ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 14, Nomor 2, Tahun 2025, Halaman 512 - 523
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

# BOOTSTRAP AGGREGATING CLASSIFICATION AND REGRESSION TREES (BAGGING CART) UNTUK KLASIFIKASI POTENSI KARYAWAN RESIGN BERDASARKAN KENYAMANAN BEKERJA

**Muhammad Fajar Syabana[1], Tatik Widiharih[2*], Masithoh Yessi Rochayani[3]**
[1,2,3] Departemen Statistika, Fakultas Sains dan Matematika, Universitas Diponegoro
***e-mail** : widiharih@live.undip.ac.id

**DOI: 10.14710/j.gauss.14.2.512-523**

**Article Info:**
Received: 2024-12-26
Accepted: 2025-12-04
Available Online: 2025-12-07

**Keywords:**
*Employee; Resignation;Working Comfort; CART ; Bagging*

**Abstract:** Phenomenon of employee resignation is a significant challenge for companies because it affect the productivity and stability of the company's operations. Every companies supposed to analyze the potential of employee resignation. This research aims to classify the potential of employee resignation based on working comfort and applies the classification modeling method from Decision Tree: Classification And Regression Trees (CART) and the ensemble Bootstrap Aggregating (Bagging) method. CART is a non-parametric method that is effective in building classification and prediction models based on decision trees, while Bagging is an ensemble method that combines several CART models to improve the accuracy and stability of predictions. The CART model provides an accuracy of 73% and f1-score of 62%, while the Bagging CART model provides an accuracy of 87% and f1-score of 88%. This research shows an increase in accuracy when using Bagging CART model of 14%. The most important variable to build the model and make predictions is the age. Age is also used as the root node in building CART model.

## 1. PENDAHULUAN

Lingkungan kerja memiliki arti segala sesuatu yang berada di lingkungan para pekerja dan dapat memengaruhi dalam menjalankan tugas (Isyandi, 2004). Lingkungan kerja yang nyaman menjadi salah satu tolok ukur yang tidak bisa dikesampingkan bagi para pekerja maupun pencari kerja khususnya di era digital saat ini. Lingkungan kerja yang baik salah satunya memiliki ciri adanya *work-life balance* (Romalla, 2021).

Dunia kerja yang kompetitif dan dinamis, adanya karyawan yang memutuskan untuk mengundurkan diri (*resign*) adalah hal yang tidak dapat dihindari. Alasan setiap individu beragam dalam membuat keputusan untuk hengkang dari perusahaan. *Resign* dapat diartikan sebagai tindakan seseorang yang secara sukarela mengundurkan diri dari pekerjaan atau jabatannya di suatu perusahaan atau organisasi. Periode setelah mendapatkan bonus maupun THR merupakan waktu yang cukup rawan karyawan melakukan *resign* (Luqman, 2023).

*Resign* merupakan hak setiap karyawan dan sudah diatur dalam UU Nomor 13 Pasal 162 Ayat 3 Tahun 2003 tentang ketenagakerjaan. Keputusan *resign* dapat disebabkan oleh berbagai faktor yang melatarbelakanginya, baik dari dalam perusahaan ataupun dari karyawan (Fergian, 2020). Menurut Mathis dan Jackson (2006), salah satu faktor yang memengaruhi *resign* adalah hubungan karyawan. Selain faktor tersebut, terdapat sejumlah faktor pendorong terjadinya *turnover* atau *resign* adalah usia, lama kerja, dan ketidakpuasan kerja.

Fenomena pengunduran diri karyawan tentu akan memengaruhi kestabilan dan produktivitas operasional perusahaan. Penting bagi suatu perusahaan untuk melakukan analisis mendalam terkait faktor-faktor maupun fenomena pengunduran diri karyawan. Berdasarkan latar belakang fenomena *resign* ini, peneliti tertarik untuk melakukan klasifikasi dan prediksi karyawan yang berpotensi *resign* berdasarkan kenyamanan bekerja.

Penelitian ini menggunakan metode klasifikasi CART (*Classification And Regression Trees*). Metode ini pertama kali ditemukan oleh Breiman *et al.* (1984). Metode CART merupakan salah satu metode dari pohon keputusan (*decision tree*) yang perhitungannya cepat dan mudah diinterpretasikan untuk model yang dihasilkan. Metode CART ini akan dipadukan dengan metode *ensemble* Bagging (*Bootstrap Aggregating*) dalam menganalisis karyawan yang berpotensi *resign* dengan melihat tingkat kenyamanan bekerja.

Penelitian terdahulu oleh Sumartini dan Purnami (2015) menggunakan metode CART untuk klasifikasi rekurensi pasien kanker serviks, menghasilkan akurasi data prediksi sebesar 69,14% dan penelitian yang dilakukan Rohmah (2018) menggunakan metode *Bootstrap Agregating* (Bagging) CART pada pengklasifikasian IPM menghasilkan ketepatan klasifikasi sebesar 83,33%, sedangkan penelitian yang dilakukan Sa'adah *et al.* (2021) menggunakan Bagging Lasso Decision Tree (CART) menghasilkan akurasi sebesar 90,29%.

## 2. TINJAUAN PUSTAKA

Metode CART (*Classification and Regression Trees*) merupakan salah satu metode non-parametrik yang digunakan dalam melakukan analisis klasifikasi dengan menggunakan teknik *decision tree* atau pohon keputusan. Metode CART dapat digunakan pada variabel respon yang berskala kategorik maupun berskala kontinu. Variabel respon yang berskala kontinu akan menghasilkan pohon regresi *(regression trees)*, sedangkan jika berskala kategorik, maka model CART akan menghasilkan pohon klasifikasi *(classification trees)*.

CART diawali dengan pemilahan biner *(binnary splitting)* pada keseluruhan pengamatan, pemilahan pertama menjadi simpul utama atau biasa disebut dengan simpul akar, yang menghasilkan pemilahan menjadi 2 bagian yang dinamakan simpul dalam. Pada tahap berikutnya setiap simpul dalam ini akan dipilah kembali menjadi 2 bagian dan begitu seterusnya sampai diperoleh simpul yang tidak dapat dipilah kembali yang disebut sebagai simpul akhir (Breiman *et al.*, 1984).

