JTERA (Jurnal Teknologi Rekayasa), Vol. 7, No. 1, Juni 2022, Hal. 149-156
Terakreditasi "**Peringkat 3**" oleh Kemenristek/BRIN, Nomor SK: 85/M/KPT/2020
DOI: 10.31544/jtera.v7.i1.2022.149-156
p-ISSN 2548-737X
e-ISSN 2548-8678

Implementasi *Data Mining* untuk Klasterisasi dan Prediksi Kelompok Keluarga

Imam Sapuan[#], Muhammad Hilmi Fauzan, Christina Juliane
Program Studi Magister Sistem Informasi, STMIK LIKMI Bandung
Jl. Ir. H. Juanda No. 96, Bandung 40132, Jawa Barat, Indonesia
[#]imam.sapuan@unpas.ac.id

Abstrak

Pengelompokkan keluarga ke dalam *cluster* mampu dan tidak mampu sangat diperlukan untuk acuan berbagai kegiatan di masa depan seperti bantuan pemerintah atau pihak terkait lainnya. *Data mining* merupakan salah satu pendekatan yang dapat digunakan untuk menyelesaikan masalah ini. Metode *data mining* yang cocok adalah *clustering* dan prediksi. Penelitian ini bertujuan untuk mengimplementasikan *data mining* untuk klasterisasi dan prediksi kelompok keluarga. Terdapat dua algoritma yang digunakan pada penelitian ini, yaitu *kModes* dan *decision tree*. Algoritma *kModes* berfungsi untuk menghasilkan *cluster* yang akan digunakan pada tahap selanjutnya, sedangkan metode *decision tree* digunakan sebagai algoritma prediksinya. Hasil penelitian menunjukkan bahwa kedua metode ini berhasil menyelesaikan masalah dengan tingkat akurasi yang sangat tinggi yaitu sebesar 95,3%, presisi sebesar 95,4%, dan *recall* sebesar 95,3%.

Kata kunci: *data mining*, *clustering*, *kModes*, *decision tree*, kelompok keluarga

Abstract

The grouping of families into rich and poor clusters is very much needed as a reference for various future activities such as government assistance or other related parties. Data mining is one approach that can be used to solve this problem. Data mining methods that are suitable are clustering and prediction. This study aims to implement data mining for clustering and predicting family groups. There are two algorithms used in this study, namely kModes and decision tree. The kModes algorithm functions to generate clusters that will be used in the next stage, while the decision tree method is used as the prediction algorithm. The results showed that these two methods succeeded in solving the problem with a very high level of accuracy, namely 95.3%, precision 95.4%, and recall of 95.3%.

Keywords: *data mining, clustering, kModes, decision tree, the grouping of families*

I. PENDAHULUAN

Pendidikan berkarakter merupakan salah satu fasilitas yang menunjang peradaban tinggi [1]. Namun status sosial antara kaya dan miskin sangat mempengaruhi luaran keduanya, karena kemiskinan akan membatasi akses seseorang pada berbagai fasilitas salah satunya pendidikan yang berkualitas, begitu juga sebaliknya [2]. Kemiskinan dinilai sebagai ketidakmampuan untuk memenuhi kebutuhan dasar makanan dan bukan makanan yang diukur dari sisi pengeluaran (*The World Bank* 2009). Garis kemiskinan merupakan tingkat minimum pendapatan yang dianggap harus dipenuhi untuk memperoleh standar yang mencukupi. Garis ini berfungsi untuk mengukur rakyat miskin dan mempertimbangkan pembaharuan sosio-ekonomi, seperti program penanggulangan kemiskinan dengan cara meningkatkan kesejahteraan [3]. Untuk menanggulangi masalah kemiskinan ini, salah satu program yang telah dilaksanakan pemerintah adalah Kartu Indonesia Pintar [4]. Sayangnya, berbagai program untuk masyarakat miskin ternyata masih banyak yang salah sasaran [5], sehingga penting untuk mengelompokkan keluarga mampu dan kurang mampu. Pengelompokkan ini juga dapat digunakan sebagai rekomendasi berbagai kegiatan di masa depan agar tepat guna dan presisi [6]. Salah satu teknik yang dapat digunakan untuk kegiatan tersebut adalah *data mining*.

