Volume 3 Issue 1 (2023) Pages 79-88 P-ISSN:2798-2696 E-ISSIN:2798-0057
JOSTECH
Journal of Science and Technology
Website: https://ejournal.uinib.ac.id/jurnal/index.php/jostech Email: jostech@uinib.ac.id

Predictive Analysis Pendidikan Menggunakan *Machine Learning* di Sumatera Barat

Fitri Rahmah Ul Hasanah ✉[1], Muhammad Kivlan Reftreka Nugraha[2]
Program Studi Matematika dan Sains Data, Universitas Andalas, Indonesia[1]
Program Studi Ekonomi, Universitas Andalas, Indonesia[2]
email: fitrirahmah26@yahoo.com[1], kivlannugraha93@gmail.com[2]

Received 30 Januari 2023, Accepted 25 Maret 2023, Published 31 Maret 2023

Abstrak

Penelitian ini menggunakan data yang besar (*big data*) yaitu data Susenas 2019. Fokus penelitian ini adalah Pendidikan. Pendidikan merupakan faktor penting dalam sebuah kehidupan, tetapi di Sumatera Barat dengan menggunakan data Susenas 2019 masih berada pada kategori rendah dan jauh dari rata-rata nasional. Pada penelitian ini akan memprediksi faktor-faktor yang mempengaruhi pendidikan di Sumatera Barat dengan metode pengklasifikasian data. Metode pengklasifikasian data dengan jumlah yang banyak telah berkembang, diantaranya adalah *Machine Learning*. *Machine Learning* merupakan bidang teknologi yang sedang marak digunakan pada masa sekarang untuk membuat algoritma dengan data yang berukuran besar (*big data*). Metode *machine learning* yang digunakan adalah *Naive Bayes* dan *Bagging*. Selanjutnya, dari dua model tersebut diuji dan menunjukkan bahwa model Naive Bayes memberikan kinerja terbaik dibadingkan model Bagging berdasarkan nilai akurasi, *sensitivity*, dan *specitivity*. Maka model *Naive Bayes* adalah model *machine learning* terbaik untuk memprediksi faktor-faktor yang mempengaruhi pendidikan yaitu anggota rumah tangga, jenis kelamin dan klasifikasi daerah.

Kata Kunci: *Machine Learning; Naïve Bayes; Bagging*

Abstract

This study uses big data, namely the socio-economic 2019 data. The focus of this research is Education. Education is an important factor in life, but in West Sumatra using the socio-economic 2019 data it is still in the low category and far from the national average. In this study will predict the factors that influence education in West Sumatra with the data classification method. Methods for classifying large amounts of data have been developed, including Machine Learning. Machine Learning is a field of technology that is currently being widely used to create algorithms with large data (big data). The machine learning method used is Naive Bayes and Bagging. Furthermore, the two models were tested and showed that the Naive Bayes model gave the best performance compared to the Bagging model based on the values of accuracy, sensitivity and specificity. So the Naive Bayes model is the best machine learning model for predicting the factors that affect education, namely household members, gender and regional classification.

Keywords: *Machine Learning; Naïve Bayes; Bagging.*

✉ Corresponding author

PENDAHULUAN

Pendidikan merupakan salah satu faktor yang dapat mengembangkan pola pikir, bakat dan spiritual seseorang. Pendidikan adalah usaha sadar dan terencana untuk mewujudkan suasana belajar dan proses pembelajaran agar peserta didik secara aktif mengembangkan potensi dirinya untuk memiliki kekuatan spiritual keagamaan, pengendalian diri, kepribadian, kecerdasan, akhlak mulia, serta keterampilan yang diperlukan dirinya, masyarakat, bangsa, dan negara. Pada dasarnya, orang tua menginginkan anak mereka dapat menempuh jenjang pendidikan yang maksimal. Jenjang Pendidikan yang diterapkan di Indonesia, di antaranya adalah pendidikan dasar Sekolah Dasar (SD, SMP dan sederjat), pendidikan menengah (SMA, SMK dan sederajat), pendidikan tinggi (diploma, sarjana, magister, spesialis dan doktor).

