Madani: Jurnal Ilmiah Multidisiplin
Volume 2, Nomor 1, 2024, Halaman 221-230

E-ISSN: 2986-6340
DOI: https://doi.org/10.5281/zenodo.10523523

Penerapan Algoritma *Decision Tree* Pada Penentuan Penerima Program Keluarga Harapan di Desa Turirejo, Kedamean Gresik

Leona Elsa Nilwanda[1], Amalia Anjani Arifiyanti[2], Rizka Hadiwiyanti[3]
[1,2,3]Universitas Pembangunan Nasional Veteran Jawa Timur, Surabaya
e-mail: 19082010004@student.upnjatim.ac.id[1], amalia_anjani.fik@upnjatim.ac.id[2], rizkahadiwiyanti.si@upnjatim.ac.id[3]

Abstrak

Berdasarkan siaran pers Badan Kebijakan Fiskal Kementerian Keuangan, angka kemiskinan di Indonesia tercatat mengalami penurunan dari 9,71% menjadi 9,54% pada Maret 2022. Meskipun terdapat tekanan terhadap harga komoditas, angka kemiskinan menunjukkan tren menurun. Selain pertumbuhan ekonomi, Indonesia memiliki beberapa program yang dicanangkan pemerintah untuk mengentaskan kemiskinan, seperti Program Keluarga Harapan. Program Keluarga Harapan merupakan program bantuan sosial pemerintah yang ditujukan kepada masyarakat miskin yang ditetapkan sebagai penerima manfaat Program Keluarga Harapan. Namun dalam ringkasan hasil pemeriksaan semester II 2021, Otoritas Pemeriksa Keuangan (BPK) menemukan adanya kesalahan pengalokasian bantuan sosial nasional (Bansos) yang mengakibatkan kerugian negara hingga Rp 6,9 triliun. Hal ini tentu merupakan permasalah yang serius, untuk itu diperlukan sebuah sistem yang dapat membantu proses klasifikasi calon penerima bantuan PKH. Untuk melakukan pengklasifikasian dapat menggunakan algoritma Decision Tree dengan model ID3, C45, dan Random forest. Untuk meningkatkan hasil akurasi model maka diperlukan metode features selection. Dari hasil proses klasifikasi, model Random forest dengan SMOTE dan Features Selection memiliki akurasi paling tinggi sebesar 91% dan uji validitas sistem sebesar 91%. Dari hasil tersebut sistem memiliki kemampuan untuk membantu proses klasifikasi penerima bantuan PKH
Kata kunci: 3-5 PKH, Klasifikasi, *Decision tree*

Abstract

Based on a press release from the Ministry of Finance's Fiscal Policy Agency, the poverty rate in Indonesia was recorded to have decreased from 9.71% to 9.54% in March 2022. Even though there is pressure on commodity prices, the poverty rate shows a downward trend. Apart from economic growth, Indonesia has several programs launched by the government to alleviate poverty, such as the Family Hope Program. The Family Hope Program is a government social assistance program aimed at poor communities who are designated as beneficiaries of the Family Hope Program. However, in a summary of the audit results for the second semester of 2021, the Financial Audit Authority (BPK) found an error in the allocation of national social assistance (Bansos) which resulted in state losses of up to IDR 6.9 trillion. This is certainly a serious problem, for this reason a system is needed that can assist in the classification process of potential PKH assistance recipients. To carry out classification, you can use the Decision Trend algorithm with the ID3, C45 and Random forest models. To improve model accuracy results, a features selection method is needed. From the results of the classification process, the Random forest model with SMOTE and Features Selection has the highest accuracy of 91% and the system validity test is 91%. From these results, the system has the ability to assist in the classification process of PKH assistance recipients
Keywords: 3-5*PKH, Classification, Decision tree*

Article Info
Received date: 20 December 2023 Revised date: 27 December 2023 Accepted date: 11 January 2024

PENDAHULUAN

Indonesia menjadi salah satu negara yang mencapai pertumbuhan ekonomi positif pada tahun 2022. Badan Pusat Statistik mengatakan pertumbuhan ekonomi Indonesia meningkat sebesar 3,72% menjadi 5,44% pada triwulan II tahun 2022. Pemulihan ekonomi ini berdampak positif terhadap peningkatan kesejahteraan Masyarakat terutama kemiskinan. Dilansir dari Siaran Pers Badan Kebijakan Fiskal Kementerian Keuangan mengatakan bahwa per Maret 2022 mengalami penurunan menjadi 9,54% dari 9,71%.

