PENERAPAN DEEP LEARNING MODEL *RANDOM FOREST* UNTUK PREDIKSI PENERIMA BANTUAN PROGRAM KELUARGA HARAPAN (PKH)

Dindin Haidar [1], Bambang Irawan [2], Agus Bahtiar [3]
[1,2] Teknik Informatika, STMIK IKMI Cirebon
[3] Sistem Informasi, STMIK IKMI Cirebon
Jl. Perjuangan No.10B, Karyamulya, Kec. Kesambi, Kota Cirebon
haidar.dindin123@gmail.com

ABSTRAK

Program Keluarga Harapan (PKH) merupakan inisiatif pemerintah Indonesia untuk memberikan dukungan kepada Rumah Tangga Sangat Miskin (RTSM). Dalam konteks globalisasi dan kemajuan teknologi informasi, optimalisasi teknologi, terutama pembelajaran mesin, menjadi krusial dalam memecahkan tantangan prediksi penerima Bantuan Program Keluarga Harapan (PKH). Dalam implementasi PKH di Kabupaten Kuningan, terdapat permasalahan dalam penentuan penerima manfaat yang disebabkan oleh ketidakakuratan dan kebingungan dalam sistem seleksi manual yang sudah ketinggalan zaman. Penelitian ini bertujuan mengembangkan model prediksi menggunakan pendekatan Machine Learning, dengan fokus pada algoritma Random Forest. Data historis terkait program PKH, seperti nominal bantuan, status sosial-ekonomi, jumlah anggota keluarga, jenis penyaluran bantuan, dan karakteristik demografi, diolah untuk membangun model prediksi. Dengan menggunakan data penerimaan bantuan PKH tahun 2023 di Kabupaten Kuningan 44.690 data, model ini mencapai ROC AUC akurasi sekitar 96%, dengan akurasi Random Forest sekitar 98,9%, dan Decision Tree sekitar 98%. Penelitian ini sebagai respons terhadap ketidakakuratan dalam penentuan penerima manfaat PKH dan kesenjangan administratif di tingkat daerah. Hasilnya diharapkan memberikan kontribusi signifikan dalam meningkatkan akurasi dan efisiensi program bantuan sosial, sesuai dengan tujuan pemerintah untuk mengurangi kemiskinan dan meningkatkan kesejahteraan masyarakat berpenghasilan rendah. Kesimpulan ini diharapkan dapat memberikan wawasan baru dalam literatur mengenai peningkatan efektivitas program bantuan sosial dalam konteks administratif daerah.

Kata kunci: *Program Keluarga Harapan, Prediksi, Random Forest, Machine Learning.*

1. PENDAHULUAN

Di era globalisasi dan pesatnya perkembangan bidang teknologi informasi, teknologi informasi telah menjadi basis perubahan sosial, ekonomi, dan budaya di seluruh dunia. Perkembangan teknologi informasi telah memberikan dampak yang signifikan terhadap berbagai aspek kehidupan manusia, antara lain perkembangan teknologi, perubahan model bisnis, inovasi di bidang pendidikan, dan perubahan interaksi sosial. Kemiskinan merupakan hal yang umum terjadi di negara-negara berkembang dan merupakan masalah yang sulit dipecahkan. Pemerintah memainkan peran yang sangat penting dalam konteks ini, namun seringkali mereka cenderung berfokus pada wilayah perkotaan, sehingga mengakibatkan peningkatan kesenjangan sosial di tempat lain. Fenomena ini menunjukkan semakin perlunya pendekatan inovatif dalam pengelolaan dan pemanfaatan data dan teknologi informasi dalam berbagai aspek kehidupan manusia.

Dalam situasi seperti ini, masalah yang muncul adalah bagaimana mengoptimalkan teknologi informasi, khususnya metode pembelajaran mesin, untuk menyelesaikan masalah yang muncul saat memprediksi penerima Bantuan Program Keluarga Harapan (PKH). Meskipun tujuan utama program adalah untuk membantu kelompok masyarakat yang membutuhkan, penentuan penerima yang efektif dan akurat merupakan masalah penting dalam pelaksanaannya. Program-program ini sangat penting untuk mendukung kesejahteraan sosial dan ekonomi masyarakat, jadi penting untuk mengatasi masalah ini. Tujuan dari penelitian ini adalah untuk memprediksi dan mengklasifikasikan data dari populasi penerima bantuan PKH di Kabupaten Kuningan. Untuk mencapai tujuan ini, penelitian ini menggunakan teknik data mining dan metode Random Forest. Dalam penelitian ini, algoritma yang digunakan model Random Forest, RapidMiner serta Python sebagai alat bantu untuk menemukan nilai yang akurat. Salah satu masalah yang dihadapi adanya ketidaktepatan dalam memilih penerima bantuan PKH menyebabkan banyak masyarakat yang benar-benar membutuhkan bantuan menjadi terabaikan dan belum adanya sistem atau model prediksi di Dinas Sosial Kuningan.

Berdasarkan referensi dari beberapa penilitian terdahulu yang terkait topik mengenai program keluarga harapan [1] dengan penelitian klasifikasi penentuan penerima manfaat Program Keluarga Harapan (PKH) dengan menggunakan algoritma C5 mengembangkan sistem pendukung keputusan untuk menentukan penerima PKH. Sistem seleksi manual yang ada saat ini dan data yang sudah ketinggalan zaman telah menyebabkan kebingungan dan ketidakakuratan dalam mengidentifikasi penerima manfaat yang layak. Sebuah studi mengenai prediksi [2] Pendeketan Deep learning dalam Memprediksi Penerima Program PKH menerapkan model

pembelajaran mesin untuk memprediksi keluarga menerima bantuan Program Keluarga Harapan (PKH). Penelitian mengenai model Random Forest yang dilakukan oleh [3] Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest mengembangkan model yang dapat memprediksi kemungkinan terjadinya penyakit diabetes pada tahap pertama dengan menggunakan klasifikasi Random Forest. Metode Random Forest adalah sebuah teknik dalam machine learning yang menggabungkan metode *Bagging* dan *Random Sub Spaces.*

