Perbandingan Kinerja Prediksi Klasifikasi Algoritma Random Forest Dan Decision Tree Menggunakan Confusion Matrix

Ellya Helmud[1]\*, Ellya Helmud[2], Fitriyani[3], Parlia Romadiana[4]

Sistem Informasi, Fakultas Teknologi Informasi ISB Atma Luhur[1], [2], [3], [4]

Pangkalpinang, Indonesia

ellyahelmud@atmaluhur.ac.id [1],ellyahelmud@students.undip.ac.id [2],fitriyani@atmaluhur.ac.id [3], parliaromadiana@atmaluhur.ac.id [4]

***Abstract— The classification method is part of data mining which is used to predict existing problems and also as predictions for the future. The form of dataset used in the classification method is supervised data. The random forest classification method is processed by forming several decision trees and then combining them to get better and more precise predictions. while a decision tree is the concept of changing a pile of data into a decision tree that presents the rules of a decision. From these two classification methods, researchers will compare the level of accuracy of predictions from both methods with the same dataset, namely the employee dataset in India, to predict the level of accuracy of employees who leave their jobs or still remain to work at their company. The number of records available is 4654 records. Of the existing data, 90% was used as training data and 10% was used as test data. From the results of testing this method, it was found that the accuracy level of the random forest method was 86.45%, while the decision tree method was 84.30% accuracy level. Then, by using the confusion matrix, you can see the magnitude of the distribution of experimental validity visually to calculate precision, recall and F1-Score. The random forest algorithm obtained precision of: 96.7%, sensitivity of: 84.7%, specificity of: 91.4%, and F1-Score of: 90.2%. Meanwhile, the decision tree algorithm obtained precision of: 95.7%, sensitivity of: 82.9%, specificity of: 88.4%, and F1-Score of: 88.8%.***

***Keywords— Data Mining, Classification, Random Forest, Decision Tree, Confusion Matrix***

***Abstrak*—** **Metode klasifikasi adalah bagian dari penggalian data yang digunakan untuk memprediksi permasalahan yang ada dan juga sebagai prediksi masa yang akan datang. Bentuk dataset yang digunakan dalam metode klasifikasi merupakan data yang diawasi. Metode klasifikasi hutan acak diproses dengan membentuk beberapa pohon keputusan kemudian disatukan untuk mendapatkan prediksi yang lebih baik dan tepat. sedangkan pohon keputusan merupakan konsep mengubah tumpukan data menjadikan pohon keputusan yang mempresentasikan aturan aturan dari sebuah keputusan. Dari kedua metode klasifikasi ini, peneliti akan mengkomparasi level ketepatan prediksi dari kedua metode dengan dataset yang sama yaitu dataset pegawai di india untuk memprediksi tingkat ketepatan pegawai yang meninggalkan pekerjaannya atau masih bertahan untuk bekerja di perusahaannya. Jumlah record yang tersedia sejumlah 4654 record. Dari data yang ada 90% dilakukan sebagai data latih data dan 10% dilakukan untuk data uji coba. Dari hasil pengujian metode ini didapatkan tingkat ketepatan dari metode hutan acak sebesar 86,45% sedangkan metode pohon keputusan sebesar 84,30% untuk tingkat ketepatannya. Kemudian dengan menggunakan confusion matrix akan dilihat besar dari sebaran validitas percobaan secara visual untuk menghitung presisi, recall dan F1-Score didapatkan algoritma hutan acak didapatkan presisi sebesar : 96.7% , sensitivity sebesar : 84,7%, spesificity sebesar : 91,4%, dan F1-Score sebesar : 90.2%. Sedangkan algoritma pohon keputusan didapatkan presisi sebesar : 95.7% , sensitivity sebesar : 82,9%, spesificity sebesar : 88,4%, dan F1-Score sebesar : 88,8%.**

