Perbandingan Algoritma *Naive Bayes* dan *Linear Discriminant Analysis* dengan *Dataset Car Evaluation*

Farhan Abie Ardandy[1], Immanuel Morries Pohan[1], Ariq Mitsal[1], Finandra Nusantara[1], Muhammad Deka Ruliansyah[1]
[1]Teknik Elektro, Fakultas Teknik
Universitas Sriwijaya
Palembang, Indonesia
correponding author(s): farhanabie14@gmail.com

***Abstrak*— Keamanan, harga, dan kemewahan adalah faktor penting yang harus dipertimbangkan saat membeli mobil. Faktor-faktor ini tergantung pada jenis, model dan merek kendaraan. Padahal, faktor-faktor tersebut sangat penting dalam hal mengurangi angka kecelakaan. Akan tetapi banyaknya variable yang harus dipertimbangkan membuat konsumen sulit menentukan mobil yg akan dibeli dan rentan terhadap *human error.* Dengan permasalahan tersebut maka perlunya sistem pengambilan keputusan yang efisian salah satunya dengan menggunakan algoritma-algoritma *Machine Learning,* penulis mencoba mengetahui perbedaan antara metode *Naïve Bayes* dan *Linier Discriminant Analysis (LDA)* pada data set *Car Evaluation*. Diharapkan dari penelitian ini dapat diketahui akurasi dari kedua metode tersebut terhadap data set *Car Evaluation.* hasil dari pengujian yang kami lakukan dengan menggunakan kedua pengklasifikasian didapatkan hasil 73,98% dan 82,23% rentang waktu 0,002 dan 0,001.**

Kata Kunci— algoritma, car evaluation, klasifikasi, LDA, naive bayes

***Abstract*— Safety, price, and luxury are important factors to consider when buying a car. These factors depend on the type, model and make of the vehicle. In fact, these factors are very important in terms of reducing the number of accidents. However, the many variables that must be considered make it difficult for consumers to determine which car to buy and are prone to human error. With these problems, the need for an efficient decision-making system, one of which uses Machine Learning algorithms, the author tries to find out the difference between the Naïve Bayes method and Linear Discriminant Analysis (LDA) on the Car Evaluation data set. It is hoped that from this research, the accuracy of the two methods on the Car Evaluation data set can be known. the results of the tests that we carried out using both classifications obtained the results of 73.98% and 82.23% for the time range of 0.002 and 0.001.**

Keywords— algorithm, car evaluation, classification, LDA, naive bayes

I. PENDAHULUAN

Keamanan, biaya, dan kemewahan adalah faktor penting yang harus dipertimbangkan saat membeli mobil. Faktor-faktor ini tergantung pada jenis, model dan merek kendaraan. Padahal, faktor-faktor tersebut sangat penting dalam hal mengurangi angka kecelakaan. Perlengkapan standar adalah salah satu faktor yang perlu dipertimbangkan saat membeli mobil. Perlengkapan standar meliputi perlengkapan kenyamanan, perlengkapan produktivitas, perlengkapan, dan perlengkapan keselamatan. Membeli mobil adalah langkah besar menuju kemerdekaan, tetapi kemerdekaan memerlukan tanggung jawab, jadi sangat penting untuk memastikan bahwa mobil yang dibeli bernilai uang. Penting untuk memahami tanggung jawab finansial yang sebenarnya dari memiliki mobil. Pada penelitian ini digunakan dataset *Car Evaluation* untuk mencari atribut-atribut harga, pemeliharaan, pintu, bagasi, penumpang dan keselamatan. Data set tersebut akan diklasifikasi dengan menggunakan *Machine Learning* metode *Naive Bayes* dan *Linear Discriminant Analysis (LDA)*, hasil pengkalisfikasian akan ditemukan jenis mobil yang ideal untuk dibeli berdasarkan dataset *Car Evaluation*.

Penelitian terdahulu yang dilakukan oleh Yudhi Ramadhani(2015) dengan menggunakan metode yang sama yaitu menggunakan *Naïve Bayes Linear Discriminant Analysis (LDA)* untuk mengklasifikasikan kanker serviks. Hasil yang diperoleh dari percobaan menunjukkan bahwa Akurasi dari algoritma LDA baik untuk mengklasifikasikan data kedalam bentuk polynomial atau memiliki klasifikasi lebih dari 2 pilihan,sedangkan naïve bayes baik untuk klasifikasi binomial atau terhadap 2 pilihan dalam hal akurasi [1].

