Gunung Djati Conference Series, Volume 3 (2021)
Mini Seminar Kelas *Data Mining* 2020
ISSN: 2774-6585
Website: https://conferences.uinsgd.ac.id/gdcs

Algoritma Decision Tree untuk Menentukan Jenis Kelamin berdasarkan Rekaman Suara

(*Decision Tree Algorithm for Determining Gender based on Sound Recording*)

Nina Nadia Syafitri Husein[1], Kamal Zaki Abdurrafi[2], Ryan Reliovani[3], Cecep Rafqi Al Husni[4], Muhammad Azka Khowarizmi[5], Deden Muhamad Furqon[6]

[1]Teknik Informatika, UIN Sunan Gunung Djati Bandung, ninanadiasyafitrihusein@gmail.com
[2]Teknik Informatika, UIN Sunan Gunung Djati Bandung, kamalzakiab@gmail.com
[3]Teknik Informatika, UIN Sunan Gunung Djati Bandung, ryan.reliovani@gmail.com
[4]Teknik Informatika, UIN Sunan Gunung Djati Bandung, ceceprafqi19@gmail.com
[5]Teknik Informatika, UIN Sunan Gunung Djati Bandung, azkakhowarizmi@gmail.com
[6]Teknik Informatika, UIN Sunan Gunung Djati Bandung, 1177050027@student.uinsgd.ac.id

Abstrak

Manusia terlahir dengan keunikan dan karakternya masing-masing, walaupun memiliki keunikan tersendiri, mengklasifikasikan manusia berdasarkan jenis kelamin merupakan hal yang tidak sulit untuk dilakukan. Pada manusia cara untuk membedakan pria dan wanita adalah dengan melihat perbedaan fisik, dan mendengarkan suara yang berbeda antara pria dan wanita. Pada komputer perbedaan jenis kelamin juga dapat dikenali dengan pengklasifikasian suara pria dan wanita menggunakan algoritma decision tree dengan dataset yang telah diproses sebelumnya berupa sampel rekaman suara suara pria dan wanita sebanyak 3168 data, sampel rekaman suara di pre-processed dengan analisis akustik dalam R menggunakan Seewave dan tuneR packages dengan interval frekuensi 0 hz - 280 hz. Meanfun digunakan sebagai prediktor pada root dataset suara, dengan ambang batas <= 0.142, dengan nilai kedalaman optimal 6 menggunakan metode cross validation, hasil yang dicapai yakni keakuratan training set sebesar 99,18809% dan akurasi test set mencapai 95,89905%.

Kata kunci: klasifikasi, decision tree, rekaman suara

Abstract

Humans are born with their own uniqueness and character, even though having a uniqueness that classifies humans by sex is something that is not difficult to do. In humans, the way to differentiate between men and women is to look at physical differences and listen to different voices between men and women. On the computer, gender differences can also be identified by classifying male and female voices using a tree decision algorithm with a previously appeared dataset in the form of a sample of 3168 male and female voice recordings, the voice recording sample is processed by acoustic analysis in R using Seewave and tuneR packages with frequency interval 0 hz - 280 hz. Meanfun is used as a predictor for the root sound dataset, with a threshold <= 0.142, with an optimal depth value of 6 using the cross validation method, the results achieved are the accuracy training set of 99.18809% and the accuracy test set reaches 95.89905%.

Keywords: *classification, decision tree, recorded voice*

Gunung Djati Converence Series
ISSN 2774-6585
Volume 3 (2021)

1 Pendahuluan

Klasifikasi jenis kelamin berdasarkan suara merupakan hal yang tidak sulit untuk diidentifikasi secara langsung dengan mendengarkan perbedaan suara pria dan wanita, para ahli menyatakan bahwa suara pria berada pada rentang 65 hingga 260 Hertz sedangkan frekuensi suara wanita terletak pada rentang 100 sampai 525 hertz [1].

