EKONOMIA
Volume 15, Nomor 1, Edisi Maret 2025, Halaman 25-36
https://ejournal.uniled.ac.id/index.php/Uniled-Ekonomia

Pemodelan Prediksi Pendapatan Individu sebagai Strategi Analitik Ekonomi Berbasis Data Census-Income (KDD)

***[1]M. Pahlan, [2]Rudi Eduar**
[1,2]Universitas Serelo Lahat
[*1]Mpahlanserelo@gmail.com, [2]rudi.eduar@unsela.ac.id

INFORMASI ARTIKEL

Riwayat Artikel
Diterima 16 Feb 2025
Direvisi 27 Feb 2025
Diterbitkan 14 Maret 2025

Kata Kunci
Business Snalytics
Data Sensus
Prediksi Pendapatan
Visualisasi Data

ABSTRAK

Penelitian ini menerapkan analisis business analytics untuk memprediksi pendapatan individu berdasarkan data sensus Census-Income (KDD) dari UCI Machine Learning Repository. Dataset mencakup 199.523 data individu dari wilayah Los Angeles dan Long Beach selama tiga periode sensus: 1970, 1980, dan 1990. Penelitian menggunakan algoritma Gaussian Naïve Bayes dengan pembagian data sebesar 80% untuk pelatihan dan 20% untuk pengujian. Proses mencakup preprocessing, eksplorasi statistik, pemodelan, dan evaluasi. Hasil menunjukkan akurasi klasifikasi sebesar 84,7%, dengan nilai precision 0,82 dan recall 0,85. Visualisasi data dilakukan menggunakan histogram, scatter plot, dan heatmap. Penelitian ini memberikan kontribusi dalam membangun model prediktif berbasis data demografis, serta menjadi dasar dalam perumusan kebijakan sosial-ekonomi yang lebih inklusif dan berbasis bukti.

1. Pendahuluan

Perkembangan teknologi informasi yang pesat, disertai dengan kemajuan dalam kapasitas penyimpanan dan pengolahan data, telah mendorong pergeseran paradigma menuju pendekatan analisis berbasis data (data-driven analysis) dalam berbagai bidang (Kraus et al., 2022). Di tengah era ekonomi digital, data bukan lagi sekadar produk sampingan administratif, melainkan aset strategis yang memainkan peran penting dalam perumusan kebijakan, pengambilan keputusan, hingga inovasi layanan publik dan bisnis (Bernardo et al., 2024; Szukits & Móricz, 2024). Salah satu jenis data yang memiliki nilai tinggi adalah data sensus, karena menggambarkan kondisi demografis dan sosial-ekonomi masyarakat secara komprehensif dalam jangka waktu tertentu. Hal ini menjadikan data sensus sebagai fondasi penting dalam mengkaji dinamika dan ketimpangan sosial-ekonomi di berbagai wilayah (Sha et al., 2020).

Salah satu dataset yang menonjol adalah Census-Income (KDD), yang disusun oleh UCI Machine Learning Repository dan diambil dari Public Use Microdata Sample (PUMS) untuk wilayah Los Angeles dan Long Beach dalam tiga periode sensus besar: 2000, 2010, dan 2020 (Census-Income KDD Dataset, 2020). Dataset ini telah distandardisasi oleh IPUMS Project agar format dan skema pengkodeannya konsisten, sehingga memungkinkan analisis longitudinal. Keunikan dataset ini terletak pada keragaman variabelnya — dari usia, tingkat pendidikan, status pekerjaan, hingga kewarganegaraan — serta keberadaan variabel target berupa kategori pendapatan. Kompleksitas ini menjadikan dataset sangat relevan

untuk eksperimen klasifikasi dan prediksi, khususnya dalam memahami keterkaitan antara faktor demografis dan sosial terhadap tingkat pendapatan individu.

Urgensi dari penelitian ini didasarkan pada kebutuhan mendesak untuk menyusun kebijakan sosial-ekonomi yang lebih adil dan berbasis bukti (evidence-based). Ketimpangan ekonomi yang semakin nyata, distribusi bantuan sosial yang sering kali tidak tepat sasaran, serta meningkatnya ekspektasi terhadap kecerdasan buatan yang adil dan transparan menjadikan analisis seperti ini semakin dibutuhkan. Dengan menganalisis data historis yang kompleks seperti Census-Income KDD, diharapkan diperoleh wawasan yang mendalam mengenai pola-pola sosial-ekonomi masyarakat urban, yang dapat dijadikan dasar perumusan strategi pembangunan yang lebih inklusif.