Penelitian ini memiliki tujuan utama untuk mengembangkan model prediksi yang dapat mengklasifikasikan dan memprediksi penerima bantuan PKH. Penelitian ini juga penting karena mengisi celah dalam literatur mengenai berbagai program bantuan, yang dapat meningkatkan efektivitas penentuan penerima manfaat dalam konteks administratif daerah seperti Dinas Sosial Kabupaten Kuningan. Oleh karena itu, tujuan dari penelitian ini adalah untuk meningkatkan akurasi dan efektivitas dalam menentukan kelayakan penerimaan bantuan sosial dari masyarakat berpenghasilan rendah serta membantu mengatasi masalah ketepatan sasaran dalam penyediaan bantuan oleh pemerintah. Penelitian ini menggunakan metode pembelajaran mesin untuk meningkatkan akurasi dan efektivitas dalam penyediaan bantuan sosial kepada masyarakat berpenghasilan rendah. Selain itu, penelitian ini dapat membantu kebijakan bantuan sosial pemerintah dan dapat mempercepat distribusi manfaat kepada kelompok masyarakat yang paling membutuhkan.

Penelitian ini menggunakan atau mengintegrasikan teknik pembelajaran mesin, khususnya algoritma Random Forest, untuk mengolah data historis terkait program PKH. Faktor-faktor seperti data nominal bantuan, status sosial ekonomi, jumlah anggota keluarga, dan karakteristik demografi dianggap sebagai karakteristik yang relevan ketika mengembangkan model prediktif. Oleh karena itu, penelitian ini mengambil pendekatan holistik untuk mengatasi tantangan dalam menentukan penerima manfaat program bantuan. Implementasi model ini dilakukan dengan bantuan alat Rapidminer atau Python yang membantu menemukan nilai pasti. Penelitian ini bertujuan untuk meningkatkan akurasi alokasi penerima bantuan Program Keluarga Harapan. Oleh karena itu, penelitian ini bertujuan untuk meningkatkan akurasi dan efisiensi dalam menentukan kelayakan penerimaan bantuan sosial bagi masyarakat berpenghasilan rendah dan membantu mengatasi permasalahan penargetan dalam pemberian bantuan pemerintah. Random Forest mempunyai beberapa kelebihannya menjadi salah satu algoritma populer dalam Deep Learning. Kelebihan utamanya mengatasi overfitting, Random Forest bisa mengatasi masalah overfitting yang sering terjadi dalam model deep learning.

Jika penelitian ini berhasil, harapannya hasilnya bisa memberikan solusi pada bidang ilmu komputer dan program bantuan sosial. Penelitian ini diharapkan dapat memberikan wawasan yang lebih mendalam mengenai potensi teknologi informasi, khususnya pembelajaran mesin, untuk mendukung kebijakan sosial ekonomi. Temuan-temuan ini dapat memberikan kontribusi penting bagi pengembangan metode pengelompokan yang lebih efisien dan akurat, tidak hanya dalam konteks kesejahteraan sosial, namun juga dalam berbagai bidang lain yang memerlukan analisis data. Hasil penelitian ini diharapkan dapat menjadi acuan pengambilan keputusan oleh pemangku kepentingan terkait alokasi sumber daya untuk layanan kesejahteraan sosial. Diharapkan dengan menggunakan pendekatan inovatif ini, program PKH akan mampu mencapai tingkat akurasi yang lebih tinggi dalam penentuan penerima manfaat, sehingga meningkatkan kesejahteraan mereka yang berhak menerima manfaat dari program tersebut, pengentasan kemiskinan akan membaik. Penggunaan model Random Forest untuk memprediksi keputusan menerima bantuan melalui program Keluarga Harapan merupakan contoh bagaimana teknologi informasi dapat dimanfaatkan untuk meningkatkan kualitas hidup masyarakat secara keseluruhan. Oleh karena itu, penelitian ini dapat memberikan landasan untuk mengembangkan solusi lebih lanjut dalam pengelolaan bantuan sosial.

2. TINJAUAN PUSTAKA

2.1. Penelitian Terdahulu

Penelitian yang dilakukan oleh [4] dengan judul penelitian "Peningkatan Sistem Prediksi Bantuan Program Keluarga Harapan Berbasis Machine Learning/Deep Learning". Penelitian ini membahas mengenai penggunaan teknologi Big Data dan pembelajaran mesin (Machine Learning) dalam mengoptimalkan program Harapan Keluarga (PKH) di Kabupaten Sidoarjo. Tujuan penelitian ini adalah untuk mengurangi penyalahgunaan program PKH dengan menggunakan pendekatan pembelajaran mesin.

Penelitian yang dilakukan oleh [5] dengan judul penelitian "E-PKH: Sistem Pendukung Keputusan Penerima Bantuan Program Keluarga Harapan Menggunakan Metode Naïve Bayes Classifier". Membahas mengenai pengembangan sistem penentuan bantuan PKH (Program Keluarga Harapan) di Karanganyar Gunung, Candisari, Semarang. Penelitian tersebut menggunakan berbagai metode penelitian, antara lain pengumpulan data, metode pemecahan masalah seperti analisis tulang ikan, dan metode pengembangan perangkat lunak. Penelitian ini juga menjelaskan jenis dan sumber data yang digunakan, seperti data primer dan sekunder, serta proses pengumpulan data melalui wawancara dan tinjauan pustaka.