II. KAJIAN LITERATUR

A. Machine Learning

Pembelajaran mesin adalah sub bidang kecerdasan buatan (AI) di mana pengenalan pola telah dikembangkan dan digunakan untuk mengeksplorasi struktur dan model data yang dapat dipahami dan digunakan pengguna. Ini menjawab pertanyaan tentang cara membuat program komputer menggunakan data historis, menyediakan solusi untuk masalah tertentu, dan secara otomatis meningkatkan efektivitas program melalui pengalaman atau pelatihan [2].

B. Feature Selection

Feature selection adalah suatu tahapan dalam data *preprocessing* untuk mendapatkan fitur yang signifikan dan menyisihkan fitur yang tidak penting. Seleksi fitur adalah salah satu tahapan penting yang dapat mempengaruhi keakuratan data.[3]

C. Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis adalah sebuah metode klasifikasi yang bekerja dengan cara melakukan analisa dari matrik penyebaran yang bertujuan untuk mendapatkan proyeksi optimal yang dapat memaksimalkan varians antara kelas dan meminimalkan varians dari kelas data wajah. Algoritma LDA memiliki karakteristik perhitungan matriks yang hampir sama dengan PCA. Perbedaan mendasar adalah bahwa LDA berusaha untuk meminimalkan perbedaan dari foto kelas. Perbedaan antara kelas diwakili oleh matriks Sb (varian antar kelas), dan perbedaan antara dalam kelas diwakili oleh matriks Sw (varian dalam kelas). Matriks kovarians diperoleh dari dua matriks. Daya pembeda digunakan untuk memaksimalkan jarak antara kelas dan meminimalkan jarak dalam kelas.[4]

D. Naive Bayes

Naive Bayes adalah algoritma pembelajaran sederhana yang menggunakan aturan Bayesian dan asumsi kuat bahwa atribut dari kelas tertentu adalah independen bersyarat. Meskipun asumsi independensi ini sering dilanggar dalam praktiknya, *Naive Bayes* masih memberikan akurasi klasifikasi yang kompetitif. Dikombinasikan dengan efisiensi komputasi dan banyak properti lain yang diinginkan, ini sebenarnya membuat Naive Bayes tersebar luas.

Teorema Bayes menyediakan cara untuk menghitung probabilitas posterior P (c | x) untuk P (c), P (x), dan P (x | c). Pengklasifikasi *Naive Bayes* mengasumsikan bahwa efek dari nilai prediktor (x) kelas tertentu (c) tidak tergantung pada nilai-nilai prediktor lainnya. Asumsi ini disebut kondisi kelas independen. Algoritma naive bayes dapat dituliskan dalam persamaan dibawah ini.[5]

$$P(c|x) = \frac{P(c|x)P(c)}{P(x)} \quad (1)$$

dimana

P(c|x) adalah probabilitas posterior kelas (target) diberikan prediktor (atribut).

P(c) adalah probabilitas kelas sebelumnya.

P(x|c) adalah peluang yang merupakan peluang kelas yang diberikan oleh prediktor.

P(x) adalah probabilitas sebelumnya dari prediktor

E. Feature Selection

Pemilihan fitur adalah terkait erat terkait erat masalah minus dimensi di mana target adalah untuk mengidentifikasi karakteristik dari kumpulan data sama pentingnya dan menolak karakteristik lain seperti informasi yang tidak relevan dan berlebihan dan akurasi pemilihan dari di masa depan dapat ditingkatkan. Seleksi fitur adalah proses memilih fitur yang relevan, atau kandidat subset fitur. Kriteria evaluasinya adalah digunakan untuk mendapatkan subset fitur yang optimal. [6]