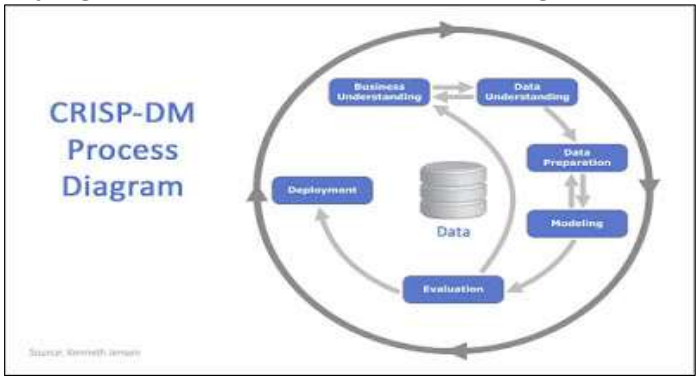
Hal ini tentu tidak lepas dari pertumbuhan ekonomi dan peran pemerintah dalam menanggulangi kemiskinan melalui program-programnya. Salah satu program yang dibuat adalah PKH. Program Keluarga Harapan adalah program buatan pemerintah yang digunakan untuk mengatasi kemiskinan di Indonesia. Dalam program tersebut masyarakat miskin mendapat akses dan manfaat pelayanan sosial seperti Kesehatan, Pendidikan, Pangan, dan Program Perlindungan Sosial lainnya. Namun pada dokumen Ikhtisar Hasil Pemeriksaan yang dikeluarkan oleh Badan Pemeriksa Keuangan (BPK) menyebutkan bahwa terdapat kesalahan dalam proses penyaluran bantuan sosial yang mengakibatkan kerugian sebesar 6,9% triliun pada negara. Dalam laporan tersebut menyebutkan bahwa PKH menjadi salah satu program yang proses penyalurannya bermasalah. Kesalahan penyaluran meliputi data penerima manfaat yang meninggal dunia yang masih masuk dalam data penerima manfaat keluarga, data Penerima Bantuan Sosial Terpadu (DTKS) tidak terdapat pada data penerima manfaat. Penerima bansos bermasalah masih teridentifikasi sebagai penerima. Penerima dengan nomor induk kependudukan yang belum terdaftar. Penerima sudah dinonaktifkan tetapi masih diberikan, dan yang terakhir penerima bansos mendapat lebih dari sekali atau ganda.

Desa Turirejo merupakan salah satu daerah yang mendapatkan Program Keluarga Harapan. Dalam proses wawancara dengan pemerintah desa disebutkan bahwa jumlah penerima manfaat PKH di Desa Turirejo kurang lebih sebanyak 412 orang. Namun terdapat kesalahan dalam pemilihan calon penerima manfaat PKH.Misalnya masyarakat yang tidak sesuai dengan kriteria penerima bantuan PKH dan sebaliknya. Oleh karena itu, diperlukan suatu metode klasifikasi yang mendukung verifikasi dan mendukung keputusan pemerintah desa dalam proses seleksi penerima Program Keluarga Harapan.

Beberapa tahun terakhir, terdapat penelitian-penelitian yang membahas penerapan klasifikasi untuk memprediksi penerima bantuan PKH. Salah satu contohnya adalah penelitian yang dilakukan oleh Abdul Rofiq Almuqorobin yang menggunakan algoritma *random forest* untuk mengklasifikasikan kelayakan penerima bantuan PKH. Penelitian Dwi Kinasih dkk. menggunakan metode ID3 untuk mengidentifikasi penerima program bantuan pemerintah daerah. Penelitian Eka Fitriani membandingkan algoritma C4.5 dengan Naive Bayes untuk mengetahui kelayakan dukungan Program Keluarga PKH. Hasil dari kedua algoritma yang digunakan: Algoritma C4. 5 memberikan nilai akurasi yang lebih tinggi (91,25%) dibandingkan dengan algoritma Naive Bayes yang hanya memberikan akurasi 87,11%. Algoritma C4.5 sendiri merupakan salah satu jenis algoritma pohon keputusan, mirip dengan algoritma *Random forest* dan ID3. Oleh karena itu jumlah tujuan dari penelitian ini adalah untuk membangun model terbaik dari algoritma *decision tree* dalam mengklasifikasikan penerima Program Keluarga Harapan.

METODE PENELITIAN

Dalam banyak hasil penelitian menunjukan bahwa CRISP-DM masih menjadi model yang masih banyak digunakan di berbagai industri dibandingkan dengan model-model yang lainnya. Mariscal, Marba dan Fernandez menyatakan CRISP-DM sebagai standar untuk pengembangan proyek data mining karena paling banyak digunakan dalam pengembangan data mining [1]. CRISP-DM memiliki enam tahapan yang dalam melakukan analisis data mining.