Penelitian yang dilakukan oleh [6] Judul penelitian ini adalah "Penerapan Data Mining untuk

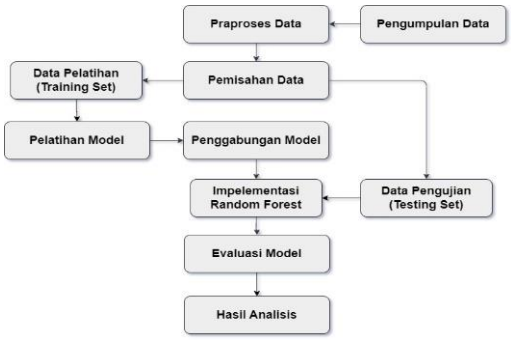
Mengklasifikasikan Penerima Bantuan Program Keluarga Harapan (PKH) di Desa Wae Jare Menggunakan Metode Naïve Bayes". Tujuan dari penelitian ini adalah untuk mengatasi kesulitan dalam mengklasifikasikan antara penerima dan bukan penerima bantuan PKH di Desa Wae Jare, Kecamatan Mbeliling, Kabupaten Manggarai Barat dengan menerapkan metode pengklasifikasian menggunakan Naïve Bayes. PKH atau Program Keluarga Harapan adalah program penanggulangan kemiskinan yang diluncurkan oleh pemerintah Indonesia.

Penelitian yang dilakukan oleh [7] judul penelitian ini adalah "Perbandingan Algoritma *NBC, KNN,* dan *C4.5* Untuk Klasifikasi Penerima Bantuan Program Keluarga Harapan". Tujuan dari penelitian ini adalah untuk membuat model klasifikasi yang paling tepat dalam mengklasifikasikan data calon penerima bantuan Program Keluarga Harapan (PKH) di Indonesia. Penelitian ini bertujuan untuk mengatasi masalah penyaluran bantuan PKH yang belum tepat sasaran. Hasil dari penelitian ini menunjukkan bahwa algoritma NBC menghasilkan nilai akurasi sebesar 77,51%, algoritma *K-NN* (K = 3) menghasilkan nilai akurasi sebesar 76,72%, dan algoritma *C4.5* menghasilkan nilai akurasi sebesar 80,16%.

3. METODE PENELITIAN

3.1. Metode Penelitian

Terdapat beberapa tahapan metode penelitian berjumlah 8 tahapan. Tahapan tersebut adalah pengumpulan data, praproes data pemisahan data, pelatihan model, penggabungan model, evaluasi model, Prediksi; dan hasil analisis. Metode penelitian yang dilakukan pada penelitian ini dapat dilihat pada gambar dan tabel.



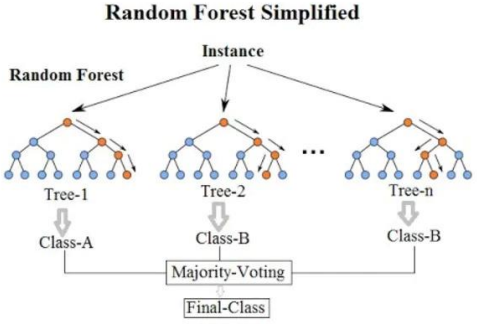
Gambar 1. Metode penelitian

Dari Gambar 1 tentang metode penelitian dapat dijelaskan sebagai berikut;

a. Pengumpulan Data
   Tahap awal ini melibatkan pemilihan data yang diperlukan sebelum memulai pencarian informasi. Data yang telah terpilih akan digunakan pada tahap selanjutnya kumpulan data mentah berupa tabel yang dapat diolah lebih lanjut.
b. Praproses Data
   Sebelum memulai data mining, langkah awalnya adalah membersihkan data. Tujuannya adalah membersihkan data dari nilai yang hilang atau anomali. Menormalisasi atau mengubah skala fitur-fitur jika diperlukan. Memisahkan data menjadi atribut (fitur) dan label (kelas/target) kekonsistenan data, dan memperbaiki kesalahan seperti kesalahan pengetikan atau cetak.
c. Pemisahan Data
   Proses ini terlibat dalam mengubah data yang sudah dipilih. Membagi dataset menjadi dua set data: set data pelatihan (*training set*) dan set data pengujian (*testing set*).
d. Pelatihan Model
   Merupakan tahap dimana membangun sejumlah pohon keputusan (decision trees) menggunakan set data pelatihan. Setiap pohon dibangun dengan memilih sampel acak dari set data pelatihan.
e. Penggabungan Model
   Merupakan tahap dimana menggabungkan hasil dari semua pohon keputusan yang telah dibangun menjadi model Random Forest. Penggabungan ini dapat dilakukan dengan mengambil hasil mayoritas dari semua pohon (klasifikasi) atau dengan mengambil rata-rata hasil (regresi).
f. Evaluasi Model
   Pola informasi yang telah diperoleh dari data menggunakan model Random Forest yang telah dibangun untuk menguji set data pengujian yang terpisah. Mengukur kinerja model dengan menggunakan metrik seperti akurasi, presisi, recall, F1-score, dan lainnya, tergantung pada jenis masalah (klasifikasi atau regresi).
g. Hasil Analisis
   Menganalisis hasil prediksi dan mengambil kesimpulan dari penelitian yang telah di uji data dengan model *random forest*.

3.2. Random Forest

Random Forest dapat menghasilkan model pohon keputusan yang akurat, namun teknik ini juga dapat menghasilkan model yang kompleks dan sulit diinterpretasikan[8]. mengatasi masalah performansi dengan menggunakan metode *Random Forest*. Hasil penelitian menunjukkan bahwa nilai akurasi untuk film 'Star Trek' berkisar 60% untuk semua metode yang digunakan, sedangkan untuk film 'Quantum of Solace' nilai akurasi berada di atas 70% untuk metode *Decision Tree*, *Naïve Bayes*, dan *Max-Entropy*, dan sekitar 60%[9].



Gambar 2. *Random Forest*

Metode Random Forest menghasilkan sejumlah pohon acak, dan kelas yang dihasilkan terbentuk melalui proses klasifikasi yang dipilih dari kelas yang paling umum (modus) yang dihasilkan oleh pohon keputusan yang telah ada**.**

3.3. Sumber Data

Dalam penelitian ini, sumber data yang digunakan berasal dari data penerima Bantuan Program Keluarga Harapan (PKH) diperoleh dari Dinas Sosial Kabupaten Kuningan. Data ini termasuk informasi tentang nominal bantuan, nama desa, nama kecamatan, kondisi sosial ekonomi berdasarkan nominal bantuan, dan jenis penyaluran bantuan untuk memudahkan penyaluran penerima bantuan, penelitian ini diharapkan dapat memberikan wawasan yang berharga tentang karakteristik penerima bantuan khususnya di Kabupaten Kuningan, yang dapat menjadi dasar untuk pengambilan keputusan yang lebih baik dalam pengembangan sistem inforasi. Data yang diambil adalah data pnerima bantuan PKH pada tahun 2023 bulan mei.