Data mining adalah ilmu untuk menemukan pengetahuan, wawasan, dan pola dalam data. Fungsi

Diterima: 25 April 2021; Direvisi: 4 Juni 2022; Disetujui: 13 Juni 2022
JTERA, Vol. 7, No. 1, Juni 2022

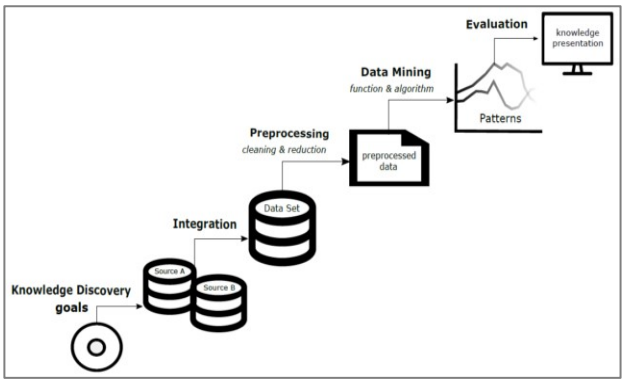
data mining adalah mengekstraksi pola yang berguna dari kumpulan data yang terorganisir. Pola yang dihasilkan harus *valid*, baru, berpotensi berguna, dan dapat dimengerti. Secara sederhana, hasil *data mining* dapat memproyeksikan masa depan menggunakan data masa lalu [7]. Beberapa tugas *data mining* adalah klasifikasi, *clustering*, prediksi, asosiasi, dan estimasi [8].

Banyak penelitian yang menunjukkan bahwa algortima *k-Means* dan *decision tree* berhasil menyelesaikan berbagai permasalahan seperti mengelompokkan data penduduk miskin di Kota Jambi [9], menerapkan metode *decision tree* untuk klasifikasi warga miskin pada desa mengandung sari dengan akurasi 100% [10], menerapkan metode *decision tree* untuk klasifikasi data peserta didik [11], menerakan metode *k-Means* untuk pemetaan penyebaran guru di provinsi Banten [12], dan pemetaan kualitas pendidikan di Indonesia [13].

Pada penelitian ini, metode yang digunakan adalah *k-Means* untuk *clustering* dan *decision tree* sebagai prediksi. Tahap pertama, *dataset* akan dikelompokkan ke dalam dua *cluster* menggunakan *k-Means* selanjutnya *cluster* tersebut akan diubah menjadi *label/target*, kemudian diprediksi menggunakan algoritma *decision tree* untuk kebutuhan di masa depan yang tepat dan akurat. *Tool* yang digunakan adalah Orange dan Jupyter Notebook. *Software* Orange yang berbasis Python ini sangat cocok karena *open source*, *drag* and *drop*, *mudah digunakan dan user friendly. Selain itu, Orange* mendukung *file* dengan ekstensi xlsx, csv, tab, data bersifat *online* seperti Google Spreadsheet, PostgreSQL, dan MSSQL [14]. Jupyter Notebook merupakan *software* yang berbasis Python yang penggunaannya lebih manual dibandingkan Orange. Jupyter Notebook digunakan untuk melakukan proses algoritma *k-Mode*, sedangkan Orange digunakan untuk melakukan proses algoritma *decision tree.*

II. METODE PENELITIAN

Ada berbagai metode yang dapat digunakan dalam proses data mining, yaitu *Knowledge Discovery on Database* (KDD), SEMMA, dan Crisp DM [15]. KDD merupakan proses mengekstraksi informasi baru dan pengetahuan dari *database* yang berukuran besar, diusulkan oleh Fayyad pada tahun 1996. Sedangkan SEMMA dan Crisp DM lebih populer digunakan pada berbagai penelitian [15]. Secara garis besar, proses ekstraksi pengetahuan dari metode-metode tersebut dapat dilihat pada Gambar 1. Tahapannya adalah *knowledge discovery goals*, *integration*, *pre-processing*, *data mining*, dan *evaluation* [16].



Gambar 1. Alur proses *data mining*

A. Knowledge Discovery Goals

Pada tahap pertama ini, tujuan dilakukannya *data mining* harus ditentukan. Tujuannya adalah melakukan *clustering* pada data kemiskinan untuk mengelompokkan keluarga mampu dan kurang mampu menggunakan algoritma *k-Means*, kemudian hasil *clustering* akan diprediksi menggunakan *decision tree* untuk kebutuhan di masa depan yang tepat dan akurat.

B. Data Integration

Dataset yang digunakan adalah data riwayat hidup siswa Yayasan Al-Muhajirin yang terdiri dari 16 atribut. Atribut yang digunakan ada 9 karena sesuai dengan proses yang dibutuhkan untuk *data mining*, terdiri dari nama siswa, saudara kandung, jenjang saat ini, jenis kelamin, pekerjaan ayah, kelompok penghasilan ayah, penghasilan ayah, pekerjaan ibu, dan penghasilan ibu. Data ini berisi 235 *record* dengan ekstensi *file* .xlsx, tidak terdapat *label* pada *dataset*, sehingga cocok untuk *clustering*.

C. Data Pre-Processing

Tahapan ini dilakukan untuk mengeksplorasi dan menganalisis data, disebut juga dengan *Exploratory Data Analysis* (EDA). Kegiatan yang dilakukan adalah mengetahui tipe data yang digunakan, memeriksa sebaran dan distribusi data, mengatasi *missing value*, mengubah nama atribut jika diperlukan, mendeteksi nilai *outlier*, melakukan normalisasi atau standarisasi data jika dibutuhkan, transformasi data, dan reduksi data. Tahapan ini sangat penting dilakukan karena hasil pembelajaran mesin sangat tergantung pada tahap ini [17].