Pada percobaan ini berfokus pada pengklasifikasian jenis kelamin berdasarkan dataset yang sebelumnya berupa sampel rekaman suara pria dan wanita yang telah di preprocessing dengan analisis akustik sehingga mendapatkan nilai-nilai statistik dari 20 variabel tunggal yakni meanfreq, sd, median, Q25, Q75, IQR, skew, kurt, sp.ent, mode, centroid, meanfun, minfun, maxfun, meandom, meandom, mindom maxdom, dfrange, dan modindx dan 1 variabel target yakni variabel label yang berisi nilai pria atau wanita pada setiap baris datanya. Algoritma Decision tree merupakan salah satu algoritma pengkasifikasi dipilih dalam penelitian ini. Variabel yang dipilih sebagai prediktor pada root adalah Meanfun(rata-rata frekuensi yang diukur), dengan nilai ambang batas <= 0.142.

Tujuan dari percobaan ini sendiri adalah untuk mengetahui apakah algoritma decision tree dapat mengklasifikasikan jenis kelamin berdasarkan dataset suara yang sebelumnya diolah dari rekaman suara pria dan wanita sebanyak 3168 data, dan dapat mengklasifikasikan jenis kelamin dengan akurasi yang tepat. Berdasarkan latar belakang diatas maka didapatkan rumusan masalah yang dibahas dalam percobaan ini adalah sebagai berikut: (1) Bagaimana implementasi algoritma decision tree untuk mengklasifikasikan jenis kelamin berdasarkan dataset suara? (2) Apakah algoritma decision tree dapat mengklasifikasikan jenis kelamin dengan akurasi yang tepat menggunakan dataset suara? Terdapat beberapa penelitian terdahulu yang terkait penelitian ini, antara lain: (1) sistem rekomendasi untuk turis dalam melakukan perjalanan menggunakan algoritma decision tree [2]; (2) disgnosis penyakit liver dengan menggunakan algoritma decision tree [3]; dan (3) klasifikasi penyakit jantung coroner dnegan membandingkan algoritma Naïve Bayes, Decision Treem Random Forest, dan K-Nearest Neighbor [4].

2 Metodologi

Data Mining adalah suatu ilmu pengetahuan yang digunakan untuk menguraikan penemuan pengetahuan didalam *database*. *Data Mining* merupakan proses yang menggunakan teknik statistik, matematika, *artificial intelligence*, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang berguna dan pengetahuan yang terkait dari berbagai *database* yang besar [5], [6]. *Data Mining* disebut juga sebagai *Knowledge Discovery in Database (KDD)* didefinisikan sebagai ekstraksi informasi potensial, implisit dan tidak dikenal dari sekumpulan data. Proses *Knowledge Discovery in Database* melibatkan hasil proses *data mining*, kemudian hasilnya diubah secara akurat dan menjadi informasi yang mudah untuk dimengerti [7]. Pada dasarnya terdapat enam elemen yang paling berpengaruh dalam teknik pencarian informasi atau pengetahuan dalam *Knowledge Discovery in Database* yaitu:

1. Mengerjakan data yang berjumlah besar,
2. Memerlukan efisiensi yang berkaitan dengan volume data,
3. Mengutamakan ketepatan atau keakuratan,
4. Membutuhkan pemakaian bahasa tingkat tinggi,
5. Menggunakan beberapa bentuk dari pembelajaran otomatis,
6. Menghasilkan hasil yang menarik.

Salah satu pendekatan dalam data mining adalah classification atau klasifikasi adalah suatu model dalam data mining yang paling umum. Tujuan dari model ini adalah untuk menganalisa data histori

yang disimpan dalam database dan secara otomatis menghasilkan suatu model yang bisa memprediksi perilaku di masa mendatang. Model ini terdiri dari generalisasi pada baris-baris data yang digunakan untuk *training* yang akan membantu membedakan *class-class.* Harapannya adalah agar model ini dapat digunakan untuk memprediksi *class-class* dari baris-baris lain yang belum diklasifikasikan, dan bisa secara akurat memprediksi peristiwa-peristiwa aktual mendatang.