Pada kenyataannya, tidak semua orang tua mampu menyekolahkan anak anak mereka pada jenjang pendidikan tinggi. Hal ini disebabkan oleh beberapa faktor, diantaranya faktor keluarga, faktor lingkungan maupun faktor diri sendiri. Faktor keluarga berupa pendapatan yang diperoleh oleh kepala keluarga tidak mampu menyekolahkan anak mereka ke jenjang lebih tinggi, jumlah anak yang ditanggung juga merupakan faktor penghambat anak melanjutkan ke jenjang berikutnya, serta jenjang pendidikan terakhir orang tua juga menjadi faktor penghalang. Faktor lingkungan dapat berupa tempat tinggal yang berada di perkotaan atau pedesaan juga menjadi faktor penentu seseorang melanjutkan pendidikan atau tidak. Faktor diri sendiri dapat dilihat dari pola pikir seseorang terhadap masa depannya.

Berdasarkan faktor-faktor tersebut, jenjang pendidikan yang ditempuh masyarakat di Indonesia diklasifikasikan menjadi SD, SMP, SMA, S1, S2 dan S3. Pada penelitian ini, akan dibahas mengenai klasifikasi jenjang pendidikan masyarakat di provinsi Sumatera Barat pada tahun 2019. Dalam hal ini, data yang diolah bersumber dari data Susenas tahun 2019. Berdasarkan data yang diperoleh, data tersebut akan diklasifikasi berdasarkan jenjang pendidikannya.Hal ini bertujuan untuk mengetahui persentase dari jenjang Pendidikan yang ada di Sumatera Barat. Klasifikasi data adalah proses untuk menemukan model atau fungsi yang dapat menggambarkan dan membedakan kelas data atau konsep, dengan tujuan agar model tersebut dapat digunakan untuk memprediksi kelas yang belum diketahui dari suatu objek pengamatan [7].

Pada era sekarang, metode pengklasifikasian data dengan jumlah yang banyak ( ) telah berkembang, diantaranya adalah *Machine Learning*. *Machine Learning* mesin yang dikembangkan untuk bisa belajar dengan sendirinya tanpa arahan dari penggunanya dimana dalam pengembangannya berdasarkan disiplin ilmu lain seperti statistika, matematika dan data mining [18]. Machine learning dapat

digunakan dalam berbagai bidang, diantaranya lalu lintas, kedokteran, Pendidikan, industri, teknologi dan lain sebagianya [13]. Berdasarkan data yang bersifat klasifikasi dan regresi, dikelompokkan dalam *machine learning* dengan kategori *supervised learning*. Dalam *supervised learning* ini, terdapat beberapa metode yang dapat digunakan untuk klasifikasi dan regresi data, diantaranya adalah n*aïve bayes* dan *boostrap aggregating* (bagging).

Tahun 2014, Bustami melakukan penelitian dengan judul penerapan algoritma *naïve bayes* untuk mengklasifikasi data nasabah asuransi dengan hasil metode tersebut dapat mengklasifikasikan kelancaran nasabah dalam membayar asuransi [3]. Tahun 2016, Syarli dan Asrul membahas tentang metode *naïve bayes* untuk prediksi kelulusan dengan hasil yang diperoleh metode naïve bayes menghasilkan keakuratan yang efektif dalm memprediksi kelulusan [15]. Satu tehun kemudian, 2017 Astrid Novita melakukan penelitian tentang penerapan *naïve bayes* untuk perangkingan kegiatan di Fakultas TIK Universitas Semarang dengan hasil metode *naïve bayes* mampu mengklasifikasi kegiatan di fakultas tersebut [10]. Tahun selanjutnya, beberapa penelitian mengkaji tentang perbandingan dari beberapa metode *machine learning*, seperti Riri Nada Devita dan kawan-kawan membandingkan metode *naïve bayes* dan *k-nearest* untuk klasifikasi artikel berbahasa Indonesia. Dari penelitian tersebut menghasilkan bahwa metode *naïve bayes* memiliki kinerja yang lebih baik dibandingkan *k-nearest* [4]. Pada tahun 2021, dilakukan penelitian dengan judul penerapan *naïve bayes* mengklasifikasikan masyrakat miskin di Desa Lepak [9] dan akhir-akhir ini metode *naïve bayes* masih digunakan dalam mengklasifikasi data. Hal ini dapat dilihat, pada tahun 2022 adanya penelitian dengan judul penerapan model klasifikasi metode *naïve bayes* terhadap penggunaan akses internet [14]. Dalam hal ini, dapat dilihat bahwa metode tersebut efektif dalam mengklasifikasikan suatu data dalam berbagai bidang.

