Nas_Jurnal #05 METODE SAMPLE BOOSTRAPING

by Purwanto Purwanto

Submission date: 29-Mar-2020 07:26AM (UTC+0700)
Submission ID: 1284134076
File name: Nas_Jurnal_05_METODE_SAMPLE_BOOSTRAPING.pdf (431.71K)
Word count: 4180
Character count: 25689

METODE *SAMPLE BOOSTRAPING* PADA *K-NEAREST NEIGHBOR* UNTUK KLASIFIKASI STATUS DESA

Eko Siswanto[1], Suprapedi[2], Purwanto[3]
[123]Pasca Sarjana Teknik Informatika Universitas Dian Nuswantoro

ABSTRACT

The Ministry of Rural Area, Remote Area Development and Transmigration divides village itself into five villages, such as, Independent Village, Advance Village, Developing Village, Remote Village and Very Extremely Remote Village. The data are based on Village Potency (Podes) in 2014 by the Ministry of Rural Area, Remote Area Development and Transmigration. It is necessary to know that the data of The Ministry of Rural Area, Remote Area Development and Transmigration can be used to predict the relationship between village development indicators and the status of villages. In this case, it means whether the indicators, which are built, can influence the status of villages or not and whether they can make the status of villages become better than before. k-Nearest Neighbor (k-NN) algorithm is a method which is used to classify toward by object based on k as the nearest neighbor. k-Nearest Neighbor (k-NN) algorithm has the strength as the effective and simple algorithm and it has been used by many problem classifications. However, it has weakness if it is used for the big dataset. It can happen because it needs higher computation time. In this research, Bootstrapping Sample method is proposed to increase the optimalization of computation accuracy and time on Boostraping Sample method. Based on this research, by using the integration of k-Nearest Neighbor (k-NN) algorithm with Bootstrapping Sample method on IPD dataset on Jepara in 2014, apparently it can increase the accuracy until 5.41 % (91.89%-97.30%) than using standard k-NN algorithm. The last, from the result of this research it can be inferred that by using the integration of K-Nearest Neighbor (k-NN) algorithm with Boostraping Sample method shows the better accuracy than using standard k-NN algorithm.

Key words: K-NN algorithm and Boostraping

1. PENDAHULUAN

Kementerian Desa, Pembangunan Daerah Tertinggal, dan Transmigrasi (Kementerian Desa PDTT) adalah kementerian dalam Pemerintahan Indonesia yang dipimpin oleh Menteri dan bertanggung jawab kepada Presiden yang membidangi urusan pembangunan desa dan kawasan perdesaan, pemberdayaan masyarakat desa, percepatan pembangunan daerah tertinggal, dan transmigrasi. Rencana Pembangunan Jangka Menengah Nasional (RPJMN) Kementerian Desa PDTT tahun 2015-2019 merupakan dokumen strategis rencana pembangunan yang harus dilakukan oleh pemerintah lima tahun kedepan. Dokumen RPJMN ini memuat sasaran pembangunan desa yang harus dicapai yaitu mengurangi jumlah Desa Tertinggal sampai 5.000 desa dan meningkatkan jumlah Desa Mandiri sedikitnya 2.000 desa di tahun 2019 . Dalam rangka pembangunan desa, Pemerintah dan Pemerintah Daerah wajib mengembangkan sistem informasi desa dan pembangunan kawasan perdesaan. Dalam mengembangkan sistem informasi desa dibutuhkan tersedianya data tentang desa. Kementerian Desa PDTT bekerjasama dengan Badan Perencanaan Pembangunan Nasional dan Badan Pusat Statistik mengeluarkan data Indeks Pembangunan Desa(IPD) tahun 2014 yang terdiri dari 74.093 desadan memiliki 42 indikator/*attribute dependent* tanpa label status desa. Data IPD 2014 merupakan cara pengukuran yang disusun berdasarkan tingkat perkembangan desa di Indonesia yang menjadikan desa sebagai unit analisis dengan mengacu pada Undang Undang Nomor 6 Tahun 2014 tentang desa, yang dimaksudkan

untuk memotret tingkat perkembangan desa di Indonesia dan dapat digunakan sebagai acuan untuk penyusunan perencanaan kebijakan dan pengawasan pembangunan desa [1].