Proses pemilahan dimulai dari menentukan kemungkinan simpul akar atau simpul utama. Menurut Breiman *et al.*(1984), terdapat aturan pemilahan untuk setiap simpul akar menjadi dua simpul dalam adalah sebagai berikut:

I. Setiap pemilahan yang dilakukan hanya bergantung pada satu nilai yang berasal dari satu variabel prediktor.
II. Apabila variabel prediktor berskala kontinu, maka pemilah ditentukan oleh $X_j \leq c_i$ dan $X_j > c_i$ , dengan $i = 1,2, \dots, n-1$ dan $c_i$ adalah nilai tengah dari dua nilai amatan sampel berurutan yang berbeda dari variabel $X_j$.
III. Apabila variabel prediktor berskala nominal dengan kategorik bertaraf $F$, akan diperoleh sebanyak $2^{F-1}-1$ pemilahan yang mungkin, sedangkan jika variabel prediktor berskala ordinal, maka diperoleh sebanyak $F-1$ pemilahan yang mungkin.

Pada setiap simpul $t$, akan diduga terdapat pemilah yang mungkin sebagai kandidat pemilah (*candidate of split*). Pemilah terbaik adalah pemilah yang memiliki keheterogenan paling tinggi. Ukuran keheterogenan diukur menggunakan nilai impuritas. Pada penelitian kali ini menggunakan indeks gini karena merupakan nilai impuritas yang sering digunakan. Indeks gini pada simpul $t$ dapat ditulis dengan persamaan berikut (Rohmah, 2018):

$$G(t) = \sum_{i=1, i \neq j}^{I} p(i|t)p(j|t) \ , \qquad (1)$$

dengan $G(t)$ adalah fungsi heterogenitas indeks gini, $p(i|t) = \frac{N_i(t)}{N(t)}$ adalah proporsi kelas i pada simpul t, $p(j|t) = \frac{N_j(t)}{N(t)}$ adalah proporsi kelas j pada simpul t, $N_i(t)$ adalah banyak pengamatan kelas i pada simpul t, $N_j(t)$ adalah banyak pengamatan kelas j pada simpul t, dan $N(t)$ adalah banyak pengamatan pada simpul t.

Jika $j = 1$ dan $i$ adalah kelas-kelas lain, Persamaan (1) dapat ditulis sebagai:

$$G(t) = \sum_{i=1,i\neq j}^{I} p(i|t)p(j|t)$$
$$= p(2|t)p(1|t) + p(3|t)p(1|t) + \cdots + p(i|t)p(1|t)$$
$$= p(1|t)[p(2|t) + p(3|t) + \cdots + p(i|t)] \,, \quad (2)$$

karena , $\sum_{j=1}^{J} p(j|t) = 1$, maka Persamaan (2) menjadi:

$$p(1|t)[p(2|t) + p(3|t) + \cdots + p(i|t)] = \mathrm{p(1|t)}\left(\sum_{j=1}^{J} p(j|t) - p(1|t)\right)$$
$$= p(1|t)(1 - p(1|t))$$
$$= p(1|t) - p^2(1|t) \,, \quad (3)$$

dan jika $j = 2$ dan $i$ adalah kelas-kelas lain, didapatkan persamaan berikut:

$$G(t) = \sum_{i=1,i\neq j}^{I} p(i|t)p(j|t) = \sum_{j=1}^{2}(p(j|t) - p^2(j|t)) \,, \quad (4)$$

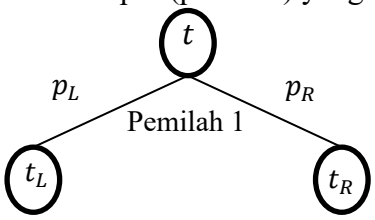
dan jika $j = 3$ dan $i$ adalah kelas-kelas lain, didapatkan persamaan berikut:

$$G(t) = \sum_{i=1,i\neq j}^{I} p(i|t)p(j|t) = \sum_{j=1}^{3}(p(j|t) - p^2(j|t)) \,, \quad (5)$$

sehingga secara umum jika hanya menggunakan kelas j atau persamaan untuk mencari indeks gini adalah sebagai berikut:

$$G(t) = \sum_{i=1,i\neq j}^{I} p(i|t)p(j|t) = \sum_{j=1}^{J}(p(j|t) - p^2(j|t))$$
$$= \sum_{j=1}^{J} p(j|t) - \sum_{j=1}^{J} p^2(j|t)$$
$$= 1 - \sum_{j=1}^{J} p^2(j|t) \quad (6)$$

Selanjutnya pemilihan pemilah terbaik pada langkah ini adalah mencari nilai pemilah yang optimal untuk membagi dataset menjadi dua atau lebih subset yang homogen atau yang memiliki kriteria *goodness of split,* yaitu evaluasi pemilahan oleh pemilah s pada simpul t. Variabel pemilah dengan nilai *goodness of split* yang tertinggi akan dipilih sebagai pemilah optimal sebagai simpul akar (*root node*) dalam pembentukan pohon keputusan untuk pemilahan pertama. Gambaran suatu simpul (pemilah) yang terbentuk ada pada Gambar 1.




Gambar 1. Contoh simpul (pemilah) pada metode CART

Nilai *goodness of split* $(\phi(s,t))$ merupakan selisih antara indeks gini pada simpul t dengan indeks gini yang terbentuk pada simpul kiri $t_L$ dan simpul kanan $t_R$ dengan masing-masing proporsi data amatan yang masuk.

$$\phi(s,t) = \Delta G(s,t)$$
$$= G(t) - (p_L G(t_L) + p_R G(t_R))$$
$$= G(t) - p_L(1 - \textstyle\sum_j^J p^2(j|t_L)) - p_R(1 - \sum_j^J p^2(j|t_R)) \quad , \quad (7)$$

dengan $\phi(s,t)$ merupakan kriteria *goodness of split*, $p_R = \frac{N(t_R)}{N}$ adalah peluang objek berada pada simpul kanan, $p_L = \frac{N(t_L)}{N}$ adalah peluang objek berada pada simpul kiri, $t_R$ adalah simpul kanan pada simpul t, dan $t_L$ adalah simpul kiri pada simpul t. Simpul akar akan dipilih berdasarkan nilai *goodness of split* yang terbesar pada pemilahan pertama.