Selain metode *naïve bayes*, metode *bagging* juga digunakan dalam beberapa penelitian untuk mengklasifikasi suatu kajian. Tahun 2015, Rizki Tri Prasetio dan Pratiwi membahas tentang penerapan Teknik bagging pada algoritma klasifikasi untuk mengatasi ketidakseimbangan kelas data set medis [12]. Tahun 2019, Ahmad Rusadi dan kawan-kawan mengkaji Teknik *bagging* dan *boosting* algoritma CART untuk klasifikasi masa studi mahasiswa [ahmad rusadi]. Tahun 2020, Andi K dan Agung juga membhas mengenai penerapan Teknik *bagging* untuk meningkatkan akurasi klasifikasi pada algoritma *naïve bayes* dalam menetukan blogger professional [8]. Satu tahun berikutnya, Istiqomatul dan Pardomuan meneliti tentang penerapan *machine learning* dalam klasifikasi risiko kejadian berat badan lahir rendah di Indonesia dengan hasil beberapa kabupaten/kota meetode *bagging* menghasilkan hasil yang lebih baik dari pada metode lain [istiqomatul]. Pada tahun yang sama, Zhafira Haura juga membahas klasifikasi angka pencurian di Riau dengan *mulrivariate adaptive regression splines (MARS)* dan *Bosstrap Agregating MARS* dengan hasil *bagging* memiliki hasil akurasi yang lebih baik [5].

Beradasarkan pemaparan di atas, metode *naïve bayes* dan *bagging* memiliki hasil akurasi yang baik dalam mengklasifikasikan data. Metode tersebut juga mampu mengklasifikasi suatu objek dalam berbgai bidang kajian. Oleh karena itu, penelitian ini ingin membahas tentang hasil akurasi dari klasifikasi dengan metode *naïve bayes* dan *bagging* terhadap jenjang Pendidikan provinsi Sumatera Barat berdasarkan hasil SUSENAS 2019 serta ingin melihat hasil akurasi mana yang lebih baik dari kedua metode yang ada.

Machine Learning

Machine Learning merupakan bidang teknologi yang sedang marak digunakan pada masa sekarang untuk membuat algoritma dengan data yang berukuran besar *(big data). Machine Learning* termasuk ke dalam bidang turunan dari *Artificial Intelegence (AI).* Pada tahun 1988, Goldberg dan Holland mendefinisikan *machine learning* sebagai aplikasi komputer dan algoritma matematika yang diadopsi dengan cara pembelajaran yang berasal dari data dan menghasilkan prediksi di masa akan datang. Tahun 2017, Geron menyatakan pendapatnya, *machine learning* adalah ilmu dan seni tentang pemograman computer yang bisa belajar dari data. Sanjaya pada tahun 2020 menyatakan bahwa *Machine Learning* adalah model statistik atau algoritma yang dapat melakukan tugas spesifik tanpa perintah dengan mengandalkan sebuah pola tertentu [2]. Dalam pengelompokkan data, machine learning secara umum di bagi menjadi 4 kelompok, yaitu [2]:

a. *Supervised learning*, data set yang memiliki label.
b. *Unsupervised learning*, data set yang tidak memiliki label.
c. *Semi supervised,* dataset untuk pelatihan yang di mana sebagian memiliki label dan Sebagian lagi tidak memiliki label.
d. *Reinforcement learning,* model belajar menggunakan sistem *reward* dan *penalties.*

NAÏVE BAYESS (NAÏVE BAYESS CLASSIFIER)

