BAB IV

HASIL DAN PEMBAHASAN

4.1 Hasil Penelitian

Pada penelitian ini dataset yang digunakan diperoleh dari *repository* Kaggle. Data *depression* diproses dengan membuang data yang tidak konsisten, kemudian di kelompokkan lalu diprediksi. Penelitian ini menggunakan jumlah data sebanyak 1.429 *record* data. Data tersebut memuat 23 atribut yaitu *survey_id* (responden), *ville_id* (identifier unik), *sex* (jenis kelamin), *age* (usia), *married* (status pernikahan), *number_children* (jumlah anak), *education_level* (tingkat pendidikan), *total_members* (jumlah anggota), *gained_asset* (aset yang diperoleh), *durable_asset* (aset tahan lama), *save_asset* (aset simpanan), *living_expenses* (biaya hidup), *other_expenses* (biaya lain), *incoming_salary* (gaji masuk), *incoming_own_fam* (pendapatan keluarga), *incoming_business* (pendapatan usaha), *incoming_agriculture* (pendapatan pertanian), *fam_expenses* (biaya keluarga), *labor_primary* (tenaga kerja utama), *lasting_investme* (investasi jangka panjang), *deressed* (depresi). Berdasarkan metodologi yang telah ditentukan untuk memprediksi tingkat depresi maka akan menggunakan algoritma KNN (K-Nearest Neighbor). Berdasarkan metodologi yang telah dipilih untuk memprediksi dan mengevaluasi dataset *depression* menggunakan algoritma KNN (K-Nearest Neighbor) maka hasil yang didapatkan adalah sebagai berikut:

4.1.1 Exploratory Data Analysis

Pada penelitian ini dataset yang digunakan diperoleh dari *repository* Kaggle. Data *depression* diproses dengan membuang data yang tidak konsisten, kemudian di kelompokkan lalu diprediksi. Penelitian ini menggunakan jumlah data sebanyak 1.429 *record* data. Data tersebut memuat 23 atribut. Berikut data yang diperoleh lalu disimpan dalam format csv.

Survey_id	Ville_id	sex	Age	Married	Number_children	education_level	total_members	gained_asset	durable_asset
926	91	1	28	1	4	10	5	28912201	22861940
747	57	1	23	1	3	8	5	28912201	22861940
1190	115	1	22	1	3	9	5	28912201	22861940
1065	97	1	27	1	2	10	4	52667108	19698904
806	42	0	59	0	4	10	6	82606287	17352654
483	25	1	35	1	6	10	8	35937466	736707
849	130	0	34	0	1	9	3	41303144	21925041
1386	72	1	21	1	2	10	4	12013633	20323505
930	195	1	32	1	7	9	9	11087568	25224208
390	33	1	29	1	4	10	5	28912201	22861940
540	52	1	84	0	0	1	5	28912201	22861940
557	93	1	59	0	2	9	3	1018915	47245342
1280	232	1	38	1	4	10	6	12390944	19186414
1195	92	1	27	1	4	10	6	16521259	37155658
603	100	1	56	1	0	12	2	93596368	21140288
729	54	1	24	1	2	10	5	1108353	12219727
770	102	1	25	1	3	10	5	37172832	75432396
76	15	1	44	1	5	12	5	28912201	22861940
1374	267	1	32	1	4	9	5	28912201	22861940
379	22	1	26	1	2	7	4	82606287	20419597

Gambar 4. 1 *Sample Data*

a. *Duplicate Data*

Duplicate data adalah data yang sama yang ditemukan lebih dari satu kali dalam dataset atau sistem penyimpanan data. Hasil pengecekan pada setiap variabel adalah:

Survey_id	0
Ville_id	0
sex	0
Age	0
Married	0
Number_children	0
education_level	0
total_members	0
gained_asset	0
durable_asset	0
save_asset	0
living_expenses	0
other_expenses	0

Gambar 4. 2 Hasil *Duplicate Data*

Setelah dilakukan pengecekan *duplicate data,* ternyata tidak ada duplikat data yang terdeteksi.

