluster 2 menunjukkan tahun-tahun dengan lonjakan pertumbuhan yang signifikan, baik dari sisi jumlah UMKM maupun kontribusi PDB.

Dengan pemetaan ini, pembuat kebijakan dapat mengenali kondisi yang mendasari pertumbuhan tinggi pada tahun-tahun tertentu, serta mengevaluasi apa yang menyebabkan pertumbuhan lambat di tahun lainnya. Clustering juga membantu dalam merancang strategi segmentasi kebijakan, misalnya menetapkan intervensi khusus untuk wilayah atau sektor yang masuk dalam cluster berisiko rendah.

3) Deteksi Anomali

Deteksi anomali bertujuan untuk mengidentifikasi data atau tahun yang menunjukkan nilai-nilai tidak biasa dibandingkan dengan tren umum. Dalam konteks UMKM, anomali dapat disebabkan oleh peristiwa luar biasa, seperti perubahan kebijakan fiskal, krisis ekonomi regional/global, atau bencana alam yang mempengaruhi aktivitas usaha. Dalam data 2015–2018, deteksi anomali mengindikasikan adanya penyimpangan yang menonjol pada:

- Pertumbuhan ekspor UMKM, yang menunjukkan fluktuasi tajam di satu atau dua tahun, kemungkinan disebabkan oleh volatilitas pasar luar negeri atau penurunan daya saing produk ekspor.

- Kontribusi terhadap PDB, yang meskipun meningkat, menunjukkan ketidaksesuaian antara kenaikan unit UMKM dan pertumbuhan nilai tambah ekonominya.

4) *Asosiasi*

Analisis asosiasi dilakukan untuk menemukan aturan atau keterkaitan antar variabel dalam dataset. Dalam penelitian ini, pendekatan asosiasi digunakan untuk menilai hubungan antara variabel-variabel seperti pertumbuhan jumlah UMKM dengan pertumbuhan tenaga kerja dan kontribusi terhadap PDB. Hasil analisis asosiasi menghasilkan pola-pola berikut:

- Jika jumlah UMKM meningkat lebih dari 5% per tahun, maka dalam sebagian besar kasus, jumlah tenaga kerja juga meningkat minimal 3%.

- Peningkatan kontribusi terhadap PDB lebih sering terjadi ketika jumlah UMKM dan ekspor tumbuh secara bersamaan.

Pola-pola ini memberikan wawasan yang berguna bagi pembuat kebijakan dan pelaku usaha. Misalnya, untuk meningkatkan kontribusi PDB dari UMKM, intervensi tidak cukup hanya meningkatkan jumlah unit UMKM, namun juga perlu mendorong produktivitas dan kemampuan ekspor. Dengan memahami hubungan-hubungan ini, strategi pengembangan UMKM dapat dirancang secara lebih terarah dan integrative

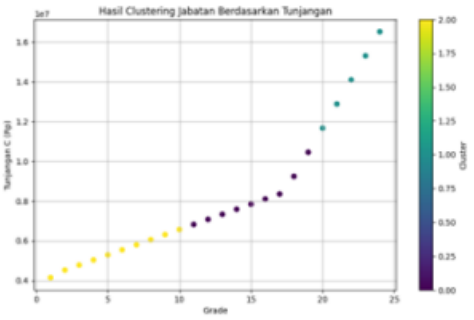
C. Struktur Gaji DJP

1) *Clustering Menggunakan K-Means*

Algoritma K-Means digunakan untuk melakukan segmentasi jabatan berdasarkan atribut numerik. K-Means merupakan metode klastering yang membagi data ke dalam K kelompok berdasarkan jarak Euclidean antara data dan pusat klaster (centroid). Data di-scaling terlebih dahulu menggunakan Standard Scaler agar seluruh kolom numerik memiliki distribusi standar dan setara dalam perhitungan jarak. Penentuan jumlah klaster optimal dilakukan dengan metode Elbow, yaitu teknik yang memplot nilai WCSS (within-cluster sum of squares) terhadap jumlah klasster. Titik siku pada grafik Elbow menunjukkan jumlah klaster terbaik. Berdasarkan visualisasi tersebut, dipilih K=2 karena data cenderung membentuk dua kelompok besar.