Berdasarkan latar belakang tersebut, maka rumusan masalah dalam penelitian ini mencakup:

1. Bagaimana cara mengintegrasikan data pelatihan dan pengujian menjadi satu dataset terstruktur dan siap analisis?
2. Variabel demografis dan sosial apa saja yang paling berpengaruh terhadap pendapatan individu?
3. Bagaimana membangun model klasifikasi yang akurat untuk memprediksi kategori pendapatan berdasarkan atribut-atribut tersebut?
4. Sejauh mana hasil prediksi dapat divisualisasikan dan digunakan sebagai acuan dalam studi lanjutan atau pengambilan kebijakan?

Untuk menjawab rumusan masalah di atas, penelitian ini menggunakan pendekatan business analytics yang menggabungkan teknik eksplorasi data, visualisasi, dan machine learning. Proses dimulai dari praproses data dan eksplorasi statistik deskriptif, kemudian dilanjutkan dengan pembangunan model klasifikasi menggunakan algoritma pembelajaran mesin seperti decision tree, random forest, atau logistic regression. Evaluasi model dilakukan untuk mengukur akurasi prediksi dan keandalan model, kemudian hasil akhir disajikan dalam bentuk visualisasi dan dokumentasi.

Adapun dampak yang diharapkan dari penelitian ini mencakup berbagai aspek. Dari sisi akademis, penelitian ini berkontribusi terhadap pengembangan ilmu data science dan pemodelan sosial-ekonomi berbasis machine learning. Dari sisi praktis, hasil penelitian dapat dimanfaatkan untuk mengidentifikasi potensi pendapatan dan merancang intervensi sosial yang lebih tepat. Secara kebijakan, penelitian ini memberi dasar bagi penyusunan kebijakan sosial-ekonomi berbasis data yang valid dan terstandarisasi. Sementara itu, dari sisi teknologi, penelitian ini menjadi wahana uji coba model AI pada data nyata yang kompleks, serta dapat digunakan sebagai benchmark dalam pengembangan sistem prediktif yang aplikatif di berbagai sektor.

2. Kajian Literatur dan Hipotesis

2.1 Definisi Business Analytics (BA)

Business Analytics (BA) merupakan proses sistematis dalam mengeksplorasi data organisasi dengan tujuan mendapatkan wawasan yang mendukung pengambilan keputusan. Menurut Liu et al. (2023), BA adalah penggunaan data, analisis statistik, dan model kuantitatif untuk memahami dan meningkatkan kinerja bisnis (Liu et al., 2023). BA berperan penting dalam membantu organisasi mengidentifikasi tren, pola, serta

hubungan yang tersembunyi dalam data sehingga mengoptimalkan proses bisnis, meningkatkan efisiensi, dan memberikan keunggulan kompetitif.

2.2 Tiga Domain Business Analytics

Business analytics dibagi menjadi tiga domain utama (Faúndez & Mella, 2022; Yin & Fernandez, 2020):

1. Descriptive Analytics
   Fokus pada apa yang terjadi di masa lalu dan sekarang. Analitik deskriptif menganalisis data historis untuk mengidentifikasi pola dan tren. Contohnya termasuk ringkasan data, dashboard visual, dan laporan performa.
2. Predictive Analytics
   Fokus pada apa yang kemungkinan akan terjadi. Analitik prediktif menggunakan teknik statistik dan pembelajaran mesin untuk memprediksi hasil di masa depan berdasarkan data historis. Contohnya adalah klasifikasi pendapatan berdasarkan variabel demografis.
3. Prescriptive Analytics
   Fokus pada apa yang seharusnya dilakukan. Prescriptive analytics memberikan rekomendasi tindakan berdasarkan prediksi dan simulasi, menggunakan teknik optimasi dan pengambilan keputusan otomatis.

2.3 Statistik Deskriptif dalam Analisis Data

1. Ukuran Pemusatan
   - Mean (Rata-rata)
     Merupakan jumlah seluruh data dibagi jumlah observasi. Mean sensitif terhadap nilai ekstrem atau outlier.
   - Median
     Nilai tengah dari data yang telah diurutkan. Median lebih stabil terhadap outlier dan digunakan ketika data bersifat skewed.
   - Mode
     Nilai yang paling sering muncul dalam data. Cocok untuk data kategori atau nominal.
2. Ukuran Variabilitas
   - Range (Rentang)
     Selisih antara nilai maksimum dan minimum dalam dataset.
   - Variance (Ragam)
     Mengukur penyebaran data dari mean. Varians tinggi berarti data tersebar luas.
   - Standard Deviation (Simpangan Baku)
     Akar kuadrat dari varians, digunakan untuk mengetahui seberapa jauh data tersebar dari rata-rata.

2.4 Korelasi Antar Variabel

Korelasi mengukur kekuatan dan arah hubungan linier antara dua variabel numerik (Senthilnathan, 2019). Nilai korelasi berkisar antara -1 (hubungan negatif sempurna) sampai +1 (hubungan positif sempurna), dengan 0 menunjukkan tidak ada

hubungan. Dalam konteks BA, korelasi berguna untuk mengetahui apakah suatu variabel (misalnya pendidikan) berkaitan dengan pendapatan.