Metode *naïve bayes* adalah suatu metode algoritma *machine learning* yang mampu mengklasifikasi data dengan menggunakan metode probabilitas dan statistik. Metode ini diperkenalkan oleh seorang ahli dari Inggris yang bernama Thomas Bayes. Menurut Bustami pada tahun 2013, klasifikasi *naïve bayes* diasumsikan bahwa ada atau tidak ciri tertentu dari sebuah kelas tidak ada hubungannya dengan ciri dari kelas lainnya. Dalam bayes (terutama *naïve bayes*) maksud independensi yang kuat adalah bahwa sebuah fitur pada sebuah data tidak berkaitan dengan ada atau tidaknya fitur lain dalam data yang sama . Selain itu, Han J dan Kamber, M menyatakan bahwa *Bayessian classifiers* mempunyai tingkat kecepatan dan akurasi yang tinggi ketika diaplikasikan dalam *database* yang besar [1]. Pengklsifikasian metode *naïve bayes* dilakukan dengan memilih probabilitas akhir (posterior) tertinggi dari masing-masing kelas [9].

Menurut Prastyo, bentuk umum *naïve bayes* dapat dilihat dalam persamaan berikut [9]:

$$P(Y|X_1,X_2,\ldots,X_p) = \frac{P(Y)P(X_1,X_2,\ldots,X_p|Y)}{P(X_1,X_2,\ldots,X_p)} \quad (1)$$

dimana:

$P(Y|X_1,X_2,\ldots,X_p)$: Probabilitas masuknya obyek dengan karakteristik variabel tertentu dalam kelompok *Y (posterior)*.

$P(X_1,X_2,\ldots,X_p|Y)$: probabilitas kemunculan variabel-variabel pada obyek yang masuk kelompok Y *(likelihood)*.

$P(Y)$: Probabilitas munculnya kelompok Y sebelum masuknya obyek *(prior)*.

$P(X_1,X_2,\ldots,X_p)$: Peluang kemunculan variabel-variabel pada obyek secara umum *(evidence)*.

Berdasarkan persamaan (1) dapat juga dituliskan secara sederhana, yaitu :

$$Posterior = \frac{Prior \times likelihood}{evidence} \quad (2)$$

Nilai posterior dapat dibandingkan dengan nilai-nilai posterior kelompok lainnya untuk menentukan kelompok suatu obyek yang diklasifikasikan. Nilai *evidence* selalu tetap untuk setiap kelompok pada satu sampel yaitu bernilai 1dan merupakan pembagi pada setiap kelompok sehingga dalam perhitungan posterior hanya cukup mengalikan nilai prior dengan *likelihood*. Nilai prior yang merupakan peluang munculnya kelompok Y sebelum masuknya obyek dapat dihitung menggunakan persamaan sebagai berikut [9]:

$$P(Y_g) = \frac{n_g}{N} \quad (3)$$

Dimana:

$P(Y_g)$: Peluang munculnya kelompok Y ke-g, $g = 1,2,3,\ldots,q$

n_g : banyak pengamatan pada kelompok ke-g

Persamaan (1) dapat dijabarkan dengan menggunakan aturan perkalian, sehingga menjadi:

$$\begin{aligned} P(Y|X_1,X_2,\ldots,X_p) &= P(Y)P(X_1|Y)P(X_2,X_3,\ldots,X_p|Y,X_1) \\ &= P(Y)\,P(X_1|Y)\,P(X_2,X_3,\ldots,X_p|Y,X_1)P(X_3,X_4,\ldots,X_p|Y,X_1,X_2) \\ &= P(Y)P(X_1|Y)\,P(X_2|Y,X_1)P(X_3|Y,X_1,X_2)\ldots P(X_4,X_5,\ldots,X_p|Y,X_1,X_2,X_3) \\ &= P(Y)P(X_1|Y)\,P(X_2|Y,X_1)P(X_3|Y,X_1,X_2)\ldots P(X_p|Y,X_1,X_2,X_3,\ldots,X_{p-1}) \end{aligned}$$