Gambar 1. Source code Clustering Menggunakan K-Means

Model K-Means kemudian dilatih pada data yang telah diskalakan. Output dari proses ini adalah penambahan kolom baru (Cluster) yang menunjukkan klaster keanggotaan setiap data. Hasil klastering digunakan sebagai dasar segmentasi antara jabatan pelaksana dan struktural.

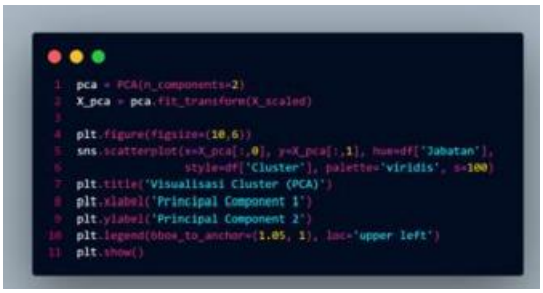


Gambar 2. Hasil clustering

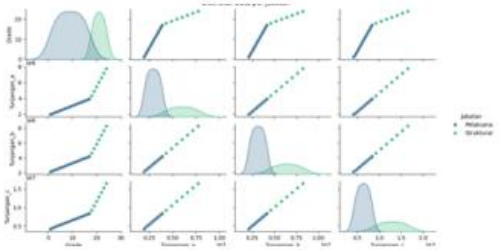
2) Visualisasi PCA

Untuk mempermudah interpretasi hasil klastering, dilakukan reduksi dimensi menggunakan PCA (Principal Component Analysis). PCA digunakan untuk mengubah data berdimensi tinggi (4 fitur numerik) menjadi dua dimensi utama. Ini memungkinkan visualisasi dalam bentuk scatterplot dua dimensi yang mudah dipahami.

PCA memproyeksikan data ke dalam komponen utama berdasarkan varians terbesar. Dua komponen pertama yang diperoleh menjelaskan lebih dari 95% variansi total dalam data, yang menunjukkan efektivitas reduksi dimensi. Hasilnya divisualisasikan dengan pewarnaan berdasarkan klaster hasil K-Means.



Gambar 3. Source Code Visualisasi PCA



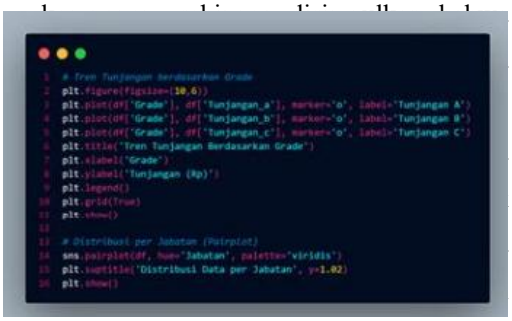
Gambar 4. Figure 2

3) Visualisasi Tren dan Distribusi

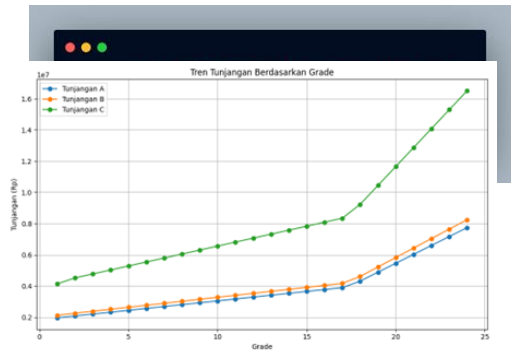
Selain klaster, penelitian juga memvisualisasikan tren nilai tunjangan terhadap grade menggunakan line chart. Grafik ini menunjukkan bahwa ketiga tunjangan mengalami kenaikan linear hingga grade 18,

lalu meningkat tajam mulai grade 19–24. Ini menunjukkan adanya "lonjakan kompensasi" untuk jabatan struktural senior. Untuk melihat distribusi per jabatan, digunakan Pairplot yang membandingkan semua pasangan kombinasi variabel. Setiap titik berwarna berdasarkan kategori jabatan. Hasilnya menunjukkan dua klaster visual alami antara pelaksana dan struktural, yang menguatkan hasil analisis K-Means dan PCA

