ANALISIS MENGGUNAKAN RANDOM FOREST DENGAN GINI INDEX ALGORITMA PADA DATA

Nuranisah
Fakultas Sosial Sains, Universitas Pembangunan Pancabudi
nuranisahasriel123@gmail.com

ABSTRACT

New Student Admission The achievement pathway has been listed in Permendikbud number 44 of 2019 in article 11 paragraphs 1 & 2, making it possible for new students who have achievements to not be able to enter the school they dream of being outside the zoning of residence. The limited quota for each school only receives 5% according to in Permendikbud to new students. Achievement should also be prioritized to encourage and motivate the interest of new students. Classification of achievement participants is carried out. Where is the dataset of prospective students at school before becoming prospective participants for the next school level, as a reference for increasing the quota of achievement pathways. Trial Random Forest method with the Gini Index Algorithm, testing using cross validation as the initial classification model, the dataset is divided into a test and training ratio of 70: 30 then test and analysis to the Random Forest method with the Gini Index Algorithm, the results obtained are 94.39% accuracy. But the need for further studies by utilizing more datasets and attributes and comparing to other methods because the test was carried out with 305 datasets, there were 5 attributes and 1 main attribute.

Keywords: Random forest, Gini Index, Algoritm

PENDAHULUAN

Tercantum pada permendikbud nomor 44 tahun 2019 , pada pasal 11 ayat 1dan 2 menyatakan proses penerimaan peserta didik baru salah satunya dilakukan secara zonasi sesuai pada keterangan permendikbud tercantum dimana kuota 50% diambil untuk jalur zonasi , 15% afirmasi dan masing-masing 5% prestasi dan pindah dari daya tampung sekolah. Kuota untuk penerimaan siswa berdasarkan zonasi sangat besar dibandingkan dengan kuota berdasarkan prestasi. Prestasi sangat di utamakan demi meningkatkan motivasi dan minat belajar peserta didik untuk kesekolah-sekolah yang diimpikannya yang berada diluar zonasi tempat tinggalnya.

Maka perlunya penyaringan untuk mengetahui peserta didik yang memiliki prestasi di setiap sekolah, sehingga menjadi acuan agar nantinya adanya perubahan peningkatan presentasi kuota untuk peserta didik yang berprestasi. Dalam prosesnya tersebut diperlukan cara untuk klasifikasi data dengan baik sebagai acuan untuk meningkatkan presentasi pada jalur prestasi. Untuk klasifikasikan data peserta didik mencoba dengan menggunakan metode *Random Forest* dengan *Algoritma Gini Index* yang memiliki kaitan erat dengan *data mining.*

Suatu teknik yang digunakan sebagai analisis dataset serta melakukan prediksi pada pola yang terkandung di suatu data merupakan *data Mining.* Didalam *data mining* dapat dicapai dengan beberapa teknik salah satunya adalah

Classification.. Sedangkan teknik pada pengumpulan data- datanya adalah klasifikasi, Ada beberapa mekanisme yang digunakan pada *data mining* salah satunya adalah *Random Forest* (FR) yang di kombinasikan dengan *Algoritma Gini Indeks*.

Adapun beberapa penelitian yang mengangkat mengenai salah satu model klasifikasi yang digunakan pada setiap penelitian mereka, seperti penelitian yang dilakukan dengan menggunakan Random Forest dan Multivariate Adaptive Regression Spline (MARS) binary response merupakan hasil penelitian dari, dimana variabel yang memiliki dominasi tertinggi pada status HIV/AIDS di Surabaya adalah usia, setelah itu jenis pekerjaan, pernah ditahan karena kasus NAPZA, status nikah, serta penggunaan jarum tertentu. Dimana hasil akurasi yang didapat adalah MARS sebesar 80,28%, lalu RF MARS dengan 91,00% serta hasil akurasi terbaik di dapat adalah metode RF dengan akurasi 97,80%.

Menggunakan Random forest untuk melakukan prediksi kelancaran kredit yang dilakukan oleh peneliti , dimana hasil akurasi yang didapat setelah melakukan beberapa meknisme skenario training menghasilkan akurasi sebesara 96,47%. Selanjutnya , peneliti menggunakan Analisis Random Forest, Multiple Regression dan metode Backpropagation dengan memprediksi indeks harga Apartemen di Indonesia menyatakan bahwa metode Backprogation menghasilkan akurasi yang lebih tinggi dari Random Forest serta Multiple Regression untuk meminimalisir kerugian investasi jual beli apartemen di masa pandemic Covid-19. Maka peneliti mencoba melakukan penelitian dengan menentukan metode Random Forest dengan Gini Indeks untuk meningkatkan dan mengetahui hasil akurasi sehingga dapat digunakan sebagai pertimbangan untuk menaikan jumlah kuota penerima peserta didik baru untuk peserta didik yang memiliki prestasi.