Gambar 1 Tahapan CRISP-DM (Kenneth Jensen, 2016)

CRISP-DM

Business Understanding

Business Understanding atau pemahaman bisnis merupakan tahapan pemahaman dari permasalahan yang dihadapi. Pada tahapan ini juga diperlukan pengetahuan dari objek bisnis, bagaimana cara mendapatkan data dan memproses data tersebut agar menjadi informasi yang dapat menyelesaikan permasalahan.

Data Understanding

Data Understanding, merupakan tahapan pemahaman, identifikasi, pengumpulan dan analisis data yang diperlukan untuk proses klasifikasi.

Data Preparation

Data Preparation merupakan tahapan mempersiapkan data yang sudah diperoleh agar dapat di proses dengan baik di tahap modelling. Beberapa contoh kegiatan *Data Preparation* seperti Pembersihan Data, pengecekan value kosong pada data set, menghapus atribut-atribut yang tidak diperlukan dalam proses modelling, dan lain-lain.

Modeling

Di tahapan ini dataset yang digunakan akan dilakukan proses pemodelan menggunakan algoritma yang sesuai dengan permasalahan. Dalam penelitian ini menggunakan algoritma *decision tree* dengan model *Random forest*, C.45, dan ID3.

Evaluation

Pada tahapan evaluation, hasil dari proses modelling akan di evaluasi dan dibandingkan antara model *decision tree* mana yang paling baik untuk diimplementasikan sebagai sistem.

Deployment

Tahapan terakhir adalah, pengimplementasian sistem sesuai dengan model yang terbaik, implementasi model dilakukan dalam bentuk website

Random forest

Algoritma *Random forest* adalah salah satu metode dalam *Decision tree*. *Random forest* dapat dijelaskan sebagai gabungan dari beberapa *decision tree*. Kelebihan metode ini antara lain dapat menghasilkan akurasi yang lebih tinggi, dapat mengatasi data dalam jumlah yang besar secara efisien, dan tidak terdapat pemangkasan variabel seperti pada algoritma pohon klasifikasi tunggal [2].

C.45

Algoritma C4.5 adalah salah satu algoritma yang digunakan dalam pembuatan pohon keputusan (*decision tree*). Algoritma ini dikembangkan oleh Ross Quinlan dan merupakan kelanjutan dari algoritma sebelumnya yang dikenal sebagai *ID3 (Iterative Dichotomiser 3).* C4.5 memiliki perbaikan dan tambahan fungsionalitas dibandingkan dengan ID3.

ID3

Algoritma ID3 (*Iterative Dichotomiser 3)* merupakan algoritma pembelajaran mesin yang digunakan untuk membangun pohon keputusan. Algoritma ini biasanya digunakan untuk tugas klasifikasi, di mana tujuannya adalah untuk memprediksi kelas atau label dari suatu data berdasarkan fitur-fitur yang ada. Cara kerja ID3 secara singkat adalah sebagai berikut: pemilihan fitur, pembentukan pohon Keputusan, rekursi, dan prediksi. Ulangi proses perhitungan information gain yang akan terus dilaksanakan sampai semua data telah termasuk dalam kelas yang sama. Variabel yang telah dipilih tidak diikutkan lagi dalam perhitungan nilai Information gain [3]

PCA

PCA *(Principal Component Analysis)* adalah teknik statistic yang diaplikasikan untuk satu kumpulan variabel ketika peneliti tertarik untuk menemukan variabel mana dalam kumpulan tersebut yang berhubungan dengan lainnya. Langkah-langkah umum dalam PCA meliputi: Standarisasi Data, Perhitungan Matrik Kovarian, Pencarian Komponen Utama, Proyeksi Data ke komponen utama, dan pemilihan jumlah komponen yang akan digunakan untuk model. [4]

SMOTE

Teknik Pengambilan Sampel Minoritas Sintetis (SMOTE) adalah cara utama untuk mengatasinya ketidakseimbangan kelas. Teknik ini mensintesis yang baru sampel dari kelas minoritas untuk menyeimbangkan kumpulan data dengan mengambil sampel ulang sampel kelas minoritas [5]. SMOTE menggabungkan sampel serupa untuk membuat sampel sintetis baru dari etnis minoritas. Berikut langkah-langkahnya: Pemilihan sampel minoritas, Pencarian Tetangga Terdekat, Pembuatan

sampel sintesis, dan pembentukan dataset baru. Smote sangat membantu untuk masalah ketidakseimbangan data dalam dataset.