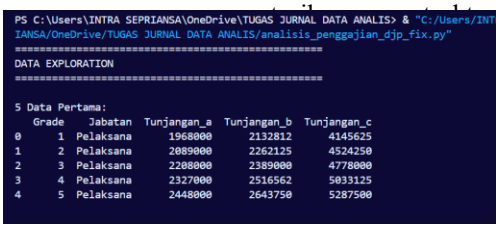
5. KESIMPULAN DAN SARAN



Gambar 5. Source Code Visualisasi Tren dan Distribusi



Gambar 6. Visualisasi_tunjangan



Gambar 7. Hasil explorasi

4) Uji t-Test

Uji t-test independen digunakan untuk mengetahui apakah terdapat perbedaan signifikan antara rata-rata nilai tunjangan kelompok jabatan pelaksana dan struktural. Uji dilakukan untuk masing-masing komponen tunjangan (A, B, dan C). Nilai p-value < 0,05 diperoleh untuk seluruh

inferensial.

5)

Setelah melalui tahapan pengolahan dan analisis data menggunakan pendekatan data mining, diperoleh sejumlah temuan penting

dan uji t-test sebagai validasi statistik inferensial.

D. Klasterisasi Saham Sektor Perbankan

1) Preprocessing dan Normalisasi

 ISSN : 2985-4768

Data saham diperoleh dari buku The Dividend Investor karya Jefferly Helianthusonfri, serta dilengkapi dengan data LQ45 edisi Oktober 2021. Dataset awal terdiri atas:

- Kode saham
- Nama emiten
- Rasio free float (dalam persentase)
- Jumlah saham beredar

Data dikonversi ke format .csv, kemudian dilakukan preprocessing:

- Penghapusan karakter non-numerik (misalnya tanda "%" dan titik)
- Konversi ke tipe data numerik
- Normalisasi menggunakan StandardScaler untuk menyamakan skala antar fitur

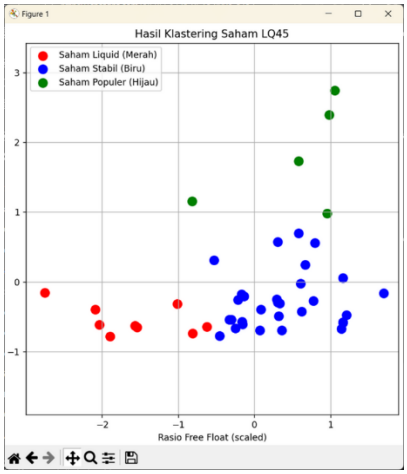
2) Proses Klasterisasi

Setelah preprocessing selesai, dilakukan proses klasterisasi menggunakan algoritma K-Means dengan jumlah klaster k = 3 (ditentukan secara eksploratif). Model kemudian diterapkan pada dua fitur utama:

- Rasio free float
- Jumlah saham

3) Hasil klasterisasi menghasilkan tiga klaster sebagai berikut:

- Klaster 0 (Merah): Free float rendah, jumlah saham kecil → kurang likuid
- Klaster 1 (Biru): Nilai sedang → saham stabil
- Klaster 2 (Hijau): Free float tinggi, jumlah saham besar → saham bluechip/BUMN