Penentuan status desa ini adalah hal yang benar-benar penting terkait penentuan lokasi desa, bisa jadi program bantuan yang seharusnya untuk desa X karena kesalahan klasifikasi status desa yang menerima adalah desa Y. Untuk itu sebisa mungkin kesalahan dapat diminimalisir dengan mengoptimalkan hasil akurasi untuk model prediksi menggunakan algioritma *k*-NN. Dalam penelitian ini mengembangkan dari penelitian sebelumnya dan fokus pada optimasi algoritma *k*- NN untuk klasifikasi status desa.

Dalam penelitian ini, peneliti menggunakan algoritma clustering untuk menentukan label dan mengusulkan metode *boostrapping* yang digunakan untuk mengoptimalkan hasil dari klasifikasi *k*-NN dalam dataset IPD tahun 2014.

2. TINJAUN PUSTAKA

2.1. Penelitian Terkait

Hermawan Prasetyo dan Ayu Purwarianti dengan judul *Comparison of Distance Measures for Clustering Data with Mix Attribute Types*[5], membahas tentang Pengelompokan status Desa dengan data set PODES II dengan cara mencari Silhouette tertinggi didalam pengelompokan status Desa menggunakan perhitungan jarak *Euclidean,* dengan hasil penelitian menunjukkan bahwa untuk beberapa nomor cluster, algoritma k-prototipe melakukan yang terbaik hasil pengelompokan oleh menerapkan rasio jarak ketidak cocokan untuk kategoris atribut.

Ayu. Klasifikasi Status Desa di Kabupaten Banyuwangi Dengan Metode Naive Bayes[14], membahas Metode yang membutuhkan beberapa data untuk menghasilkan keakuratan dalam menentukan status desa, yaitu desa swakarya dan swasembada di Kabupaten Banyuwangi menggunakan metode Naive Bayes Clasifikasi, dengan hasil data yang sama pada setiap pengujian tidak menentukan bahwa tingkat akurasi yang dihasilkan juga akan sama. Pengujian yang dilakukan sebanyak empat kali dengan jumlah data training dan data testing yang sama, yaitu 54 untuk data training dan 18 untuk data testing membuktikan bahwa tingkat akurasi yang dihasilkan berbeda, yaitu tingkat akurasi tertinggi mencapai 94,4% dan terendah 77,7%. Rata–rata akurasi yang diperoleh dari pengujian adalah 84,4%.

Religia Pengelompokan status desa menggunakan Algoritma K-means[10], membahas tentang perhitungan jarak k-mean manakah yang paling efektif untuk mengelompokkan data Potensi Desa ke dalam 5 status Desa menggunakan metode peritungan jarak Manhattan, Euclidean dan Chebyshev pada algoritma k- mean, dengan hasil penelitian bahwa penentuan jarak yang paling optimum menggunakan Chebyshev dan yang paling efisien dalam waktu menggunakan Euclidean.

Tamrin Klasifikasi status perkembangan desa menggunakan algoritma k-nearest neighbor[13], membahas tentang pengklasifikasian status desa pada data potensi desa belum dapat di tentukan nilai akurasinya karena masih mengunakan penilaian pembagian sejumlah variabel menggunakan metode k-NN dan Decision Tree dengan hasil dari penelitian ini adalah bahwa akurasi dari pengklasifikasi *k*-NN sebesar 90.18 % lebih baik dibandingkan dengan akurasi *Decision Tree* sebesar 79.50 %.

2.2. Landasan Teori

2.2.1 *Data Mining*

Data mining merupakan salah satu bidang paling penting dalam penelitian yang bertujuan untuk memperoleh informasi dari data set. *Data mining* mulai ada sejak 1990-an sebagai cara yang efektif untuk mengambil pola dan informasi yang sebelumnya tidak diketahui dari suatu data set [5].

Clustering merupakan pekerjaan yang memisahkan data/vektor ke dalam sejumlah kelompok (*cluster*) menurut karakteristiknya masing-masing. Data-data yang mempunyai kemiripan karakteristik akan berkumpul dalam *cluster* yang sama, dan data-data dengan karakteristik berbeda akan terpisah dalam *cluster* yang berbeda. Tidak diperlukan label kelas untuk setiap data yang diproses dalam *clustering* karena nantinya label baru bisa diberikan ketika cluster sudah terbentuk. Karena tidak adanya target

label kelas untuk setiap data, maka *clustering* sering juga disebut juga pembelajaran tidak terbimbing (*unsupervised learning*) karena tidak ada label kelas yang digunakan dalam prosesnya, clustering sangat cocok untuk melakukan clustering data yang label kelasnya memang sulit didapatkan pada saat pembangkitan fitur. Pada clustering, segera setelah cluster terbentuk, maka label kelas untuk setiap data dapat diberikan dengan mengamati hasil cluster.