D. Seleksi Atribut Manual

Tahapan ini berfungsi untuk menghilangkan atribut yang tidak sesuai dengan tujuan data mining. Hal ini dilakukan untuk memaksimalkan proses belajar algortima dan mengurangi biaya komputasi agar lebih cepat dan hemat. Perbedaan sebelum dan sesudah dilakukan seleksi atribut dapat dilihat pada Tabel 1.

Tabel 1. Seleksi atribut manual

Atribut sebelum seleksi	Atribut setelah seleksi
Anak Ke	Anak Ke
Saudara Kandung	Saudara Kandung
Jenjang Saat Ini	Jenjang Saat Ini
Pekerjaan Ayah	Pekerjaan Ayah
Kelompok Penghasilan Ayah	Kelompok Penghasilan Ayah
Penghasilan Ayah	Penghasilan Ayah
Pekerjaan Ibu	Pekerjaan Ibu
Penghasilan Ibu	Penghasilan Ibu
No.	-
Nama Siswa	-
Nama Ayah	-
Alamat	-
Jenis Kelamin	-

Tabel 2. Tipe *data dataset*

Nama Atribut	Tipe Data
Anak Ke	*Categorical*
Saudara Kandung	*Categorical*
Jenjang Saat Ini	*Categorical*
Pekerjaan Ayah	*Categorical*
Kelompok Penghasilan Ayah	*Categorical*
Penghasilan Ayah	*Categorical*
Pekerjaan Ibu	*Categorical*
Penghasilan Ibu	*Categorical*