Features Importance

Features Importance adalah salah satu jenis Features Selection. Features Importance memiliki pengaruh yang besar untuk meningkatkan kualitas suatu model dengan menghilangkan fitur yang tidak diperlukan sehingga hanya menyisakan fitur yang paling penting dan berpengaruh terhadap atribut kelas [6]. Konsep ini dapat mengukur seberapa penting setiap fitur atau atribut dalam prediksi yang dibuat oleh model pembelajaran mesin tertentu. Hal ini juga tergantung pada algoritma yang digunakan. Beberapa algoritme, seperti pohon keputusan dan ansambel seperti hutan acak dan peningkatan gradien, dapat memberikan informasi tentang pentingnya fitur.

Confusion Matrix

Pengukuran kinerja akurasi sebaiknya dilakukan untuk setiap algoritma yang digunakan, dengan tujuan untuk menunjukkan tingkat kinerja algoritma pada dataset yang digunakan.Matriks konfusi adalah metode yang dapat digunakan untuk mengukur kinerja suatu metode klasifikasi. Pada dasarnya konfusi matriks memberikan informasi dengan membandingkan hasil klasifikasi yang dilakukan oleh sistem model dengan hasil klasifikasi sebenarnya [7]. Matriks Konfusi adalah tabel yang memiliki empat kombinasi kolom yang berbeda. Keempat kolom tersebut sebut merepresentasikan nilai prediksi dan nilai aktual, ke empat hasil proses klasifikasi pada empat tabel itu terdiri dari True positive, False positive, False negative, dan True negative.

HASIL DAN PEMBAHASAN

Business Understanding

Pada tahapan ini peneliti melakukan proses observasi untuk mengidentifikasi permasalahan yang berada pada objek penelitian. Permasalah yang terjadi di Desa Turirejo adalah kesalahan penyaluran PKH, masih terjadi kesalahan verifikasi penerima bantuan PKH.

Data Understanding

Pada tahapan data understanding peneliti melakukan proses pemahaman data apa saja yang dibutuhkan untuk proses klasifikasi penerima PKH, dalam salah satu program pemerintah yaitu program registrasi sosial dan ekonomi menggambarkan keadaan sosial, ekonom dan tingkat kesejahteraan tiap-tiap keluarga termasuk informasi keluarga penerima dan yang tidak menerima bantuan dari pemerintah, salah satunya PKH. Atribut yang diperoleh Nama, Status Kepemilikan Rumah, Luas Lantai, Lantai Rumah, Dinding, Atap,Sumber Air, Sumber Penerangan, Daya Listrik, Bahan Bakar, Umur, Pendidikan Terakhir, Jumlah Tanggungan, Pekerjaan, Penghasilan, dan Program Bantuan. Pada data ini data kelas target dibagi menjadi 2 yaitu Layak dan Tidak Layak.

Data Preparation

Pada tahapan ini data mentah yang didapat dari proses data understanding diolah kembali menjadi dataset yang siap diolah untuk proses modelling. Tahapan *prepocessing* meliputi cleansing. visualization dan encoding. Tahapan cleansing data yang memiliki nilai kosong atau missing value akan diisi dengan metode mean untuk data numerik, sementara modus untuk data kategorikal. Kemudian menghapus kolom yang hanya memiliki 1 nilai value karena tidak berpengaruh pada atribut kelas target.

Modeling

Pada tahapan modelling akan digunakan 5 macam Skenario.

***Random forest*, ID3, C.45**

Skenario pertama yang dilakukan dengan melakukan pengujian model dengan tiga algoritma yang berbeda yaitu *Random forest*, ID3, dan C45 tanpa bantuan metode yang lain. Hasil dari skenario ini dapat dilihat pada tabel 1 berikut:

Table 1 Skenario Random Forest, ID3, C.45

Skenario	*Accuracy*	*Precision*	*Recall*	F1 Score	AUC
Random forest	**88**	**58**	**70**	**52**	**72**
ID3	78	33	50	48	67
C4.5	73	27	50	50	64

***Random forest*, ID3, C.45 + PCA**

Skenario kedua dilakukan dengan melakukan pengujian model dengan tiga algoritma yang berbeda yaitu *Random forest*, ID3, dan C45 dengan metode PCA. Penggunaan PCA dimaksudkan untuk mereduksi dimensi dalam dataset dan melakukan interpretasi data dengan cepat. Hasil dari skenario ini dapat dilihat pada tabel 2 berikut:

Table 2 Skenarion Random Forest, ID3, C45 dan PCA

Skenario	*Accuracy*	*Precision*	*Recall*	F1 Score	AUC
Random forest + 7 PCs	82	41	50	45	69
ID3 + 7 PCs	**84**	**46**	**70**	**56**	**78**
C4.5 + 7 PCs	81	40	60	48	72
Random forest + 8 PCs	86	55	50	52	72
ID3 + 8 PCs	85	50	60	54	75
C4.5 + 8 PCs	82	42	60	50	73
Random forest + 9 PCs	86	55	50	52	72
ID3 + 9 PCs	82	40	40	40	65
C4.5 + 9 PCs	79	33	40	36	63
Random forest + 10 PCs	85	50	60	54	75
ID3 + 10 PCs	82	40	40	40	65
C4.5 + 10 PCs	82	41	50	45	69
Random forest + 11 PCs	85	50	50	50	71
ID3 + 11 PCs	85	50	50	50	71
C4.5 + 11 PCs	79	35	50	41	67
Random forest + 12 PCs	85	50	50	50	71
ID3 + 12 PCs	86	55	50	52	72
C4.5 + 12 PCs	85	50	50	50	71

***Random forest*, ID3, C.45 + SMOTE**

Berdasarkan hasil perbandingan data latih orang yang layak dan tidak layak dengan data latih, diperoleh data tidak seimbang pada angka 175 dan 76. Data yang tidak seimbang mengakibatkan nilai akurasi yang kurang optimal pada kelas minoritas. Untuk itu, skenario ketiga dilakukan dengan menggunakan metode SMOTE dan menguji model dengan tiga algoritma berbeda: *Random forest*, ID3, dan C45. Hasil dari skenario ini dapat dilihat pada tabel 3 berikut:

Table 3 Skenarion Random Forest, ID3, C45 dan SMOTE

Skenario	*Accuracy*	*Precision*	*Recall*	F1 Score	AUC
Random forest	**89**	**91**	**88**	**82**	**90**
ID3	82	84	81	75	82
C4.5	85	88	81	70	85

***Random forest*, ID3, C.45 + PCA + SMOTE**

Untuk mencapai hasil yang maksimal, skenario keempat dilakukan dengan menguji model dengan tiga algoritma berbeda: *Random forest*, ID3, dan C45 menggunakan teknik PCA dan SMOTE. Kedua metode ini bertujuan untuk mengatasi ketidakseimbangan data sekaligus mengurangi dimensi data untuk mencapai hasil yang lebih akurat. Hasil dari skenario ini dapat dilihat pada tabel 4 berikut:

Table 4 Skenario Random Forest, ID3, C45 dengan SMOTE dan PCA

Skenario	*Accuracy*	*Precision*	*Recall*	F1 Score	AUC

RF + 7PCS	87	84	89	87	88
ID3+ 7PCS	75	69	83	75	76
RF + 8PCS	88	86	89	87	89
ID3+ 8PCS	74	68	81	74	75
C4.5+ 8PCS	78	75	79	77	78
RF + 9PCS	89	88	89	88	90
ID3+ 9PCS	75	69	83	75	76
C4.5+ 9PCS	77	71	83	77	78
RF + 10PCS	**90**	**86**	**93**	**90**	**91**
ID3+ 10PCS	78	73	83	78	78
C4.5+ 10PCS	76	70	85	77	77
RF + 11PCS	89	86	86	91	89
ID3+ 11PCS	71	65	81	72	72
C4.5+ 11PCS	76	70	85	77	77
RF + 12PCS	89	83	95	89	90
ID3+ 12PCS	76	72	79	75	76
C4.5+ 12PCS	78	73	83	78	78

Random forest, ID3, C.45 + Features Importance

Skenario kelima dilakukan dengan melakukan pengujian model dengan tiga algoritma yang berbeda yaitu *Random forest*, ID3, dan C45 dengan metode Feature Importance. Penggunaan Feature Importance dimaksudkan untuk mereduksi dimensi dalam dataset dan melakukan intrepetasi data dengan cepat Hasil dari skenario ini dapat dilihat pada tabel 5 berikut:

Table 5 Skenario Random Forest, ID3, C45 dan Features Importance

Skenario	***Accuracy***	***Precision***	***Recall***	**F1 Score**	**AUC**
Random forest	**92**	**99**	**85**	**92**	**93**
ID3	89	94	85	89	90
C4.5	91	92	91	91	91