Gambar 1. Visualisasi hasil clustering saham sector perbankan menggunakan K-Means

```python
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt

# 1. Baca dataset
df = pd.read_csv("data_lq45.csv")
df.columns = df.columns.str.strip()
print("Kolom yang terbaca:", df.columns.tolist())  # debug

# 2. Bersihkan kolom Jumlah_Saham
df["Jumlah_Saham"] = df["Jumlah_Saham"].astype(str).str.replace(".", "", regex=False)
df["Jumlah_Saham"] = df["Jumlah_Saham"].str.lstrip(".")
df["Jumlah_Saham"] = df["Jumlah_Saham"].astype("int64")
```

```python
# 3. Ambil fitur untuk clustering
X = df[["Rasio_Free_Float", "Jumlah_Saham"]]

# 4. Normalisasi
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# 5. Klastering dengan KMeans
kmeans = KMeans(n_clusters=3, random_state=42)
df["Cluster"] = kmeans.fit_predict(X_scaled)

# 6. Tambahkan kolom label klaster yang mudah dipahami
label_mapping = {
    0: 'Saham Liquid',
    1: 'Saham Stabil',
    2: 'Saham Populer'
}
df["Kategori_Klaster"] = df["Cluster"].map(label_mapping)
```

```
plt.title("Hasil Klastering Saham LQ45")
plt.xlabel("Rasio Free Float (scaled)")
plt.ylabel("Jumlah Saham (scaled)")
plt.legend()
plt.grid(True)
plt.tight_layout()
plt.savefig("klaster_lq45.png")
plt.show()
```

Gambar 2. Kode Python clustering saham

4) Interpretasi dan Diskusi

Hasil klasterisasi ini menunjukkan bahwa free float dan jumlah saham dapat digunakan untuk mengelompokkan saham berdasarkan likuiditas dan kapitalisasi:

- Saham dalam klaster biru memiliki distribusi data yang "aman" dan bisa dipertimbangkan sebagai saham stabil
- Saham dalam klaster hijau cocok untuk investor institusi karena termasuk bluechip dengan distribusi saham luas
- Saham dalam klaster merah rentan kurang likuid dan cenderung tidak menarik bagi investor konservatif

Model ini dapat digunakan oleh investor untuk filtering saham secara cepat berdasarkan distribusi, terutama saat memilih sektor perbankan yang kompetitif.

V. KESIMPULAN

a. Pendekatan data mining dapat memetakan entitas publik secara lebih akurat berdasarkan performa dan pola data.
b. KMeans efektif dalam membagi unit kerja atau sektor berdasarkan karakteristik dominan.
c. Decision Tree memberikan wawasan logis terhadap penentuan efektivitas kebijakan.

DAFTAR PUSTAKA

[1] Halim, A. (Ed.). (2012). Manajemen Keuangan Sektor Publik: Problematika Penerimaan dan Pengeluaran Pemerintah. Jakarta: Salemba Empat.

[2] Sekretariat Nasional FITRA. (2012). Kinerja Anggaran Kementerian/Lembaga Tahun 2012. Jakarta: Seknas FITRA.

[3] Han, J., Pei, J., & Kamber, M. (2011). Data Mining: Concepts and Techniques (3rd ed.). San Francisco: Morgan Kaufmann.

[4] Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann.

[5] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651–666.

[6] BPS. (2017). Statistik UMKM Indonesia. Kementerian Koperasi dan UKM Republik Indonesia. Buku: Manajemen Keuangan untuk Wirausaha Mula. Han, J., Pei, J., & Kamber, M. (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann**.**

[7] Subroto, G. (2022). *Pajak dan Pendanaan Peradaban Indonesia*. Jakarta: Penerbit DJP.

[8] Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.

[9] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.

[10] Junaedi, D. (2020). Penerapan Algoritma K-Means pada Pengelompokan Data Gaji Pegawai. *Jurnal Informatika dan Sistem Informasi*, 16(2), 88–97.

[11] ugiyono. (2017). Metode Penelitian Kuantitatif, Kualitatif dan R&D. Bandung: Alfabeta.

[12] McKinney, W. (2022). *Python for Data Analysis* (3rd ed.). O'Reilly Media.

[13] Scikit-learn Developers. (2024). scikit-learn: Machine Learning in Python. https://scikit-learn.org

[14] Helianthusonfri, J. (2021). The Dividend Investor: Cara untuk selalu cuan dari dividen saham. Jakarta: Elex Media Komputindo.

[15] Referensi Teknologi (jika kamu pakai):

[16] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[17] Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90–95.

[18] McKinney, W. (2010). Data structures for statistical computing in Python. In Proceedings of the 9th Python in Science Conference (pp. 51–56).