Decision Tree yang merupakan salah satu algoritma atau metode dengan pendekatan klasifikasi, mengklasifikasikan sampel secara *topdown*, mulai dari simpul akar dengan menjaga jarak dengan hasil dari tes *node* internal, sampai simpul daun yang dicapai oleh kelas label. Keuntungan paling signifikan dari pohon keputusan adalah kenyataan bahwa pengetahuan dapat diekstraksi dan direpresentasikan dalam bentuk aturan klasifikasi *if-then* [8]. *Decision tree* juga berguna untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan sebuah variabel target. *Decision tree* memadukan antara eksplorasi data dan pemodelan, sehingga sangat bagus sebagai langkah awal dalam proses pemodelan bahkan ketika dijadikan sebagai model akhir dari beberapa teknik lain.

3 Hasil dan Pembahasan

3.1 Data awal

Data yang berbentuk .csv dimasukkan ke dalam dataset. Data berupa angka statistik rekaman data suara. Tiap baris data memiliki label yang berikan jenis kelamin.

	meanfreq	sd	median	Q25	Q75	IQR	skew	kurt	sp.ent	sfm	mode	centroid	meanfun	minfun	maxfun	meandom	mindom	maxdom	dfrange	modindx	label
0	0.059781	0.064241	0.032027	0.015071	0.090193	0.075122	12.863462	274.402906	0.893369	0.491918	0.000000	0.059781	0.084279	0.015702	0.275862	0.007812	0.007812	0.007812	0.000000	0.000000	male
1	0.066009	0.067310	0.040229	0.019414	0.092666	0.073252	22.423285	634.613855	0.892193	0.513724	0.000000	0.066009	0.107937	0.015826	0.250000	0.009014	0.007812	0.054688	0.046875	0.052632	male
2	0.077316	0.083829	0.036718	0.008701	0.131908	0.123207	30.757155	1024.927705	0.846389	0.478905	0.000000	0.077316	0.098706	0.015656	0.271186	0.007990	0.007812	0.015625	0.007812	0.046512	male
3	0.151228	0.072111	0.158011	0.096582	0.207955	0.111374	1.232831	4.177296	0.963322	0.727232	0.083878	0.151228	0.088965	0.017798	0.250000	0.201497	0.007812	0.562500	0.554688	0.247119	male
4	0.135120	0.079146	0.124656	0.078720	0.206045	0.127325	1.101174	4.333713	0.971955	0.783568	0.104261	0.135120	0.106398	0.016931	0.266667	0.712812	0.007812	5.484375	5.476562	0.208274	male
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3163	0.131884	0.084734	0.153707	0.049285	0.201144	0.151859	1.762129	6.630383	0.962934	0.763182	0.200836	0.131884	0.182790	0.083770	0.262295	0.832899	0.007812	4.210938	4.203125	0.161929	female
3164	0.116221	0.089221	0.076758	0.042718	0.204911	0.162193	0.693730	2.503954	0.960716	0.709570	0.013683	0.116221	0.188980	0.034409	0.275862	0.909856	0.039062	3.679688	3.640625	0.277897	female
3165	0.142056	0.095798	0.183731	0.033424	0.224360	0.190936	1.876502	6.604509	0.946854	0.654196	0.008006	0.142056	0.209918	0.039506	0.275862	0.494271	0.007812	2.937500	2.929688	0.194759	female
3166	0.143659	0.090628	0.184976	0.043508	0.219943	0.176435	1.591065	5.388298	0.950436	0.675470	0.212202	0.143659	0.172375	0.034483	0.250000	0.791360	0.007812	3.593750	3.585938	0.311002	female
3167	0.165509	0.092884	0.183044	0.070072	0.250827	0.180756	1.705029	5.769115	0.938829	0.601529	0.267702	0.165509	0.185607	0.062257	0.271186	0.227022	0.007812	0.554688	0.546875	0.350000	female