***Kata Kunci—*** **Penggalian Data, Klasifikasi, Hutan acak, Pohon Keputusan, Confusion Matrix**

#  PENDAHULUAN

Penambangan data merupakan suatu proses penemuan pengetahuan dalam basis data. Penambangan data yang akan yang dilakukan ekstraksi informasi atau pola penting dalam data yang berukuran besar [1]. Pengerjaan penambangan data bisa dilakukan dengan menggunakan metode prediksi, asosiasi dan sekmentasi [2] , dimana metode prediksi dibagi dalam tiga bagian yang meliputi klasifikasi, regresi, dan rangkaian waktu. Dalam Klasifikasi algoritma yang digunakan antara lain : Pohon keputusan, Hutan acak, jaringan saraf tiruan, Pendukung vector mesin, K-Nearest Neighbor, Naïve Bayes, dan GA. Klasifikasi mempunyai karakteristik pengelompokan data berdasarkan keterikatan data terhadap data sampel, dimana dalam pengelompokan data tersebut harus mempunyai atribut label atau target . Dari beberapa algoritma yang ada dalam metode klasifikasi, kami akan membandingkan 2(dua) algoritma dalam metode klasifikasi yaitu pohon keputusan dan hutan acak. Pohon keputusan adalah pengklasifikasi dasar yang memiliki dua langkah pembelajaran dan klasifikasi [3]. Dalam fase pembelajaran, pohon keputusan belajar menghasilkan pohon keputusan dari sekumpulan sampel petrainingan yang telah diklasifikasikan [3]. Dalam tahap klasifikasi, pohon keputusan yang diperoleh dari pembelajaran fase digunakan untuk mengklasifikasikan data yang tidak diklasifikasikan. Hutan acak yang diusulkan oleh [4] Ada beberapa pohon keputusan dalam Pengklasifikasi hutan acak, dan ketepatan rata-rata dari pohon keputusan ini adalah digunakan untuk meningkatkan ketepatan dalam kumpulan dari beberapa pohon keputusan, ambil perkiraan dari setiap pohon dan memprediksi hasil akhir berdasarkan suara terbanyak. Berdasarkan dari beberapa penelitian terdahulu :

a. **“Metode Pembelajaran Mesin untuk Memprediksi Persetujuan Pinjaman dengan Membandingkan Algoritma hutan acak dan pohon keputusan”** mengatakan bahwa Metode Hutan acak ditemukan memiliki tingkat keberhasilan 79,4490 persen dalam hal ini investigasi. Jika dibandingkan dengan Pohon keputusan

b **“Komparasi Tingkat Ketepatan Hutan acak dan Pohon keputusan C4.5 Pada Klasifikasi Data Penyakit Infertilitas Agung”** mengatakan bahwa dari hasil uji kedua algoritma ini diperoleh hasil yang berbeda dengan perbedaan akurasi yang signifikan pada klasifikasi fertility data. Hasil dari pengujian Hutan acak dan *Pohon keputusan* C4.5 dalam melakukan prediksi dari tingkat keberhasilan dan dapat diambil kesimpulan bahwa hasil pengujian menggunakan Hutan acak memperoleh ketepatan sebesar 87.20 %, kemudian *Pohon keputusan* C4.5 memperoleh ketepatan sebesar 85.90 %. [5]

c. **“Komparasi Algoritma Hutan acak Dan Pohon keputusan Untuk Memprediksi Keberhasilan Immunotheraphy”** [5]mengatakan bahwa dengan adanya pengujian dengan menggunakan validasi silang maka di dapatkan nilai ketepatan 84,4 % dengan menggunakan metode Pohon keputusan kemudian nilai ketepatan 85,5% dengan menggunakan metode Hutan acak.

Dengan menggunakan dataset pegawai di India untuk memprediksi pegawai yang masih tetap bekerja ataupun keluar dari pekerjaanya. Setelah didapatkan hasil prediksi tingkat ketepatan dari kedua algoritma ini, selanjutkan akan dilakukan perhitungan ketepatan, presisi, recall dan F1-Score untuk melihat sebaran validitas percobaan secara visual dengan menggunakan confusion matrix. Dataset yang digunakan sebanyak 4654 record, dan kemudian dipilah untuk data Latihan dan data uji coba dengan rasio perbandingan 9:1

# METODOLOGI PENELITIAN

Penelitian ini dilakukan untuk membandingkan tingkat kinerja algoritma hutan acak dan pohon keputusan dengan menggunakan dataset pegawai di India yaitu memprediksi pegawai yang meninggalkan pekerjaannya atau masih bertahan dalam suatu instansi/perusahaan. Tahapan yang dilakukan untuk membandingkan dua metode ini dapat digambarkan pada gambar berikut.