Pembentukan pohon klasifikasi dilakukan dengan pengulangan langkah pemilahan simpul secara rekursif sampai terbentuk pohon klasifikasi maksimal. Simpul-simpul yang terbentuk akan dilakukan proses pelabelan kelas (*Class Assignment*). Pelabelan kelas

dilakukan mulai dari awal pemilahan simpul sampai simpul akhir terbentuk. Pelabelan setiap simpul akhir berdasarkan aturan jumlah anggota kelas terbanyak, yaitu jika:

$$p(j_0|t) = max_j\, p(j|t) = max_j\, \frac{N_j(t)}{N(t)}\,, \quad (8)$$

Berdasarkan Persamaan (8), label kelas untuk simpul $t$ adalah $j_0$ yang merupakan proporsi kelas terbesar dari simpul t. Setelah proses pelabelan kelas pada pemilahan simpul, perlu dilakukan proses penghentian pemilahan (*stop splitting*). Suatu simpul $t$ akan dikatakan menjadi simpul akhir jika sudah berada pada kondisi maksimal pemilahan dan adanya batasan jumlah kedalaman pohon maksimal yang ditentukan peneliti (Lewis, 2000). Jika hal tersebut terpenuhi, maka pembentukan pohon dihentikan dan diperoleh pohon klasifikasi maksimal (*maximal tree*).

Pemangkasan (*pruning*) pohon merupakan langkah penting dalam proses pembangunan pohon keputusan dalam mengurangi risiko *overfitting*. Pemangkasan dilakukan untuk menghilangkan bagian-bagian dari pohon yang tidak signifikan atau yang tidak dapat memberikan peningkatan yang signifikan dalam kinerja pada *data testing*. Menurut Breiman *et al.* (1984), ukuran pemangkasan menggunakan persamaan berikut:

$$R(t) = \sum_{t\in T} r(t)p(t)\,, \quad (9)$$

$$r(t) = 1 - max_j p(j|t)\,, \quad (10)$$

dengan $R(t)$ merupakan probabilitas terjadinya kesalahan klasifikasi yang disebabkan oleh pohon klasifikasi, T adalah jumlah seluruh himpunan simpul pada pohon klasifikasi, $p(t)$ adalah proporsi amatan yang masuk ke dalam sampul $t$, dan $r(t)$ merupakan probabilitas terjadinya kesalahan klasifikasi pada simpul $t$.

Secara umum, proses ini melibatkan penghapusan atau penggabungan simpul akhir yang kurang signifikan atau tidak memberikan peningkatan yang cukup dalam kinerja pohon. Prosedur pemangkasan pohon klasifikasi dimulai dari simpul dalam yang memuat simpul akhir pertama atau pemangkasan pohon klasifikasi dilakukan pada simpul dalam yang memuat simpul akhir kiri $t_L$ dan simpul akhir kanan $t_R$ yang pertama. Simpul kiri $t_L$ dan simpul kanan $t_R$ dipangkas apabila diperoleh dua simpul akhir dan simpul dalamnya memenuhi persamaan berikut:

$$R(t) = R(t_L) + R(t_R) \quad (11)$$

Pada proses pembentukan model pohon klasifikasi tentu perlu adanya proses seleksi pohon optimal. Proses seleksi pohon klasifikasi optimal menggunakan parameter kedalaman pohon (*max_depth*) yang bernilai {3, 4, 5, 6, 7, 8, 9, 10} dengan menggunakan nilai akurasi pada masing-masing kedalaman pohon. Parameter yang digunakan akan dilakukan *trial and error* untuk mendapatkan nilai akurasi tertinggi.

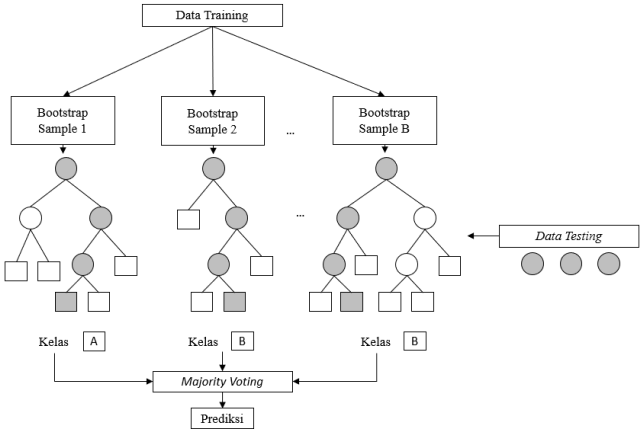
Metode bagging terdiri dari teknik *bootstrap* dan *aggregating* yang pertama kali diperkenalkan oleh Breiman (1996). Teknik ini muncul sebagai salah satu solusi untuk meningkatkan kinerja model prediksi, terutama pada kasus-kasus di mana model tunggal cenderung *overfitting* atau memiliki varians yang tinggi. Beberapa model prediksi yang dibuat berbeda dan kemudian digabungkan dengan meningkatkan akurasi prediksi secara keseluruhan. Pendekatan ini didasarkan pada konsep *resampling* dengan pengembalian (*bootstrap*) dan penggunaan agregasi (mayoritas suara) untuk menghasilkan prediksi *ensemble* yang baik.

Pada penelitian ini, parameter yang digunakan adalah *n_estimator* dengan nilai {10, 20, 30, 40, 50, 75, 100, 150, 200}, yaitu pengambilan sampel sebanyak B kali yang optimal berdasarkan nilai akurasi dan *max_depth* yang menyesuaikan dari kedalaman pohon optimal dari model CART yang telah dibangun.

