Klasifikasi Kemiskinan di Indonesia Menggunakan Algoritma K-Means, SVM, DBSCAN, Random Forest, dan Gaussian Mixture Models

1st Muhamad Ibnu Rizky
UBP Karawang
Karawang, Indonesia
If21.muhamadrizky@mhs.ubpkarawang.ac.id

2nd Bariz Akhdan Faisal
UBP Karawang
Karawang, Indonesia
If21.muhamadrizky@mhs.ubpkarawang.ac.id

3rd Ferdi Arnanda Putra
UBP Karawang
Karawang, Indonesia
If21.ferdiputra@mhs.ubpkarawang.ac.id

4th Rizki Masharikul
UBP Karawang
Karawang, Indonesia
if21.rizkianwar@mhs.ubpkarawang.ac.id

5th Muhammad Lukman
UBP Karawang
Karawang, Indonesia
if21.muhammadhakim@mhs.ubpkarawang.ac.id

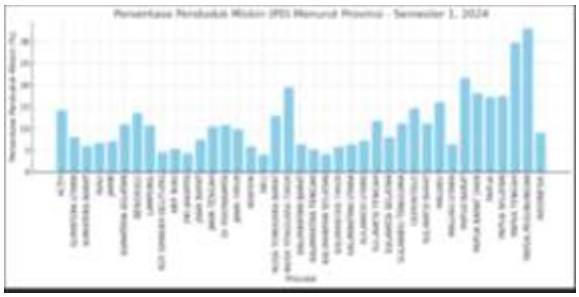
6th Deden Wahiddin M.Kom
UBP Karawang
Karawang, Indonesia
deden.wahiddin@ubpkarawang.ac.id

***Abstract*—Poverty classification is a crucial step in understanding the distribution and factors contributing to poverty in Indonesia. This study evaluates the performance of machine learning algorithms, namely K-means, Support Vector Machine (SVM), DBSCAN, Random Forest, and Gaussian Mixture Models (GMM), in clustering and classifying poverty data based on socio-economic indicators such as income, education level, healthcare access, and housing conditions. The analysis process includes data preprocessing to ensure optimal data quality, the application of K-means and DBSCAN for clustering to identify patterns in unlabeled data, and the use of SVM, Random Forest, and GMM for labeled data classification to predict poverty status.**

***Keywords*— Kemiskinan, Klasifikasi, *Clustering*, *K-means*, *Support Vector Machine* (SVM), *DBSCAN*, *Random Forest*, *Gaussian Mixture Models* (GMM).**

I. PENDAHULUAN

Permasalahan yang dihadapi dan mengkhawatirkan dunia adalah kemiskinan. Saat ini, negara-negara yang kurang beruntung justru menghadapi masalah distribusi pendapatan dan pembangunan yang tidak konsisten, untuk sementara banyak negara non-industri yang mengalami perkembangan finansial yang tinggi namun tidak memberikan keuntungan bagi masyarakat miskin[1]. Kemiskinan menjadi salah satu permasalahan negara indonesia yang terus meningkat dan belum bisa diselesaikan hingga saat ini, namun kebutuhan masih menjadi masalah yang harus dipikirkan. Kemiskinan adalah suatu kondisi di mana individu tidak dapat mengatasi masalah-masalah penting seperti makanan, pakaian, pelatihan, dan berbagai kebutuhan dasar lainnya [2]. Permasalahan yang menyebabkan kemiskian adalah kondisi suatu negara dan ituasi global. Kemiskinan bukan hanya sekedar tidak cukupnya ekonomi, namun juga gagalnya kewajiban untuk memenuhi hak dan perlakuan bermartabat dari setiap orang [3]. Hal ini menunjukkan bahwa Indonesia termasuk dalam kategori salah satu negara dengan jumlah penduduk miskin yang tinggi di dunia. Tingkat kemiskinan penduduk Indonesia di setiap provinsi dapat dilihat pada gambar 1 di bawah ini.



Gambar. 1 Peresentase Tingkat Kemiskinan Penduduk di Indonesia Menurut Provinsi
(Sumber: Badan Pusat Statistik, 2024)

Pada gambar 1. Menunjukkan nilai rata-rata tingkat kemiskinan di Indonesia sangat tinggi. Provinsi Papua menempati peringkat pertama tingkat kemiskinan tertinggi dengan jumlah persentase 32,97%. Begitu pula dengan Bali yang jumlah kemiskinannya berada di peringkat terakhhir mencapai 4,45%. Hal ini disebabkan karena berbagai faktor seperti tingginya pertumbuhan penduduk, meningkatnya angka pengangguran, rendahnya pertumbuhan ekonomi dan taraf Pendidikan yang rendah. Mengingat kemiskinan yang dialami masyarakatdi Indonesia, maka pemerintah perlu melakukan upaya pengurangan persentase penduduk miskin dengan berbagai bantuan[4]. Oleh karena itu, diperlukan rencana provinsi mana yang harus dijadikan prioritas. Pemerintah mengalami kesulitan untuk menentukan gambaran wilayah miskin dengan angka tertinggi. Masalah tersebut dapat diatasi dengan menggunakan pengelompokan data mining[5].