Gambar 1 Metode *Research*

2.1. Studi Literatur

*Research* harus merujuk pada suatu studi literatur. Studi literatur yang merupakan rangkaian aktivitas yang mencakup koleksi data pustaka, membaca dan mencatat, serta mengelola bahan penelitian (Zed, 2008:3). Studi literatur merupakan suatu aktivitas yang harus dilakukan dalam penelitian, lebih khusus penelitian yang berhubungan dengan akademik dimana tujuannya adalah mengeksplorasi sisi dari konseptual maupun dari sisi teknis. Studi pustaka dilakukan oleh para peneliti dengan maksud adalah untuk memperoleh teori dasar, framework of thinking, dan mendapatkan perkiraan temporer yang disebut sebagai hipotesis *research*. Dengan demikian para peneliti dapat membuat group, allokasi melakukan pengorganisasian, serta berbagai pengembangan pustaka untuk reset penelitiannya. Adanya studi pustaka membuat para peneliti dapat secara mendalam dan luas untuk menggali lebih dalam masalah yang akan diteliti. Dengan studi literatur peneliti dapat menentukan topik dalam penelitan, masalah yang dirumuskan kemudian mereka dapat langsung ke tempat riset yang dituju untuk koleksi data yang dibutuhkan (Darmadi, 2011)

2.2 Pengumpulan Data

Pada saat dilakukan kegiatan pengumpulan data yaitu melakukan pencarian dataset yang merupakan sumber data yang digunakan klasifikasi. Dataset yang diambil untuk *research* ini adalah berupa data publik bersumber dari website https://www.kaggle.com/ yaitu employee dataset. Dataset ini berisi data data pegawai yang pada saat dilakukan pengujian akan mempengaruhi pegawai tersebut meninggalkan pekerjaannya atau masih tetap bekerja . Data tersebut sebanyak 4654 record dimana terdapat 8 atribut dan 1 atribut yang dijadikan label sasaran. Sebelum dataset diproses untuk diuji, maka perlu dilihat ( data preparation ). Data Preparasi adalah suatu tahap untuk melakukan cleaning data serta transformasi data mentah sebelum diproses dan dianalisis yang mencakup semua langkah yang diperlukan untuk mendapatkan, persiapan, ketepatan, dan mengawasi sumber daya data di perusahaan/organisasi . Penting untuk dilakukan para peneliti sebelum data diproses dan melakukan format terhadap data, data diperiksa terlebih dahulu, dan melakukan kombinasi data koleksi untuk ekplorasi data. Sebelum dilakukan analisis terhadap data, para peneliti perlu mengumpulkan data dari berbagai sumber, menghapus, atau mengisi data yang Null, menduplikasi data, atau melakukan update data menjadi suatu aturan yang benar. Berikut contoh dataset yang sudah dipersiapkan untuk melakukan uji akurasi prediksi terhadap metode random forest dan decision tree, dapat dilihat pada gambar dibawah :





Gambar 2

Dataset Employee

## 2.3 Hutan acak

*Hutan acak* adalah algoritma pembelajaran supervisi dimana terdapat data training dan juga data testingyang dikemukakan oleh Breiman pada tahun 2001 (Louppe, 2014) [6]. *Hutan acak* biasa digunakan penyelesaian suatu masalah yang ada hubugannya degan klasifikasi, regresi, dan yang lainnya. Mengapa algoritma ini disebut acak disebabkan karena :

* + 1. Setiap tree tumbuh pada contoh *bootstrap* yang tidak sama, diambil dari data training dilakukan secara random.
		2. Dalam setiap *node split* selama pembentukan *pohon keputusan*, sebagian sampel dari *m* variabel dipilih dari koleksi data aslinya setelah itu yang paling baik akan digunakan dalam *node* tersebut.