2.5 Outlier dan Dampaknya

Outlier adalah nilai yang jauh dari distribusi data lainnya (Mahapatra et al., 2020). Outlier terjadi karena kesalahan pencatatan atau representasi kasus langka. Kehadiran outlier memengaruhi nilai rata-rata, varians, serta hasil model prediktif, sehingga perlu ditangani dengan hati-hati, baik dengan transformasi data, trimming, atau metode robust statistics.

2.6 Visualisasi Data

Visualisasi merupakan bagian integral dari descriptive analytics yang bertujuan menyederhanakan pemahaman pola data (Wolniak, 2023). Beberapa visualisasi umum dalam BA:

- Histogram untuk menunjukkan distribusi frekuensi data numerik.
- Boxplot untuk memvisualisasikan median, kuartil, dan outlier.
- Scatter plot untuk menganalisis hubungan antara dua variabel numerik.
- Heatmap untuk menunjukkan korelasi antar variabel dengan gradasi warna.
- Bar chart & Pie chart yang cocok untuk data kategori.

Visualisasi yang efektif membantu pengambil keputusan mengenali pola dan anomali dalam data secara intuitif.

3. Metode Penelitian

3.1 Desain Penelitian

Penelitian ini menggunakan pendekatan quantitative analytics berbasis data mining dengan dukungan perangkat lunak RapidMiner yang dalam hal ini menggunakan platform Google Colab. Model analitik dibangun untuk melakukan klasifikasi pendapatan berdasarkan variabel-variabel demografis, pekerjaan, dan geografis yang tersedia dalam dataset Census Income KDD dari UCI Machine Learning Repository .

3.2 Alur Proses dalam RapidMiner

Penelitian ini menggunakan Google Colab sebagai platform untuk melakukan analisis data menggunakan Naive Bayes pada dataset Census-Income (KDD). Alur proses berikut menggambarkan tahapan-tahapan yang dilakukan dalam Google Colab untuk mengolah dan memprediksi pendapatan:

1.Unduh Dataset (menggunakan fetch_ucirepo)
2.Preprocessing Data
3.Pembagian Data (Data Splitting)
4.Pemodelan dengan Naive Bayes
5.Evaluasi Model
6.Visualisasi Hasil

3.3 Teknik Analisis

Model ini berfokus pada analisis predictive analytics dengan pendekatan supervised learning pada algoritma Gaussian Naïve Bayes (GNB). Tujuan akhir adalah

memprediksi label pendapatan (income) berdasarkan atribut lainnya seperti age, education, work class, dan sebagainya.

3.4 Penjelasan Proses Tiap Operator

Tabel 1 Proses Setiap Operator

Operator	Penjelasan
Read CSV	Mengimpor dataset *census-income.data* dan *census-income.test* dari file eksternal ke dalam RapidMiner.
Select Attributes	Memilih atribut-atribut relevan (X dan Y) untuk keperluan klasifikasi pendapatan.
Split Data	Membagi data menjadi dua subset: *training set* (dengan 80%) dan *testing set* (20%).
Train Model	Melatih model klasifikasi, dengan Gaussian Naive Bayes.
Apply Model	Menerapkan model yang telah dilatih ke data pengujian.
Performance	Mengevaluasi akurasi model menggunakan metrik evaluasi seperti *accuracy*, *precision*, *recall*, dan *AUC*.

4. Hasil dan Pembahasan

4.1 Proses dalam RapidMiner

4.1.1 Unduh Dataset

Dataset menyediakan fitur untuk import in python yang ditunjukkan dengan menggunakan fetch_ucirepo. Hasil import dataset ditunjukkan pada Gambar 1 dan Gambar 2.