Adapun tujuan dari penelitian ini adalah untuk

1. Menganalisis Tingkat Kemiskinan: Mengidentifikasi dan menganalisis tingkat kemiskinan di berbagai provinsi di Indonesia serta faktor-faktor yang mempengaruhinya.
2. Pengelompokan Data Mining: Menerapkan teknik data mining untuk mengelompokkan provinsi berdasarkan tingkat kemiskinan, sehingga memudahkan identifikasi wilayah yang membutuhkan intervensi.
3. Solusi Berbasis Data: Menyediakan solusi yang berbasis data untuk pengentasan kemiskinan yang dapat digunakan oleh pembuat kebijakan.

Dan manfaat dari penelitian ini adalah

1. Informasi yang Akurat: Memberikan data dan informasi yang akurat mengenai tingkat kemiskinan di berbagai provinsi di Indonesia, yang dapat digunakan sebagai dasar untuk pengambilan keputusan.
2. Prioritas Kebijakan: Membantu pemerintah melakukan kepentingan dalam menentukan prioritas wilayah yang membutuhkan perhatian lebih dalam program pengentasan kemiskinan.
3. Pendekatan Berbasis Data: Mendorong penggunaan pendekatan berbasis data dalam merumuskan kebijakan dan program, sehingga intervensi yang dilakukan lebih efektif dan efisien.

II. METODOLOGI

Metode dalam penelitian ini, dibagi dalam lima tahap yaitu : pengambilan data, pengolahan data, menentukan jumlah cluster, klustering data dan analisis hasil dan evaluasi. Metode penelitian dapat dilihat pada Gambar 1, berikut :

Gambar. 2 Alur Penelitian

A. *Pengumpulan Data*

Pada tahap pengumpulan data, dataset yang akan digunakan yaitu dari public domain Kaggle. Data berjumlah 515, Pengumpulan sumber data utama dilakukan dengan melihat data penduduk miskin di website kaggle.com pada tahun 2014 hingga 2024. Seluruh data diproses agar mendapatkan daftar penduduk miskin di Indonesia dari setiap provinsi. Data ini dapat membantu memberikan informasi kepada pemerintah untuk penyebaran bantuan sosial kepada masyarakat. Untuk proses klasterisasi data, dibutuhkan 9 atribut dari data bps yang telah didapat sebelumnya. Atributnya yaitu provinsi dan tahun 2014-2024. Atribut-atribut ini menjadi acuan untuk pengelompokan. Data yang telah didapat bisa dilihat ditabel 1

Tabel 1. Data Penduduk Miskin di Indonesia.

B. *Pre-Processing Data*

Preprocessing adalah tahap awal dalam pengolahan data yang bertujuan untuk mempersiapkan data mentah agar siap digunakan dalam analisis atau model pembelajaran mesin. Tahap ini meliputi pembersihan data (menghapus data duplikat, menangani nilai yang hilang, dan mengatasi outliers), transformasi data (normalisasi untuk menyelaraskan rentang nilai, standarisasi untuk distribusi normal, dan encoding untuk mengubah data kategorikal menjadi numerik), serta reduksi dimensi untuk menyederhanakan data tanpa kehilangan informasi penting. Proses ini sangat penting untuk meningkatkan kualitas data, sehingga model yang dibangun lebih akurat dan efisien.

C. *Clustering*

Clustering merupakan teknik dalam data mining yang berguna untuk mengelompokkan sekumpulan objek ke dalam beberapa cluster dengan karakteristik yang sama, sehingga objek sebuah cluster mirip tetapi tidak mirip dengan objek dalam cluster yang berbeda. Cluster adalah kumpulan objek yang serupa tetapi berbeda dalam suatu grup dengan benda-benda milik kelompok lain. Ada dua metode clustering, yaitu hirarkis clustering dan pengelompokan partisi. Data dikelompokkan berdasarkan grafik hierarkis dalam metode pengelompokan hierarkis, dimana dua grup terdekat digabung atau semua data dikelompokkan menjadi cluster. Dalam pengelompokan partisi pengelompokan data tidak memiliki hierarki apapun, setiap cluster memiliki centroid, tujuannya adalah untuk meminimalkan jarak semua data ke centroid [6].