3.4. Split Data

Split data berguna untuk membagi data menjadi dua bagian atau lebih yaitu data latih (training data) dan data uji (testing data). Hal ini sangat penting dilakukan karena dalam pemodelan data machine learning, model harus dibangun dan diuji dengan data yang berbeda. Data latih digunakan untuk membangun model, sementara data uji digunakan untuk mengevaluasi kinerja model.

Peneliti membagi data menjadi 2 bagian, yaitu model A dengan perbandingan 90% data train: 10% data test, model B dengan perbadingan 80% data train: 20% data test. Dengan pembagian beberapa dataset yang berbeda akan didapatkan nilai ketepatan prediksi terbaik. Peneliti menggunakan teknik split data Cross-Validation atau biasa dikenal dengan teknik K-fold cross Validation. K-fold Cross Validation yaitu metode yang membagi dataset menjadi beberapa bagian, lalu menggunakan beberapa bagian untuk training dan beberapa bagian untuk testing.

3.5. Teknik Pengumpulan Data

Langkah dalam teknik pengumpulan data meliputi:

a. Wawancara dengan kepala sekertariat PKH Kabupaten Kuningan untuk mengumpulkan informasi.
b. Wawancara kepada penanggung jawab data untuk mengumpulkan informasi tambahan tentang kondisi terbaru penerima bantuan.
c. Pengumpulan data sekunder dari catatan Dinas Sosial dan sumber data lainnya pada website Kementrian Sosial. Kemudian untuk mengklasifikasikan produksi perikanan budidaya menurut jenis ikan unggulan, data yang diperoleh akan diolah berdasarkan metode *K-Means* yang mengambil beberapa nilai atribut yang terdapat pada data tersebut.

3.6. Teknik Analisis Data

Analisis data melibatkan langkah-langkah sistematis untuk mengumpulkan dan mengatur informasi dari berbagai sumber, mengelompokkan data ke dalam kategori, menggambarkannya dalam unit-unit yang relevan, menyusunnya secara keseluruhan, mencari pola, menyoroti informasi penting, menentukan fokus belajar, dan merangkumnya untuk mendapatkan pemahaman lebih dalam mengenai data yang akan diuji.

4. HASIL DAN PEMBAHASAN

4.1. Pengumpulan Data

Penelitian ini menggunakan dataset penerima bantuan Program Keluarga Harapan di Dinas Sosial Kabupaten Kuningan dengan rentang data tahun 2023 bulan mei sebanyak 44.690 data langsung diambil langsung dari Dinas Sosial Kabupaten Kuningan.

	NOMINAL	PENERIMA	SD	SMP	SMA	LANSIA	BALITA	PROVINSI	KABUPATEN	KECAMATAN	KELURAHAN
0	875000	AAM DAMAYANTI	0	1	1	0	0	JAWA BARAT	KAB. KUNINGAN	CIAWIGEBANG	CIAWIGEBANG
1	600000	YUYUN YUNIAH	0	0	0	1	0	JAWA BARAT	KAB. KUNINGAN	CIAWIGEBANG	CIAWIGEBANG
2	725000	SAETI	1	0	1	0	0	JAWA BARAT	KAB. KUNINGAN	CIAWIGEBANG	CIAWIGEBANG
3	975000	NURHAYATI	1	0	0	0	1	JAWA BARAT	KAB. KUNINGAN	CIAWIGEBANG	CIAWIGEBANG
4	450000	SISKA MELIYA PERTIWI	1	0	0	0	0	JAWA BARAT	KAB. KUNINGAN	CIAWIGEBANG	CIAWIGEBANG
...	...	...	...	...	...	...	...	...	...	...	...
44685	600000	DATI	0	0	0	1	0	JAWA BARAT	KAB. KUNINGAN	SUBANG	TANGKOLO
44686	600000	KARSA	0	0	0	1	0	JAWA BARAT	KAB. KUNINGAN	SUBANG	TANGKOLO
44687	375000	SANI	0	1	0	0	0	JAWA BARAT	KAB. KUNINGAN	SUBANG	TANGKOLO
44688	600000	KARSAH	0	0	0	1	0	JAWA BARAT	KAB. KUNINGAN	SUBANG	TANGKOLO
44689	600000	ASIAH	0	0	0	1	0	JAWA BARAT	KAB. KUNINGAN	SUBANG	TANGKOLO

44690 rows × 11 columns

Gambar 3. Data PKH

4.2. Praproses Data

Pada tahapan data preprocessing proses pengolahan data menjadi lebih terstruktur dari data yang awalnya mentah. Tujuannya, untuk optimalisasi pemodelan *machine learning* algoritma dalam melakukan*training* dan evaluasi model secara efisien [10]. Penelitian ini melakukan tahapan *preprocessing* sebagai berikut:

1. Mengklasifikasikan kategori nilai nominal bantuan per-triwulan berdasarkan nominal bantuan pada dataset akan ada perubahan pada beberapa kolom berikut kategori info bantuan PKH yang diambil hanya katgeori per-triwulan dan beberapa tahap1-2:

KATEGORI	PER-TRIWULAN	TAHAP 1-2	1 TAHUN	JULI-AGST	SEPT-OKT	NOV-DES
				PER-2 BULAN	PER-2 BULAN	PER-2 BULAN
SD	225.000	450.000	900.000	150.000	150.000	150.000
SMP	375.000	750.000	1.500.000	250.000	250.000	250.000
SMA	500.000	1.000.000	2.000.000	333.333	333.333	333.333
DSB	600.000	1.200.000	2.400.000	400.000	400.000	400.000
LANSIA	600.000	1.200.000	2.400.000	400.000	400.000	400.000
HAMIL	750.000	1.500.000	3.000.000	500.000	500.000	500.000
BALITA	750.000	1.500.000	3.000.000	500.000	500.000	500.000