III. METODOLOGI PENELITIAN

A. Desain Penelitian

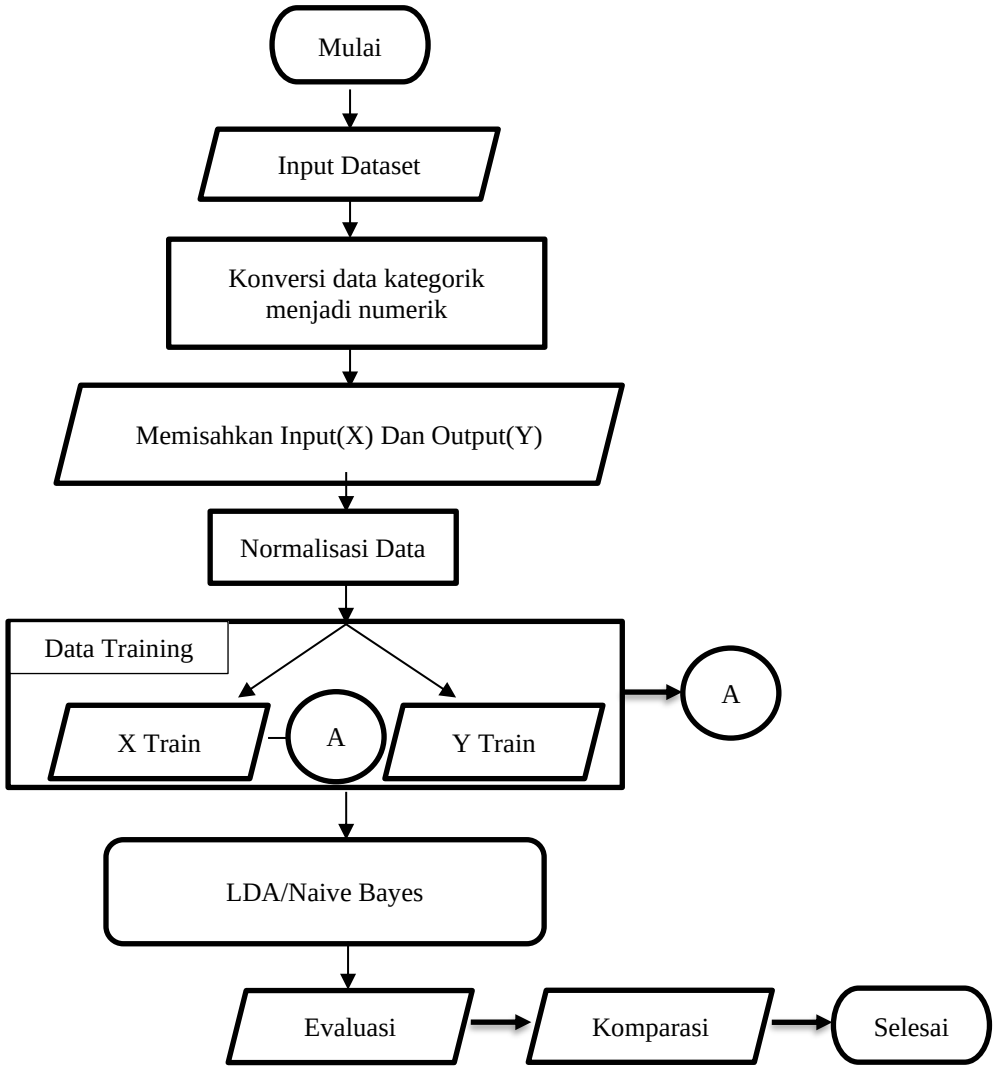
Desain penelitian adalah metode atau tahapan yang ditentukan oleh peneliti sebagai pedoman atau acuan untuk mendapatkan hasil dari tujuan penelitian yang sedang dilakukan.

B. Pengumpulan dataset

Penelitian kali ini kami mengambil dataset dari publik yang tersebar di internet untuk melakukan data *testing* dan data *training* yaitu dataset dari *Car Evaluation* yang diperoleh melalui halaman *website:* https://archive.ics.uci.edu/ml/datasets/Car+Evaluation. Basis data *Car Evaluation* adalah model data yang mengevaluasi mobil sesuai dengan struktur konsep diantaranya : harga keseluruhan,harga beli,harga perawatan,karakteristik teknis,kenyamanan,jumlah pintu,kapasitas penumpang,ukuran bagasi,dan keselamatan mobil.[7]

C. Metode yang dilakukan

Pada penelitian kali ini kami menggunakan beberapa tahapan untuk mendapatkan hasil yang optimal. Tahap pertama dimulai dengan melakukan input dataset. Tahap selanjutnya mengubah data kategorik menjadi data numerik. Lalu data diklasifikasikan menggunakan algoritma *Linear Discriminant Analysis (LDA)* atau *Naive Bayes.*Hasil terbaik dari klasifikasi akan dibandingkan dengan menggunakan nilai akurasi.