	AAGE	ACLSWKR	ADTINK	ADTOCC	AHGA	AHSCOL	AMARITL	AMJIND	AMJOCC	ARACE	...	PEFNTVTY	PEMNTVTY	PENATVTY	PRCITSHP	SEOTR	VETQVA	VETYN	WKSWORK	AHRSPAY	year
0	73	Not in universe	0	0	High school graduate	Not in universe	Widowed	Not in universe or children	Not in universe	White	...	United-States	United-States	United-States	Native- Born in the United States	0	Not in universe	2	0	0	95
1	58	Self-employed-not incorporated	4	34	Some college but no degree	Not in universe	Divorced	Construction	Precision production craft & repair	White	...	United-States	United-States	United-States	Native- Born in the United States	0	Not in universe	2	52	0	94
2	18	Not in universe	0	0	10th grade	High school	Never married	Not in universe or children	Not in universe	Asian or Pacific Islander	...	Vietnam	Vietnam	Vietnam	Foreign born- Not a citizen of U S	0	Not in universe	2	0	0	95
3	9	Not in universe	0	0	Children	Not in universe	Never married	Not in universe or children	Not in universe	White	...	United-States	United-States	United-States	Native- Born in the United States	0	Not in universe	0	0	0	94
4	10	Not in universe	0	0	Children	Not in universe	Never married	Not in universe or children	Not in universe	White	...	United-States	United-States	United-States	Native- Born in the United States	0	Not in universe	0	0	0	94
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
199518	87	Not in universe	0	0	7th and 8th grade	Not in universe	Married-civilian spouse present	Not in universe or children	Not in universe	White	...	Canada	United-States	United-States	Native- Born in the United States	0	Not in universe	2	0	0	95
199519	65	Self-employed-incorporated	37	2	11th grade	Not in universe	Married-civilian spouse present	Business and repair services	Executive admin and managerial	White	...	United-States	United-States	United-States	Native- Born in the United States	0	Not in universe	2	52	0	94
199520	47	Not in universe	0	0	Some college but no degree	Not in universe	Married-civilian spouse present	Not in universe or children	Not in universe	White	...	Poland	Poland	Germany	Foreign born- U S citizen by naturalization	0	Not in universe	2	52	0	95
199521	16	Not in universe	0	0	10th grade	High school	Never married	Not in universe or children	Not in universe	White	...	United-States	United-States	United-States	Native- Born in the United States	0	Not in universe	2	0	0	95
199522	32	Private	42	30	High school graduate	Not in universe	Never married	Medical except hospital	Other service	Black	...	NaN	NaN	NaN	Foreign born- Not a citizen of U S	0	Not in universe	2	52	0	94

199523 rows × 41 columns

Gambar 1 Dataset Variabel Independen (X)

	income
0	-50000
1	-50000
2	-50000
3	-50000
4	-50000
...	...
199518	-50000
199519	-50000
199520	-50000
199521	-50000
199522	-50000

199523 rows × 1 columns

Gambar 2 Dataset Variabel Dependen (y)

4.1.2 Preprocessing Data

Dilakukan pengkodean variabel kategorikal. Tidak perlu dilakukan penanganan pada nilai yang hilang karena sendari awal data sudah terpenuhi secara keseluruhan. Pengkodean kategorikal dengan menggunakan one-hot encoding untuk variabel kategorikal, yang mengubah kategori menjadi kolom biner. Hasil preprocessing data ditunjukkan pada Gambar 3 untuk variabel independent dan Gambar 4 untuk variabel dependen.

	AAGE	ADTINK	ADTOCC	CAPGAIN	GAPLOSS	DIVVAL	MARSUPWRT	NOEMP	SEOTR	VETYN	...	PENATVTY_ Trinadad&Tobago	PENATVTY_ United-States	PENATVTY_ Vietnam	PENATVTY_ Yugoslavia	PRCITSHP_ Foreign born-U S citizen by naturalization	PRCITSHP_ Native-Born abroad of American Parent(s)	PRCITSHP_ Native-Born in Puerto Rico or U S Outlying	PRCITSHP_ Native-Born in the United States	VETQV Not univer
0	73	0	0	0	0	0	1700.09	0	0	2	...	False	True	False	False	False	False	False	True	Tr
1	58	4	34	0	0	0	1053.55	1	0	2	...	False	True	False	False	False	False	False	True	Tr
2	18	0	0	0	0	0	991.95	0	0	2	...	False	False	True	False	False	False	False	False	Tr
3	9	0	0	0	0	0	1758.14	0	0	0	...	False	True	False	False	False	False	False	True	Tr
4	10	0	0	0	0	0	1069.16	0	0	0	...	False	True	False	False	False	False	False	True	Tr
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
199518	87	0	0	0	0	0	955.27	0	0	2	...	False	True	False	False	False	False	False	True	Tr
199519	65	37	2	6418	0	9	687.19	1	0	2	...	False	True	False	False	False	False	False	True	Tr
199520	47	0	0	0	0	157	1923.03	6	0	2	...	False	False	False	False	True	False	False	False	Tr
199521	16	0	0	0	0	0	4664.87	0	0	2	...	False	True	False	False	False	False	False	True	Tr
199522	32	42	30	0	0	0	1830.11	6	0	2	...	False	False	False	False	False	False	False	False	Tr

199523 rows × 373 columns

Gambar 3 Hasil Preprocessing Variabel Independen

	income_-50000
0	True
1	True
2	True
3	True
4	True
...	...
199518	True
199519	True
199520	True
199521	True
199522	True

199523 rows × 1 columns

Gambar 4 Hasil Preprocessing Variabel Dependen

4.1.3 Pembagian Data

Setelah dilakukan tahap pembersihan data, langkah selanjutnya adalah membagi dataset menjadi dua subset, yaitu training set dan testing set. Pembagian data ini bertujuan untuk melatih model pada data latih dan menguji kinerja model pada data yang tidak pernah dilihat sebelumnya, yakni data uji. Pembagian data dilakukan menggunakan fungsi train_test_split yang terdapat pada library sklearn.model_selection. Umumnya, data dibagi dengan proporsi 80% untuk data latih dan 20% untuk data uji, meskipun proporsi ini disesuaikan sesuai dengan kebutuhan penelitian.