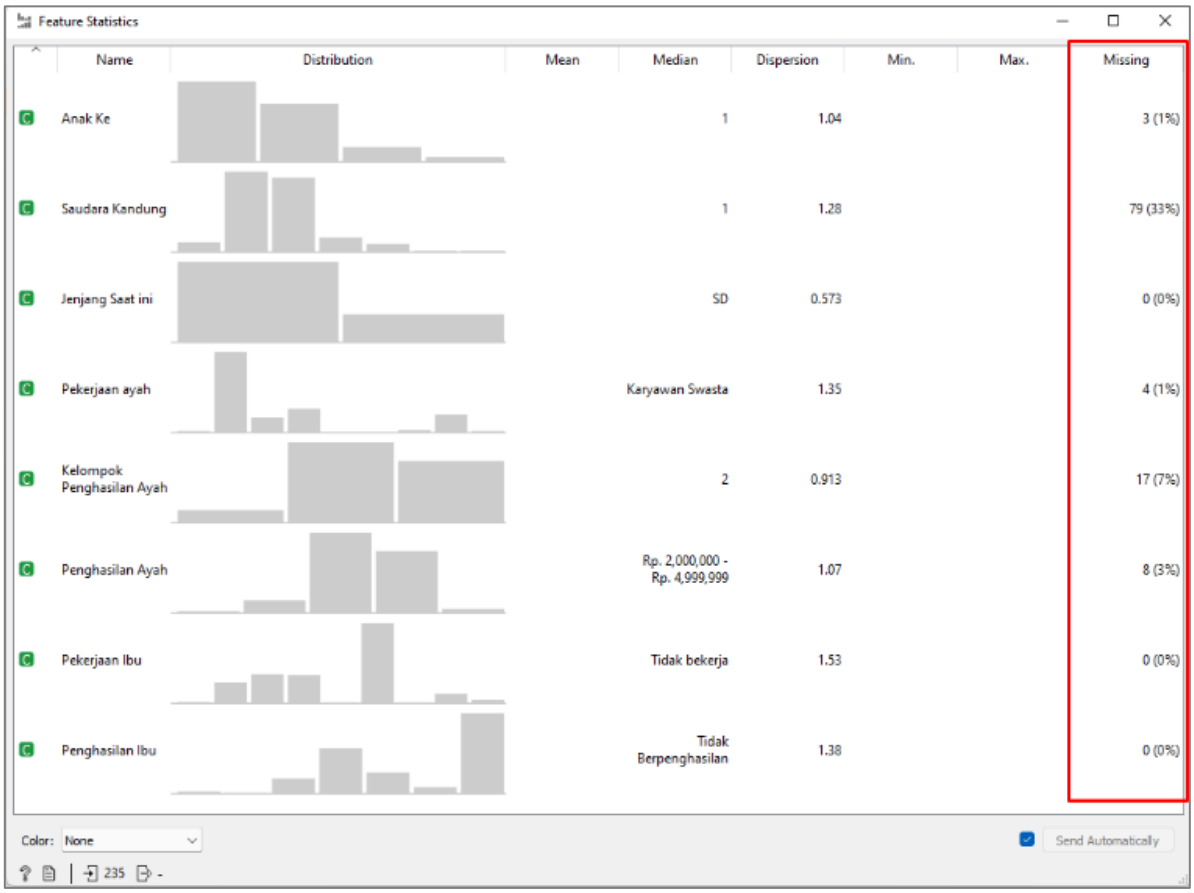
E. Tipe Data

Tipe data sangat penting diketahui karena setiap algoritma *data mining* memiliki cara kerja yang berbeda sesuai dengan tipe data yang diprosesnya. Tipe data pada dataset ini dapat diuraikan pada Tabel 2. Berdasarkan Tabel 2, algoritma *k-Mode* sangat cocok untuk menanggulangi data yang berjenis kategorik.

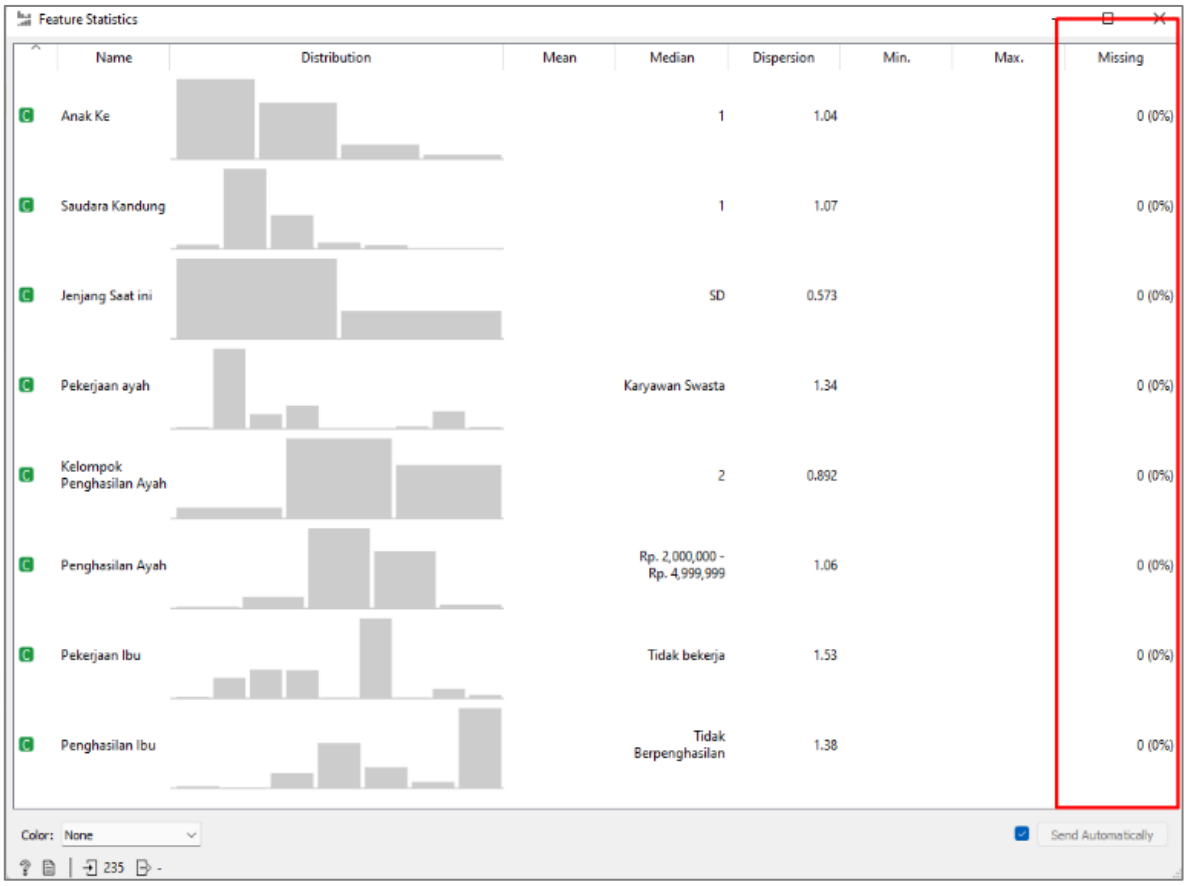
F. Imputasi Missing Value dan Distribusi Data

Distribusi data dilakukan untuk melihat sebaran data pada tiap *record* yang ada. Imputasi artinya mengisi nilai yang kosong. Terdapat beberapa teknik yang bisa digunakan untuk menangani *missing value*, seperti menghapus baris data, mengisi secara acak, dan mengisinya dengan nilai tengah (*mean, median*, atau *modus*) [18]. Pada penelitian ini, teknik yang digunakan untuk melakukan imputasi adalah dengan mengisinya menggunakan nilai tengah.