- K-Means

K-Means adalah sebuah metode pengelompokan data menjadi dua atau lebih dari kelompok[7]. Algoritma K-Means merupakan metode analisis kelompok yang membagi subjek penelitian ke dalam kelompok-kelompok, dimana setiap subjek yang akan diamati berada dalam satu kelompok data dengan rata-rata yang berdekatan satu sama lain. Seperti yang kita inginkan, K digunakan untuk konstanta clustering total dan Means berarti mean dari dataset, dalam hal ini sebagai cluster, jadi K-means Clustering adalah metode analisis data yang menggunakan sistem partisi untuk pengelompokan data[8].

- Support Vector Machine

Support Vector Machine (SVM) adalah algoritma pembelajaran mesin yang digunakan untuk klasifikasi dan regresi[9]. Dalam konteks klasifikasi kemiskinan di Indonesia, SVM berfungsi untuk memisahkan data berdasarkan fitur-fitur yang relevan, seperti pendapatan, pendidikan, akses terhadap pelayanan kesehatan, dan faktor sosial-ekonomi lainnya. Metode Support Vector Machine (SVM) digunakan dalam klasifikasi kemiskinan di Indonesia untuk memisahkan data berdasarkan hyperplane. SVM efektif dalam menangani data non-linier dengan

menggunakan kernel, sehingga dapat mengidentifikasi pola kemiskinan yang kompleks dalam dataset yang beragam[10].

- Random Forest

Algoritma Random Forest (RF) merupakan penyempurnaan dari algoritma decision tree dengan merestrukturisasi tree karena algoritma decision tree tidak dapat bekerja secara maksimal pada data dengan dimensi yang sangat besar karena strukturnya akan sangat kompleks dan menyebabkan terjadinya overfitting [11]. Model RF dibangun dengan konsep bagging (bootstrap aggregation) yaitu pengumpulan secara acak sampel observasi ke dalam suatu penampung yang disebut bag, kemudian data pada bag diambil secara WR (with replacement) yang memungkinkan sebuah hasil observasi memiliki peluang untuk terpilih kembali.

- Gaussian Mixture Models

Algoritma Gaussian Mixture Mode (GMM) merupakan metode analisis cluster non-hierarchical yang bekerja untuk memodelkan beberapa data dalam distribusi Gaussian dengan parameter mean dan varians tertentu [(12]. Dalam melakukan clustering ,Algoritma Gaussian Mixture Model adalah salah satu jenis soft clustering dimana satu data point bisa berada pada dua atau lebih cluster. Menurut [13]. Algoritma Gaussian Mixture Model sendiri merupakan model statistik yang sangat populer dan umum digunakan.

- DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) merupakan Metode yang digunakan untuk mengelompokkan data berdasarkan kepadatan. Dalam konteks klasifikasi kemiskinan di Indonesia, DBSCAN efektif dalam mengidentifikasi pola kemiskinan dengan membentuk cluster yang tidak teratur dan mengatasi noise atau outlier yang mungkin ada dalam data algoritma yang dapat mendeteksi outlier atau noise dan menghasilkan cluster lebih akurat dan baik untuk data dalam jumlah besar serta tidak perlu menentukan jumlah cluster di awal. Algoritma ini lebih efektif dan lebih baik dalam menentukan parameter daripada algoritma Dynamic Method Density Based Spatial Clustering of Application with Noise (DMDBSCAN) dan memiliki kemampuan yang berbeda dari algoritma K-Means dan K-Medoids[13].

III. HASIL DAN PEMBAHASAN

Hasil penelitian yang dilakukan berupa klasifikasi kemiskinan di Indonesia dengan menggunakan perbandingan hasil dari algoritma K-means, DBSCAN, Gaussian Mixture Models (GMM), Support Vector Machines (SVM) dan Random Forest terhadap sehingga didapatkan hasil algoritma terbaik.

A. Import Data

Informasi yang terkandung dalam data ini meliputi provinsi dan kabupaten/kota, persentase penduduk miskin (P0), rata-rata lama sekolah penduduk usia 15 tahun ke atas, pengeluaran per kapita yang disesuaikan, indeks pembangunan manusia (IPM), serta umur harapan hidup di setiap daerah. Selain itu, data ini juga memuat persentase rumah tangga yang memiliki akses terhadap sanitasi layak, persentase rumah tangga yang memiliki akses terhadap air minum layak, tingkat pengangguran terbuka, tingkat partisipasi angkatan kerja, dan PDRB atas dasar harga konstan. Klasifikasi kemiskinan juga disertakan dalam dataset ini untuk membantu mengkategorikan tingkat kemiskinan berdasarkan berbagai indikator tersebut. Dengan data ini, analisis lebih lanjut dapat dilakukan untuk memahami faktor-faktor yang mempengaruhi kemiskinan di Indonesia.