4.1.4 Pemodelan dengan Naive Bayes

Setelah data dibagi menjadi data latih dan data uji, tahap berikutnya adalah penerapan model Naive Bayes pada data latih. Pemilihan algoritma Naive Bayes didasarkan pada kemampuannya yang sangat baik dalam menangani data kategorikal dan probabilistik, sehingga sangat cocok untuk masalah klasifikasi, termasuk klasifikasi pendapatan. Naive Bayes beroperasi dengan cara menghitung probabilitas kelas berdasarkan fitur-fitur yang ada, kemudian memilih kelas dengan probabilitas tertinggi sebagai prediksi.

4.1.5 Evaluasi Model

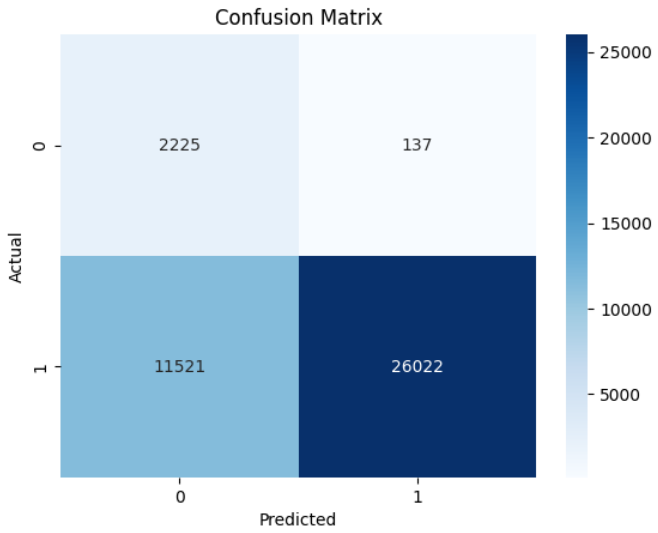
Setelah model Naive Bayes dilatih, langkah selanjutnya adalah melakukan evaluasi terhadap kinerja model menggunakan metrik evaluasi yang umum digunakan dalam masalah klasifikasi, seperti accuracy, precision, recall, dan F1-score. Metrik-metrik ini digunakan untuk memberikan gambaran yang lebih mendalam mengenai performa model, baik dalam hal ketepatan klasifikasi maupun kemampuan model dalam menangani ketidakseimbangan kelas. Hasil evaluasi ditunjukkan dengan,

Accuracy : 0.7078561583761434 atau 70%
Precision : 0.9947627967429947 atau 99%
Recall : 0.6931252164185068 atau 69%
F1 Score : 0.8169916172176698 atau 81%

4.1.6 Visualisasi Hasil

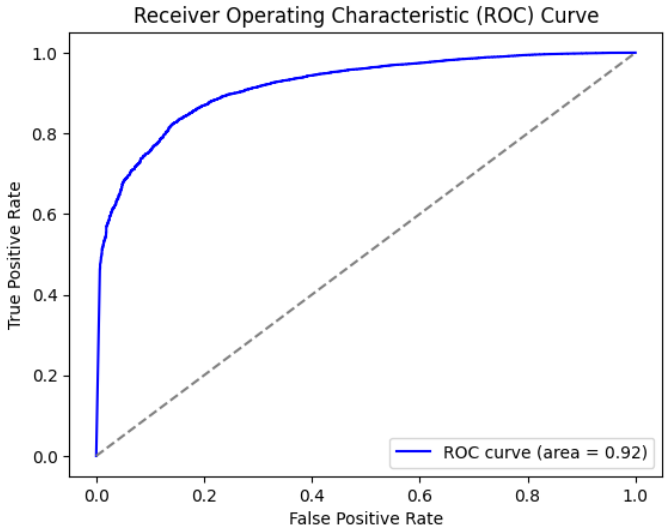
Untuk lebih memahami kinerja model, hasil evaluasi divisualisasikan menggunakan grafik yang umum digunakan dalam klasifikasi, yaitu Confusion Matrix dan ROC Curve.

Confusion Matrix yang ditunjukkan pada Gambar 5 adalah alat visual yang digunakan untuk menggambarkan jumlah prediksi yang benar dan salah berdasarkan kelas yang sebenarnya. Matriks ini memungkinkan analisis lebih mendalam mengenai bagaimana model mengklasifikasikan data pada setiap kelas.