Berdasarkan data yang ditampilkan pada Gambar 2 terlihat banyaknya *missing value* pada setiap atribut berbeda-beda. Setelah dilakukan proses imputasi maka hasilnya dapat dilihat pada Gambar 3.



Gambar 2. Visualisasi *missing value* pada *dataset*



Gambar 3. Visualisasi *missing value* setelah diimputasi

G. Mengubah Nama Atribut, Mendeteksi Nilai Outlier, dan Normalisasi Data

Pada penelitian ini, nama atribut tidak diubah karena penamaan dapat dipahami dengan mudah sehingga tidak perlu melakukan proses ini. Selain itu, karena tipe data bersifat kategorik sehingga tidak terlihat nilai outlier dan tidak perlu melakukan normalisasi (penskalaan) data.

H. Data Mining

Ini merupakan tahapan inti dari penelitian. Ada dua kegiatan *data mining* yang dilakukan yaitu *clustering* dengan *k-Mode* dan prediksi menggunakan *decision tree*. *Clustering k-Mode* dilakukan menggunakan *software* Jupyter Notebook dengan nilai k = 2, artinya akan menghasilkan dua *cluster* yaitu *cluster* mampu dan kurang mampu. Tampilan baris kode yang dilakukan pada Jupyter Notebook dan hasilnya dapat dilihat pada Gambar 4.

```
km=KModes(n_clusters=2,max_iter=5)
clusters=km.fit_predict(data_x)
clusters
```

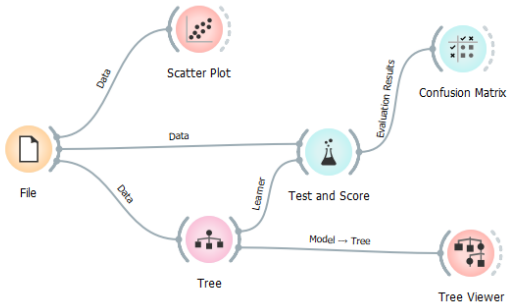
```
array([1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0,
       0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0,
       0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0,
       1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1,
       0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0,
       0, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0,
       0, 1, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0,
       1, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0], dtype=uint16)
```

```
[82] print(km.cluster_centroids_)
```

```
[['1.0' '1.0' 'SD' 'L' 'Karyawan Swasta' '2.0'
  'Rp. 2,000,000 - Rp. 4,999,999' 'Tidak bekerja' 'Tidak Berpenghasilan']
 ['2.0' '1.0' 'SD' 'P' 'Karyawan Swasta' '3.0'
  'Rp. 5,000,000 - Rp. 20,000,000' 'Lainnya'
  'Rp. 5,000,000 - Rp. 20,000,000']]
```

Gambar 4. *Clustering* menggunakan *k-Modes* pada Jupyter Notebook

Setelah proses *clustering*, *dataset* tersebut akan dilakukan prediksi menggunakan *decision tree*. Label atau target yang digunakan pada *dataset* berasal dari *cluster* yang dihasilkan pada kegiatan sebelumnya. Proses kegiatan prediksi ini dilakukan menggunakan *software* Orange dan hasilnya dapat dilihat pada Gambar 5.



Gambar 5. Prediksi menggunakan *decision tree* pada Orange

Tabel 3. *Confusion matrix*

	Actual Positive	***Actual Negative***
Predic. Positive	TP	FP
Predic. Negative	FN	TN

I. Evaluation

Tahapan ini berfungsi untuk mengetahui evaluasi hasil belajar yang dilakukan oleh algoritma *decision tree*. *Confusion matrix* merupakan salah satu metode evaluasi yang dapat digunakan dalam kegiatan berbasis klasifikasi dan prediksi [19]. *Confusion matrix* menghasilkan nilai yang dapat dilihat pada Tabel 3.

TP artinya *True Positive*, yaitu hasil prediksi benar dan sesuai dengan aktual benar. FP adalah *False Positive*, yaitu hasil prediksi benar tidak sesuai dengan aktualnya salah. FN adalah *False Negative*, yaitu hasil prediksi salah tidak sesuai dengan aktualnya benar. Terakhir, TN adalah *True Negative*, yaitu hasil prediksi salah sesuai dengan aktual salah.