Langkah-langkah penerapan Bagging (*Botstrap Aggregating*) pada pembangunan model CART (Breiman, 1996):

1. *Bootstrap Sampling*, membuat dataset *bootstrap* dengan langkah yang dilakukan yaitu:
   a. Mengambil data sampel asli *data training* sebanyak *n* sampel $(x_1, x_2, x_3, \dots, x_n)$.
   b. Membuat sampel atau dataset *bootstrap* pada *data training* $(x_1^*, x_2^*, x_3^*, \dots, x_n^*)$.
   c. Ulangi proses resampling sebanyak B kali.
2. Pembuatan model CART, yaitu membuat dan melatih model CART pada setiap dataset *bootstrap* yang dihasilkan.
3. Prediksi model, yaitu membuat prediksi pada *data testing* dengan menggunakan model-model CART.
4. Agregasi model, yaitu menggabungkan prediksi dari model-model CART untuk menghasilkan prediksi akhir dengan metode suara terbanyak *(majority vote)*.

Konsep Bagging pada CART dapat dilihat pada Gambar 2.




Gambar 2. *Bootstrap Aggregating* pada CART

Evaluasi model digunakan dalam rangka mengukur ketepatan klasifikasi dan untuk mengetahui apakah model melakukan klasifikasi dengan baik atau tidak. Cara yang digunakan untuk mengukur ketepatan klasifikasi adalah menggunakan *confusion matrix*. *Confusion matrix* merupakan salah satu metode yang digunakan untuk melakukan perhitungan akurasi pada konsep data mining (Rosandy, 2016). Selain dari akurasi, terdapat nilai *f1-score* yang digunakan sebagai evaluasi model klasifikasi pada penelitian kali ini.

*Confusion matrix* berisikan informasi mengenai hasil klasifikasi aktual yang telah dilakukan prediksi oleh sistem klasifikasi. Matriks yang akan digunakan pada penelitian ini adalah matriks berukuran 3×3, dengan 3 kelas klasifikasi berdasarkan variabel respon potensi *resign* berdasarkan kenyamanan bekerja yang terdiri dari rendah, sedang, dan tinggi. Berikut tampilan *confusion matrix* yang dapat digunakan dalam evaluasi model klasifikasi:

Tabel 1. *Confusion Matrix*

| | | **Nilai Prediksi** | | |
|---|---|---|---|---|
| | | Kelas 1 | Kelas 2 | Kelas 3 |
| **Nilai Aktual** | Kelas 1 | $N_{11}$ | $N_{12}$ | $N_{13}$ |
| | Kelas 2 | $N_{21}$ | $N_{22}$ | $N_{23}$ |
| | Kelas 3 | $N_{31}$ | $N_{32}$ | $N_{33}$ |

$$Akurasi = \frac{N11+N22+N33}{N11+N12+N13+N21+N22+N23+N31+N32+N33} = \frac{N11+N22+N33}{\sum N} \qquad (12)$$

*F1-score* dibentuk dari nilai *recall* dan *precision*, dengan *precision* menunjukkan jumlah objek dalam suatu kelas secara model prediksi benar (Harits *et al*., 2018), sedangkan

*recall* menunjukkan jumlah objek yang secara aktual benar. *F1-score* dapat dihitung untuk setiap kelas secara terpisah atau dapat dihitung secara rata-rata semua kelas.

Tabel 2. *Confusion matrix* kelas 1

| | | **Nilai Prediksi** | | |
|---|---|---|---|---|
| | | Kelas 1 | Kelas 2 | Kelas 3 |
| **Nilai Aktual (Kelas 1)** | Kelas 1 | *TP* | *FN* | *FN* |
| | Kelas 2 | *FP* | *TN* | *TN* |
| | Kelas 3 | *FP* | *TN* | *TN* |

Berdasarkan Tabel 2, persamaan yang digunakan untuk menghitung *f1-score* adalah:

$$F1 - score_i = 2 \times \frac{precision_i \times recall_i}{precision_i + recall_i}, \quad (13)$$

$$recall_i = \frac{TP_i}{TP_i + FN_i}, \quad (14)$$

$$precision_i = \frac{TP_i}{TP_i + FP_i}, \quad (15)$$

Kurva Receiving Operating Characteristic (ROC) merupakan gambaran dari hubungan antara sensitivity dan specificity secara grafis (Arian R & Peter, 1998). ROC *(Receiver Operating Characteristic) Curve* adalah grafik yang digunakan untuk mengevaluasi kinerja model klasifikasi. ROC menampilkan hubungan antara *True Positive Rate* (TPR) dan *False Positive Rate* (FPR) pada berbagai *threshold* prediksi yang berbeda. ROC curve dibentuk dengan memvariasikan *threshold* (ambang batas) pada model. Biasanya, model prediksi memberikan probabilitas dan *threshold* digunakan untuk mengklasifikasikan hasilnya.

Area dibawah kurva ROC merupakan wilayah yang menunjukkan tingkat keakuratan dari model empirik dan dapat dihitung dengan Area Under Curve (AUC). AUC adalah metriks yang mengukur luas di bawah kurva ROC AUC sendiri merupakan daerah berbentuk persegi yang nilainya selalu berada diantara 0 dan 1. Jika AUC yang dihasilkan < 0,5, maka model statistik yang dievaluasi memiliki tingkat keakuratan yang sangat rendah dan mengindikasikan bahwa model tersebut sangat buruk jika digunakan (Zou et al., 2007).

*Feature importance* merupakan metode atau langkah dalam pembentukan model yang melibatkan perhitungan skor semua fitur atau variabel untuk menetapkan pentingnya setiap fitur dalam proses pengambilan keputusan. Semakin tinggi skor (nilai *importance*) suatu variabel, semakin besar pengaruh terhadap model dalam melakukan prediksi. *Feature importance* sangat berguna untuk memahami hubungan antar variabel, mengetahui variabel apa yang tidak relevan bagi model, dan menentukan variabel mana yang paling memengaruhi daya prediksi pada model (Shin, 2023).

## 3. METODE PENELITIAN

Penelitian ini menggunakan data primer dengan pengambilan data menggunakan kuesioner yang disebarkan melalui metode *purposive sampling* dan *quota sampling* pada karyawan perusahaan PT.XYZ, perusahaan yang bergerak di Hutan Tanaman Industri (HTI). Kuesioner dibuat menggunakan bantuan *Microsoft Form.* Pengumpulan data dilakukan selama 6 bulan dengan jumlah responden sebanyak 100 responden.