b. *Missing Value*

Missing value adalah kondisi di mana informasi dalam dataset tidak lengkap atau tidak ada. Data yang hilang ini dapat berupa angka, kategori, atau jenis

data lainnya. Dilakukannya untuk menghapus data yang tidak konsisten, atau menghapus atribut yang tidak diperlukan. Hasil pengecekan pada setiap variabel adalah:

durable_asset	0
save_asset	0
living_expenses	0
other_expenses	0
incoming_salary	0
incoming_own_farm	0
incoming_business	0
incoming_no_business	0
incoming_agricultural	0
farm_expenses	0
labor_primary	0
lasting_investment	0
no_lasting_investmen	20
depressed	0

Gambar 4. 3 Hasil *Missing Value*

Setelah dilakukan pengecekan *missing value,* ada 20 *missing value* yang terdeteksi sehingga jumlah dari 1.429 data menjadi 1.409 data.

c. Outlier

Outlier adalah data yang nilainya sangat berbeda jauh dari nilai-nilai observasi lainnya dalam suatu set data. Sederhananya, outlier merupakan nilai ekstrem atau anomali dalam data. Dalam analisis data menggunakan algoritma KNN (K-Nearest Neighbor), keberadaan outlier dapat memberikan dampak yang signifikan terhadap hasil prediksi. Outlier, sebagai titik data yang secara signifikan berbeda dari sebagian besar data lainnya, dapat memengaruhi perhitungan jarak dan identifikasi tetangga terdekat. Ketika sebuah titik query berada dekat dengan outlier dalam ruang fitur, outlier tersebut berpotensi menjadi salah satu dari K tetangga terdekat, meskipun karakteristiknya tidak representatif dari kelompok data yang sebenarnya. Hal ini dapat menyebabkan misklasifikasi atau prediksi nilai yang tidak akurat untuk titik query tersebut. Selain itu, outlier dalam data pelatihan juga dapat mendistorsi batas keputusan dalam klasifikasi atau nilai rata-rata dalam regresi, terutama ketika nilai K yang dipilih kecil, karena

pengaruh setiap tetangga menjadi lebih besar. Oleh karena itu, penting untuk mengidentifikasi dan mempertimbangkan penanganan outlier dalam tahap pra-pemrosesan data sebelum menerapkan KNN, seperti melalui teknik visualisasi, metode statistik, atau transformasi data, untuk memitigasi potensi dampak negatifnya terhadap kinerja model.

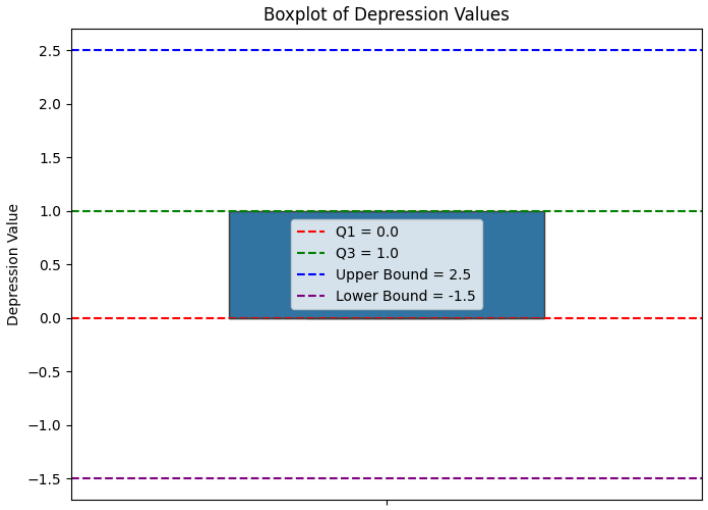
```
"Kolom sex memiliki 114 outlier\n",
"Kolom Age memiliki 46 outlier\n",
"Kolom Married memiliki 317 outlier\n",
"Kolom Number_children memiliki 22 outlier\n",
"Kolom education_level memiliki 208 outlier\n",
"Kolom total_members memiliki 31 outlier\n",
"Kolom gained_asset memiliki 170 outlier\n",
"Kolom durable_asset memiliki 275 outlier\n",
"Kolom save_asset memiliki 464 outlier\n",
"Kolom living_expenses memiliki 120 outlier\n",
"Kolom other_expenses memiliki 113 outlier\n",
"Kolom incoming_salary memiliki 248 outlier\n",
"Kolom incoming_own_farm memiliki 0 outlier\n",
"Kolom incoming_business memiliki 154 outlier\n",
"Kolom incoming_no_business memiliki 0 outlier\n",
"Kolom incoming_agricultural memiliki 107 outlier\n",
"Kolom farm_expenses memiliki 78 outlier\n",
"Kolom labor_primary memiliki 295 outlier\n",
"Kolom lasting_investment memiliki 109 outlier\n",
"Kolom no_lasting_investmen memiliki 97 outlier\n",
"Kolom depressed memiliki 234 outlier\n"
```