METODE PENELITIAN

Pada pengkajian untuk penelitian memerlukan langkah-langkah yang perlu dilakukan untuk mendapatkan hasil output yang di inginkan menggunakan metode Random Forest dengan Algoritma Gini Indeks , Maka memerlukan alur Flowchart untuk menjelaskan keseluruhan rangkaian proses yang dilakukan. Gambar 1 merupakan rangkaian Flowchart dari rancangan sistem yang dibuat.

Dataset yang digunakan pada penelitian ini adalah menggunakan dataset pada SMP Negeri 22 Medan. Pada dataset ini terdapat beberapa atribut-atribut , namun diantara dataset tersebut hanya 5 atribut yang akan digunakan berikut tertera tabel 1.

Tabel 1. Dataset

Atribut	Variabel
X1	Nilai rata-rata
X2	Kategori
X3	Absensi
X4	Prilaku

Penerapan metode random forest dapat meningkatkan hasil akurasi, dimana simpul anak untuk setiap node-node dilakukan dengan acak. Umumnya metode ini digunakan untuk membangun pohon keputusan yang terdiri dari root node, internal node dan leaf node dengan mengambil atribut dan data secara acak

sesuai ketentuan yang diberlakuan. Root node adalah simpul terletak paling atas, atau biasa disebut sebagai akar dari pohon keputusan. Internal node adalah impul percabangan, dimana node ini mempunyai output minimal 1 atau 2 input. Sedangkan leaf node atau terminal node merupakan simpul akhir memiliki satu input dan tidak mempunyai output. Pohon keputusan dimulai dari perhitungan nilai enropy sebagai penentuan tingkat ketidakmurnian atribut dan nilai information gain. Untuk menghitung nilai entropy digunakan rumus seperti pada persamaan 1:

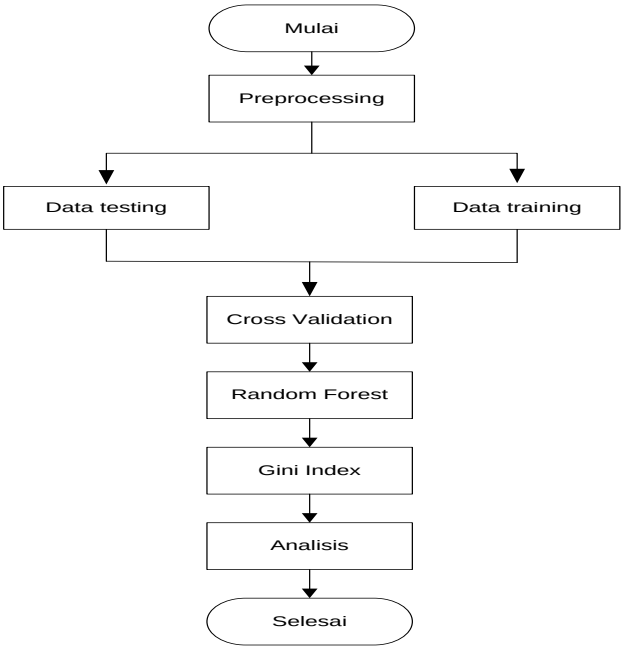
$$\boldsymbol{Entropy}\ (\boldsymbol{Y}) = -\boldsymbol{\Sigma_i}\ \mathrm{p}\ (\mathrm{c}|\mathrm{Y})\ \boldsymbol{log_2}\ \mathrm{p}\ (\mathrm{c}|\mathrm{Y}) \qquad (1)$$

PENERAPAN ALGORITMA GINI INDEX

Selanjutnya, penerapan Gini Index merupakan probabilitas dari dua data yang berbeda. Gini Index digunakan oleh Breiman, Friedman and Olshen (1948) [8]Agar mendapatkan hasil dari pohon klasifikasi pada decision tree. Misalkan S adalah 1 set dari sejumlah s data. Data ini memiliki sejumlah m class yang berbeda ($C_i, = i = 1, \dots, m$) . Berdasarkan pada class tersebut, kita bisa membagi dimana S di proses ke dalam jumlah m subset (S_i, i= 1, …,m) misalkan Si adalah dataset yang digabungkan ke dalam class C_i, S_i adalah jumlah daripada S_i maka dari itu Gini Index dapat dirumuskan sebagai berikut :

$$\textbf{Gini Index}\ (\mathbf{S}) = \mathbf{1} - \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \left(\frac{\boldsymbol{S_i}}{\boldsymbol{s}}\right)^{\mathbf{2}} \qquad (2)$$

Langkah awal dilakukan pada untuk memproses data di tunjukkan pada Gambar.1, adalah melakukan preprocessing data terlebih dahulu.