Penelitian ini menggunakan variabel terikat tentang beberapa faktor terkait dengan keputusan karyawan *resign* di suatu perusahaan. Variabel-variabel tersebut sudah dipertimbangkan berdasarkan kondisi dari perusahaan terkait yaitu PT.XYZ. Potensi *resign* menjadi variabel respon penelitian yang merupakan variabel amatan dari kenyamanan dalam

bekerja pada 100 karyawan PT.XYZ. Terdapat 8 variabel prediktor (X) dan satu variabel respon (Y). Selengkapnya pada Tabel 3.

Tabel 3. Variabel data penelitian

| Variabel | Nama | Keterangan | Skala |
|---|---|---|---|
| Y | POTENSI RESIGN | 1 = Kurang Nyaman<br>2 = Cukup Nyaman/Nyaman<br>3 = Sangat Nyaman | Ordinal |
| X1 | USIA | Usia | Rasio (tahun) |
| X2 | RIWAYAT PERUSAHAAN | Perusahaan yang ke berapa selama berkarier | Rasio (bilangan asli) |
| X3 | KENAIKAN JABATAN | 0 = Tidak<br>1 = Ya | Nominal |
| X4 | LAMA BEKERJA | 1 = < 1 Tahun<br>2 = 1 – 3 Tahun<br>3 = > 3 Tahun | Ordinal |
| X5 | LAMA PERJALANAN KANTOR | 1 = < 20 menit<br>2 = 20 - 45 menit<br>3 = > 45 menit | Ordinal |
| X6 | LEMBUR | 1 = Tidak pernah lembur<br>2 = 1-3 kali lembur<br>3 = > 3 kali lembur | Ordinal |
| X7 | AKTIVITAS REKAN KERJA | 1 = Tidak pernah<br>2 = 1-2 kali<br>3 = > 2 kali | Ordinal |
| X8 | PROYEK PEKERJAAN | 1 = < 3 proyek/pekerjaan<br>2 = 3-5 proyek/pekerjaan<br>3 = > 5 proyek/pekerjaan | Ordinal |

Penelitian yang dilakukan menggunakan metode klasifikasi CART (*Classification And Regression Trees*) dengan metode *ensemble* yang digunakan adalah metode Bagging (*Bootstrap Aggregating*) dalam menganalisis karyawan yang berpotensi *resign* dengan melihat tingkat kenyamanan dalam bekerja.

Tahapan analisis yang digunakan dalam penelitian ini sebagai berikut:

1. Melakukan *data preprocessing* dengan melakukan pembersihan data, transformasi data, dan pembagian data kedalam *data training* dan *data testing*.
2. Membuat pohon klasifikasi (CART) dan Bagging CART dengan menggunakan *data training.*
   a. CART
      i. Menentukan kemungkinan pemilah pada setiap peubah prediktor dan menghitug indeks gini pada semua kemungkinan pemilah dengan menggunakan Persamaan (6).
      ii. Memilih pemilah terbaik menggunakan Persamaan (7) (*Goodness of split*) berdasarkan nilai yang terbesar.
      iii. Menentukan simpul dalam dengan pengulangan Langkah (i) dan Langkah (ii) hingga membentuk pohon klasifikasi maksimal.
      iv. Pemangkasan pohon klasifikasi maksimal jika memenuhi Persamaan (11)
      v. Pohon klasifikasi optimal dan model klasifikasi CART terbentuk.
   b. Bagging CART
      i. Melakukan *resampling bootstrap* sebanyak B optimal pada *data training.*

ii. Membentuk pohon klasifikasi CART pada masing-masing sampel *bootstrap* dan model klasifikasi Bagging CART terbentuk.

3. Melakukan pemodelan pada masing-masing model klasifikasi yang terbentuk menggunakan *data testing.*
4. Evaluasi model dengan menghitung masing-masing ketepatan klasifikasi CART dan Bagging CART dan membandingkan tingkat ketepatan klasifikasi kedua model.
5. Mengidentifikasi *feature importance.*