2.2.2 Algoritma *k*-Means

Algoritma k-Means merupakan algoritma pengelompokan iteratif yang melakukan partisi dataset ke dalam sejumlah k cluster yang sudah ditetapkan di awal. Algoritma k-Means sederhana untuk diimplementasikan dan dijalankan, relatif cepat, mudah beradaptasi, umum penggunaannya dalam praktek. Secara historis, k-Means menjadi salah satu algoritma yang paling penting dalam bidang data mining[7].

Algoritma k-Means berusaha untuk meminimalkan fungsi obyektif yaitu meminimalkan total jarak kuadrat (*squared distance*) seperti dinyatakan oleh persamaan berikut:

$$J = \sum_{i=1}^{N} \sum_{l=1}^{K} a_{ic}\, d(x_i, c_l)^2$$

Adapun tahapan dari algoritma k-Means adalah sebagai berikut:

a. Inisialisasi: tentukan nilai k sebagai jumlah cluster yang diinginkan dan metrik ketidak miripan (jarak) yang diinginkan. Jika perlu, tetapkan ambang batas perubahan fungsi obyektif dan ambang batas perubahan posisi centroid.
b. Pilih k data dari dataset x sebagai centroid.
c. Alokasikan semua data ke centroid terdekat dengan metrik jarak yang sudah ditetapkan (memperbarui cluster ID setiap data).

2.2.3 *k-Nearest Neighbor* (k-NN)

Algoritma k-Means merupakan algoritma pengelompokan iteratif yang melakukan partisi dataset ke dalam sejumlah k cluster yang sudah ditetapkan di awal. Algoritma k-Means sederhana untuk diimplementasikan dan dijalankan, relatif cepat, mudah beradaptasi, umum penggunaannya dalam praktek. Secara historis, k-Means menjadi salah satu algoritma yang paling penting dalam bidang data mining[7].

Untuk mengukur jarak dari atribut yang mempunyai nilai besar, seperti atribut pendapatan, maka dilakukan normalisasi. Normalisasi bisa dilakukan dengan *min-max normalization* atau *Z-score standardization* [18]. Jika data *training* terdiri dari atribut campuran antara numerik dan kategori, lebih baik gunakan *min-max normalization* [18]. Untuk menghitung kemiripan kasus, digunakan rumus:

$$Similarity\ (p,q) = \frac{sas \sum_{i=1}^{n} f\ (pi,qi\ ) X\ wi}{wi}$$

Keterangan :

P = Kasus baru
q = Kasus yang ada dalam penyimpanan
n = Jumlah atribut dalam tiap kasus
i = Atribut individu antara 1 sampai dengan n
f = Fungsi *similarity* atribut i antara kasus p dan kasus q
w = Bobot yang diberikan pada atribut ke-i

Adapun tahapan dari algoritma *k*-NN adalah sebagai berikut:

a. Menentukan jumlah k, yaitu jumlah observasi terdekat yang akan digunakan.

b. nghitung jarak antar observasi pada variabel yang bersesuaian berdasarkan formula 2.4.
c. Mencari k observasi terdekat berdasarkan jarak terkecil.

2.2.4 *Sample Bootstrapping*

Suatu metode untuk menderivasikan estimasi yang kuat dari error standar dan interval kepercayaan untuk mengestimasi proporsi, rerata,median, odds ratio, koefisien korelasi atau koefisien regresi. *Bootstrapping* juga dapat digunakan untuk mengembangkan uji hipotesis. *Bootstrapping* sangat berguna sebagai alternatif untuk estimasi parameter ketika peneliti merasa ragu dapat memenuhi asumsi pada data mereka. Misalnya kasus heteroskedastisitas muncul pada analisis regresi karena ukuran sampel yang kita miliki kecil. *Bootstrapping* juga berguna ketika inferensi parametrik tidak mungkin dilakukan atau memerlukan rumus yang sangat rumit untuk menghitung eror standar untuk median, kuartil, persentil dan lainnya.