Evaluation

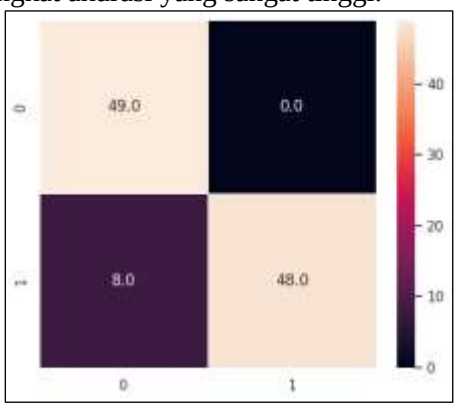
Setelah melakukan pemodelan dengan kelima scenario, diambil tiap-tiap scenario model mana yang memiliki Tingkat evaluasi paling baik. Berikut hasilnya pada tabel dibawah ini:

Table 6 Hasil *Perbandingan Skenario*

No	**Skenario**	***Accuracy***	***Precision***	***Recall***	**F1 Score**	**AUC**
1	*Random forest* (RF)	81%	64%	59%	61%	58%
2	ID3 + 7 PCs	84	46	70	56	78
3	*Random forest+SMOTE*	85%	87%	84%	81%	86%
4	RF + SMOTE + 10PCS	90%	86%	93%	90%	91%
5	**RF + SMOTE + *Features Importance***	**92%**	**99%**	**85%**	**92%**	**93%**

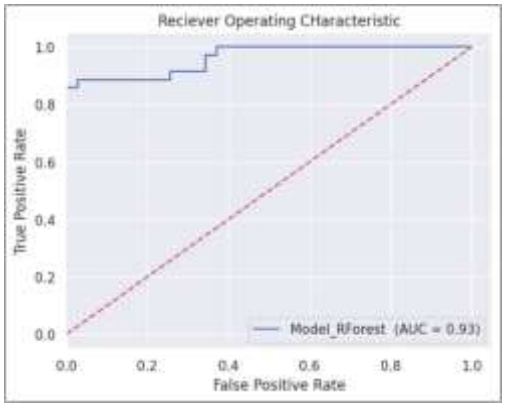
Berdasarkan hasil perbandingan skenario diatas, dapat disimpulkan bahwa algoritma yang menunjukan performa terbaik ditunjukan pada model Algoritma *Random forest* dengan SMOTE dan *Features Importance*. Dengan hasil accuracy 92%, precision 99%, recall 85%, f1-score 92%, dan AUC 93%. Dengan acuan f1-score yang paling tinggi dibandingkan dengan algoritma yang lain. F1-

score digunakan sebagai acuan karena dataset memiliki ketidakseimbangan kelas, Metrik ini memadukan precision dan recall, yang berguna untuk keadaan kelas minoritas lebih penting untuk di identifikasi dengan benar. Didapatkan juga nilai Nilai AUC sebesar 93 menunjukan performa model secara keseluruhan memiliki Tingkat akurasi yang sangat tinggi.



Gambar 1 Confusion Matrix

Pada gambar 2 confusion matrix di atas kelas 0 menggambarkan kelas Layak menerima bantuan, kelas 1 menggambarkan kelas Tidak Layak mendapat bantuan. Dari hasil confusion matrix model *Random forest* dengan SMOTE dan Features Importance didapatkan nilai True positive 49 data, False negative sebanyak 0 data, False positive 8 data, dan True negative sebanyak 48 data

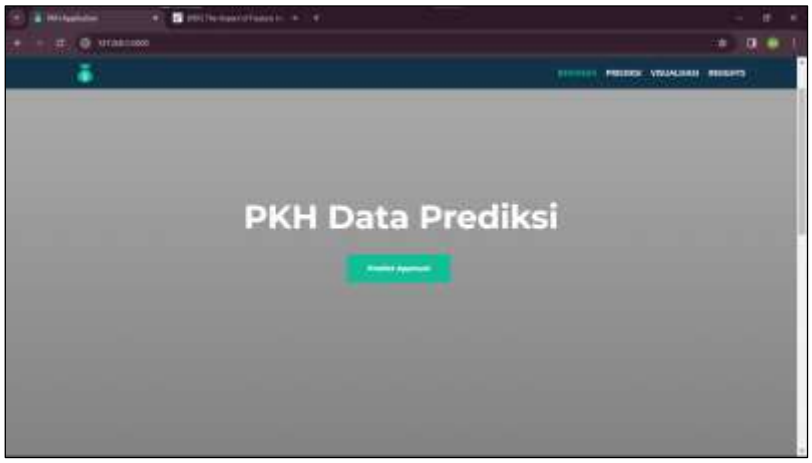


Gambar 2 ROC/AUC

Hasil dari perhitungan nilai Kurva AUC pada gambar 4.34 untuk model *Random forest* +SMOTE + Features Importance adalah sebesar 0.93. Berdasarkan penelitian yang dilakukan Gorunescu tahun 2011 terkait panduan untuk mengklasifikasi nilai keakuratan test nilai AUC, maka klasifikasi dengan model *Random forest* + SMOTE + Features Importance termasuk kedalam klasifikasi yang memiliki tingkat akurasi sangat baik.