Persamaan di atas memiliki faktor-faktor yang kompleks untuk dianalisa. Oleh karena itu, digunakan asumsi independensi yang sangat tinggi *(naïve)* dengan $(X_1, X_2, ..., X_p)$ saling bebas *(independent)* satu sama lain sehingga berlaku kesamaan sebagai berikut [wiwit]:

$$P(X_a|X_b) = \frac{P(X_a \cap X_b)}{P(X_b)} = \frac{P(X_a)P(X_b)}{P(X_b)} = P(X_a) \qquad (4)$$

untuk $a \neq b$ sehingga $P(X_a|Y, X_b) = P(X_a|Y)$

Pada persamaan (4) dapat disimpulkan bahwa asumsi independensi *naïve* tersebut membuat syarat peluang menjadi sederhana sehingga perhitungan menjadi dapat dilakukan. Penjabaran tersebut dapat disederhanakan menjadi :

$$\begin{aligned}\left(Y\middle|X_{1,}X_2, ..., X_p\right) &= P(Y)P(X_1|Y)P(X_2|Y)P(X_3|Y)...P\left(X_p\middle|Y\right) \\ &= P(Y)\prod_{k=1}^{p} P(X_k|Y)\end{aligned} \qquad (5)$$

Persamaan (5) merupakan model dari teorema *naïve bayes* yang digunakan dalam klasifikasi.

BOOTSTRAP AGGREGATING (Bagging)

Tahun 1996, Breiman merancang algoritma *bagging*, yang menyatakan *bagging* adalah sebuah metode yang menghasilkan berbagai versi model dan melakukan prediksi atau estimasi dengan merata-ratakan berbagai versi model tersebut. *Bagging* merupakan algoritma pertama yang dimodelkan untuk meningkatkan stabilitas dan akurasi algoritma pada data yang bersifat klasifikasi dan regresi. *Bagging* adalah (Alypadin, 2010) sebuah algoritma pembelajaran yang stabil pada perubahan kecil dalam training set menyebabkan perbedaan besar dalam peserta didik yang dihasilkan, yaitu algoritma belajar pada data yang memiliki varians tinggi (noise). *Bagging* dikenal sangat efektif ketika pengklasifikasi tidak stabil, yaitu ketika perturbing set belajar dapat menyebabkan perubahan signifikan dalam perilaku klasifikasi, karena *bagging* meningkatkan kinerja generalisasi karena pengurangan varians *(noise)* tetap terjaga atau hanya sedikit meningkatkan bias (Kim & Kang, 2012). Pada dasarnya, *bagging* menerapkan konsep meminimalkan variansi dan *overfitting*.

Algoritma yang digunakan pada *bagging* menurut Friedman sebagai berikut:

1. Buat sampel *boostrap* $\{(X_1^*, Y_1^*), (X_2^*, Y_2^*), ..., (X_n^*, Y_n^*)\}$ dengan penggantian secara acak dari data *training* $\{(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)\}$ mencocokkan dengan *classifier* $C_b$dinyalakan pada sampel yang sesuai *bootstrap*.
2. Output *classifier* akhir :

$$C(x) = B^{-1} \sum_{b=1}^{B} C_b(x) \quad (6)$$

EVALUASI KLASIFIKASI

Evaluasi klasifikasi bertujuan untuk memilih metode algoritma yang lebih dari beberapa metode yang ada dengan melihat kinerja evaluasi klasifikasi. Pada penelitian ini, ukuran kinerja kalasifikasi dilihat berdasarkan *confussion matrix. Confussion matrix* adalah alat yang berguna untuk menganalisis seberapa baik atau seberapa akurat metode klasifikasi dapat mengenali objek pengamatan dari kelas yang berbeda [17]. Berikut ini tabel yang menyajikan *confussion matrix* :

Tabel 1. *Confussion matrix*

Confussion matrix		Kelas Aktual		Total
		Yes	No	
Kelas Prediksi	Yes	TP	FP	P′
	No	FP	TN	N′
Total		P	N	

Ukuran kinerja klasifikasi yang dapat diperoleh dari *Confussion matrix* adalah seperti berikut:

$$Akurasi = \frac{TP + TN}{TP + FP + FN + TN} \quad (7)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (8)$$

$$Specifity = \frac{TN}{TN + FP} \quad (9)$$

$$G - mean = \sqrt{Sensitivity \times Specifity} \quad (10)$$

Sensitivity dan *specifity* merupakan ukuran statistic dalam pengklasifikasian biner yang digunakan untuk memilih model yang paling efisien.