Algoritma ini berupa kombinasi dari beberapa predictor pohonatau bisa disebut pohon keputusandimana setiap pohonbergantung pada nilai random vector untuk dijadikan contoh bebas dan meratapada semua pohon dalam hutantersebut [7]. Hasil prediksi dari *Hutan acak* diperoleh dari hasil paling banyak dari masing masing pohon keputusan(voting untuk klasifikasi dan rata-rata untuk regresi). Untuk RF yang terdiri dari *N trees* dirumuskan sebagai:

## Pohon keputusan

Pohon keputusan digunakan untuk membentuk pohon keputusan dengan prediksi yang sangat kuat. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang merepresentasikan aturan yang dapat dengan mudah dipahami [8]. Tahapan dalam Pohon keputusan yaitu :

1. Kalkulasi nilai entropipada setiap atribut :

1. Kalkulasi value perolehan informasi pada setiap atribut :

1. Kalkulasi value informasi terpisah untuk masing masing atribut :

1. Menghitung nilai rasio keuntungan untuk masing masing atribut
2. Atribut yang memiliki rasio keuntunganpaling tinggi dipilih menjadi akar (*splitting attribute)* dan atribut yang memiliki nilai rasio keuntunganlebih rendah dari akar (*root*) dipilih menjadi cabang (*branches*),
3. Menghitung lagi nilai rasio keuntungantiap - tiap atribut dengan tidak mengikutsertakan atribut yang terpilih menjadi akar (*root*) di tahap sebelumnya,
4. Atribut yang memiliki *Gain Ratio* paling tinggi akan dijadikan cabang *(branches).* Kemudian dilakukan iterasi untuk langkah ke 4 dan kelima didapatkan value dari *Gain* sebesar 0 dari semua sisa atribut yang ada

## Confusion Matrix

Kalkulasi dari performa klasifikasi dari masingl-lmasing pengujian metode dengan *Confusion Matrix* untuk memperoleh hasil akurasi, presisi, dan *Recall.* matriks kebingungandigunakan perolehan perkiraan seberapa bagus penggolongan pendeteksian kelas yang tidak sama [19]. Tabel *Confusion Matrix* dapat dilihat pada Tabel



Tabel 1.Tabel *Confusion Matrix*

Pendapat Sokolova (2009) didapatkan beberapa metrik untuk dikalkulasi dalam matriks kebingunganyaitu akurasi,presisi, *recall, spesifisiti, dan* F1*-Score*. Dalam perhitungannya didapatkan rumus :

# PEMBAHASAN

3.1. Proses pengujian dan hasil dari pengujian algoritma hutan acak dan pohon keputusan menggunakan tool rapid miner adalah sebagai berikut :



Gambar 3 Proses Pengujian dari algoritma Random Forest dan Decision Tree

 

Gambar 4 Hasil Prediksi dari Algoritma Random Forest



Gambar 5 Hasil Prediksi dari Algoritma Decison Tree

Dari data pegawai yang terdapat di India hasil dari prediksi klasifikasi algoritma menggunakan hutan acak dengan jumlah record sebanyak 4654 di spit untuk data training sebesar 90% dan data testing sebesar 10%. menunjukkan tingkat ketepatan dari algoritma hutan acak sebesar : **86,45%.** Sedangkan tingkat ketepatan dari algoritma dicision tree sebesar **84,30%.**

**3.2** Kemudian hasil dari besar sebaran validitas algoritma hutan acak dan pohon keputusan menggunakan confusion matrix untuk menentukan presisi, recall dan F1-Score untuk masing masing metode adalah sebagai berikut :



Gambar 6 Confusion Matrix dari Algoritma Random Forest



Gambar 7 Confusion Matrix dari Algoritma Decision Tree

##### KESIMPULAN

Dari hasil pengujian dari 2 algoritma klasifikasi ini didapatkan bahwa

1. Tingkat keakuratan dari algoritma hutan acak sebesar 86,45% lebih tinggi dibandingkan dengan algoritma pohon keputusan yaitu sebesar 84,30%
2. Presisi : Model memiliki kemampuan dalam mengklasifikasikan data pegawai yang meninggalkan pekerjaannya dengan benar saat model diprediksi untuk algoitma hutan acak sebesar 96,7 % dan algoritma pohon keputusan sebesar 95,73%
3. Recall : Model memiliki kemampuan dalam mengklasifikasikan data pegawai yang meninggalkan pekerjaannya dengan benar untuk algoitma hutan acak sebesar 84,7 % dan algoritma pohon keputusan sebesar 82,9%
4. Spesificity : Model memiliki kemampuan dalam mengklasifikasikan data pegawai yang tetap bekerja di perusahaannya dengan benar untuk algoritma hutan acak sebesar 91,4% dan algoritma pohon keputusan sebesar 88,4%
5. F1-Score : Ukuran keseimbangan antara presisi dan recall untuk algoritma hutan acak sebesar 90,2% dan algoritma pohon keputusan sebesar 88,8%