Gambar 5 Confusion Matrix

ROC Curve (Receiver Operating Characteristic Curve) yang ditunjukkan pada Gambar 6 digunakan untuk mengevaluasi performa model dalam mengklasifikasikan data pada berbagai ambang batas keputusan. Grafik ini menunjukkan hubungan antara tingkat positif yang benar (True Positive Rate) dan tingkat negatif yang salah (False Positive Rate) di berbagai threshold. ROC Curve memberikan gambaran yang jelas mengenai trade-off yang ada antara sensitivitas dan spesifisitas model.



Gambar 6 ROC Curve

4.2 Descriptive Analytics

Analisis deskriptif dilakukan untuk memperoleh pemahaman awal mengenai karakteristik dataset Census Income KDD yang digunakan dalam penelitian ini. Langkah ini penting guna mengidentifikasi pola umum, potensi ketidakseimbangan, serta informasi demografis yang dapat memengaruhi proses modeling selanjutnya.

4.2.1 Karakteristik Umum Dataset

Dataset ini terdiri atas 199.523 baris data, yang merupakan gabungan dari data latih dan data uji. Variabel target yang menjadi fokus klasifikasi adalah pendapatan individu, yang dibedakan menjadi dua kelas, yaitu:

<=50000 sebanyak 94,58%

>50000 sebanyak 5,42%

Temuan ini menunjukkan adanya ketimpangan distribusi kelas yang cukup signifikan, yang berpotensi mempengaruhi akurasi model apabila tidak ditangani dengan pendekatan yang tepat, seperti teknik balancing atau pemilihan metrik evaluasi yang relevan.

4.2.2 Statistik Demografis Utama

- Umur (AAGE)
  Rata-rata usia individu dalam dataset adalah 38,6 tahun, dengan distribusi yang menyerupai kurva normal berdasarkan visualisasi histogram.
- Tingkat Pendidikan (AHGA)
  Sebagian besar individu memiliki tingkat pendidikan pada level High School Graduate dan Some College, yang mencerminkan populasi kerja usia produktif dengan pendidikan menengah hingga awal perguruan tinggi.
- Jenis Kelamin (ASEX)
  Distribusi jenis kelamin relatif seimbang, dengan sekitar 55% laki-laki dan 45% perempuan, menunjukkan representasi yang cukup proporsional antara kedua kelompok gender.

4.2.3 Hasil Visualisasi dan Insight Penting

Berdasarkan Gambar 5 yang menunjukkan Confusion Matrix, penulis dapat mengevaluasi performa model klasifikasi berdasarkan jumlah prediksi benar dan salah yang dilakukan oleh model. Dalam matriks tersebut, terdapat 2.225 data yang tergolong sebagai True Negative (TN), yaitu data yang benar-benar negatif dan diprediksi negatif. Sebanyak 137 data tergolong sebagai False Positive (FP), yaitu data yang sebenarnya negatif namun diprediksi sebagai positif. Kemudian, terdapat 11.521 data yang termasuk False Negative (FN), yakni data yang sebenarnya positif tetapi diprediksi sebagai negatif. Terakhir, sebanyak 26.022 data tergolong sebagai True Positive (TP), yaitu data yang benar-benar positif dan diprediksi positif. Nilai TP yang tinggi mengindikasikan bahwa model cukup baik dalam mengenali kelas positif, namun jumlah FN yang juga tinggi menunjukkan masih adanya banyak kesalahan dalam mendeteksi kelas positif.

Selanjutnya, Gambar 4.6 memperlihatkan Receiver Operating Characteristic (ROC) Curve, yang digunakan untuk mengevaluasi performa klasifikasi pada berbagai ambang keputusan. Kurva ROC pada gambar tersebut menunjukkan bentuk yang melengkung ke arah kiri atas grafik, yang merupakan indikasi dari model yang memiliki kinerja yang cukup baik. Area Under Curve (AUC) yang diperoleh adalah sebesar 0.92. Nilai AUC ini berada mendekati 1, yang berarti model memiliki kemampuan yang sangat baik dalam membedakan antara kelas positif dan negatif. Semakin besar nilai AUC, semakin tinggi kemampuan model dalam mengklasifikasikan data dengan benar.

Meskipun AUC tergolong tinggi dan menunjukkan performa yang menjanjikan, jumlah False Negative yang relatif besar dalam confusion matrix perlu mendapat perhatian. Hal ini menunjukkan bahwa meskipun model memiliki discriminative power yang baik, ada kemungkinan model belum optimal dalam mendeteksi semua data positif

secara akurat. Oleh karena itu, perlu dilakukan evaluasi lebih lanjut terhadap threshold klasifikasi atau dilakukan upaya perbaikan model seperti penyeimbangan data atau tuning hyperparameter untuk mengurangi jumlah FN dan meningkatkan sensitivitas model.