B. Pre-Processing Data

- Mengubah Data Menjadi Float

Mengubah data menjadi tipe float adalah langkah penting dalam pre-processing. Proses ini melibatkan penggantian koma (,) dengan titik (.) sebagai pemisah desimal, lalu konversi menggunakan .astype(float). Ini memastikan data kompatibel dengan algoritma statistik dan pembelajaran mesin, serta memfasilitasi perhitungan dan analisis yang akurat. Dengan tipe float, data siap untuk analisis lebih lanjut dan mengurangi risiko kesalahan.

- Menghapus Missing Value

Menghapus missing value adalah langkah penting dalam pre-processing untuk menangani data kosong (NaN). Dengan fungsi dropna(how='all') untuk menghapus baris kosong atau dropna(subset=['Nama Kolom']) untuk kolom tertentu, proses ini menjaga integritas dataset. Menghilangkan missing value mencegah error, meningkatkan akurasi model, dan memastikan kualitas data untuk analisis dan algoritma pembelajaran mesin.

Gambar. 3 Missing Value

- Mengubah Tipe Data dari Float menjadi integer

Mengubah tipe data dari float ke integer penting untuk data yang tidak memerlukan desimal, seperti kategori atau jumlah bulat. Proses ini dilakukan dengan .astype(int) untuk mengurangi kompleksitas dan menghemat ruang penyimpanan. Namun, perlu diperhatikan bahwa konversi ini menghilangkan bagian desimal, sehingga harus dipastikan tidak mempengaruhi analisis. Data integer siap untuk algoritma yang memerlukan data diskrit.

Identify applicable funding agency here. If none, delete this text box.

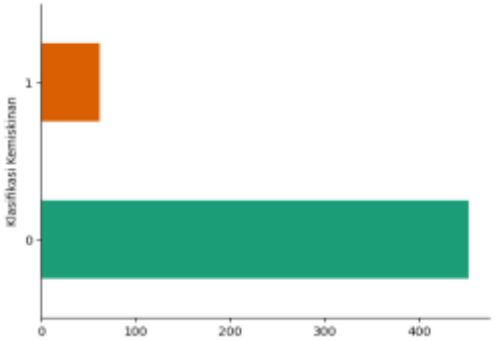
```
<class 'pandas.core.frame.DataFrame'>
Index: 514 entries, 0 to 513
Data columns (total 13 columns):
 #   Column                                                  Non-Null Count  Dtype
---  ------                                                  --------------  -----
 0   Provinsi                                                514 non-null    object
 1   Kab/Kota                                                514 non-null    object
 2   Persentase Penduduk Miskin (P0) Menurut Kabupaten/Kota (Persen)  514 non-null  float64
 3   Rata-rata Lama Sekolah Penduduk 15+ (Tahun)             514 non-null    float64
 4   Pengeluaran per Kapita Disesuaikan (Ribu Rupiah/Orang/Tahun)  514 non-null  float64
 5   Indeks Pembangunan Manusia                              514 non-null    float64
 6   Umur Harapan Hidup (Tahun)                              514 non-null    float64
 7   Persentase rumah tangga yang memiliki akses terhadap sanitasi layak  514 non-null  float64
 8   Persentase rumah tangga yang memiliki akses terhadap air minum layak  514 non-null  float64
 9   Tingkat Pengangguran Terbuka                            514 non-null    float64
 10  Tingkat Partisipasi Angkatan Kerja                      514 non-null    float64
 11  PDRB atas Dasar Harga Konstan menurut Pengeluaran (Rupiah)  514 non-null  float64
 12  Klasifikasi Kemiskinan                                  514 non-null    int64
dtypes: float64(10), int64(1), object(2)
memory usage: 56.2+ KB
```

Gambar. 4 Data Float to Integer

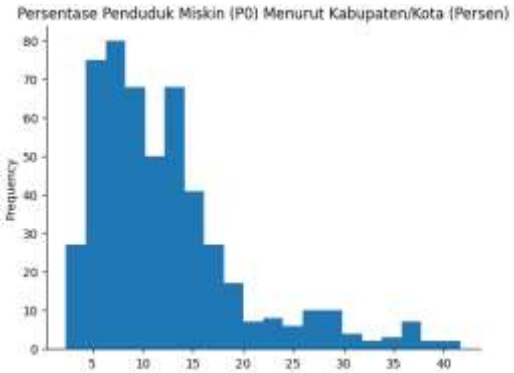
C. *Exploratory Data*

- Distribusi Data

Diagram kemiskinan dan klasifikasi penduduk miskin memvisualisasikan sebaran kemiskinan di berbagai wilayah. Diagram ini menunjukkan proporsi penduduk miskin dan membandingkan tingkat kemiskinan antar kabupaten dan kota. Dalam machine learning, visualisasi ini membantu mengidentifikasi pola dan ketidakseimbangan data, serta memberikan wawasan untuk perencanaan kebijakan dan intervensi berbasis data.