Gambar 1. Desain alur prediksi

Penjelasan pada Gambar.1. menjelaskan jika rancangan analisis yang dilakukan melewati beberapa tahapan diantaranya yaitu membaca dataset terlebih dahulu, tahapan preprocessing lalu data di agi menjadi dua rasio yaitu perbandingan 70 : 30 sebagai data testing dan data training setelah itu sebelum melakukan prediksi dengan menggunakan Random Forest dengan algortitma Gini Index terlebih dahulu melawati tahapan Cross Validation yang akan digunakana sebagai evaluasi kinerja prediktif dari model

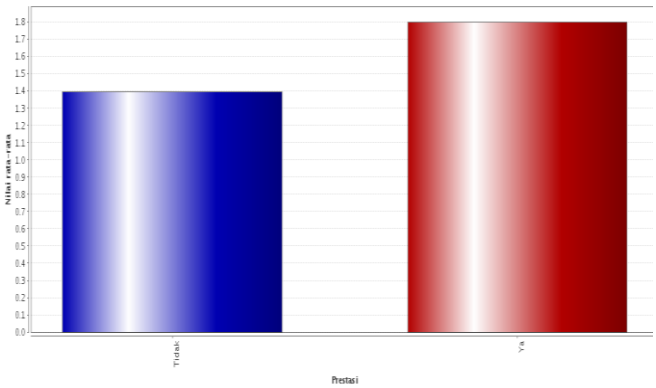
Pada penelitian ini penulis melakukan beberapa hal diantaranya :

1. Mencari node dan leaf , serta rule yang dihasilkan dari Random Forest
2. Tidak semua atribut pada kasus ini akan digunakan, dimana hanya menggunakan 5 atribut dan 1 diantarnya menjadi atribut target.
3. Uji data input dilakukan dengan split data perbandingan ratio 70 : 30
4. Proses dilakukan dari tahap preprocessing, tahap testing , training , lalu melakukan dengan proses cross validation sebelum masuk ke tahapan proses menggunakan Random forestdengan Gini Index.
5. Akurasi data.

Data yang digunakan merupakan data yang bersumber dari SMP Negeri 22 Medan. Data yang terdiri dari 8 atribut sebelum dilakukan pengolahan data untuk di uji.

HASIL PENELITIAN DAN PEMBAHASAN

Melakukan Split data dengan memisahkan dataset menjadi dua bagian dengan masing – masing ratio sebesar 70 : 30 secara acak. Maka didapatkan hasil perbanding jumlah persentasi melalui Gambar.2.



Gambar 2. Bagan hasil persentasi peserta didik yang berprestasi dan tidak

Confusion Matrix merupakan metode pengukur dari kinerja suatu metode klasifikasi yang mengandung sebuah informasi sebagai hasil klasifkasi yang dilakukan oleh sistem dengan hasil klasifikasi.