Accuracy (A) mendeskripsikan seberapa akurat model algortima dapat memprediksi dengan benar, sehingga *accuracy* merupakan rasio prediksi benar (baik benar secara positif maupun negatif) dengan keseluruhan data. Dengan kata lain, *accuracy* merupakan tingkat kedekatan nilai prediksi dengan nilai aktual. Rumus *accuracy* dapat dilihat sebagai berikut:

$$A = \frac{(TP+TN)}{(TP+TN+FP+FN)} x\ 100\% \qquad (1)$$

Precision (P) menggambarkan tingkat keakuratan antara data yang diminta dengan hasil prediksi yang diberikan oleh model. Maka, *precision* merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif. Dari semua kelas positif yang telah di prediksi dengan benar, berapa banyak data yang benar-benar positif.

$$P = \frac{TP}{FP+TP} x100\% \qquad (2)$$

Recall (R) menggambarkan keberhasilan model dalam menemukan kembali sebuah informasi. Maka, *recall* merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif.

$$R = \frac{TP}{TP+FN} x100\% \qquad (3)$$

F1 Score merupakan perbandingan rata-rata presisi dan *recall* yang dibobotkan.

$$F1 = 2 * \frac{(Recall*Precision)}{(Recall+Precision)} \qquad (4)$$

III. HASIL DAN PEMBAHASAN

Bagian ini merupakan hasil yang sudah dikerjakan pada tahapan sebelumnya. Algortima *k-Modes* yang sudah dijalankan pada Jupyter Notebook menghasilkan dua *cluster*, yaitu mampu dan kurang mampu. *Cluster* 1 mewakili kelompok mampu dan *cluster* 0 mewakili kelompok tidak mampu. Hasilnya dapat dilihat pada Gambar 6.

Label yang akan digunakan pada proses prediksi adalah hasil *cluster* dari kegiatan sebelumnya. Setelah model dibuat dengan *software* Orange, hasil pohon keputusan *decision tree* dapat dilihat pada Gambar 7.

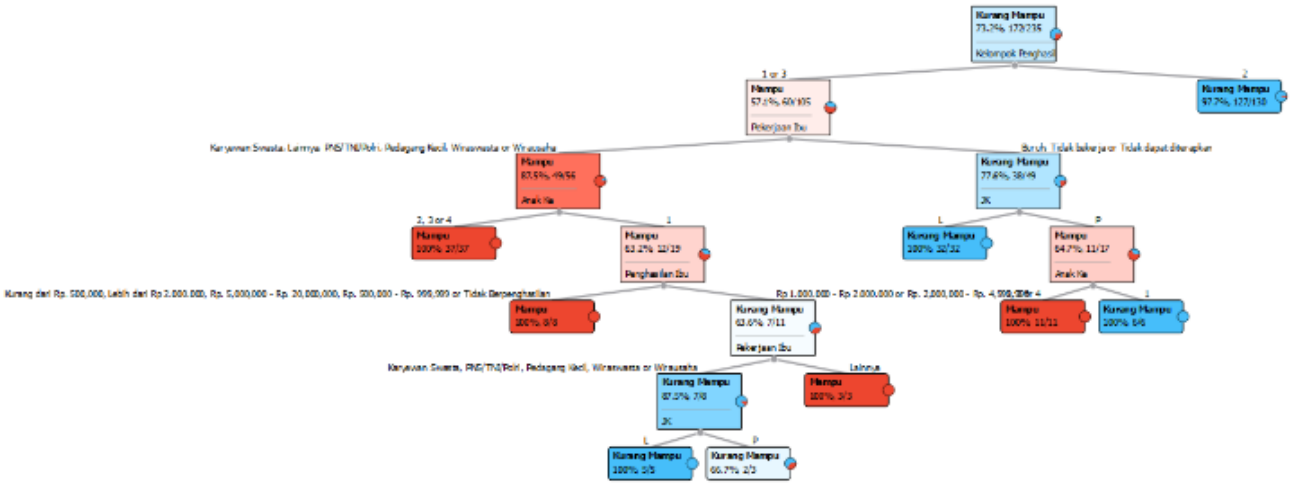
Kemudian, proses pengujian dilakukan dengan 10 *cross fold validation* yang memiliki hasil lebih baik dari *split validation* [20]. Proses metode ini pada *software* Orange dapat dilihat pada Gambar 8.

Dari proses pengujian tersebut didapatkan hasil *confusion matrix* yang dapat dilihat pada Gambar 9. *Confusion matrix* dapat digunakan untuk menghitung akurasi, presisi dan *recall*. Nilai akurasi yang didapatkan sebesar 95,3%, presisi sebesar 95,4%, dan *recall* sebesar 95,3%.