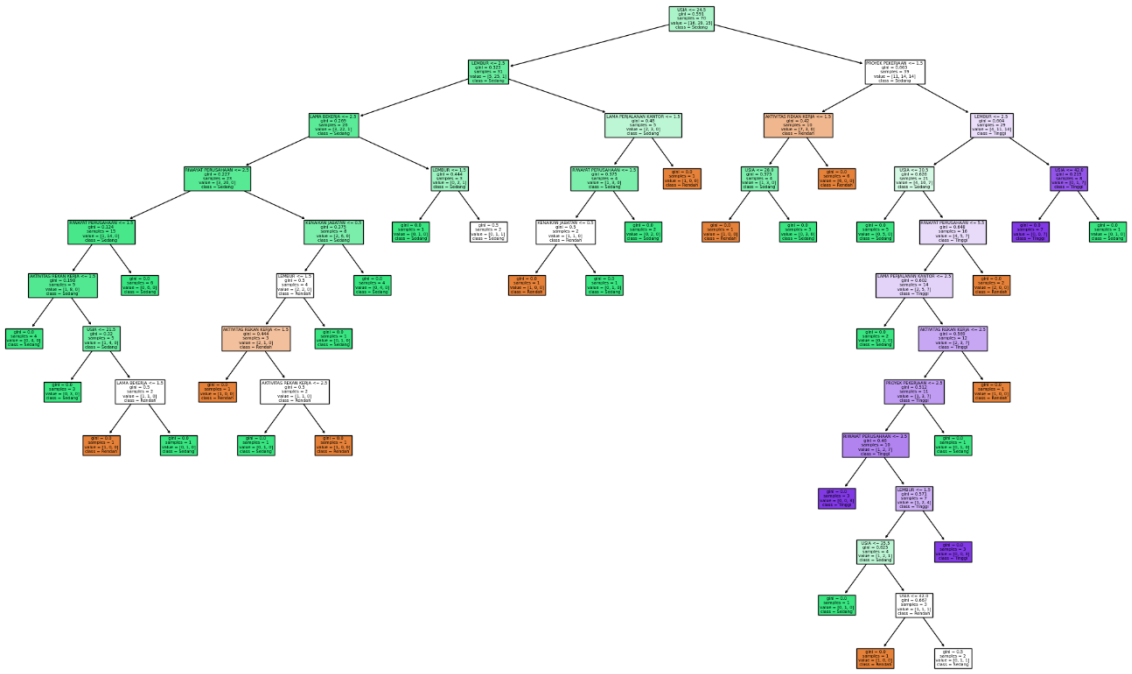
## 4. HASIL DAN PEMBAHASAN

Variabel respon yang digunakan adalah potensi *resign* yang didasarkan pada kenyamanan bekerja. Proporsi kenyamanan bekerja pada 100 karyawan cenderung merata, tidak adanya ketimpangan kenyamanan berdasarkan data yang diperoleh dari karyawan PT.XYZ. Data menunjukkan bahwa dari 100 karyawan, terdapat 17% orang yang merasa kurang nyaman, 26% karyawan merasa cukup nyaman, 31% karyawan merasa nyaman, dan 26% karyawan merasa sangat nyaman bekerja di PT.XYZ.

*Data preprocessing* diawali dengan melakukan pembersihan data, transformasi data, dan pembagian data ke dalam *data training* dan *data testing*. Pada semua data variabel tidak terdapat data yang hilang sehingga dapat dilanjutkan transformasi data. Transformasi data yang dilakukan menggunakan 2 jenis, yaitu *label encoding* pada data nominal { Ya=1 dan Tidak =0} dan *ordinal encoding* pada data ordinal {Rendah = 1, Sedang = 2, dan Tinggi = 3}. Pembagian data yang dilakukan pada penelitian ini menggunakan proporsi sebesar 70% untuk *data training* dan 30% untuk *data testing* dengan jumlah *data training* sebanyak 70 data dan *data testing* sebanyak 30 data.

Pemilah paling baik adalah pemilah yang memiliki nilai heterogenitas yang paling tinggi. Ukuran keheterogenitasan dicari menggunakan indeks gini dan pemilah yang terbaik adalah pemilah yang memiliki nilai *goodness of split* yang tertinggi. Perhitungan nilai *goodness of split* menghasilkan variabel Usia yang memiliki nilai *goodness of split* terbesar dengan indeks gini sebesar 0,591. Calon simpul ini dipilih menjadi simpul akar pada pembentukan pohon klasifikasi dengan simpul kiri {≤ 24,5} dan simpul kanan {> 24,5}. Langkah pemilihan simpul ini akan tetap dilakukan pengulangan secara rekursif pada setiap simpul yang terbentuk dalam proses pembentukan pohon klasifikasi.

Pembentukan pohon klasifikasi dimulai berdasarkan pada simpul akar. Simpul akar yang terbentuk adalah variable Usia. Pada simpul akar maupun simpul-simpul yang nantinya terbentuk akan dilakukan proses pelabelan kelas. Label kelas simpul akar yang terbentuk adalah kelas 2 atau termasuk kedalam kelas Sedang. Proses pembentukan pohon klasifikasi tentu akan mencapai tingkat maksimal pohon. Berdasarkan simpul akar yang diperoleh yaitu Usia, hasil dari simpul kiri dan simpul kanan akan dipilah kembali pada masing-masing simpul sehingga terbentuk simpul dalam yang baru dan begitu seterusnya hingga mencapai tingkat maksimal pohon. Pohon klasifikasi maksimal diperoleh jika sudah tidak ada lagi simpul yang dapat dipilah dengan kata lain simpul tidak dapat dipilah kembali apabila memiliki variabel respon identik. Pohon klasifikasi maksimal CART terbentuk pada Gambar 3. Pohon klasifikasi maksimal terdiri dari 28 simpul dalam, 30 simpul akhir, dan kedalaman pohon maksimal sebesar 12. Pada pohon maksimal memuat semua variabel prediktor dan semua kelas pada variabel respon.



Gambar 3. Pohon klasifikasi maksimal CART

Pemangkasan pohon pertama kali dilakukan pada simpul dalam yang memuat simpul akhir pertama dan kedua. Berdasarkan pada Persamaan (11), jika persamaan tersebut terpenuhi maka simpul akhir tersebut harus dipangkas. Berdasarkan pohon klasifikasi maksimal yang terbentuk, simpul 9 merupakan simpul dalam yang memuat simpul akhir pertama dan kedua. Simpul 9 akan dilakukan proses pemangkasan. Simpul ini memiliki pemilah variabel Kenaikan Jabatan dengan label kelas {Sedang} dan total data sebanyak 3 data amatan. Proporsi kelas pada simpul 9 yang paling besar atau $max_j\ p(j|t) = 0{,}6667$, sehingga untuk hasil dari,

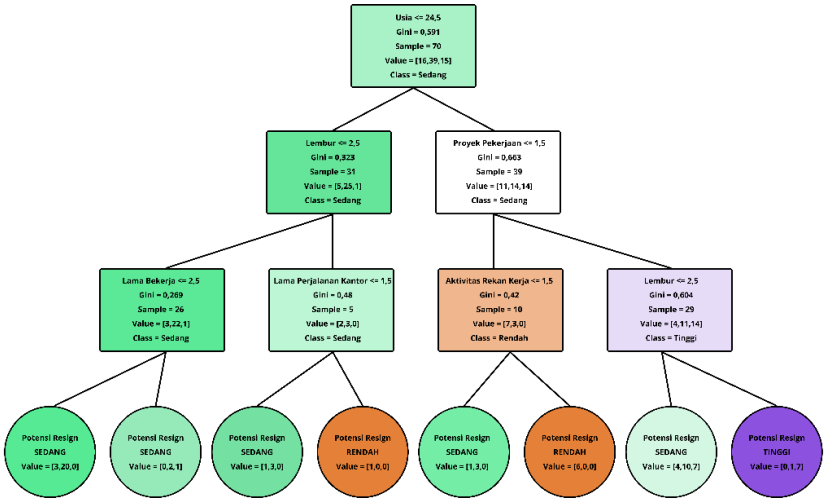
$r(t) = 1 - max_j p(j|t) = 1 - 0{,}6667 = 0{,}3333$