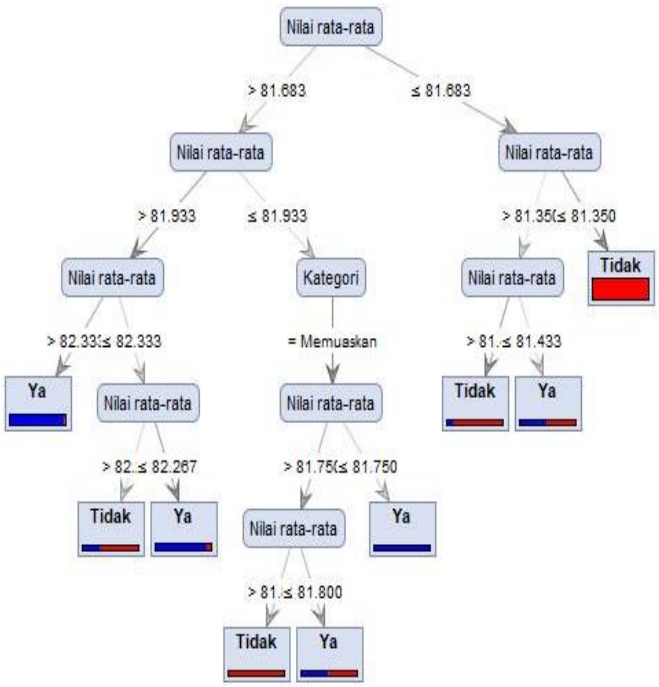
Tabel.2. Confusion Matrix

Kinerja Klasifikasi	Predicted Class	
Actual Class	Predicted.Ya	Predicted.Tidak
Actual.Ya	59 (True Positive)	7 (False Negative)
Actual.Tidak	5 (False Positive)	143 (True Negative)

Berdasarkan tabel.2. maka dilanjutkan dengan menghitung nilai Accuracy klasifikasi dari model klasifikasi Random Forest dengan Gini Index menggunakan dataset. Berikut hasil perhitungannya :

Accuracy : $\frac{TP+TN}{TP+TN+FP+FN} = \frac{59+143}{59+143+7+5} = \frac{202}{214}$ = 0.9439*100% = 94.39%

Pengetahuan yang menghasilkan oleh Random Forest dengan Algoritma Gini Index di presentasikan dengan pohon keputusan seperti pada gambar.3.



Gambar 3. Model Pohon Random Forest

Model Pohon dapat dibaca dari atas hingga ke bawah atau dari akar (simpul pertama paling atas) lalu ke daun (simpul terluar yang tidak lagi memiliki cabang). Berikut penjelasan sedikit mengenai gambar 3.

Apabila nilai > 81,083 , Karena menyatakan nilai lebih tinggi dari 81,083 maka langsung dinyatakan prestasi "YA" tetapi jika nilai lebih kecil dari 81,083 masuk ke simpul selanjutnya dengan atribut yang lain jika hasil nya "Memuaskan" dengan nilai > 81,933 maka hasilnya adalah prestasi " Ya". Dan begitu selanjutnya pada simpul-simpul yang terlampir lainnya.

KESIMPULAN DAN SARAN

Berdasarkan hasil percobaan yang dilakukan pada *Random Forest* dengan gini index yang di uji coba menunjukkan hasil yang cukup memuaskan. Dimana terdapat 94,39% hasil akurasi yang didapat. Model dan rule yang dihasilkan oleh *Random Forest* dengan algoritma *Gini Index* digunakan sebagai acuan dasar pengembangan. Untuk penelitian selanjutnya dapat dilakukan dengan dataset jumlah record yang lebih banyak dan melakukan perbandingan dengan metode lain , karena penulis hanya menggunakan 1 metode yang digabungkan untuk lakukan uji coba.

DAFTAR PUSTAKA

D. Larose, "Discovering Knowledge in Data : An Introduction to Data Mining," *Jhon Willey & Sons*, pp. Inc 129-240, 2005.

I. Breiman, J. Friedman, C. Shone and R. Oslen, "Classification And Regression Tress," 1984.

Kemendikbud, " Tentang Penerimaan Peserta Didik Baru," in *Permendikbud Nomor 44*, Tahun 2019.

M. I. Putra, A. Yusuf and N. Yalina, "Klasifikasi Kelancaran Kredit dengan Metode Random Forest," *Systemic : Information System and Informatics Journal*, vol. No. 5, pp. 7-12, 2 Desember 2019.

N. Y. Saputra, S. Saadah and P. E. Yunanto, "Analysis of Random Forest,Multiple Regresstion and Backprogation Methods in Predicting Apartment price index in Indonesia," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEK)*, vol. 7 No.2, no. ISSN : 2338-3070,001:10:26555/JITEKI.V712.20997, pp. 238-248, August 2021.

Nidhomuddin and B. Otok, "Random Forest and Multivariate Adaptive Regression Spline (Mars) Binary Response Untuk Klasifikasi Penderita HIV/AIDS Disurabaya," *Statistika*, vol. 3 (1), 2 Desember 2015.

P. Pumpuang, A. Srivihok and P. P, "Comparisons of Classifier Algorithms : Bayesian Network, C4.5 Decision Forest and NBTree for Course Registration Planning Model of Under," *2008 IEEE International Conference on Systems, Man and Cybernetics*, pp. IEEE. 3647-240, 2008.

S. JK, F. F and R.Dekker, "An-Information Gain-Driven Feature Study for Aspect-Based Sentiment Analysis," *In International Conference on Application of Natural Language to Information Systems*, pp. 48-59, 2016.