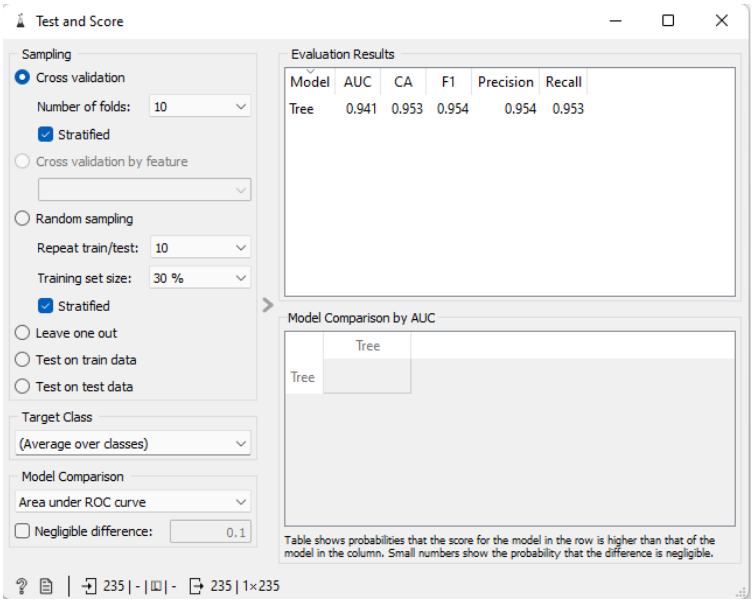
	Nama	Anak Ke	Saudara Kandung	Jenjang Saat ini	JK	Pekerjaan ayah	Kelompok Penghasilan Ayah	Penghasilan Ayah	Pekerjaan Ibu	Penghasilan Ibu	CLUSTER
0	Salma Nufah Maulida	3.0	4.0	SD	P	PNS/TNI/Polri	3.0	Rp. 5,000,000 - Rp. 20,000,000	Wiraswasta	Rp. 5,000,000 - Rp. 20,000,000	1
1	Rizki Anggraha Mahmudin	1.0	4.0	SMP	L	PNS/TNI/Polri	2.0	Rp. 2,000,000 - Rp. 4,999,999	Wiraswasta	Rp. 2,000,000 - Rp. 4,999,999	0
2	AHMAD KAMIL ASSEGAF	1.0	1.0	SD	L	Wiraswasta	2.0	Rp. 2,000,000 - Rp. 4,999,999	Tidak bekerja	Tidak Berpenghasilan	0
3	Ali Zainal Abidin Assegaf	2.0	1.0	SD	L	Wiraswasta	3.0	Rp. 5,000,000 - Rp. 20,000,000	Tidak bekerja	Tidak Berpenghasilan	0
4	Bima Alvaro Maulana	3.0	2.0	SD	L	Karyawan Swasta	3.0	Rp. 5,000,000 - Rp. 20,000,000	Tidak bekerja	Tidak Berpenghasilan	0
...	...	...	...	...	...	...	...	...	...	...	...
230	Davina Aleana Firghi	1.0	2.0	SD	P	Lainnya	3.0	Rp. 5,000,000 - Rp. 20,000,000	Tidak bekerja	Tidak Berpenghasilan	0
231	DEVANO FABIAN FIRGHI	3.0	1.0	SD	L	PNS/TNI/Polri	1.0	Rp 1.000.000 - Rp 2.000.000	Tidak bekerja	Kurang dari Rp. 500,000	0
232	DIANDRA RAGIL FEBRIYANTO	2.0	2.0	SD	L	Wiraswasta	1.0	Rp 1.000.000 - Rp 2.000.000	Tidak bekerja	Tidak Berpenghasilan	0
233	MUHAMMAD EGA ALFARO	2.0	1.0	SD	L	Wiraswasta	3.0	Rp. 5,000,000 - Rp. 20,000,000	Tidak bekerja	Tidak Berpenghasilan	0
234	ARYA DINATA EKA AZHAR SAPUTRA	1.0	1.0	SMP	L	Wirausaha	2.0	Rp. 2,000,000 - Rp. 4,999,999	Karyawan Swasta	Rp. 2,000,000 - Rp. 4,999,999	0

235 rows × 11 columns

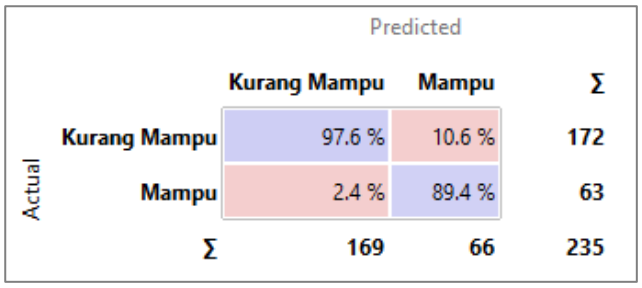
Gambar 6. Hasil *clustering* dengan *k-Mode* pada Jupyter Notebook



Gambar 7. Pohon keputusan yang dihasilkan dengan *software* Orange



Gambar 8. Sepuluh *cross fold validation* dengan *software* Orange



Gambar 9. *Confusion matrix* dengan *software* Orange

IV. KESIMPULAN

Clustering dengan algoritma *k-Modes* dan prediksi menggunakan algoritma *decision tree* terbukti dapat menyelesaikan masalah pada penelitian ini dengan tingkat akurasi yang sangat tinggi. Pada penelitian selanjutnya dapat dibandingkan algoritma lain seperti *Support Vector Machine* (SVM), kNN, *Neural Network*, dan *Random Forest* untuk dibandingkan kinerjanya dan dianalisis algoritma mana yang memiliki kinerja yang baik dalam kasus ini. Selain itu, beberapa algoritma *booster* seperti *adaboost* dan *xgboost* dapat dipakai untuk meningkatkan performa dan akurasinya.