Gambar 4. 4 Hasil *Outlier*

Semua nama kolom tersedia dari data frame, kecuali kolom *incoming_own_farm* dan *incoming_no_business* yang tidak dimuat. Alasan kedua kolom ini tidak diikutsertakan karena terlalu mirip dengan kolom lain dan menyederhanakan analisis. Intinya, selain dua kolom yang disebutkan bertujuan untuk analisis serta pemodelan data lebih lanjut. Maka dilakukan capping outlier yaitu membatasi nilai ekstrem (outlier) dalam sebuah dataset tanpa menghapusnya secara keseluruhan. Sebelum penanganan outlier melalui metode capping, terdapat 1.409 entri data dengan 23 atribut. Namun, setelah dilakukan proses capping terdapat 1.364 data dan memiliki 21 atribut yang berbeda, seperti usia, jenis kelamin, tingkat pendidikan, status pernikahan, jumlah anak, pendapatan, dan lain – lain. Output dari capping outlier ini telah menghapus 45 baris dari dataset yang ada karena data tersebut dianggap sebagai outlier. Jumlah kolom tidak berubah karena tidak menghapus kolom, tetapi hanya data individu yang dianggap ekstrem. Output ini membantu akan memahami ukuran dataset sebelum dan sesudah menangani outlier.

d. Boxplot

Boxplot atau diagram kotak dan garis, merupakan bentuk visualisasi data yang digunakan untuk menggambarkan distribusi suatu kumpulan data berdasarkan lima ukuran statistik utama.

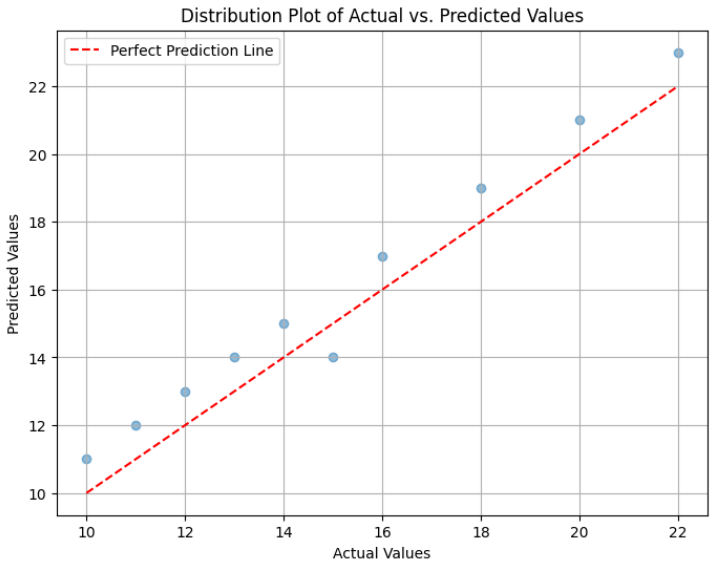


Gambar 4. 5 Boxplot

Kotak pada grafik ini, meskipun tampak terisi penuh dan berada di antara nilai 0 dan 1 pada sumbu vertikal, sebenarnya adalah representasi visual dari distribusi nilai depresi. Garis putus-putus berwarna merah menandai kuartil pertama (Q1) pada nilai 0, garis putus-putus hijau menandai kuartil ketiga (Q3) pada nilai 1. Karena Q1 dan Q3 memiliki nilai yang berbeda, kotak yang terbentuk seharusnya memiliki tinggi. Namun, visualisasi ini mungkin terdistorsi atau mewakili kasus khusus di mana sebagian besar data terpusat di antara nilai 0 dan 1. Garis putus-putus biru di atas (2.5) dan ungu di bawah (-1.5) menunjukkan batas atas dan bawah yang mungkin digunakan untuk mengidentifikasi outlier, meskipun tidak ada whisker atau outlier yang terlihat jelas dalam visualisasi ini.