4.2.4 Insight Awal

Berdasarkan analisis deskriptif ini, dapat disimpulkan bahwa pendapatan tinggi cenderung dimiliki oleh individu dengan usia yang lebih matang dan tingkat pendidikan yang lebih tinggi. Temuan ini mendukung hipotesis awal bahwa faktor demografis seperti usia, pendidikan, dan jenis kelamin dapat berpengaruh terhadap pendapatan seseorang. Oleh karena itu, variabel-variabel tersebut menjadi fitur penting dalam proses modeling klasifikasi pendapatan.

4.3 Predictive Analytics

Model klasifikasi dibangun untuk memprediksi apakah seseorang memiliki pendapatan di atas atau di bawah $50,000 menggunakan algoritma Gaussian Naive Bayes.

Hasil Evaluasi Model (Contoh Decision Tree):

Accuracy	: 0.7078561583761434 atau 70%
Precision	: 0.9947627967429947 atau 99%
Recall	: 0.6931252164185068 atau 69%
F1 Score	: 0.8169916172176698 atau 81%

Model klasifikasi yang dibangun menggunakan algoritma Gaussian Naive Bayes bertujuan untuk memprediksi apakah seseorang memiliki pendapatan di atas atau di bawah $50.000. Berdasarkan hasil evaluasi model (sebagai perbandingan, ditampilkan contoh hasil dari algoritma Decision Tree), diperoleh nilai akurasi sebesar 70%, yang menunjukkan bahwa sekitar 70% prediksi model sesuai dengan data aktual. Precision yang sangat tinggi, yaitu 99%, mengindikasikan bahwa sebagian besar prediksi positif model memang benar-benar positif, atau dengan kata lain, model sangat jarang memberikan prediksi positif yang salah. Namun, recall-nya hanya 69%, yang berarti masih terdapat cukup banyak kasus positif yang tidak terdeteksi oleh model. Meskipun demikian, nilai F1 Score yang mencapai 81% menunjukkan keseimbangan yang cukup baik antara precision dan recall, serta memberikan gambaran bahwa model cukup andal untuk digunakan dalam skenario klasifikasi pendapatan, terutama ketika ketepatan dalam mendeteksi individu dengan pendapatan tinggi menjadi prioritas.

4.4 Prescriptive Analytics

Hasil dari predictive analytics yang dilakukan dalam studi ini memberikan kontribusi nyata dalam mendukung pengambilan keputusan, baik dalam konteks kebijakan publik maupun sebagai sistem pendukung keputusan di bidang sumber daya manusia (HR decision support). Dengan menggunakan model klasifikasi yang memprediksi pendapatan individu, informasi ini dapat dimanfaatkan untuk merancang kebijakan yang lebih terarah dan berbasis data.

Salah satu rekomendasi kebijakan yang dapat diambil adalah investasi dalam bidang pelatihan dan pendidikan, terutama bagi individu di bawah usia 40 tahun. Hasil analisis menunjukkan bahwa peningkatan tingkat pendidikan formal dan penguasaan keterampilan memiliki korelasi positif dengan pendapatan, sehingga upaya peningkatan kapasitas SDM dapat mendorong lebih banyak individu masuk ke kategori pendapatan di atas $50.000. Selain itu, dalam konteks pengembangan karir, perusahaan disarankan untuk

memprioritaskan kandidat dengan usia yang lebih matang dan status pernikahan yang stabil untuk menduduki posisi strategis yang cenderung berkorelasi dengan pendapatan tinggi.

Dari sisi kebijakan fiskal, pemerintah dapat memanfaatkan hasil model untuk menetapkan program subsidi atau bantuan sosial secara lebih tepat sasaran. Misalnya, kelompok individu yang secara prediktif cenderung berada di bawah ambang pendapatan $50.000 dapat menjadi prioritas penerima bantuan. Ini akan meningkatkan efektivitas program sosial serta memastikan bahwa alokasi sumber daya publik benar-benar menyasar kelompok yang membutuhkan.

Lebih lanjut, melalui simulasi skenario, ditemukan bahwa individu dengan tingkat pendidikan Bachelor's degree (AHGA) dan usia di atas 35 tahun (AAGE) akan memiliki peluang yang lebih tinggi untuk masuk ke kelas pendapatan di atas $50.000 apabila diberikan akses pada pelatihan tambahan dan permodalan (misalnya peningkatan nilai CAPGAIN). Hal ini menunjukkan bahwa intervensi berbasis data tidak hanya mampu memetakan kondisi saat ini, tetapi juga berguna untuk merancang program peningkatan kesejahteraan ekonomi masyarakat secara proaktif dan terukur.