Gambar 4. Kategori bantuan PKH

2. Mengecek dan menghapus nilai *null*. Pada *dataset* penelitian tidak terdapat nilai kosong, pada *dataset* terdapat nilai *non-null* dengan jumlah sebagai berikut:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 44690 entries, 0 to 44689
Data columns (total 11 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   NOMINAL    44690 non-null  int64
 1   PENERIMA   44690 non-null  object
 2   SD         44690 non-null  int64
 3   SMP        44690 non-null  int64
 4   SMA        44690 non-null  int64
 5   LANSIA     44690 non-null  int64
 6   BALITA     44690 non-null  int64
 7   PROVINSI   44690 non-null  object
 8   KABUPATEN  44690 non-null  object
 9   KECAMATAN  44690 non-null  object
 10  KELURAHAN  42057 non-null  object
dtypes: int64(6), object(5)
memory usage: 3.8+ MB
```

Gambar 5. Data info

dtypes: int64(6), object(5) Setelah pengecekan tidak adanya nilai null maka pengecekan tetap harus dilakukan sebagai optimalisasi pengembangan model *machine learning*. Menganalisis kategori tanggungan dari hasil analisis nominal bantuan terhadap kategori per-triwulan dan beberapa data masuk ke kategori tahap 1-2. Berikut gambar kode python:

```
# Menghitung jumlah kemunculan setiap nilai nominal
nominal_counts = my_data['NOMINAL'].value_counts()
print(nominal_counts)

# Menghitung jumlah jenis nominal yang berbeda
print("Jumlah jenis nominal:", len(nominal_counts))
```

Gambar 6. Pengkategorian nominal

```
975000     3438
1200000    3076
500000     2756
725000     2570
1125000    1424
750000     1376
1100000    1304
825000      979
450000      853
875000      728
1250000     560
1575000     370
1325000     283
1700000     229
1350000     221
950000      210
1500000     138
1475000     129
1725000     122
1425000     120
1800000     118
1000000     115
1850000      68
1050000      65
1225000      54
1600000      45
1375000      33
675000       33
1925000      20
2100000      18
2175000      17
1550000      16
2075000      14
1950000      13
2300000      12
2325000      10
1650000       8
2400000       8
1825000       7
2200000       5
1975000       5
187000        4
1275000       4
2700000       3
125000        3
75000         2
2550000       2
2450000       2
1175000       2
2025000       2
1525000       1
200000        1
150000        1
1450000       1
625000        1
1875000       1
Name: NOMINAL, dtype: int64
Jumlah jenis nominal: 59
```

Gambar 7. Pengkategorian PKH

Setelah dilakukan analisis berikut hasil dari pengkategorian data nominal bantuan berdasarkan ketegori penerima bantuan PKH:

4.3. Pemisahan Data

Pembagian data untuk model A dan B yang disebutkan dalam permasalahan sebelumnya dapat dihitung sebagai berikut:

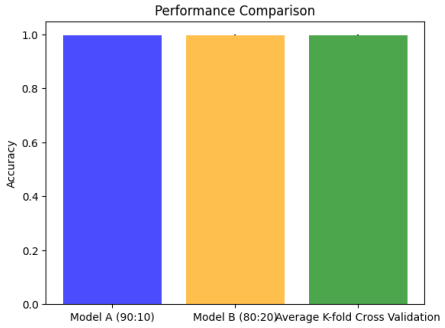
1. Model A: 90% data train, 10% data test
   - Data train: 90% dari 44.690 = 40.221
   - Data test: 10% dari 44.690 = 4.469
2. Model B: 80% data train, 20% data test
   - Data train: 80% dari 44.690 = 35.752
   - Data test: 20% dari 44.690 = 8.938

Berikut gambar kode pembagian data latih dan data uji untuk Model A dan model B:

```
from sklearn.model_selection import train_test_split,
cross_val_score
from sklearn.ensemble import RandomForestClassifier
import numpy as np

# Convert 'NOMINAL' column into a binary classification
problem
median_nominal = data ['NOMINAL'].median()
data['NOMINAL_above_median'] = (data['NOMINAL'] >
median_nominal).astype(int)

# Define the features and the target
X = data[['SMA', 'SD', 'SMP','SMA', 'LANSIA', 'BALITA']]
y = data['NOMINAL_above_median']

# Model A: 90% train data, 10% test data
X_train_A, X_test_A, y_train_A, y_test_A =
train_test_split(X, y, test_size=0.1, random_state=42)
rf_classifier_A =
RandomForestClassifier(n_estimators=100, random_state=42)
rf_classifier_A.fit(X_train_A, y_train_A)
accuracy A = rf classifier A.score(X test A, y test A)

# Model B: 80% train data, 20% test data
X_train_B, X_test_B, y_train_B, y_test_B =
train_test_split(X, y, test_size=0.2, random_state=42)
rf_classifier_B =
RandomForestClassifier(n_estimators=100, random_state=42)
rf_classifier_B.fit(X_train_B, y_train_B)
accuracy_B = rf_classifier_B.score(X_test_B, y_test_B)

# K-fold Cross Validation
k_fold_scores = cross_val_score(rf_classifier_B, X, y,
cv=10)

# Output the performance
print('Accuracy of Model A (90:10):', accuracy_A)
print('Accuracy of Model B (80:20):', accuracy_B)
print('Average K-fold Cross Validation Score:',
np.mean(k_fold_scores))
 print('Standard Deviation of K-fold Cross Validation
 Score:', np.std(k_fold_scores))
```

Gambar 8. Split data