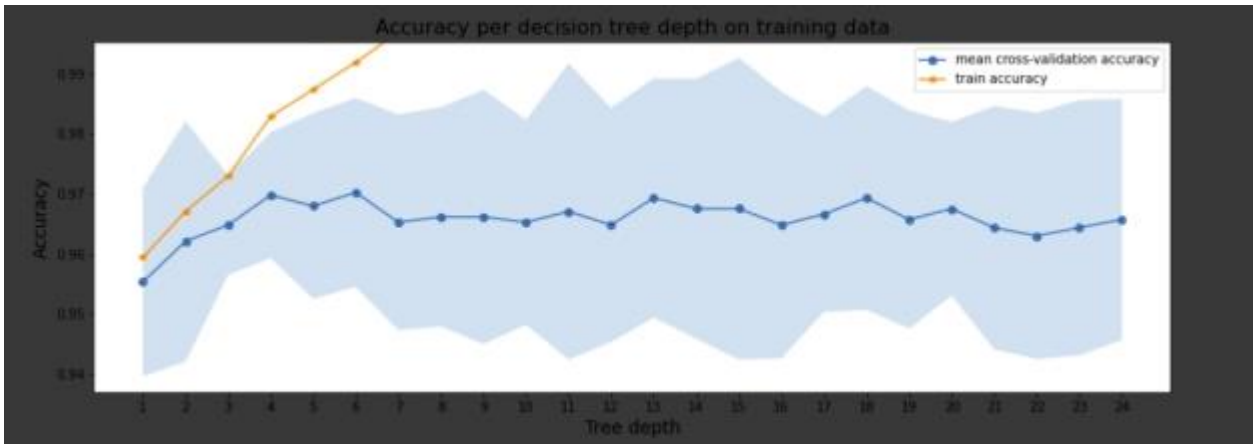
3168 rows × 21 columns

Gambar 1. Dataset yang didapatkan dari file .csv

3.2 Pre-processing

Menentukan kolom X dan Kolom Y, kolom X berisikan dari isi data yang akan diolah sedangkan Kolom Y berisikan label data yang menandai suatu baris data. Kami menggunakan cross-validation untuk mencari nilai kedalaman pohon yang optimal pada Algoritma decision tree yang akan digunakan. sehingga hasil tidak akan menjadi overfitting. Kedalaman pohon terbaik dinilai dari nilai akurasi dari tiap level kedalaman yang jadikan percobaan. Kedalaman pohon yang memiliki nilai akurasi tertinggi akan diambil sebagai nilai kedalaman pohon optimal. Disini kami mencoba dari kedalaman 1 sampai dengan kedalaman 24. Nilai kedalaman yang paling optimal dari hasil metode cross-validation yang telah dilakukan akan ditampilkan. disini ditemukan bahwa pohon dengan nilai kedalaman 6 menghasilkan nilai rata rata cross-validation terbaik.

Gunung Djati Converence Series
ISSN 2774-6585
Volume 3 (2021)



Gambar 2. Grafik tingkat akurasi per kedalaman pohon

```
[ ] idx_max = sm_cv_scores_mean.argmax()
    sm_best_tree_depth = sm_tree_depths[idx_max]
    sm_best_tree_cv_score = sm_cv_scores_mean[idx_max]
    sm_best_tree_cv_score_std = sm_cv_scores_std[idx_max]
    print('The depth-{} tree achieves the best mean cross-validation accuracy {} +/- {}% on training dataset'.format(
          sm_best_tree_depth, round(sm_best_tree_cv_score*100,5), round(sm_best_tree_cv_score_std*100, 5)))

    The depth-6 tree achieves the best mean cross-validation accuracy 97.02316 +/- 0.78586% on training dataset
```

Gambar 3. Format dataset sampel yang telah dicari nilai kedalamannya dengan cross validation