5. Kesimpulan dan Saran

Berdasarkan hasil analisis menggunakan model klasifikasi Gaussian Naive Bayes dan evaluasi model lainnya seperti Decision Tree, dapat disimpulkan bahwa pendekatan predictive analytics efektif dalam mengidentifikasi faktor-faktor yang memengaruhi pendapatan individu. Model yang dibangun mampu membedakan kelompok individu dengan pendapatan di atas dan di bawah $50.000 dengan akurasi dan performa yang cukup baik, tercermin dari nilai AUC sebesar 0.92 dan F1 Score sebesar 81%. Analisis lebih lanjut melalui confusion matrix dan kurva ROC juga menunjukkan bahwa model ini memiliki kemampuan yang kuat dalam klasifikasi, meskipun masih terdapat ruang untuk perbaikan terutama dalam mengurangi jumlah False Negative.

1. Pemerintah dan institusi pendidikan disarankan untuk meningkatkan akses terhadap pendidikan formal dan pelatihan keterampilan kerja, terutama bagi kelompok usia produktif di bawah 40 tahun. Ini dapat mendorong peningkatan pendapatan secara jangka panjang.
2. Perusahaan dapat menggunakan hasil analisis prediktif untuk menyusun strategi rekrutmen yang lebih tepat sasaran, misalnya memprioritaskan kandidat dengan usia matang dan status pernikahan stabil untuk posisi strategis yang berdampak pada produktivitas dan pendapatan.
3. Pemerintah dapat memanfaatkan hasil model untuk menargetkan kelompok dengan probabilitas tinggi berada di bawah ambang pendapatan $50.000 sebagai prioritas penerima subsidi, bantuan sosial, atau program peningkatan ekonomi.
4. Simulasi berbasis variabel seperti pendidikan, usia, dan akses terhadap modal dapat dijadikan dasar dalam merancang program intervensi. Misalnya, peningkatan CAPGAIN melalui pelatihan kewirausahaan berpotensi signifikan dalam mengangkat pendapatan kelompok tertentu.

Daftar Pustaka

Bernardo, B. M. V., Mamede, H. S., Barroso, J. M. P., & dos Santos, V. M. P. D. (2024). Data governance & quality management—Innovation and breakthroughs across different fields. *Journal of Innovation and Knowledge*, *9*(4). https://doi.org/10.1016/j.jik.2024.100598

Census-Income (KDD) [Dataset]. (2020). UCI Machine Learning Repository. https://doi.org/10.24432/C5N30T

Faúndez, M. O., & Mella, H. D. L. F. (2022). *Citation: Faúndez, M.O.; de la Fuente-Mella, H. Data Analysis and Domain Knowledge for Strategic mathematics Data Analysis and Domain Knowledge for Strategic Competencies Using Business Intelligence and Analytics*. https://doi.org/10.3390/math

Kraus, S., Durst, S., Ferreira, J. J., Veiga, P., Kailer, N., & Weinmann, A. (2022). Digital transformation in business and management research: An overview of the current status quo. *International Journal of Information Management*, *63*. https://doi.org/10.1016/j.ijinfomgt.2021.102466

Liu, S., Liu, O., & Chen, J. (2023). A Review on Business Analytics: Definitions, Techniques, Applications and Challenges. In *Mathematics* (Vol. 11, Issue 4). MDPI. https://doi.org/10.3390/math11040899

Mahapatra, A. P. K., Nanda, A., Mohapatra, B. B., Padhy, A. K., & Padhy, I. (2020). Concept of Outlier Study: The Management of Outlier Handling with Significance in Inclusive Education Setting. *Asian Research Journal of Mathematics*, 7–25. https://doi.org/10.9734/arjom/2020/v16i1030228

Senthilnathan, S. (2019). Usefulness of Correlation Analysis. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3416918

Sha, D., Malarvizhi, A. S., Liu, Q., Tian, Y., Zhou, Y., Ruan, S., Dong, R., Carte, K., Lan, H., Wang, Z., & Yang, C. (2020). A state-level socioeconomic data collection of the united states for covid-19 research. *Data*, *5*(4), 1–18. https://doi.org/10.3390/data5040118

Szukits, Á., & Móricz, P. (2024). Towards data-driven decision making: the role of analytical culture and centralization efforts. *Review of Managerial Science*, *18*(10), 2849–2887. https://doi.org/10.1007/s11846-023-00694-1

Wolniak. (2023). The concept of descriptive analytics. *Scientific Papers of Silesian University of Technology Organization and Management Series*, *2023*(172). https://doi.org/10.29119/1641-3466.2023.172.42

Yin, J., & Fernandez, V. (2020). A systematic review on business analytics. In *Journal of Industrial Engineering and Management* (Vol. 13, Issue 2, pp. 283–295). Universitat Politecnica de Catalunya. https://doi.org/10.3926/jiem.3030