4.1.2 Distribution Plot

Distribution plot adalah alat visual untuk memahami konsep dari distribusi data. Tujuannya untuk membantu melihat dan menganalisis pola sebaran data dengan lebih mudah dan intuitif.



Gambar 4. 6 Distribution Plot

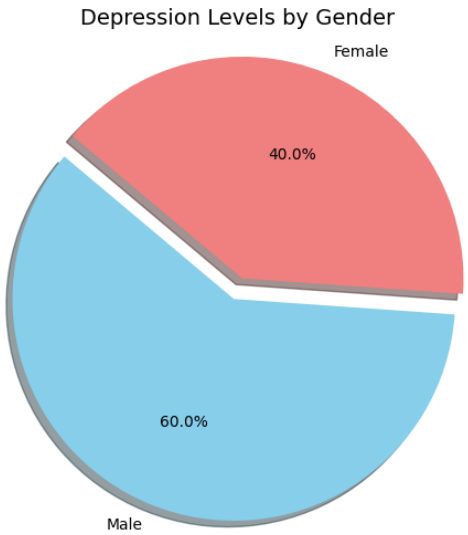
Grafik ini membandingkan nilai aktual dengan nilai prediksi. Setiap titik biru mewakili satu data poin, dengan posisi horizontalnya menunjukkan nilai aktual dan posisi vertikalnya menunjukkan nilai yang diprediksi oleh suatu model. Garis putus-putus berwarna merah adalah garis prediksi sempurna, di mana nilai aktual akan sama dengan nilai prediksi. Semakin dekat titik-titik biru ke garis merah, semakin baik kinerja model dalam memprediksi nilai aktual. Dalam grafik ini, sebagian besar titik biru terletak cukup dekat dengan garis merah, menunjukkan bahwa model memiliki kemampuan prediksi yang cukup baik, meskipun terdapat beberapa perbedaan antara nilai aktual dan prediksi.

4.2 Visualisasi Model

Visualisasi model dalam data mining menggunakan algoritma KNN (K-Nearest Neighbor) sangat penting untuk memahami bagaimana algoritma bekerja dan bagaimana data dikelompokkan.

a. *Nearest Neigbors for Gender*

Data ini ditampilkan dalam bentuk diagram lingkaran *(pie chart)* dengan dua warna berbeda untuk membedakan kedua jenis kelamin. Gambar dibawah ini menyajikan data tingkat depresi berdasarkan jenis kelamin, dengan perempuan *(female)* dan laki-laki *(male)* sebagai kategori utama.

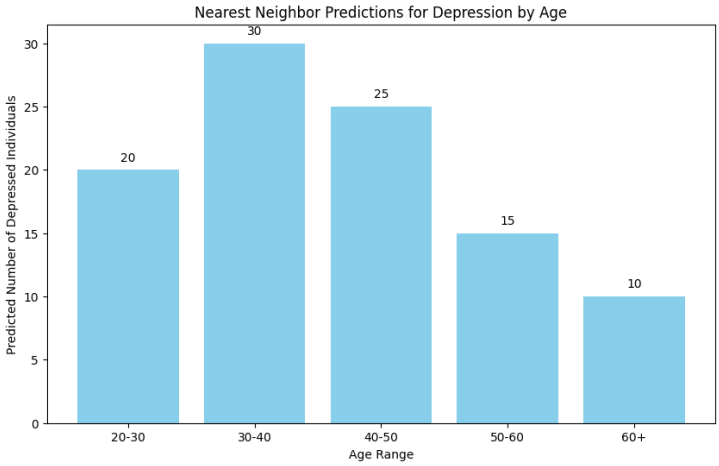


Gambar 4. 7 ***Nearest Neighbors for Gender***

Berdasarkan gambar tersebut, tingkat depresi pada perempuan adalah 40%, sedangkan pada laki-laki sebesar 60%. Artinya, dari total populasi yang disajikan, 40% perempuan mengalami depresi dan 60% laki-laki juga mengalami depresi. Data ini memberikan gambaran bahwa tingkat depresi pada laki-laki lebih tinggi dibandingkan pada perempuan berdasarkan data yang disajikan.

b. Nearest Neigbors for Marital Age

Data ini ditampilkan dalam bentuk bentuk grafik batang *(bar graph)* dengan sumbu horizontal (x-axis) menunjukkan kelompok usia dan sumbu vertikal (y-axis*)* menunjukkan *"Average Depression Level"* atau rata - rata tingkat depresi dengan skala dari 0 hingga 7. Gambar dibawah ini menyajikan data tingkat depresi berdasarkan kelompok usia, yaitu 20-30, 30-40, 40-50, 50-60, dan 60 keatas.