$p(t) = p(9) = \frac{N(9)}{N} = \frac{3}{70} = 0{,}0429$ , maka

$R(t) = r(t)p(t) = (0{,}3333)(0{,}4286\ ) = 0{,}0143$

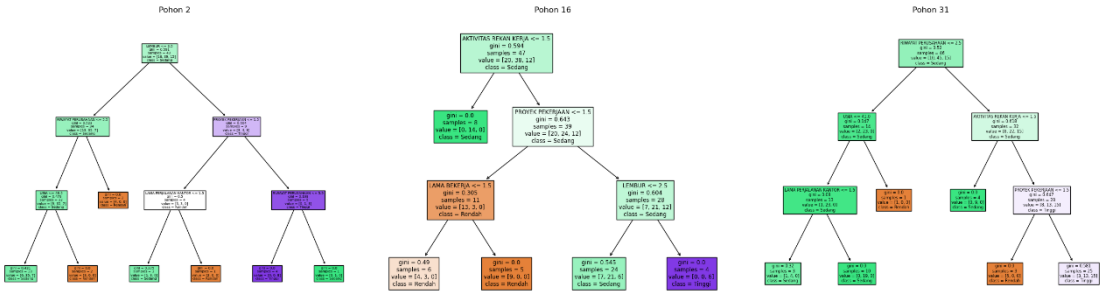
Selanjutnya dengan cara yang sama, menghitung $R(t)$ pada simpul kiri $R(t_L)$ dan simpul kanan $R(t_R)$. Hasil yang diperoleh $R(t_L) = 0{,}0000$ dan $R(t_R) = 0{,}0143$. Pada perhitungan ini dapat dilihat bahwa $R(t) = R(t_L) + R(t_R)$, sehingga pada simpul 9 dilakukan pemangkasan. Langkah pemangkasan pohon ini dilanjutkan pada simpul lain dengan kriteria dan cara yang sama sampai tidak terdapat pemangkasan pohon lagi yang mungkin.

Pohon klasifikasi maksimal yang terbentuk belum tentu merupakan pohon klasifikasi yang optimal sehingga perlu dilakukan seleksi terhadap pohon klasifikasi maksimal untuk mendapatkan pohon kasifikasi yang optimal. Parameter yang digunakan dalam menentukan pohon klasifikasi optimal adalah *max_depth* yang didasarkan pada nilai akurasi tertinggi. Pohon klasifikasi optimal memiliki kedalaman maksimal atau *max_depth* = 3. Berikut pohon klasifikasi optimal terbentuk pada Gambar 4. Pada pohon optimal memuat semua kelas pada variabel respon, tetapi hanya memuat variabel prediktor Usia, Lembur, Proyek Pekerjaan, Lama Bekerja, Lama Perjalanan Kantor, dan Aktivitas Rekan Kerja. Model CART pada pohon maksimal memiliki tingkat akurasi 67% dengan kedalaman pohon maksimal (*max_depth*) sebanyak 12, sedangkan untuk model CART pada pohon optimal memiliki tingkat akurasi sebesar 73% dengan kedalaman maksimal pohon (*max_depth*) sebesar 3.




Gambar 4. Pohon klasifikasi optimal CART

Parameter yang digunakan untuk metode Bagging CART dalam penelitian ini adalah *max_depth* yang menyesuaikan dari model CART optimal, yaitu *max_depth = 3* dan *n_estimator* terbaik adalah *bootstrap* sebanyak 40. Pengulangan yang dilakukan akan mendapatkan 40 pohon klasifikasi yang berbeda, beberapa pohon yang terbentuk terdapat pada Gambar 5, kemudian selanjutnya dilakukan agregasi dengan *majority vote* atau pengambilan suara terbanyak.



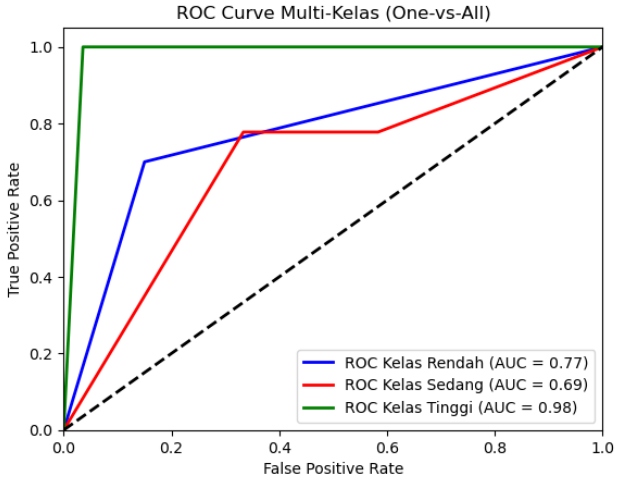

Gambar 5. Contoh pohon klasifikasi dari model Bagging CART

Model CART dan Bagging CART belum bisa memberikan akurasi 100%, namun ada peningkatan tingkat ketepatan model pada Bagging CART dibandingkan dengan model CART itu sendiri. Peningkatan yang dilakukan model Bagging CART sebesar 14%. Berikut detail evaluasi model CART dan Bagging CART:

Tabel 4. Perbandingan Evaluasi Model CART dan Bagging CART

| **Model** | **Matrik Evaluasi Model** | | | |
|---|---|---|---|---|
| | ***Akurasi*** | ***Recall*** | ***Presisi*** | ***F1-Score*** |
| CART | 73% | 63% | 69% | 62% |
| Bagging CART | 87% | 87% | 94% | 88% |

Hasil ROC-AUC yang didapatkan terlampir pada Gambar 6. Nilai AUC kelas Rendah sebesar 0,77, kelas Sedang sebesar 0,69, dan kelas Tinggi sebesar 0,98. Rata-rata nilai AUC untuk seluruh kelas sebesar 0,82. Berdasarkan nilai rata-rata AUC yang didapat, dapat disimpulkan bahwa model cukup bisa diandalkan untuk melakukan prediksi setiap kelas.




Gambar 6. Kurva ROC yang terbentuk

*Feature importance* merupakan analisis dalam menentukan fitur atau variabel mana yang paling berperan dalam pemilahan pohon keputusan hingga pembuatan keputusan prediksi. Fitur atau variabel yang memiliki nilai *importance* tertinggi atau yang paling berpengaruh biasanya marupakan variabel yang menjadi simpul akar pada proses pembentukan pohon klasifikasi.