Gambar 1. Diagram alir penelitian

IV. PEMBAHASAN

Percobaan yang dilakukan dengan menggunakan 2 algoritma yaitu Naïve Bayes dan Linear Discriminant Analysis. Hal ini dimaksudkan untuk mengetahui mana pengklasifikasi yang paling sesuai dengan dataset dalam hal mengklasifikasikan data sebelum diolah, data *training*, data *testing*, dan pembuatan data *prediction* menggunakan model yang diperoleh dari proses *training*. Rincian prosedur dalam percobaan ini adalah:

A. Data Cleaning

Data yang diperoleh dari *UCI Dataset Repository* harus dibersihkan untuk memastikan bahwa data tersebut sesuai kualitas standar sebelum dimulainya pembuatan model. Pembersihan data dilakukan pada *dataset* seperti yang terlihat pada Tabel 1 yang terlihat sebagai konversi atribut nominal ke bentuk atribut numerik. Dan hal ini dilakukan dengan tujuan untuk memudahkan dalam proses menganalisa.[8]

B. Normalisasi Data

Normalisasi data menjadi sebuah proses yang sangat penting dalam *preprocessing data*. Karena melibatkan normalisasi dan agregasi. Normalisasi data adalah salah satu tahapan *preprocessing data* dimana data akan diskalakan atau diubah untuk membuat kontribusi yang sama dari setiap *feature.* Tahap normalisasi dapat dilakukan dengan banyak cara, dalam *paper* ini menggunakan *MinMaxScaler* untuk menormalkan *dataset*. Dan sebagai acuan, hasil normalisasi *min-max* selalu berada dalam rentang 0-1.[9]. Tabel 1 merupakan konversi data nominal ke numerik.

C. Dataset Split

Pre-processed dataset dibagi menjadi dua bagian dari berbagai ukuran pada waktu yang berbeda untuk digunakan sebagai *training dataset* dan *testing dataset* di berbagai algoritma klasifikasi *data mining* yang

berbeda untuk pembuatan model dan pengamatan dari model mana yang menghasilkan hasil yang terbaik.[10]

TABEL 1. KONVERSI DATA NOMINAL KE NUMERIK.

Attribute	Nominal	New Numeric Value
Class Values	Unaccepted	0
	Accepted	1
	Good	2
	Very good	3
Buying	Very high	3
	High	2
	Medium	1
	Low	0
Maintenance	Very high	3
	High	2
	Medium	1
	Low	0
Doors	Two	2
	Three	3
	Four	4
	Five or More	5
Persons	Two	2
	Four	4
	More	5
Lug Boot	Small	0
	Medium	1
	Big	2
Safety	Low	0
	Medium	1
	High	2

1. Training dan Testing

Dataset yang digunakan untuk *training* sebagian besar berasal dari *dataset* dari algoritma yang digunakan untuk mempelajari kelas atau hasil model yang dibuat dari setiap model, dan *dataset split* yang digunakan dalam penelitian ini dapat dilihat pada Tabel 2. Metode pembelajarannya adalah berdasarkan atribut atau *feature* dari *dataset* dalam perbandingan hasil atau kelas. Dan terakhir *output* adalah sebuah model yang digunakan untuk membandingkan separuh *dataset* lainnya, yaitu *testing data.*

TABEL 2. *DATASET SPLIT* UNTUK PEMBUATAN MODEL

Precentage Split	
Training	***Testing***
%	%
90	10
66	44

2. Classification

Model pelatihan menggunakan semua atribut termasuk *class attribute*. *Classification* dianggap sebagai *supervised model creation*, karena model dibangun berdasarkan *class values* yang terkorelasi dengan nilai-nilai atribut masing-masing.