Oleh sebab itu kita dapat disimpulkan dari hasil proses pengujian bahwa algoritma menggunakan hutan acak lebih unggul dibandingkan dari algoritma pohon keputusan

##### References

[1] P. Han, Kamber, *Data Mining Concepts and Techniques*. 2012.

[2] F. A. Hermawati, “Data Mining,” no. January, 2018.

[3] T. Lan, H. Hu, C. Jiang, G. Yang, and Z. Zhao, “ScienceDirect A comparative study of decision tree , random forest , and convolutional neural network for spread-F identification,” *Adv. Sp. Res.*, vol. 65, no. 8, pp. 2052–2061, 2020, doi: 10.1016/j.asr.2020.01.036.

[4] L. Breiman, “Random Forest,” pp. 1–33, 2001.

[5] A. Prabowo, S. Wardani, R. W. Dewantoro, and W. Wesly, “Komparasi Tingkat Akurasi Random Forest dan Decision Tree C4 . 5 Pada Klasifikasi Data Penyakit Infertilitas,” vol. 4, no. 1, pp. 218–224, 2023, doi: 10.30865/klik.v4i1.1115.

[6] C. Science, “U niversity of L iège,” no. July, 2014.

[7] C. Curtis, C. Liu, T. J. Bollerman, and O. S. Pianykh, “Machine Learning for Predicting Patient Wait Times and Appointment Delays,” *J. Am. Coll. Radiol.*, no. Ml, pp. 1–7, 2017, doi: 10.1016/j.jacr.2017.08.021.

[8] P. Bhargav and K. Sashirekha, “A Machine Learning Method for Predicting Loan Approval by Comparing the Random Forest and Decision Tree Algorithms .,” vol. 10, pp. 1803–1813, 2023.

[9] N. Sunanto and G. Falah, “Penerapan Algoritma C4.5 Untuk Membuat Model Prediksi Pasien Yang Mengidap Penyakit Diabetes,” Rabit J. Teknol. dan Sist. Inf. Univrab, vol. 7, no. 2, pp. 208–216, 2022, doi: 10.36341/rabit.v7i2.2435.

[10] R. Estian Pambudi, Sriyanto, and Firmansyah, “Klasifikasi Penyakit Stroke Menggunakan Algoritma Decision TreeC.45,” Ijccs, vol. x, No.x, no. x, pp. 1–5, 2022.

[11] M. Ardiansyah, A. Sunyoto, and E. T. Luthfi, “Analisis Perbandingan Akurasi Algoritma Naïve Bayes Dan C4.5 untuk Klasifikasi Diabetes,” Edumatic J. Pendidik. Inform., vol. 5, no. 2, pp. 147–156, 2021, doi: 10.29408/edumatic.v5i2.3424.

[12] U. Erdiansyah, A. I. Lubis, and G. Syahputra, “Klasifikasi Penyakit Diabetic Retinopathy Menggunakan Multilayer Perceptron,” JAISE J. Artif. Intell. Softw. Eng., vol. 2, no. 1, pp. 1–6, 2022

[13] Svetnik V 2003 Random forest: a classification and regression tool for compound classification and QSAR modeling J. Journal of Chemical Information & Computer Sciences 1 43

[14] Zhong F and Zhang J 2013 Linear discriminant analysis based on l1-norm maximization IEEE Trans Image Process 22 3018-27

[15] Siddiqui M K, Moralesmenendez R and Ahmad S 2020 Application of receiver operating characteristics (roc) on the prediction of obesity Brazilian Archives of Biology and Technology 63 2

[16] Zhang G Y, Zhang C X and Zhang J S 2010 Out-of-Bag Estimation of the Optimal Hyperparameter in SubBag Ensemble Method Communications in Statistics - Simulation and Computation 39 1877-92