HASIL DAN PEMBAHASAN

Sebelum dilakukan pemodelan spesifikasi maka ditampilkan statistik deskriptif terhadap keseluruhan variabel yang digunakan pada penelitian ini

Tabel 2. Frekuensi Variabel Pendidikan

Pendidikan	Jumlah	Persentase
Rendah	**26685**	**73,1 %**

Tinggi	**9844**	**26,9 %**

Pada variabel pendidikan pada data Susenas 2019 terlihat bahwa pendidikan di Sumatera Barat masih berada pada kategori rendah yaitu sekitar 26685 rumah tangga atau 73,1 % di Sumatera Barat masih berpendidikan rendah sedangkan rumah tangga yang berpendidikan tinggi di Sumatera Barat berjumlah 9844 atau 26,9 %. Hal ini menandakan bahwa di Sumatera Barat masih memiliki pendidikan yang masih jauh dari kata baik.

Tabel 3. Frekuensi Variabel Anggota Rumah Tangga

Anggota Rumah Tangga	Jumlah	Persentase
Banyak	**27002**	**73,9 %**
Sedikit	**9527**	**26,1 %**

Pada variabel anggota rumah tangga pada data Susenas 2019 terlihat bahwa di Sumatera Barat 27022 atau 73,99 % masih memiliki anggota rumah tangga yang banyak atau tanggungan yang banyak. Sedangkan 9527 atau 26,1 % memiliki anggota rumah tangga yang sedikit atau memiliki tanggungan yang sedikit.

Tabel 4. Frekuensi Variabel Jenis Kelamin

Jenis Kelamin	Jumlah	Persentase
Laki-Laki	**18076**	**49,5 %**
Perempuan	**18453**	**50,5 %**

Pada variabel jenis kelamin pada data Susenas 2019 terlihat bahwa di Sumatera Barat terdapat 18076 responden atau 49,5 % yang berjenis kelamin laki-laki. Sedangkan yang berjenis kelamin perempuan berjumlah 18453 atau 50,5 %. Maka dapat disimpulkan bahwa responden yang terbanyak adalah perempuan.

Tabel 5. Frekuensi Variabel Klasifikasi Daerah

Klasifikasi Daerah	Jumlah	Persentase
Desa	**20904**	**57,2 %**
Kota	**15625**	**42,8 %**

Pada variabel klasifikasi daerah pada data Susenas 2019 terlihat bahwa di Sumatera Barat masih banyak rumah tangga yang tinggal di desa yaitu berjumlah 20904 rumah tangga atau 57,2 %. Sedangkan rumah tangga yang tinggal di kota berjumlah 15625 atau 42,8%.

Pada data Susenas 2019 di Sumatera Barat masih tertinggal jauh dalam segi Pendidikan yang di mana faktor-faktor yang mempengaruhi Pendidikan tersebut pada penelitian ini adalah banyak anggota rumah tangga (banyaknya tanggungan), jenis kelamin dan klasifikasi daerah tempat tinggal.

Model *Machine Learning* yang digunakan mampu untuk menjelaskan faktor-faktor yang mempengaruhi pendidikan tersebut. Akan tetapi sangat banyak model *Machine Learning* yang dapat digunakan untuk menjelaskan hal tersebut. Model *Machine Learning* yang digunakan pada penelitian ini adalah *Bagging* dan *Naïve Bayes*. Untuk melihat model yang terbaik pada metode *Machine Learning* adalah melihat Akurasi, *Sensitivity* dan *Specitivity* yang paling baik atau paling besar nilainya.