TABEL 3. KLASIFIKASI AKURASI DARI *NAÏVE BAYES*

Naive Bayes					
Precentage Split		***Time in Seconds***		***Naive Bayes***	
Training	*Testing*	*Build*	*Test*	*Correct*	*Incorrect*
%	%			%	%
90	10	0,004	0,003	73,98	26,02
66	44	0,001	0,002	77,36	22,64

TABEL 4. KLASIFIKASI AKURASI DARI *LDA*

LDA					
Precentage Split		***Time in Seconds***		***LDA***	
Training	*Testing*	*Build*	*Test*	*Correct*	*Incorrect*
%	%			%	%
90	10	0,012	0,001	83,23	16,77
66	44	0,001	0,001	82,5	17,5

Pada kedua Tabel 3 dan 4 di atas didapatkan bahwa akurasi dari algoritma LDA lebih tinggi dibandingkan algoritma *Naïve Bayes* dengan nilai 83.23% untuk *percentage split* 90:10, dan untuk *percentage split* 66:44 dengan nilai 82.5%.

V. KESIMPULAN

Hasil klasifikasi dataset dari perbandingan dua pengklasifikasian menunjukan bahwa LDA dan Naive bayes memiliki tingkat akurasi yang berbeda *naive baye*s sebesar (73,98% dan 77,36 %) dan LDA sebesar (82,23% dan 82,5%) dengan perbandingan split data *training* dan *testing* sebesar (90:10 dan 66:44). Naive bayes memiliki dimensi yang lebih banyak daripada LDA, sehingga akurasi yang dihasilkan oleh algoritma LDA lebih tinggi, Jumlah data juga akan mempengaruhi persentase akurasi pada algoritma. Dalam melakukan *build* dan *test* hasil menunjukan bahwa pengklasifikasi *Naive Bayes* yang tercepat. Juga, diamati semakin dalam penelitian ini semakin kecil jumlah dimensi kelas dari suatu data set, maka semakin tinggi akurasinya.

DAFTAR PUSTAKA

[1] Y. Ramdhani, “Komparasi LDA dan Naïve Bayes Dengan Optimasi Fitur Untuk Klasifikasi,” vol. II, no. 2, pp. 434–441, 2015.

[2] G. I. Webb, “Encyclopedia of Machine Learning and Data Mining,” Encycl. Mach. Learn. Data Min., no. April, 2016, doi: 10.1007/978-1-4899-7502-7.

[3] V. Kumar, “Feature Selection: A literature Review,” Smart Comput. Rev., vol. 4, no. 3, 2014, doi: 10.6029/smartcr.2014.03.007.

[4] A. Tharwat, T. Gaber, A. Ibrahim, and A. E. Hassanien, “Linear discriminant analysis: A detailed tutorial,” AI Commun., vol. 30, no. 2, pp. 169–190, 2017, doi: 10.3233/AIC-170729.

[5] K. Vembandasamy, R. Sasipriya, and E. Deepa, “Heart Diseases Detection Using Naive Bayes Algorithm,” Int. J. Innov. Sci. Eng. Technol., vol. 2, no. 9, pp. 441–444, 2015.

[6] H. Ariesta and M. A. Kartawidjaja, “Feature Selection pada Azure Machine Learning untuk Prediksi Calon Mahasiswa Berprestasi,” TESLA J. Tek. Elektro, vol. 20, no. 2, p. 166, 2019, doi: 10.24912/tesla.v20i2.2993.

[7] Z. Rehman et al., “Performance evaluation of MLPNN and NB: A Comparative Study on Car Evaluation Dataset Smart Health View project IMAGE PROCESSING View project Performance evaluation of MLPNN and NB: A Comparative Study on Car Evaluation Dataset,” IJCSNS Int. J. Comput. Sci. Netw. Secur., vol. 18, no. 9, p. 144, 2018, [Online]. Available: https://www.researchgate.net/publication/332465875.

[8] Nurdiawan, O., & Salim, N. “Penerapan Data Mining Pada Penjualan Barang Menggunakan Metode Naive Bayes Classifier Untuk Optimasi Strategi Pemasaran.Jurnal Teknologi Informasi dan Komunikasi STMIK Subang, April 2014 ISSN: 2252-4517,” no. April, pp. 1–15, 2014.

[9] D. Singh and B. Singh, “Investigating the impact of data normalization on classification performance,” Appl. Soft Comput., vol. 97, p. 105524, 2020, doi: 10.1016/j.asoc.2019.105524.

[10] H. Trier, “Fordele ved at strukturere mundtlig kommunikation,” Ugeskr. Laeger, vol. 174, no. 26, p. 1796, 2012.