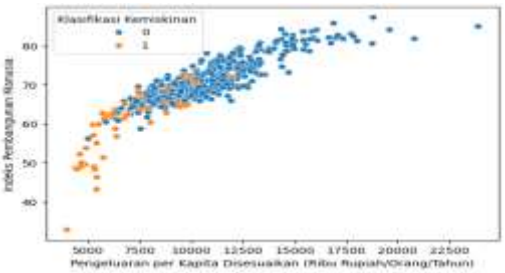
Gambar. 5 Klasifikasi Kemiskinan



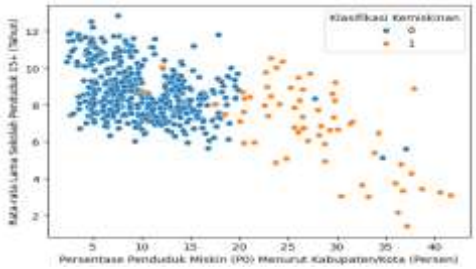
Gambar. 6 Peresentase Penduduk Miskin Menurut Kota

- Hubungan Variabel

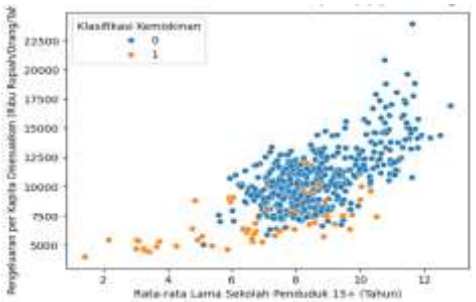
Proses analisis hubungan variabel dalam data mining dimulai dengan pengumpulan data yang relevan dari berbagai sumber, diikuti dengan pembersihan data untuk menghapus duplikasi dan menangani nilai yang hilang, kemudian dilakukan eksplorasi data untuk memahami distribusi dan karakteristiknya, serta pemilihan variabel yang akan dianalisis berdasarkan tujuan penelitian.



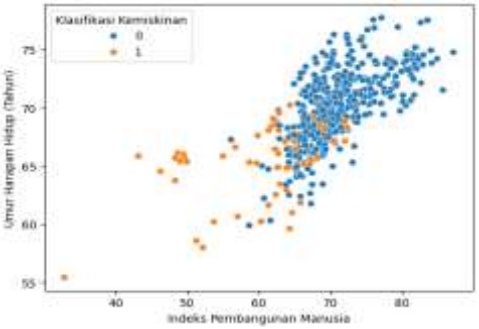
Gambar. 7 Pengeluaran Per Kapita



Gambar. 8 Peresentase Penduduk



Gambar. 9 Rata-rata lama sekolah



Gambar. 10 Indexs Pembangunan

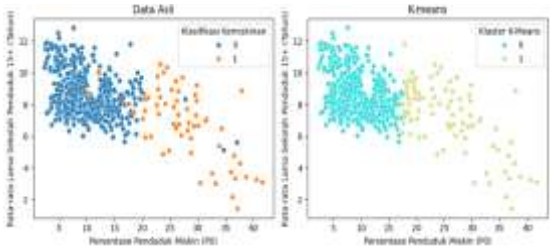
D. *Uji Menggunakan Algoritma*

Dalam penelitian ini, dilakukan perbandingan 5 algortima menggunakan K-means, DBSCAN, Gaussian Mixture Models (GMM), Support Vector Machines (SVM) dan Random Forest dengan data hasil normalisasi sebelumnya. Beberapa percobaan klastering menggunakan algoritma K-Means, Support Vector Machines, Random Forest, Gaussian Mixture Models dan Dbscan. Hal ini bertujuan untuk mendapatkan nilai cluster yang tepat.