Tabel 6. Perbandingan Model *Machine Learning*

Model	Akurasi	*Sensitivity*	*Specitivity*
Bagging	0,7310	1,000	0,000
Naive Bayes	0,7313	1,000	0,000

Setelah dilakukan pengolahan pada aplikasi R terhadap model *Machine Learning* yaitu model *Bagging* dan *Naïve Bayes* di dapatkan nilai akurasi, sensitivity dan specitivity. Pada model *Bagging* didapatkan nilai akurasi sebesar 0,7310, nilai sensitivity sebesar 1,000 dan specitivity sebesar 0,000. Sedangkan model *Naïve Bayes* didapatkan nilai akurasi sebesaar 0,7313, nilai *sensitivity* sebesar 1,000 dan nilai specitivity sebesar 0,000. Dari hasil ini maka didapatkan nilai *sentivity* dan *specitivity* yang sama antara model Bagging dan *Naïve Bayes*, tapi tingkat akurasi model antara keduanya berbeda. Model Machine Learning *Naïve Bayes* lebih memiliki tingkat akurasi yang tinggi dari pada model Bagging. Maka dapat dikatakan model *Naive Bayes* lebih baik dari pada model *Bagging* untuk memprediksi faktor yang mempengaruhi Pendidikan yaitu anggota rumah tangga, jenis kelamin dan klasifikasi daerah pada data Susenas 2019

Pada model *Naïve Bayes* merupakan metode yang cocok untuk klasifikasi biner dan *multiclass*. Metode yang juga dikenal sebagai *Naïve Bayes Classifier* ini menerapkan teknik *supervised* klasifikasi objek di masa depan dengan menetapkan label kelas ke *instance*/catatan menggunakan probabilitas bersyarat. Probabilitas bersyarat adalah ukuran peluang suatu peristiwa yang terjadi berdasarkan peristiwa lain yang telah (dengan asumsi, praduga, pernyataan, atau terbukti) terjadi. Istilah *supervised* merujuk pada klasifikasi *training data* yang sudah diberi label dengan kelas. Pada model *Naïve Bayes* ini digunakan data *training* dan data testing. 80% dari data digunakan untuk data training dan 20% digunakan untuk data testing. *Naive Bayes* juga dikenal dengan metode Machine Learning yang klasifikasi data berdasarkan faktor-faktor probabilitas.

Melihat struktur data yaitu dummy dengan 2 faktor analisis (1 dan 0), hasil akurasi, *sensitivity* dan *specitivity* dan melihat deskripsi dari Model *Naïve Bayes*, maka untuk memprediksi faktor yang mempengaruhi Pendidikan yaitu anggota rumah tangga, jenis kelamin dan klasifikasi daerah pada data Susenas 2019 model Naive Bayes sangat cocok untuk digunakan.

SIMPULAN

Penelitian ini melakukan perbandingan pemodelan *Machine Learning* dengan menggunakan model *Naïve Bayes* dan *Bagging*. Penelitian ini menggunakan data yang besar (*big data*) yaitu data Susenas 2019. Pada kasus *imbalanced* data dan set data besar terbukti mampu meningkatkan ketepatan klasifikasi yang dapat diihat dari nilai *sensitivity* yang tinggi dibandingkan data asli (tanpa *treatment*). Selanjutnya, dari kedua model *Machine Learning* yang diuji menunjukkan bahwa model *Naive Bayes* memberikan kinerja terbaik berdasarkan nilai akurasi, *sensitivity*, dan *specitivity* tertinggi. Maka pada model *Naive Bayes* lebih baik dari pada model *Bagging* untuk memprediksi faktor-faktor yang mempengaruhi Pendidikan yaitu anggota rumah tangga, jenis kelamin dan klasifikasi daerah. Pada penelitian selanjutnya, pemodelan klasifikasi dengan skema *both sampling* dapat dibandingkan dengan teknik *resample*

lain dalam penanganan kasus *imbalanced data* agar kinerja klasifikasi lebih optimal.

DAFTAR PUSTAKA

[1]. Arrahimi, Ahmad Rusadi, dkk. 2019. Teknik *Bagging* dan *Boosting* pada Algoritma *CART* untuk Klasifikasi Masa Studi Mahasiswa. Jurnal Sains dan Informatika Vol. 5 No. 1.