```
Accuracy of Model A (90:10): 0.9966435444170956
Accuracy of Model B (80:20): 0.996755426269859
Average K-fold Cross Validation Score: 0.9973819646453345
Standard Deviation of K-fold Cross Validation Score:
0.001025658235521682
```



Gambar 9. Hasil akurasi split data

Keduanya menunjukkan akurasi yang sangat tinggi. Validasi silang K-fold memberikan skor rata-rata di beberapa pemisahan, yang menunjukkan stabilitas model dan konsistensi kinerja.

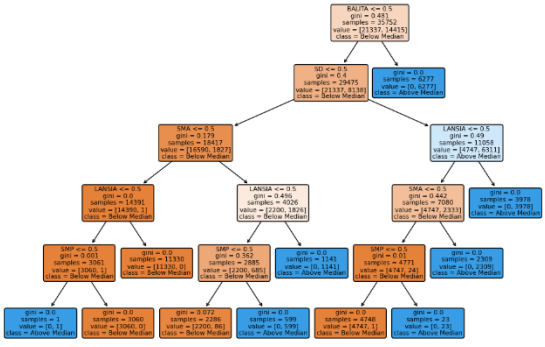
4.4. Pelatihan Model

Membangun sejumlah pohon keputusan (decision trees) menggunakan set data pelatihan. Setiap pohon dibangun dengan memilih sampel acak dari set data pelatihan. Setiap pemilihan split pada setiap node pohon menggunakan subset acak dari fitur-fitur. Berikut gambar kodenya:

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split,
cross_val_score
import numpy as np

# Convert 'NOMINAL' column into a binary classification
problem
median_nominal = data ['NOMINAL'].median()
data['NOMINAL_above_median'] = (data['NOMINAL'] >
median_nominal).astype(int)
```

```
# Define the features and the target
X = data[['SMA', 'SD', 'SMP','SMA', 'LANSIA', 'BALITA']]
y = data['NOMINAL_above_median']

# Split the data into training and testing sets for Model
B (80:20)
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

# Initialize the Decision Tree Classifier
decision_tree = DecisionTreeClassifier(random_state=42)

# Perform K-fold Cross Validation
k_fold_scores = cross_val_score(decision_tree, X, y,
cv=10)

# Train the model
decision_tree.fit(X_train, y_train)

# Evaluate the model
accuracy = decision_tree.score(X_test, y_test)

# Output the performance
print('Accuracy of the Decision Tree Classifier:',
accuracy)
print('Average K-fold Cross Validation Score:',
np.mean(k_fold_scores))
print('Standard Deviation of K-fold Cross Validation
Score:', np.std(k_fold_scores))

Accuracy of the Decision Tree Classifier:
0.996755426269859
Average K-fold Cross Validation Score:
0.9973819646453345
```

Gambar 10. Pelatihan model *decion tree*



Gambar 11. Visualisasi model *decion tree*

Model ini mencapai akurasi sempurna pada set pengujian dan akurasi rata-rata yang sangat tinggi di seluruh validasi silang K-fold, yang menunjukkan performa dan konsistensi yang kuat.

4.5. Penggabungan Model

Implementasi Random Forest dan penerapan Decision Tree beserta Bootstrap Sampling.

```
# Convert 'NOMINAL' column into a binary classification
problem
median_nominal = data ['NOMINAL'].median()
data['NOMINAL_above_median'] = (data['NOMINAL'] >
median_nominal).astype(int)

# Define the features and the target
X = data[['SMA', 'SD', 'SMP','SMA', 'LANSIA', 'BALITA']]
y = data['NOMINAL_above_median']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

# Initialize the models
rf_classifier = RandomForestClassifier(n_estimators=100,
random_state=42)
decision_tree = DecisionTreeClassifier(random_state=42)

# Bootstrap sampling
bootstrap_samples = 100
rf_accuracies = []
dt_accuracies = []

for _ in tqdm(range(bootstrap_samples)):
    X_resampled, y_resampled = resample(X_train, y_train)
    rf_classifier.fit(X_resampled, y_resampled)
    dt_classifier =
DecisionTreeClassifier(random_state=42,
max_features='sqrt')
    dt_classifier.fit(X_resampled, y_resampled)
    rf_accuracies.append(rf_classifier.score(X_test,
y_test))
    dt_accuracies.append(dt_classifier.score(X_test,
y_test))

# Calculate the average accuracy over all bootstrap
samples
rf_average_accuracy = np.mean(rf_accuracies)
dt_average_accuracy = np.mean(dt_accuracies)

# Output the performance
print('Average accuracy of Random Forest over bootstrap
samples:', rf_average_accuracy)
```

```
100%|██████████| 100/100 [01:38<00:00, 1.10it/s]
Average accuracy of Random Forest over bootstrap samples: 0.9967554262698589
Average accuracy of Decision Tree over bootstrap samples: 0.9967554262698589
```

Gambar 12. Akurasi model *random forest decion tree*

Akurasi rata-rata *Random Forest* terhadap sampel *bootstrap*: 0,9998444842246588 Akurasi rata-rata *Decision Tree* dibandingkan sampel *bootstrap*: 0,999904900425151

4.6. Evaluasi Model

Mengukur akurasi model, metrik evaluasi yang digunakan adalah *ROC AUC (Area Under the Receiver Operating Characteristic Curve).*

```
# Pisahkan data menjadi data latih dan data uji
X_train, X_test, y_train, y_test =
train_test_split(fitur, label, test_size=0.2,
random_state=42)

# Membuat dan melatih model Random Forest
rf_model = RandomForestClassifier(n_estimators=100,
random_state=42)
rf_model.fit(X_train, y_train)

# Melakukan prediksi pada data uji
prediksi_rf = rf_model.predict(X_test)

# Evaluasi kinerja model
print("Classification Report:")
print(classification_report(y_test, prediksi_rf))
```

Gambar 13. Evalusi model *ROC AUC*

Classification Report:

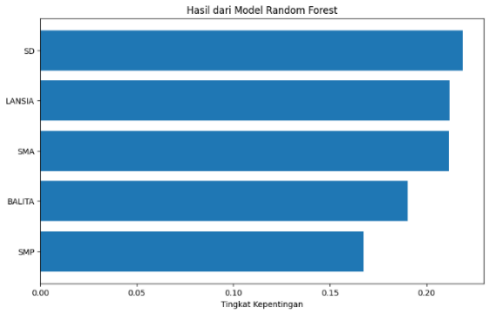
	precision	recall	f1-score	support
187000	0.00	0.00	0.00	1
225000	0.87	1.00	0.93	1036
375000	1.00	1.00	1.00	809
450000	0.00	0.00	0.00	160
500000	0.95	1.00	0.97	556
600000	1.00	1.00	1.00	2811
675000	0.91	1.00	0.95	10
725000	0.91	1.00	0.95	478
750000	0.90	1.00	0.95	275
825000	0.90	1.00	0.95	189
875000	0.96	1.00	0.98	156
950000	0.00	0.00	0.00	38
975000	1.00	1.00	1.00	718
1000000	0.00	0.00	0.00	29
1050000	0.00	0.00	0.00	16
1100000	1.00	1.00	1.00	263
1125000	1.00	1.00	1.00	286
1175000	0.00	0.00	0.00	1
1200000	0.92	1.00	0.96	578
1225000	0.00	0.00	0.00	9
1250000	1.00	1.00	1.00	119
1275000	0.00	0.00	0.00	1
1325000	0.98	1.00	0.99	51
1350000	1.00	1.00	1.00	35
1375000	0.00	0.00	0.00	6
1425000	0.00	0.00	0.00	25
1450000	0.00	0.00	0.00	1
1475000	1.00	1.00	1.00	29
1500000	0.00	0.00	0.00	30
1550000	0.00	0.00	0.00	6
1575000	1.00	1.00	1.00	74
1600000	1.00	1.00	1.00	7
1700000	0.93	1.00	0.96	38
1725000	1.00	1.00	1.00	27
1800000	0.00	0.00	0.00	26
1825000	1.00	1.00	1.00	1
1850000	1.00	1.00	1.00	17
1925000	0.00	0.00	0.00	3
1950000	1.00	1.00	1.00	2
2075000	1.00	1.00	1.00	3
2100000	0.00	0.00	0.00	6
2175000	0.22	1.00	0.36	4
2200000	0.00	0.00	0.00	1
2300000	0.00	0.00	0.00	1
2400000	0.00	0.00	0.00	4
2550000	0.00	0.00	0.00	1
2700000	0.00	0.00	0.00	1
accuracy			0.96	8938
macro avg	0.52	0.55	0.53	8938
weighted avg	0.92	0.96	0.94	8938

Gambar 14. Hasil akurasi model *ROC AUC*

Model menunjukkan kinerja baik dengan akurasi sekitar 96%. Metrik precision, recall, dan F1-score mencerminkan kemampuan model mengidentifikasi, menemukan, dan mencapai keseimbangan antara precision dan recall.

4.7. Hasil Analisis

Interpretasi hasil berdasarkan kode Python yang telah diimplementasikan dan visualisasi yang dihasilkan. *Random Forest* adalah salah satu dari sedikit model yang dapat digunakan untuk menentukan peringkat pentingnya fitur dalam kumpulan data. Pada *random forest* memiliki atribut yang disebut feature_importances_, yang merupakan perkiraan nilai normalisasi kekuatan prediksi fitur berdasarkan gabungan keduanya.

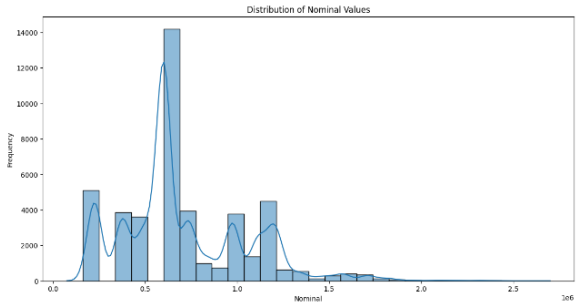


Gambar 15. Visualisasi model *random forest*

Fitur terpenting dalam kumpulan data ini adalah SD Dalam visualisasi model Random Forest, terlihat bahwa fitur paling krusial dalam kumpulan data ini adalah penrima bantuan tingkat (SD), diikuti oleh tingkat lanjut usia lansia, tingkat pendidikan SMA dan ibu hamil balita, dan jumlah penerima bantuan di tingkat pendidikan terakhir (SMP). Tingkat akurasi dari model mencapai sekitar 96%, dengan SD ROC AUC menjadi faktor terpenting dalam prediksi penerima bantuan PKH. Peringkat fitur tersebut diperkuat oleh nilai feature_importances_ sekitar 0,9998444842246588 pada Random Forest dan sekitar 0,999904900425151 pada Decision Tree. Temuan ini menegaskan bahwa Random Forest mampu secara efektif menentukan dan memberikan peringkat pada fitur-fitur yang paling berpengaruh, memberikan wawasan berharga dalam konteks prediksi penerima bantuan Program Keluarga Harapan (PKH).

4.8. Pengembangan Model Prediksi

Setelah memasukkan dan memeriksa data, peneliti mulai memodelkan dataset. Dengan menggunakan algoritma *random forest* melalui *google colab python*. Langkah awal adalah memasukkan data ke dalam *google colab*, diikuti dengan pemrosesan data dan split data dengan dengan teknik *K-fold cross Validation*. Dengan pelatihan model *decision tree* kemudian menggabungkan dan mengimplementasikan *machine learning* model *random forest.*



Gambar 16. Visualisasi distribusi data

Nilai ROC *AUC* mendekati 1 menunjukkan kinerja yang baik, sedangkan nilai mendekati 0.5 menunjukkan kinerja acak. Begitupun dengan model *decision tree* dan *random forest. ROC AUC* menghasilkan akurasi sekitar 96%, *decision tree* menghasilkan akurasi 98% dan *random forest* menghasilkan akurasi 99%.

Model ini mencapai akurasi sempurna pada set pengujian dan akurasi rata-rata yang sangat tinggi di seluruh validasi silang K-fold serta dalam pelatihan dan penggabungan model, yang menunjukkan performa dan konsistensi yang kuat.