- K-Means

Algoritma K-Means untuk mengelompokkan data berdasarkan dua variabel, yaitu "Persentase Penduduk Miskin (P0)" dan "Rata-rata Lama Sekolah Penduduk 15+". Proses klasterisasi dilakukan dengan membagi data menjadi dua klaster, hasilnya disimpan dalam kolom baru

`Klaster K-Means`. Data kemudian divisualisasikan dalam dua scatterplot: plot pertama menampilkan data asli dengan pewarnaan berdasarkan kategori "Klasifikasi Kemiskinan", sedangkan plot kedua menampilkan hasil klasterisasi dengan pewarnaan berdasarkan klaster yang dihasilkan oleh K-Means. Visualisasi ini mempermudah perbandingan pola antara data asli dan hasil klasterisasi.



Gambar. 11 Rata-rata lama sekolah

Setelah itu, dilakukan perhitungan akurasi untuk mengevaluasi hasil klasterisasi, dan diperoleh nilai akurasi sebesar 0.9396887159533074, menunjukkan bahwa model memiliki performa yang baik.

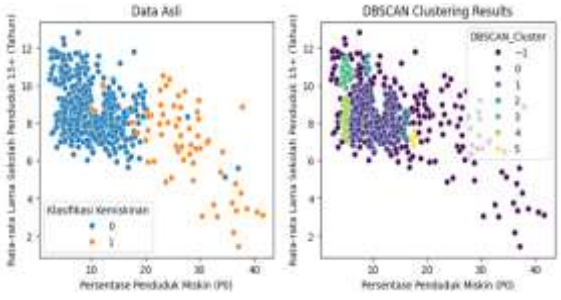
```
from sklearn.metrics import accuracy_score
# Menghitung akurasi K-Means
accuracy_nb = accuracy_score(df['Klasifikasi Kemiskinan'], df['Klaster K-Means'])
print(f"Akurasi K-Means: {accuracy_nb}")
```

```
Akurasi K-Means: 0.9396887159533074
```

Gambar. 12 Hasil Klasterisasi K-Means 93%

- DBSCAN

Algoritma DBSCAN untuk melakukan klasterisasi pada data berdasarkan dua variabel, yaitu "Persentase Penduduk Miskin (P0)" dan "Rata-rata Lama Sekolah Penduduk 15+". DBSCAN bekerja dengan mendeteksi klaster berdasarkan kepadatan data, menggunakan parameter epsilon (eps=0.5) sebagai radius maksimum dan min_samples=5 sebagai jumlah minimum titik dalam suatu klaster. Hasil klasterisasi disimpan dalam kolom baru DBSCAN_Cluster. Data kemudian divisualisasikan dalam dua scatterplot: plot pertama menunjukkan data asli dengan pewarnaan berdasarkan kategori "Klasifikasi Kemiskinan", sedangkan plot kedua menunjukkan hasil klasterisasi DBSCAN dengan pewarnaan berdasarkan klaster yang dihasilkan. Visualisasi ini membantu menganalisis pola klaster yang teridentifikasi oleh DBSCAN.



Gambar. 13 Visualisasi DBSCAN

Perhitungan akurasi terhadap hasil klasterisasi menunjukkan nilai sebesar 0.50, yang mengindikasikan performa model kurang optimal dalam mengelompokkan data tersebut.

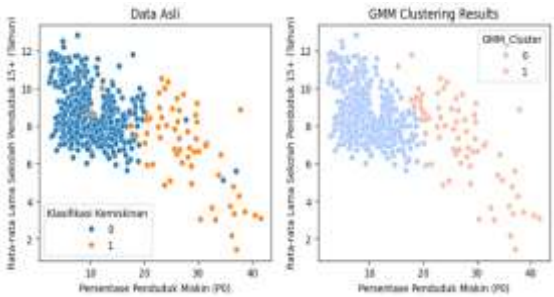
```
# Calculate accuracy
accuracy_dbscan = accuracy_score(df['Klasifikasi Kemiskinan'], df['DBSCAN_Cluster'])
print(f"Accuracy of DBSCAN: {accuracy_dbscan:.2f}")
```

```
Accuracy of DBSCAN: 0.50
```

Gambar. 14 Hasil Klasterisasi DBSCAN 50%

- Gaussian Mixture Models

Algoritma Gaussian Mixture Model (GMM) untuk melakukan klasterisasi pada data berdasarkan dua variabel, yaitu "Persentase Penduduk Miskin (P0)" dan "Rata-rata Lama Sekolah Penduduk 15+". GMM bekerja dengan memodelkan data sebagai campuran distribusi Gaussian, dengan jumlah klaster yang ditentukan (n_components=2). Hasil klasterisasi disimpan dalam kolom baru GMM_Cluster. Data kemudian divisualisasikan dalam dua scatterplot: plot pertama menampilkan data asli dengan pewarnaan berdasarkan kategori "Klasifikasi Kemiskinan", sedangkan plot kedua menunjukkan hasil klasterisasi GMM dengan pewarnaan berdasarkan klaster yang dihasilkan. Visualisasi ini memungkinkan perbandingan pola antara data asli dan hasil klasterisasi.