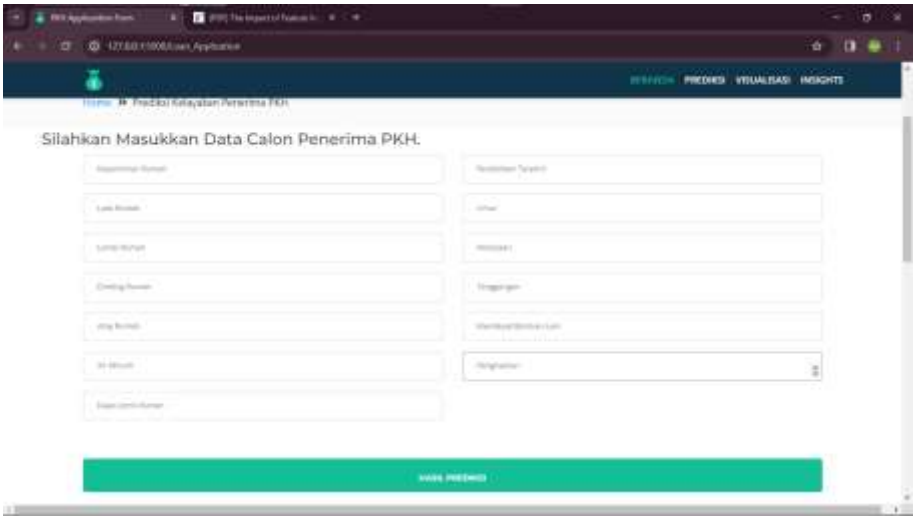
Deployment

Proses deployment dilakukan dengan menerapkan model terbaik tersebut kedalam aplikasi web. Software yang akan digunakan adalah bahasa pemrograman Pyhton, web browser untuk menjalankan system, dan Flask sebagai kerangka kerja untuk pengembangans website. Kebutuhan fungsional dari sistem ini adalah website dapat melakukan klasifikasi penerima pkh dan menampilkan hasil klasifikasi penerima pkh.



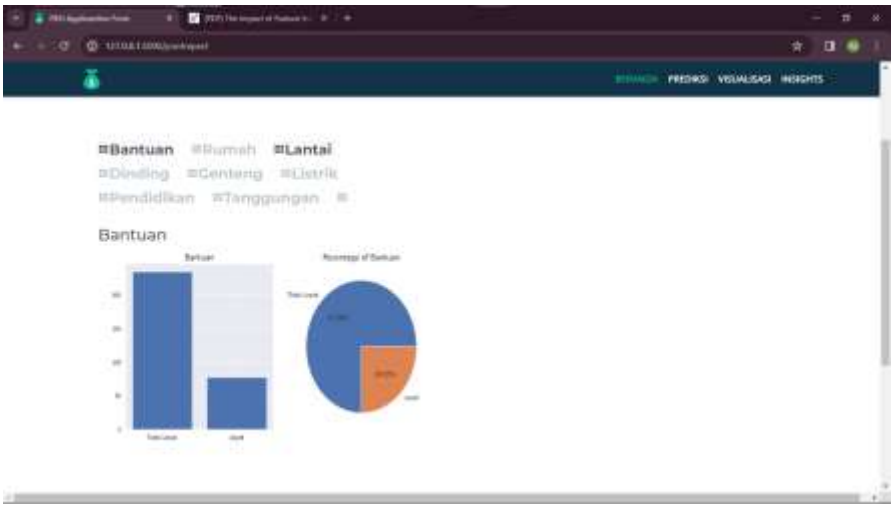
Gambar 3 Beranda Aplikasi

Pada gambar 4 menampilkan halaman awal menampilkan informasi menu apa saja yang ada pada website aplikasi PKH, terdiri dari menu Prediksi, Visualisasi dan Beranda.



Gambar 4 Halaman Prediksi

Pada gambar 5 menampilkan halaman prediksi pengguna diminta untuk memasukan data calon penerima mulai dari kepemilikan rumah, luas, pekerjaan, penghasilan, tanggungan, dan lain-lain. Dari hasil yang dimasukan kemudian akan di prediksi oleh sistem.