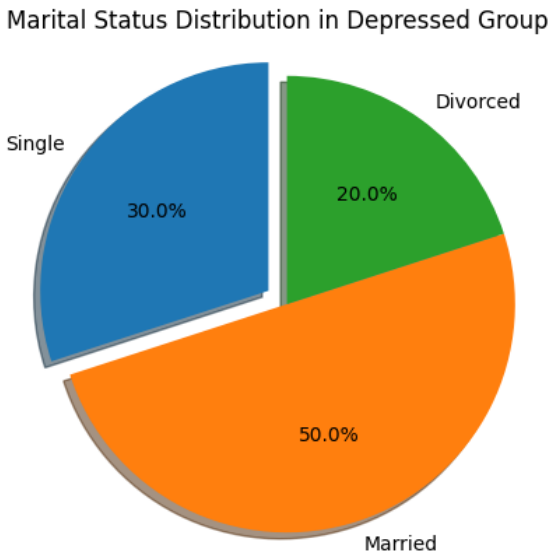


Gambar 4. 8 ***Nearest Neighbors for Age***

Grafik batang ini menunjukkan prediksi jumlah individu yang mengalami depresi berdasarkan kelompok usia, menggunakan metode Nearest Neighbor. Sumbu horizontal dibagi menjadi rentang usia: 20-30, 30-40, 40-50, 50-60, dan 60+. Sumbu vertikal menunjukkan perkiraan jumlah individu yang diprediksi mengalami depresi dalam setiap kelompok usia. Terlihat bahwa kelompok usia 30-40 diprediksi memiliki jumlah individu dengan depresi tertinggi (30 orang), diikuti oleh kelompok usia 40-50 (25 orang), kemudian 20-30 (20 orang), 50-60 (15 orang), dan yang paling sedikit diprediksi mengalami depresi adalah kelompok usia 60+ (10 orang).

c. *Nearest Neigbors for Marital Status*

Gambar berikut ini menyajikan data tingkat depresi berdasarkan status pernikahan, dengan tiga kategori utama yaitu *Single* (belum menikah), *Married* (menikah), dan *Divorced* (bercerai). Data ini ditampilkan dalam bentuk diagram donat *(donut chart)* dengan tiga warna berbeda untuk membedakan ketiga kategori tersebut.

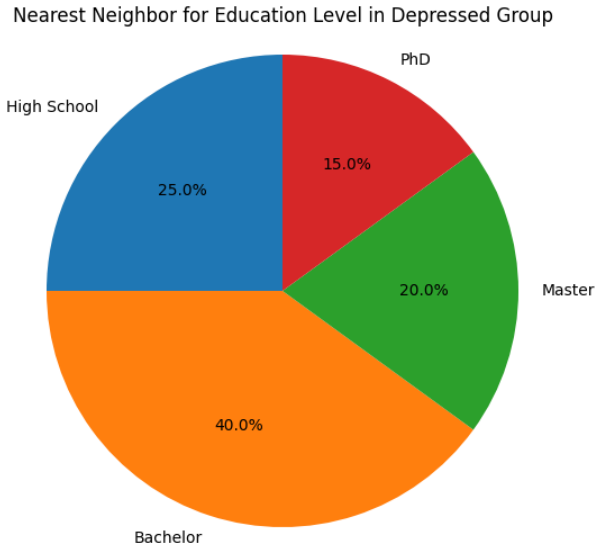


Gambar 4. 9 ***Nearest Neighbors for Marital Status***

Grafik pai ini menggambarkan distribusi status pernikahan dalam kelompok individu yang mengalami depresi. Terlihat bahwa kelompok terbesar adalah mereka yang berstatus menikah, mencakup 50% dari total. Kelompok kedua terbesar adalah individu yang berstatus lajang, dengan proporsi 30%. Sisanya, sebanyak 20%, adalah individu yang berstatus cerai. Grafik ini menunjukkan bahwa dalam kelompok yang mengalami depresi, mayoritas berstatus menikah, diikuti oleh lajang, dan kemudian bercerai.

d. *Nearest Neigbors for Education Level*

Gambar berikut ini menyajikan data tingkat depresi berdasarkan tingkat pendidikan, dengan empat kategori utama yaitu PhD, Master, Bachelor, dan High School. Data ini ditampilkan dalam bentuk diagram lingkaran *(pie chart)* dengan empat warna berbeda untuk membedakan keempat kategori tersebut.