Gambar. 15 Visualisasi Gaussian

Setelah dilakukan perhitungan akurasi terhadap hasil klasterisasi, diperoleh nilai akurasi sebesar 0.95, menunjukkan bahwa algoritma GMM memiliki performa yang sangat baik dalam mengelompokkan data tersebut.

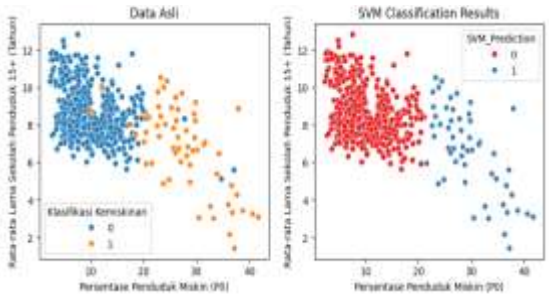
```
# Calculate accuracy
accuracy_gmm = accuracy_score(df['Klasifikasi Kemiskinan'], df['GMM_Cluster'])
print(f'Accuracy of GMM: {accuracy_gmm:.2f}')
```

```
Accuracy of GMM: 0.95
```

Gambar. 16 Hasil Klasterisasi GMM 95%

- Support Vector Machine

Algoritma Support Vector Machine (SVM) untuk melakukan klasifikasi data berdasarkan dua variabel, yaitu "Persentase Penduduk Miskin (P0)" dan "Rata-rata Lama Sekolah Penduduk 15+". Model SVM dilatih menggunakan kernel linear (`kernel='linear'`) dengan data pelatihan yang terdiri dari variabel independen `X_train` dan target `y_train`, yaitu kategori "Klasifikasi Kemiskinan". Hasil prediksi disimpan dalam kolom baru `SVM_Prediction`. Data kemudian divisualisasikan dalam dua scatterplot: plot pertama menampilkan data asli dengan pewarnaan berdasarkan kategori "Klasifikasi Kemiskinan", sedangkan plot kedua menunjukkan hasil klasifikasi oleh SVM dengan pewarnaan berdasarkan prediksi model. Visualisasi ini memungkinkan analisis efektivitas model dalam memisahkan data berdasarkan fitur yang digunakan.



Gambar. 17 Visualisasi SVM

Setelah dilakukan perhitungan akurasi, diperoleh nilai akurasi sebesar 0.97, menunjukkan bahwa model SVM mampu melakukan klasifikasi dengan performa yang sangat baik.

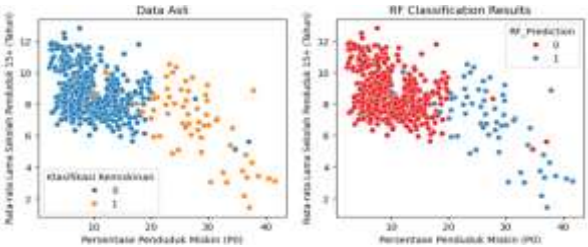
```
[ ] # Calculate accuracy
    accuracy_SVM = accuracy_score(df['Klasifikasi Kemiskinan'], df['SVM_Predic
    print(f'Accuracy of SVM: {accuracy_SVM:.2f}')

→ Accuracy of SVM: 0.97
```

Gambar. 18 Hasil Klasterisasi SVM 97%

- Random Forest

Algoritma Random Forest Classifier untuk melakukan klasifikasi data berdasarkan dua variabel, yaitu "Persentase Penduduk Miskin (P0)" dan "Rata-rata Lama Sekolah Penduduk 15+". Model ini dilatih dengan menggunakan 100 pohon keputusan (`n_estimators=100`) dan parameter `random_state=42` untuk menjaga konsistensi hasil. Data pelatihan terdiri dari variabel independen `X_train` dan target `y_train`, yaitu kategori "Klasifikasi Kemiskinan". Hasil prediksi disimpan dalam kolom baru `RF_Prediction`. Data kemudian divisualisasikan dalam dua scatterplot: plot pertama menunjukkan data asli dengan pewarnaan berdasarkan kategori "Klasifikasi Kemiskinan", sedangkan plot kedua menampilkan hasil klasifikasi Random Forest dengan pewarnaan berdasarkan prediksi model. Visualisasi ini memberikan gambaran tentang efektivitas model dalam memisahkan kategori berdasarkan fitur yang digunakan.