Metode *bootstrap* bukan cara untuk mengurangi error, akan tetapi hanya mencoba untuk memperkirakan error sehingga didapatkan error standar (SE) pada dataset. Formula untuk memperkirakan error standar yang digunakan adalah:

$$SE = \frac{\sigma}{\sqrt{n}}$$

dimana σ = standar deviasi

n = banyaknya record

Semakin besar nilai n, maka semakin kecil nilai error yang didapatkan dan semakin kecil nilai standar deviasi maka semakin menurun nilai error standar.

2.2.5 *plit Validation*

Split Validation adalah sebuah tek yang digunakan untuk sebuah validasi yang membagi data menjadi dua bagian secara acak, sebagian data untuk data training dan sebagian data lainnya untuk data testing. Dengan menggunakan Split Validation maka akan dilakukan sebuah percobaan training berdasarkan split ratio yan lah ditentukan sebelumnya, untuk selanjutnya sisa dari split ratio data training akan dianggap bagai data testing. Data training adalah data yang akan dipakai dalam melakukan pembelajaran sedangkan data testing adalah data yang belum pernah dipakai sebagai pembelajaran dan akan berfungsi sebagai data pengujian keakurasian sebuah hasil pembelajaran [19].

2.2.6 *K-Fold Cross Validatio*

Cross Validation merupakan salah satu metode yang digunakan untuk memperoleh parameter terbaik menggunakan cara pengujian besarnya *error* pada data testing. *Cross Validation* membagi data secara acak kedalam *k* bagian dengan ukuran yang sama dan masing-masing bagian akan dilakukan proses klasifikasi. Secara umum pengujian nilai dilakukan sebanyak 10 kali untuk memperkirakan akurasi estimasi. Dalam penelitian yang digunakan berjumlah 10 atau *10-fold Cross Validation*. Penggunaan 10 *fold* ini dianjurkan karena merupakan jumlah *fold* terbaik untuk uji validitas. Tiap percobaan akan menggunakan 1 data *testing* dan *k*-1 bagian akan menjadi *data training*, kemudian *data testing* itu akan ditukar dengan atu buah *data training* sehingga untuk tiap percobaan akan didapatkan *data testing* yang berbedabeda. Misalnya ada 10 subset data maka akan menggunakan 9 subset untuk training dan 1 subset untuk testing dilakukan untuk semua kemungkinan [20].

2.2.7 *Confusion Matrix*

Confusion matrix adalah tabel *matrix* yang terdiri i dua kelas, yaitu kelas yang dianggap sebagai positif dan kelas yang dianggap sebagai negatif [22]. *Confusion matrix* berisi informasi aktual (*actual*) dan prediksi (*predicted*) pada sistem klasifikasi.

Tabel 1. Model *Confusion Matrix*

Classification		Predicted Class	
		Class = Yes	Class = No.
Observed Class	Class = Yes	A True Positif – tp	B False Negatif – tn
	Class = No	C False Positif – fp	D True Negatif – tn

Keterangan dari tabel 1:

True Positive (tp) = proporsi positif dalam data set yang diklasifikasikan positif
True Negative (tn) = proporsi negatif dalam data set yang diklasifikasikan negatif
False Positive (fp) = proporsi negatif dalam data set yang diklasifikasikan potitif
False Negative (fn) = proporsi negatif dalam data set yang diklasifikasikan negatif