Gambar 5 Halaman Visualisasi

Pada gambar 6 menampilkan halaman visualisasi akan menampilkan gambar diagram dari *feature-feature* yang digunakan dalam pemodelan, seperti Bantuan, Rumah, Lantai, dan lainnya.

Uji Validitas Sistem

Setelah berhasil mengembangkan sistem maka perlu dilakukan proses uji validitas sistem, untuk mengetahui sistem tersebut bisa digunakan atau tidak untuk proses klasifikasi, berikut adalah hasil uji validitas sistem

Table 7 Confusion Matrix Uji Validitas Sistem

	Layak	**Tidak Layak**
Layak	6 (TP)	1 (FP)
Tidak Layak	1 (FN)	16 (TN)

Setelah mendapat nilai *confusion matrix*, kemudian menghitung nilai *accuracy* sebagai berikut :

$$accuracy = \frac{\text{TP} + \text{TN}}{TP+TN+FP+FN}$$

$$accuracy = \frac{6+16}{6+16+1+1}$$

$$accuracy = \frac{22}{24} = 0.91$$

$$accuracy = 0.91\ x\ 100\% = 91\ \%$$

Dari hasil uji validitas sistem berhasil mendapatkan akurasi sebesar 91%, dari hasil tersebut dapat disimpulkan model yang sudah di implementasikan pada sistem sesuai dengan hasil evaluasi *modelling*.

SIMPULAN

Berdasarkan penelitian yang sudah dilakukan dalam merancang model klasifikasi menggunakan metode CRISP-DM algoritma terbaik adalah Random Forest dengan SMOTE dan Features Importance. Penggunaan SMOTE berhasil dalam meningkatan evaluasi performa setiap model yang memiliki dataset tidak seimbang. Penggunaan PCA dan Featues Importance sama-sama meningkatkan evaluasi performa model. Dalam hal ini Features Importance memiliki nilai evaluasi yang lebih tinggi daripada PCA. Random Forest dengan SMOTE dan Features Importance berhasil menghasilkan accuracy sebesar 92%, precision 99%, recall 85%, f1 score 92%, dan nilai AUC sebesar 93%. Nilai ini lebih tinggi sedikit dari model Random Forest dengan SMOTE dan 10 PCs yang menghasilkan nilai accuracy sebesar 90%, precision 86%, recall 93%, f1 score 90% dan nilai AUC 91%. Dari hasil uji validitas model Random Forest+Smote+Features Importance memiliki akurasi sebesar 91%, hal ini menunjukan bahwa model berhasil melakukan klasifikasi warga calon penerima Program Keluarga Harapan.

UCAPAN TERIMA KASIH

Penulis mengucapkan puji syukur kehadirat Allah SWT yang telah memberikan kelancaran sehingga penelitian ini dapat terselesaikan dengan baik. Penulis mengucapkan terima kasih kepada Bu Amalia Anjani dan Bu Rizka Hadiwiyanti, selaku dosen yang telah membimbing hingga penelitian ini berhasil. Penulis juga mengucapkan terima kasih kepada orang tua dan adik yang telah memberikan semangatnya sehingga penelitian ini dapat berjalan dengan cepat.

REFERENSI

[1] T. Mauritsius and F. Binsar, "Cross-Industry Standard Process for Data Mining (CRISP-DM)," Binus University Graduate Program, 18 09 2020. [Online]. Available:

https://mmsi.binus.ac.id/2020/09/18/cross-industry-standard-process-for-data-mining-crisp-dm/. [Accessed 15 11 2022]

[2] L. Breiman, "Random Forests," *Machine Learning*, pp. 5-32, 2001.

[3] O. Kristanto, Penerapan Algoritma Klasifikasi Data Mining ID3 Untuk Menentukan Penjurusan Siswa SMAN 6 Semarang, Semarang: Universitas Dian Nuswantoro Semarang, 2014.

[4] H. B. Umar, "PRINCIPALCOMPONENTANALYSIS(PCA) DAN APLIKASINYA DENGAN PCA," *Jurnal Kesehatan Masyarakat*, pp. 97-101, 2009.

[5] R. Siringoringo, "KLASIFIKASI DATA TIDAK SEIMBANG MENGGUNAKAN SMOTE DAN K-NEAREST NEIGHBOR," 2018.

[6] K. R. Rub G, "“Feature Selection for Wheat Yield," *Research and Development in Intelligent* , pp. 465-478, 2010.

[7] K. S. Nugroho, "Medium," 13 November 2019. [Online]. Available: https://ksnugroho.medium.com/confusion-matrix-untuk-evaluasi-model-pada-unsupervised-machine-learning-bc4b1ae9ae3f.