3.3 Implementasi Algoritma

Nilai kedalaman pohon optimal yang sudah didapatkan akan digunakan pada algoritma Decision Tree. Dilakukan percobaan untuk menjalankan satu pohon dengan data train dan data test. fungsi ini menghasilkan nilai akurasi data dari masing masing percobaan.

```
[ ] # function for training and evaluating a tree
    def run_single_tree(X_train, y_train, X_test, y_test, depth):
        model = DecisionTreeClassifier(max_depth=depth).fit(X_train, y_train)
        accuracy_train = model.score(X_train, y_train)
        accuracy_test = model.score(X_test, y_test)
        print('Single tree depth: ', depth)
        print('Accuracy, Training Set: ', round(accuracy_train*100,5), '%')
        print('Accuracy, Test Set: ', round(accuracy_test*100,5), '%')
        return accuracy_train, accuracy_test, model


    # train and evaluate a 6-depth tree
    sm_best_tree_accuracy_train, sm_best_tree_accuracy_test, ds_model = run_single_tree(X_train, Y_train,
                                                                        X_test, Y_test,
                                                                        sm_best_tree_depth)

    Single tree depth:  6
    Accuracy, Training Set:  99.18809 %
    Accuracy, Test Set:  95.89905 %
```

Gambar 4. Nilai akurasi dari masing-masing training set dan test set dari implementasi algoritma Decision Tree

Model yang sudah dibuat dari fungsi diatas kemudian divisualisasikan menggunakan pydotplus, matplotlib, dan IPython.

Gambar 5. Visualisasi dari model Decision Tree yang sudah dibuat

4 Simpulan

Berdasarkan percobaan kami mengenai implementasi algoritma decision tree. Dapat dilihat bahwa Algoritma decision tree dapat mengklasifikasikan dataset suara dengan baik menggunakan variabel - variabel yang telah ditentukan. Yakni dengan menenetukan kolom X dan Y terlebih dahulu, lalu mencari nilai kedalaman optimal menggunakan metode cross validation, kemudian implementasi algoritma decision tree. Dari percobaan yang telah dilakukan didapatkan bahwa nilai kedalaman optimal untuk dataset suara adalah 6, dengan akurasi training set mencapai 99,18809% dan akurasi test set mencapai 95,89905% pada algoritma decision tree.

Referensi

[1] I. T. Handoko and S. Suyanto, "Klasifikasi Gender dan Usia berdasarkan Suara Pembicara Menggunakan Hidden Markov Model," *Indones. J. Comput.*, vol. 4, no. 3, pp. 99–106, 2020.

[2] P. Thiengburanathum, S. Cang, and H. Yu, "A decision tree based recommendation system for tourists," in *2015 21st International Conference on Automation and Computing: Automation, Computing and Manufacturing for New Economic Growth, ICAC 2015*, 2015, doi: 10.1109/IConAC.2015.7313958.

[3] I. Setiawati, A. P. Wibowo, and A. Hermawan, "Implementasi Decision Tree Untuk Mendiagnosis Penyakit Liver," *JOISM J. Inf. Syst. Manag.*, vol. 1, no. 1, pp. 13–17, 2019.

[4] A. B. Wibisono and A. Fahrurozi, "PERBANDINGAN ALGORITMA KLASIFIKASI DALAM PENGKLASIFIKASIAN DATA PENYAKIT JANTUNG KORONER," *J. Ilm. Teknol. dan Rekayasa*, vol. 24, no. 3, pp. 161–170, 2019, doi: 10.35760/tr.2019.v24i3.2393.

[5] L. MeLian and A. Nursikuwagus, "Prediction Student Eligibility in Vocation School with Naïve-Byes Decision Algorithm," in *IOP Conference Series: Materials Science and Engineering*, 2018, vol. 407, no. 1, doi: 10.1088/1757-899X/407/1/012140.

[6] A. Nursikuwagus and T. Hartono, "IMPLEMENTASI ALGORITMA APRIORI UNTUK ANALISIS PENJUALAN DENGAN BERBASIS WEB," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 7, no. 2, p. 701, 2016, doi: 10.24176/simet.v7i2.784.

[7] S. Andayani, "Penggalian Informasi Potensial dari Basis Data di Perguruan Tinggi," *Data Min. Ann. Inf. Syst.*, p. 8, 2010, [Online]. Available: http://citeseerx.ist.psu.edu/

[8] H. Yu, X. Huang, X. Hu, and H. Cai, "A comparative study on data mining algorithms for individual credit risk evaluation," in *Proceedings - 2010 International Conference on Management of e-Commerce and e-Government, ICMeCG 2010*, 2010, pp. 35–38, doi: 10.1109/ICMeCG.2010.16.