REFERENSI

[1] S. L. Dewi, "Membangun Peradaban Bangsa Dalam Era Globalisasi Pendidikan Karakter," *PENDAS MAHAKAM: Jurnal Pendidikan dan Pembelajaran Sekolah Dasar*, vol. 4, no. 1, pp. 48-54, 2019.

[2] R. Al-Hamdi, "Ketika Sekolah Menjadi Penjara: Membongkar Dilema Pendidikan Masyarakat Modern," *J Soc Media*, vol. 1, no. 1, pp. 11-34, 2017.

[3] Y. C. Pratama, "Analisis Faktor-Faktor Yang Mempengaruhi Kemiskinan Di Indonesia," *Esensi*, vol. 4, no. 2, pp. 210–223, 2015.

[4] Kemendikbud RI, "Program Indonesia Pintar," https://indonesiapintar.kemdikbud.go.id/, 2021.

[5] I. A. Arbi, "Dari Pungli hingga Salah Sasaran Penerima, Ini Ragam Masalah Bansos di Jabodetabek," https://megapolitan.kompas.com/read/2021/07/30/08055781/dari-pungli-hingga-salah-sasaran-penerima-ini-ragam-masalah-bansos-di?page=all, 2021.

[6] E. Fammaldo and L. Hakim, "Penerapan Algoritma *K-Means Clustering* Untuk Pengelompokan Tingkat Kesejahteraan Keluarga Untuk Program Kartu Indonesia Pintar," *J Ilm Teknol Infomasi Terap*, vol. 5, no. 1, pp. 23–31, 2019.

[7] A. K. Maheshwari, *Business Intelligence and Data Mining*. Ferguson M, editor. Business Expert Press, LLC, 2015.

[8] P. S. Hasugian, "Penerapan *Data Mining* untuk Klasifikasi Produk Menggunakan Algoritma *K-Means* (Studi Kasus : Toko Usaha Maju Barabai)," *J Mantik Penusa*, vol. 2, no. 2, pp. 191–198, 2018.

[9] D. Sunia and P. A. J. Kurniabudi, "Penerapan *Data Mining* untuk *Clustering* Data Penduduk Miskin Menggunakan Algoritma *K-Means*," *J Ilm Mhs Tek Inform*, vol. 1, no. 2, pp. 121–134, 2016.

[10] C. E. Purnomo and Rikendry, "Penerapan metode c4.5 untuk klasifikasi warga miskin pada desa mengandung sari," vol. 2, no. 3, pp. 14–25, 2021.

[11] I. Sutoyo, "Implementasi Algoritma *Decision Tree* Untuk Klasifikasi Data Peserta Didik," *J Pilar Nusa Mandiri*, vol. 14, no. 2, pp. 217, 2018.

[12] Y. A. Priambodo, S. Y. J. Prasetyo, "Pemetaan Penyebaran Guru di Provinsi Banten dengan Menggunakan Metode *Spatial Clustering K-Means* (Studi kasus : Wilayah Provinsi Banten)," *Indones J Comput Model*, vol. 1, no. 1, pp. 18–27, 2018.

[13] G. S. Nugraha and Hairani, "Aplikasi Pemetaan Kualitas Pendidikan Di Indonesia Menggunakan Metode *K-Means*," *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 17, no. 2, pp. 13-23, 2018.

[14] Orange: *Data Mining Fruitful and Fun*. https://orangedatamining.com/, 2021.

[15] A. Azevedo and M. F. Santos MF. *KDD, SEMMA AND CRISP-DM : A PARALLEL OVERVIEW*. June 2014.

[16] H. Akbar, Ingin Terapkan *Data Mining*? Ini Tahapannya. https://mti.binus.ac.id/2017/12/05/ingin-terapkan-data-mining-ini-tahapannya/, 2017.

[17] S. A. Alasadi and W. S. Bhaya "Review of data preprocessing techniques in data mining," *J Eng Appl Sci*., vol. 12, no. 16, pp. 4102–4107, 2017.

[18] S. Kumar S, *7 Ways to Handle Missing Values in Machine Learning*. https://towardsdatascience.com/7-ways-to-handle-missing-values-in-machine-learning-1a6326adf79e, 2017.

[19] X. Deng, Q. Liu, Y. Deng, and S. Mahadevan, "An improved method to construct basic probability assignment based on the confusion matrix for classification problem," *Inf Sci (Ny)*, vol. 340–341, pp. 250–261, 2016.

[20] D. Shulga, "*Reasons why you should use Cross-Validation in your Data Science Projects*," https://towardsdatascience.com/5-reasons-why-you-should-use-cross-validation-in-your-data-science-project-8163311a1e79, 2018.