[2] Astuti, Fitri Andri. 2021. Pemanfaatan Teknologi *Artifficial Intelligence* untuk Penguatan Kesehatan dan Pemulihan Ekonomi Nasional. Jurnal Sistem Cerda, vol. 4 No.1.

[3] Bustami. 2014. Penerapan Algoritma *Naïve Bayes* Untuk Mengklasifikasi Data Nasabah Asuransi. Jurnal Informatika, vol. 8, No. 1.

[4] Devita, Riri Nada dkk. 2017. Perbandingan Kinerja Metode Naïve Bayes dan K-Nearest Neighbor untuk Klasifikasi Artikel Berbahasa Indonesia. JTIIK, vol. 5 No. 4.

[5] Haura, Zhafira dan Hariso. 2021. Klasifikasi Angka Pencurian di Riau dengan *Multivariate Adaptive Regression Splines* (MARS) dan *Boostrap Aggregaring* MARS. Jurusan Matematika dan Ilmu Pengetahuan Alam Universitas Riau.

[6] Irawan, Devi dkk. 2020. Perbandingan Klasifikasi SMS Berbasis Support Vector Machine, Naïve Bayes Classifier, Random Forest, dan Bagging Classifier. Jurnal SISFOKOM, vol 10 No. 3.

[7] J . Han, M. Kamber, and J. Pei, "Data Mining Concepts and Techniques," Third Edit., Elsevier, 2012.

[8] Kurniawan, Andi dan Prihandono, Agung. 2020. Penerapan Teknik *Bagging* untuk Meningkatkan Akurasi Klasifikasi Pada Algoritma *Naïve Bayes* dalam Menentukan Blogger Profesional. Jurnal Bisnis Digitasi dan Sistem Informasi, Vol. 1 No. 1

[9] Nurmayanti, Wiwit Pura. 2021. Penerapan *Naïve Bayes* dalam Mengklasifikasikan Masyarakat Miskin di Desa Lepak. Jurnal Kajian Ilmu dan Pendidikan Geografi, vol. 5 No.1.

[10] Putri, Astrid Novita. 2017. Penerapan *Naïve Bayesian* untuk Perangkingan Kegiatan Fakultas TIK Universitas Semarang. Jurnal SIMETRIS vol 8 No. 2.

[11] Putro, Hakam M., Vulandari, Retno Tri., Saptomo, Wawan L. 2020. Penerapan Metode Naïve Bayes untuk Klasifikasi Pelanggan. Jurnal TikomSIN, vol. 8 No. 2.

[12] Prasetio, Riski Tri. 2015. Penerapan Teknik Bagging Pada Algoritma Klasifikasi Untuk Mengatasi Ketidakseimbangan Kelas Dataset Medis. Informatika, vol 2 No. 2.

[13] Roihan, Ahmad dkk. 2020. Pemanfaatan *Macine learning* dalam Berbagai Bidang. IJCIT, vol 5 No. 1.

[14] Susana, Heliyanti, Suarna, Nana, Fathurrohman, Kaslani. 2022. Penerapan Model Klasifikasi Metode *Naïve Bayes* terhadap Penggunaan Akses Internet. JURSISTEKNI Vol. 4 No. 1.

[15] Syarli dan Muin, A. Ashari. 2016. Metode *Naïve Bayes* untuk Memprediksi Kelulusan. Jurnal Ilmiah Ilmu Komputer, vol. 2 No 1.

[16] Taufiq, Nuri dan Mariyah, Siti. 2021. Pendekatan Model *Machine Learning* dalam Pemeringkatan Status Sosial Ekonomi Rumah Tangga di Indonesia. Seminar Nasional Official Statistics 2021.

[17] Wibowo, Ari. 2015. Analisis Pebandingan Kinerja Metode Klasifikasi dalam Data Mining. Jurnal Integrasi Vol. 7 No. 1.

[18] Yuliati, Istiqomatul F dan Sihombing, P. Robinson. 2021. Penerapan Metode Machine Learning dalam Klasifikasi Risiko Kejadian Berat Badan Lahir Rendah di Indonesia. Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer, vol. 20 No. 2.