2.3. Metode *Sample Boostreping* pada k-Nearest Neighbor untuk Klasifikasi Satus Desa

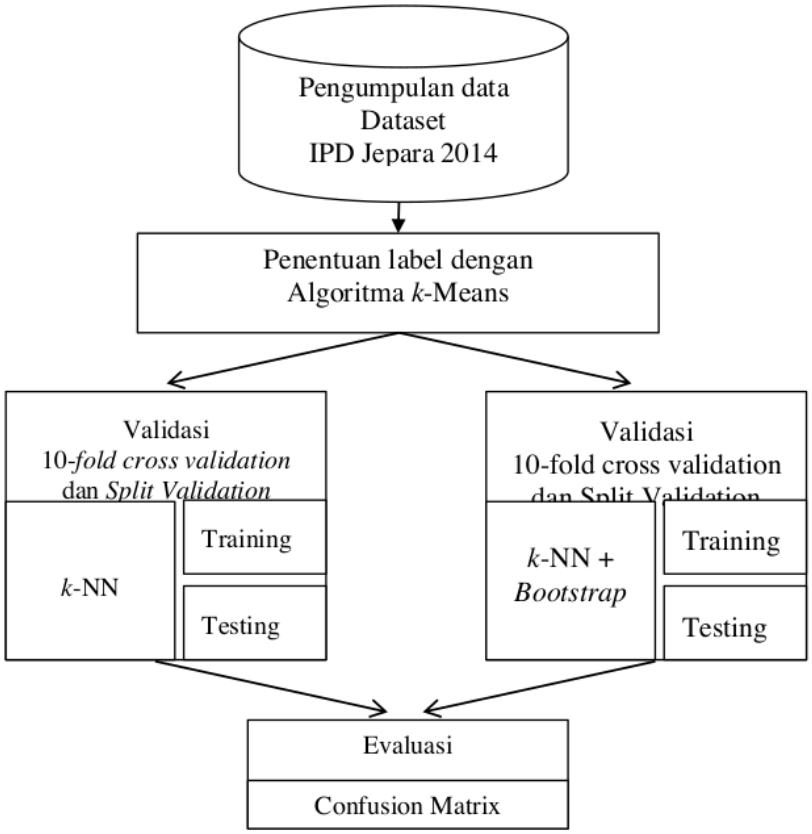
Berdasarkan penelitian yang dilakukan oleh Tamrin [13]. Membahas tentang pengelompokan status desa pada data IPD 2014, data yang ada terlebih dahulu dilakukan *prepocesing*, sehingga didapatkan dari data IPD yang akan diuji menurut *cluster*. Cluster atau kelompok yang dimaksud adalah penentuan kelompok dari tiap jenis status desa menggunakan *algorthm K-means* yang merupakan salah satu algoritma *clustering* dengan tujuan algoritma ini untuk membagi data menjadi beberapa kelompok yang menerima masukan data tanpa label atau kelas *(unsupervised)* yang dirubah menjadi cluster yang memiliki label *(supervised)*. Algoritma ini akan mengelompokkan data atau objek ke dalam *k* buah kelompok tersebut. Pada setiap *cluster* terdapat titik pusat (*centroid*) yang merepresentasikan *cluster* tersebut.

Setelah cluster terbentuk data diolah menggunakan algoritma K-NN yang merupakan salah satu algoritma supervised dari metode klasifikasi dimana hasil dari sampel uji yang baru diklasifikasikan berdasarkan mayoritas dari kategori pada *K-NN*. Tujuan dari algoritma ini adalah mengklasifikasi objek baru berdasakan atribut data training dan sampel data testing. Dimana ditemukan sejumlah *K* objek (*centroid*) yang paling dekat dengan titik testing. Klasifikasi menggunakan pengelompokan terbanyak di antara klasifikasi dari *K* objek. Algoritma K-NN memiliki bebrapa kekurangan dalam menentukan nilai parameter *K* dengan adanya kekurangan tersebut, peneliti menggunakan *distance matric* sebagai nilai prediksi dari sample data testing yang baru untuk mencari data yang terdekat dengan titik cluster dengan perhitungan jarak *Eucledian*, data diuji mengunkan confusion matrix dapat disimpulkan bahwa algoritma data mining *K-nearest neighbor* memiliki kinerja terbaik untuk klasifikasi 5 status desa, yaitu : ***Desa sangat tertinggal, Desa tertinggal, Desa berkembang, Desa maju dan Desa Mandiri*** dengan nilai accuracy yaitu 90,18% sedangkan pengujian dengan menggunakan Decision Tree didapatkan nilai accuracy 79,50%.