Gambar. 19 Visualisasi Random Forest

Setelah dilakukan perhitungan akurasi, diperoleh nilai akurasi sebesar 1.00, menunjukkan bahwa model Random Forest mampu melakukan klasifikasi dengan performa yang sempurna pada data tersebut.

```
[ ] # Calculate accuracy
    accuracy_RF = accuracy_score(df['Klasifikasi Kemiskinan'], df['RF_Prediction'])
    print(f'Accuracy of RF: {accuracy_RF:.2f}')

→ Accuracy of RF: 1.00
```

Gambar. 20 Hasil Klasterisasi Random Forest 100%

IV. KESIMPULAN

Dalam klasifikasi kemiskinan di Indonesia, algoritma Random Forest Classifier merupakan pilihan terbaik karena mampu memberikan hasil yang sempurna. SVM dan GMM juga menjadi alternatif yang sangat baik dengan akurasi tinggi. Pemilihan algoritma yang tepat bergantung pada karakteristik data dan kebutuhan analisis, tetapi algoritma berbasis pohon keputusan seperti Random Forest lebih disarankan untuk kasus klasifikasi kemiskinan yang melibatkan pola data kompleks.

[1] S. R. Dinata, M. Romus, and Yanti, "Faktor Faktor Yang Mempengaruhi Tingkat Kemiskinan Di Provinsi Riau Tahun 2003-2018," Al-Iqtishad, vol. 2, no. 16, pp. 116–137, 2020.

[2] S. R. Dinata, M. Romus, and Yanti, "Faktor Faktor Yang Mempengaruhi Tingkat Kemiskinan Di Provinsi Riau Tahun 2003-2018," Al-Iqtishad, vol. 2, no. 16, pp. 116–137, 2020.

[3] G. Dwilestari, Mulyawan, Martanto, and I. Ali, "Analisis Clustering menggunakan K-Medoid pada Data Penduduk Miskin Indonesia," JURSIMA J. Sist. Inf. dan Manaj., vol. 9, no. 3, pp. 282–290, 2021.

[4] D. V. Ferezagia, "Analisis Tingkat Kemiskinan di Indonesia," J. Sos. Hum. Terap., vol. 1, no. 1, pp. 1–6, 2018, doi: 10.7454/jsht.v1i1.6.

[5] N. Normah, S. Nurajizah, and A. Salbinda, "Penerapan Data MiningMetode K-Means Clustering Untuk Analisa Penjualan Pada Toko Fashion Hijab Banten," J. Tek. Komput. AMIK BSI, vol. 7, no. 2, pp. 158–163, 2021

[6] H. N. Putri and D. R. S. Saputro, "Clustering Data Campuran Numerik dan Kategorik Menggunakan Algoritme Ensemble Quick RObust Clustering using linKs (QROCK)," in PRISMA, Prosiding Seminar Nasional Matematika, 2022, vol. 5, pp. 716–720.

[7] R. Adha, N. Nurhaliza, U. Sholeha, and M. Mustakim, "Perbandingan Algoritma DBSCAN dan K-Means Clustering untuk Pengelompokan Kasus Covid-19 di Dunia," SITEKIN J. Sains, Teknol. dan Ind., vol. 18, no. 2, pp. 206–211, 2021.

[8] S. Regina, E. Sutinah, and N. Agustina, "Clustering Kualitas Kinerja Karyawan Pada Perusahaan Bahan Kimia Menggunakan Algoritma K-Means," J. MEDIA Inform. BUDIDARMA, vol. 5, no. 2, pp. 573–582, 2021

[9] DP Indini, SR Siburian, N Nurhasanah, DP Utomo, M Mesran ESCAF, 1328–1335-1328–1335 "Implementasi algoritma Dbscan untuk clustering seleksi penentuan mahasiswa yang berhak menerima yayasan"

[10] Betha Nurina Sari, Aji Primajaya, "Penerapan Clustering Dbscan untuk pertanian padi dikabupaten karawang" , vol 4 no 1 pp. 28-30, 2019.

[11] Nurkhaliza, Ayu A., and Arie W. Wijayanto. "Perbandingan Algoritma Klasifikasi Support Vector Machine dan Random Forest pada Prediksi Status Indeks Mitigasi dan Kesiapsiagaan Bencana (IMKB) Satuan Kerja BPS di Indonesia Tahun 2020." *Jurnal Informatika Universitas Pamulang*, vol. 7, no. 1, 2022, pp. 54-59, doi:10.32493/informatika.v7i1.16117.

[12] Ilham Kurniawan, Duwi Cahya Putri Buani, Abdussomad Abdussomad ,Widya Apriliah ,Rizal Amegia Saputra, "Implementasi Algoritma Random Forest Untuk Menentukan Penerima Bantuan Raskin" Vol 10 No. 2 : April 2023.