Tabel 5. Tingkat *Importance* Setiap Variabel

| **Variabel** | **Tingkat *Importance*** |
|---|---|
| USIA | *34.7%* |
| LEMBUR | *14.5%* |
| AKTIVITAS REKAN KERJA | *13.7%* |
| PROYEK PEKERJAAN | *13.1%* |
| RIWAYAT PERUSAHAAN | *8.9%* |
| LAMA PERJALANAN KANTOR | *6.3%* |
| KENAIKAN JABATAN | *5.1%* |
| LAMA BEKERJA | *3.7%* |

Tabel 5 menunjukkan variabel mana yang paling berperan atau memiliki kontribusi besar dalam pembuatan model dan keputusan prediksi. Variabel usia memiliki kontribusi paling besar dibandingkan variabel lainnya, sedangkan variabel yang paling kecil pengaruh atau kontribusinya adalah variabel lama bekerja. Tidak hanya berdasarkan tingkat importance, variabel usia juga merupakan simpul akar dalam pembentukan model pohon klasifikasi, sehingga variabel ini merupakan variabel yang memiliki pengaruh terbesar dalam pembuatan model prediksi.

## 5. KESIMPULAN

Hasil analisis pada pembentukan model pohon keputusan dengan menggunakan metode CART menunjukkan bahwa variabel usia menjadi pemilah utama atau simpul akar yang didasarkan pada perhitungan indeks gini dan pemilihan nilai *goodness of split* terbesar. Hal ini mengindikasikan bahwa variabel usia merupakan variabel penting dalam pembentukan model pohon klasifikasi atau keputusan prediksi. Indikasi ini juga dibuktikan dengan analisis *feature importance* yang menunjukkan bahwa variabel usia memiliki pengaruh terbesar dalam pembentukan model CART.

Model klasifikasi CART menunjukkan nilai akurasi sebesar 73% dan *f1-score* sebesar 62%, sedangkan model Bagging CART menunjukkan nilai akurasi sebesar 87% dengan *f1-score* sebesar 88%. Dari hasil penelitian ini, dapat disimpulkan juga bahwa adanya peningkatan nilai akurasi pada model Bagging CART dibandingkan dengan model CART sebesar 14% dalam mengidentifikasi karyawan PT.XYZ yang berpotensi *resign*.

## DAFTAR PUSTAKA


Arian R, V. E., & Peter, M. (1998). Receiver Operating Characteristic (ROC) Analysis: Basic Principlesn and Application in Radiologi. European Journal of Radiology, 27(2).

Breiman L. 1996. Bagging Predictors. *Machine Learning, 24:* 123-140.

Breiman, L., Friedman, J.H., Olshen, R.A., dan Stone, C.J. 1984. *Classification and Regression Trees.* New York: Chapman and Hall.

Fergian, A. 2020. Faktor-Faktor Resign Karyawan Marketing Terhadap Kinerja Bank Syariah. *Skripsi.* Jurusan Perbankan Syariah Fakultas Ekonomi dan Bisnis Islam InstitutAgama Islam Negri Metro.

Harits, A., Robin, D., dan Arief, P.M. 2018. *Evaluasi Performa Metode Deep Learning untuk Klasifikasi Citra Lesi Kulit The HAM10000*. SNIKO, Bandung.

Isyandi, B. 2004. *Manajemen Sumber Daya Manusia Dalam Perspektif Global*. Pekanbaru: Unri Press.

Lewis, R.J. 2000. An Introduction to Classification and Regression Trees (CART) Analysis. *Presented at the 2000 Annual Meeting of Society for Academic Emergency Medicine of Sanfransisco*. California.

Luqman, H. 2023. Mengidentifikasi Fenomena Resign Setelah THR.Talentics.id Tersedia: https://talentics.id/resources/blog/resign-adalah/ (diakses pada 24 Januari 2024).

Mathis, R. L. dan Jackson, J. H. 2006. Manajemen Sumber Daya Manusia. Jakarta: Salemba Empat.

Republik Indonesia, Undang Undang Nomor 13 Tahun 2003 Tentang Ketenagakerjaan

Rohmah, L. 2018. *Bootstrap Aggregating Classification and Regression Tree* (Bagging CART) pada Pengklasifikasian Indeks Pembangunan Manusia Kabupaten/Kota di Indonesia Tahun 2016. *Skripsi*. Program Studi Statistika Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Brawijaya Malang.

Rommalla, S., 2021. 7 Ciri Lingkungan Kerja Positif yang Ideal bagi Milenial. Glints.com Tersedia: https://employers.glints.com/id-id/blog/7-ciri-lingkungan-kerja-positif-yang-ideal-bagi-milenial/ (diakses pada tanggal 19 Desember 2023).

Rosandy, T. 2016. Perbandingan Metode Naïve Bayes Classifier dengan Metode Decission Tree (C4.5) untuk Menganalisa Kelancaran Pembiayaan. *Jurnal TIM Darmajaya,* 2(1), pp. 52-62.

Sa'adah, U., Rochayani, M.Y., dan Astuti, A.B. 2021. Knowledge discovery from gene expression dataset using bagging lasso decisiton tree. *Indonesian Journal of Electrical Engineering and Computer Science.* Vol.21, No. 2, pp. 1151-1159.

Shin, T. 2023. Understanding Feature Importance in Machine Learning. *Builtin.com.* Tersedia: https://builtin.com/data-science/feature-importance (diakses pada tanggal 10 Agustus 2024).

Sumartini, S.H. dan Purnami, S.W. 2015. Penggunaan Metode Classification and Regression Trees (CART) untuk Klasifikasi Rekurensi Pasien Kanker Serviks di RSUD Dr. Soetomo Surabaya. *Jurnal Sains dan Seni ITS,* Vol. 4, No. 2, Hal: D211-D216.

Zou, K. H., A.James, O., & Mauri, L. (2007). Receiver Operating Characteristic (ROC) Analysis For Evaluating Diagnostic Test and Predictive Models. Circulation, 115(5).