Untuk meningkatkan performa klasifikasi menggunakan algoritma *K-NN*, penulis mengusulkan metode *sample bootstrapping* yang dapat melakukan pengambilan sampel data asli dengan penggantian (*sampling with replacement*). Dalam *sampling* dengan penggantian, di setiap langkah semua sampel memiliki peluang yang sama untuk terpilih. Setelah sampel telah dipilih, sampel diseleksi dan dapat dipilih kembali pada langkah-langkah berikutnya. Dengan demikian sampel dengan penggantian dapat memiliki sampel yang sama dalam beberapa kali. Jumlah sampel dapat ditentukan secara *absolute* atau *relative* tergantung setting parameter sampel, tujuan dari penelitian ini yaitu menganalisis peningkatan akurasi pada algoritma *k*-NN pada saat diterapkan metode *sample bootstrapping* pada klasifikasi status desa maka dapat diambil kesimpulan bahwa penggunaan metode *Bootstrap* terbukti dapat meningkatkan *accuracy* dari pengklasifikasi sebesar 5.41 %, dari pengujian *k-NN* mendapatkan accuracy 91,89% dan mengunakan Bootstap dengan *k*-NN 97,30%.

3. METODE PENELITIAN

Metode dilaksanakan dengan tahapan sebagai berikut.



Gambar 1. Diagram Alir Metode

3.1. Pengumpulan Data

Penelitian ini menggunakan data sekunder yang diperoleh dari Badan Pusat Statistik berupa data Indeks Pembangunan Desa tahun 2014 (IPD 2014). Data IPD 2014 merupakan data yang disusun berdasarkan Undang-Undang Nomor 6 Tahun 2014 tentang desa, dengan menjadikan desa sebagai unit analisis. Data IPD 2014 terdiri dari 42 indikator dan memiliki 74.093 *instances*. Data ini dipilih karena belum banyak peneliti yang menggunakan untuk data penelitian dan daerah yang dipilih adalah daerah Jepara.

Tabel 2. Dataset yang Digunakan dalam Eksperimen

Dataset	Jumlah *Instances*	Jumlah Atribut	Jumlah Atribut Nominal	Jumlah Atribut Numerik
Indek Pembangunan Daerah (IPD) Jepara Tahun 2014	184	42	-	42

Dari 42 indikator yang ada, belum tersedia label dari tiap *instance*, sedangkan jumlah data pada dataset adalah 184 data. Agar data IPD 2014 memiliki label, data IPD 2014 akan di kelompokkan kedalam 5 *cluster* menggunakan algoritma *k-Means*.

3.2. Eksperimen

Tahapan eksperimen pada penelitian ini adalah sebagai berikut:

a. Menyiapkan dataset untuk ekperimen.
b. Melakukan proses clustering dengan menggunakan algoritma k-Means sehingga didapatkan 5 buah cluster (berdasarkan formula 2.2 dan 2.3).
c. Menentukan label dari hasil clustering yang telah didapatkan.
d. Melakukan training dan testing terhadap algoritma *k*-NN (berdasarkan formula 2.4).
e. Melakukan training dan testing terhadap algoritma *k*-NN yang di integrasikan dengan *sample bootstrapping* (berdasarkan formula 2.4 dan 2.7).
f. Mencatat hasil kinerja pengklasifikasi diantaranya *accuracy, precission* dan *recall* (berdasarkan formula 2.8, 2.9 dan 2.10)
g. Membandingkan hasil kinerja antara *k*-NN standar dengan *k*-NN + *Bootstrap*.
h. Menganalisa metode yang terbaik antara *k*-NN klasik dengan K-NN + *Bootstrap*.

3.3.1 Algoritma *k*-Means

Algoritma *k*-Means digunakan untuk melakukan pengelompokkan pada dataset IPD Jepara 2014 sehingga dapat ditentukan labelnya. Jumlah *k* yang digunakan adalah 5 sehingga terbentuk 5 buah cluster sebagai penentu label.

3.3.2 Sample Bootstrapping

Pada penelitian ini diterapkan metode *sample bootstrapping* yaitu pengambilan sampel data asli dengan penggantian (*sampling with replacement*). Dalam *sampling* dengan penggantian, di setiap langkah semua sampel memiliki peluang yang sama untuk terpilih. Setelah sampel telah dipilih, sampel diseleksi dan dapat dipilih kembali pada langkah-langkah berikutnya. Dengan demikian sampel dengan penggantian dapat memiliki sampel yang sama dalam beberapa kali. Jumlah sampel dapat ditentukan secara *absolute* atau *relative* tergantung setting parameter sampel. Dalam penelitian ini parameter sampel yang digunakan adalah *absolute* 100 dan *relative* dengan *ratio* 1.0.