4.9. Implikasi Penggunaan Machine Learning

Dalam prediksi berdasarkan nominal bantuan, beberapa nominal dengan label smp menunjukkan akan adanya perubahan setelah 3 tahun mendatang dimana berdasarkan kategori bantuan PKH dari

nominal bantuan Rp. 375.000 akan berubah atau bertambah sekitar Rp. 125.000, untuk nominal label SMA dan Balita dalam 3 tahun masa mendatang kemungkinan akah habis masa bantuannya sehinga nilai nominal bantuan tidak ada. Penggunaan teknologi Machine Learning memberikan dampak positif dalam meningkatkan efisiensi dan akurasi penentuan penerima bantuan. Hal ini terbukti dari hasil analisis data historis pada program PKH di tingkat daerah dan Kabupaten.

5. KESIMPULAN DAN SARAN

Hasil analisis menunjukkan nilai ROC AUC yang mendekati 1 pada model Machine Learning, terutama pada decision tree dan random forest, mengindikasikan kinerja yang baik dengan akurasi sekitar 96%, 98%, dan 99% masing-masing. Performa yang sangat baik terlihat pada uji set pengujian, validasi silang K-fold, serta pada tahap pelatihan dan penggabungan model. Selama lebih dari enam bulan penelitian, model prediksi Machine Learning dalam klasifikasi penerima bantuan PKH di Kabupaten Kuningan menunjukkan efisiensi dan akurasi yang memadai. Evaluasi model menegaskan ketepatan dalam mengidentifikasi karakteristik dan kriteria penerima bantuan. Penggunaan teknologi Machine Learning memberikan dampak positif dalam meningkatkan efisiensi dan akurasi penentuan penerima bantuan, dengan hasil analisis data historis program PKH di tingkat daerah menunjukkan kemampuan Machine Learning dalam mengidentifikasi pola kompleks dan meramalkan nilai pada masa yang akan datang.

Berdasarkan hasil penelitian ini, tentu masih terdapat banyak kekurangan yang dapat ditemui pada prediksi penerima bantuan ini. Maka dari itu, penulis memberikan beberapa saran guna memperbaiki dan meningkatkan kekurangan yang didapati dalam prediksi penerima bantuan ini. Memperbanyak feature atau atribut yang tidak dapat penulis capai guna hasil prediksi yang lebih kompleks pada algoritma random forest. Penambahan feature yang dimaksud dapat meliputi ‘jumlah tanggungan’, ‘jumlah pendapatan’, ‘kepemilikan rumah / tanah’, dan sebagainya. Melakukan perbandingan metode yang lain dengan metode random forest untuk prediksi dalam menentukan rekomendasi kelayakan program penerima bantuan.

DAFTAR PUSTAKA

[1] D. Wintana, H. Hikmatulloh, N. Ichsan, J. J. Purnama, and A. Rahmawati, “KLASIFIKASI PENENTUAN PENERIMA MANFAAT PROGRAM KELUARGA HARAPAN (PKH) MENGGUNAKAN ALGORITMA C5.0 (Studi kasus: Desa Sukamaju, Kec.Kadudampit),” *Klik - Kumpul. J. Ilmu Komput.*, vol. 6, no. 3, p. 254, 2019, doi: 10.20527/klik.v6i3.206.

[2] R. A. Z. Irwan Agus Sobari, “Pendekatan Machine Learning dalam Memprediksi Keluarga Penerima Program PKH,” *J. Tek. Komput. AMIK BSI*, vol. 8, no. 2, pp. 174–180, 2022, doi: 10.31294/jtk.v4i2.

[3] W. Apriliah, I. Kurniawan, M. Baydhowi, and T. Haryati, “Informasi Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest,” *J. Sist. Inf.*, vol. 10, no. 1, pp. 163–171, 2021, [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[4] I. W. S. Supriana, M. A. Raharja, and I. M. S. Bimantara, “Pengembangan Sistem Prediksi Bantuan Program Keluarga Harapan (PKH) Berbasis Machine Learning,” *SINTECH (Science Inf. Technol. J.*, vol. 6, no. 1, pp. 26–36, 2023, doi: 10.31598/sintechjournal.v6i1.1297.

[5] Amin Abdullah Sidiq and Febrian Wahyu Christanto, “Algoritma Naive Bayes Untuk Penentuan Pkh (Program Keluarga Harapan) Berbasis Sistem Pendukung Kepu-Tusan (Studi Kasus: Kelurahan Karanganyar Gunung Se-Marang),” *J. Riptek*, vol. 14, no. 1, pp. 65–71, 2020.

[6] A. I. Purnama, A. Aziz, and A. S. Wiguna, “Penerapan Data Mining Untuk Mengklasifikasi Penerima Bantuan Pkh Desa Wae Jare Menggunakan Metode Naïve Bayes,” *Kurawal - J. Teknol. Inf. dan Ind.*, vol. 3, no. 2, pp. 173–180, 2020, doi: 10.33479/kurawal.v3i2.348.

[7] A. Dina, I. Permana, F. Muttakin, and ..., “Perbandingan Algoritma NBC, KNN, dan C4. 5 Untuk Klasifikasi Penerima Bantuan Program Keluarga Harapan,” *J. Media …*, vol. 7, pp. 1079–1087, 2023, doi: 10.30865/mib.v7i3.6316.

[8] G. A. Sandag, “Prediksi Rating Aplikasi App Store Menggunakan Algoritma Random Forest,” *CogITo Smart J.*, vol. 6, no. 2, pp. 167–178, 2020, doi: 10.31154/cogito.v6i2.270.167-178.

[9] M. A. A. Jihad, Adiwijaya, and W. Astuti, “Analisis Sentimen Terhadap Ulasan Film Menggunakan Algoritma Random Forest,” *e-Proceeding Eng.* , vol. 8, no. 5, pp. 10153–10165, 2021.

[10] H. E. Abdelkader, A. G. Gad, A. A. Abohany, and S. E. Sorour, “An Efficient Data Mining Technique for Assessing Satisfaction Level With Online Learning for Higher Education Students during the COVID-19,” *IEEE Access*, vol. 10, pp. 6286–6303, 2022, doi: 10.1109/ACCESS.2022.3143035.