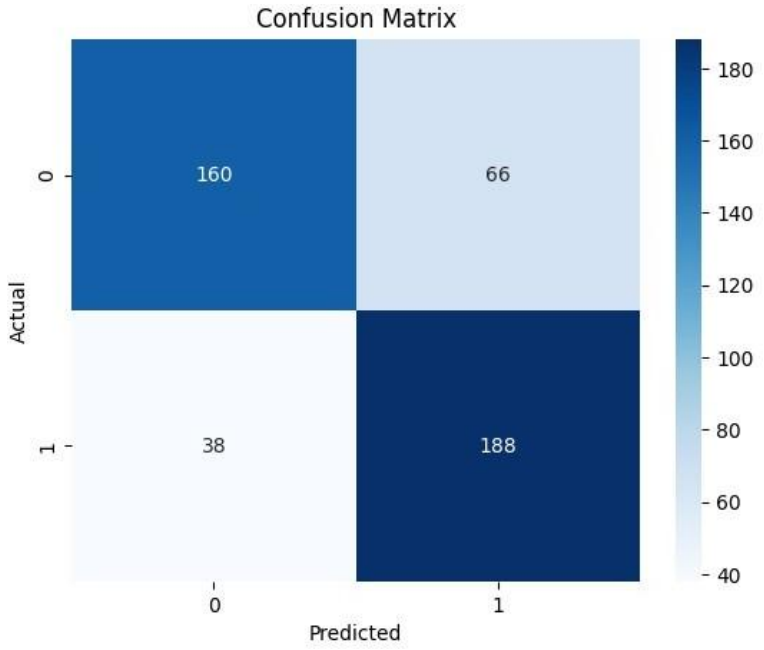


Gambar 4. 10 ***Nearest Neighbors for Education Level***

Grafik ini memperlihatkan proporsi tingkat pendidikan dalam kelompok individu yang mengalami depresi, berdasarkan analisis Nearest Neighbor. Mayoritas dari kelompok ini memiliki tingkat pendidikan Sarjana (Bachelor), mencakup 40%. Tingkat pendidikan tertinggi kedua adalah Sekolah Menengah Atas (High School) dengan 25%. Kemudian, sebanyak 20% memiliki gelar Magister (Master), dan sisanya, 15%, memiliki gelar Doktor (PhD). Dengan demikian, dalam kelompok yang diprediksi mengalami depresi menggunakan metode Nearest Neighbor, lulusan Sarjana merupakan kelompok terbesar.

4.3 Confusion Matrix

Confusion matrix digunakan untuk mengevaluasi kinerja model klasifikasi dengan membandingkan hasil prediksi model dengan data sebenarnya. Dalam klasifikasi biner, confusion matrix memiliki empat komponen utama: *True Positive* (TP), yaitu data yang sebenarnya positif dan diprediksi positif; *True Negative* (TN), data yang sebenarnya negatif dan diprediksi negatif; *False Positive* (FP), data yang sebenarnya negatif tetapi diprediksi positif; dan *False Negative* (FN), data yang sebenarnya positif tetapi diprediksi negatif.



Gambar 4. 11 *Confusion Matrix*

Angka-angka dalam kotak menunjukkan jumlah data yang termasuk dalam kategori tertentu. Misalnya, angka 160 di pojok kiri atas berarti ada 160 data yang sebenarnya berlabel 0 dan diprediksi juga berlabel 0 oleh model *(True Negative)*. Angka 66 di pojok kanan atas menunjukkan ada 66 data yang sebenarnya berlabel 0, tetapi diprediksi 1 *(False Positive)*. Angka 38 di pojok kiri bawah menunjukkan ada 38 data yang sebenarnya berlabel 1, tetapi diprediksi 0 *(False Negative)*. Terakhir, angka 188 di pojok kanan bawah menunjukkan ada 188 data yang sebenarnya berlabel 1 dan diprediksi juga berlabel 1 *(True Positive)*. Data dibagi menjadi empat kategori yaitu *True Negative* (TN), *False Positive* (FP), *False Negative* (FN), dan *True Positive* (TP). Berdasarkan angka dalam matrix, total data adalah 452, dengan 160 TN, 66 FP, 38 FN, dan 188 TP. Akurasi model dihitung dengan menjumlahkan TN dan TP, lalu dibagi total data. Hasilnya, akurasi model adalah sekitar 77%, yang berarti model berhasil memprediksi dengan benar sekitar 77% dari keseluruhan data. Sama halnya dengan proses metrik evaluasi yang telah dilakukan sebelumnya.

4.4 Pembahasan

Hasil evaluasi *Confusion Matrix* terhadap model klasifikasi KNN (K-Nearest Neighbor) dengan tingkat akurasi 77% memberikan gambaran awal mengenai

kemampuannya dalam mengklasifikasikan data terkait depresi. Tingkat akurasi ini secara spesifik mengindikasikan bahwa dari keseluruhan data yang dievaluasi, proporsi prediksi yang tepat adalah sebesar 77%. Dengan kata lain, dari setiap 100 unit data yang diuji, model berhasil mengidentifikasi status depresi secara benar pada 77 kasus.

Namun, penting untuk menggarisbawahi bahwa 23% sisa dari data mengalami misklasifikasi. Kesalahan ini terbagi menjadi dua jenis yang memiliki implikasi berbeda dalam konteks prediksi depresi. *False positive*, atau kesalahan prediksi positif, terjadi ketika model memprediksi seorang individu mengalami depresi padahal kenyataannya tidak. Dalam aplikasi klinis, *false positive* dapat menyebabkan kecemasan yang tidak perlu, stigma, serta potensi intervensi yang tidak dibutuhkan. Sebaliknya, *false negative*, atau kesalahan prediksi negatif, terjadi ketika model memprediksi seorang individu tidak mengalami depresi padahal kenyataannya ia mengalaminya. Tipe kesalahan ini lebih problematik karena dapat mengakibatkan keterlambatan diagnosis dan penanganan yang krusial bagi individu yang membutuhkan bantuan.

Meskipun tingkat akurasi 77% dapat dianggap sebagai indikator awal yang menjanjikan bahwa model KNN memiliki kemampuan untuk menangkap pola-pola yang relevan dalam data terkait depresi, keberadaan tingkat kesalahan sebesar 23% menunjukkan adanya keterbatasan dan potensi risiko jika model ini langsung diimplementasikan tanpa validasi lebih lanjut. Faktor - faktor seperti kualitas dan representasi data latih, pemilihan nilai K yang optimal dalam algoritma KNN, serta potensi adanya *noise* atau fitur yang kurang relevan dalam dataset dapat berkontribusi terhadap tingkat kesalahan ini.

Oleh karena itu, langkah-langkah pengembangan dan validasi lanjutan menjadi krusial. Pengembangan dapat meliputi upaya untuk meningkatkan kualitas data, melakukan *feature engineering* untuk menghasilkan fitur-fitur yang lebih informatif, atau bahkan membandingkan kinerja model KNN dengan algoritma klasifikasi lainnya. Validasi yang komprehensif, melalui teknik seperti *cross-validation* pada dataset yang berbeda dan lebih besar, diperlukan untuk menguji

generalisasi model dan memastikan bahwa kinerja yang baik tidak hanya terbatas pada data evaluasi awal.

Implikasi dari hasil evaluasi ini adalah bahwa meskipun model KNN menunjukkan potensi yang cukup baik dalam tugas prediksi depresi, implementasi efektifnya memerlukan kehati-hatian dan tahapan pengembangan serta validasi yang rigorous. Tujuannya adalah untuk meminimalkan risiko kesalahan prediksi, baik *false positive* maupun *false negative*, sehingga model dapat menjadi alat yang akurat dan bermanfaat dalam membantu proses identifikasi dan penanganan depresi. Dengan demikian, penelitian lebih lanjut yang fokus pada peningkatan akurasi dan validasi model KNN ini sangat dianjurkan sebelum dapat dipertimbangkan untuk implementasi dalam skala yang lebih luas.