Tabel 3. Rencana Eksperimen

Setting Parameter	***Parameter Sample***
Absolute	100
Relative	Ratio 1.0

3.3.3 Algoritma k-Nearest Neighbor (k-NN)

Algoritma *k*-NN dengan jumlah *k* yang digunakan menyatakan jumlah tetangga terdekat yang dilibatkan dalam penentuan prediksi kelas pada data uji. Dari *k* tetangga terdekat yang terpilih, kemudian dilakukan voting kelas dari *k* tetangga terdekat tersebut. Kelas dengan jumlah suara tetangga terbanyaklah yang kemudian diberikan sebagai label kelas hasil prediksi pada data uji tersebut. Nilai *k* yang digunakan dalam penelitian ini adalah 1, 3 dan 5.

3.3. Evaluasi

Evaluasi dilakukan dengan mengamati hasil dari klasifikasi menggunakan integrasi antara *Sample Bootstrapping* dengan Algoritma *k*-NN untuk klasifikasi status desa dengan menggunakan pengujian *split validation* dengan perbandingan 80:20 dan 10-*Fold cross validation*. Sedangkan pengukuran tingkat akurasi dilakukan dengan menganalisa hasil dari evaluasi model *confusion matrix* yang mana bertujuan untuk mengetahui seberapa meningkat akurasi dari algoritma tersebut.

4. HASIL DAN PEMBAHASAN

4.1. Hasil

Pada sub bab ini akan dicontohkan perhitungan *pre processing* untuk menentukan label dari setiap data yang terdapat pada dataset Potensi Desa tahun 201[illegible] Metode yang digunakan adalah metode *k*-means. Selanjutnya di proses dengan cara melakukan klasifikasi menggunakan *k-NN* dan dibandingkan menggunakan *k-NN* + *Boostraping*.

Setelah semua tahapan prediksi dari metode *k*-NN dan *k*-NN dengan *Bootstrap* telah dilakukan, kemudian dicatat hasil validasi menggunakan 10-*Fold Cross Validation* dan *Split Validation* serta pengukuran metode berdasarkan hasil *confusion matrix* meliputi *accuracy, precision* dan *recall*. Pada tabel 4 menunjukkan rekap hasil pengukuran terhadap dataset IPD Jepara 2014 sebagai berikut:

Tabel 4. Hasil Pengukuran Pengklasifikasi pada Dataset IPD Jepara

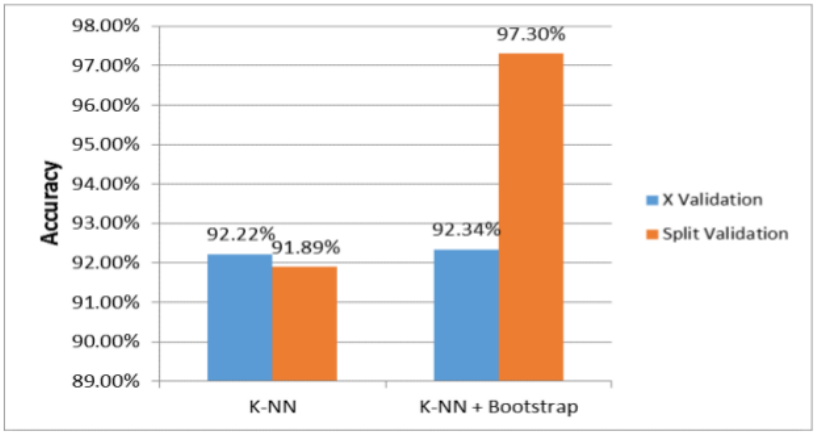
Metode		10-*Fold Cross Validation*			*Split Validation*		
		Accuracy	***Preccision***	***Recall***	***Accuracy***	***Preccision***	***Recall***
k-NN	*k* = 1	90.18 %	99.00 %	86.08 %	86.49 %	85.71 %	75.00 %
	k = 3	92.22 %	76.47 %	97.50 %	91.89 %	88.89 %	100 %
	k = 5	87.78 %	78.00 %	97.50 %	89.19 %	87.50 %	100 %
k-NN + *Bootstrap*	*k* = 1	95.64 %	91.43 %	91.43 %	89.19 %	100 %	90.00 %
	k = 3	92.34 %	77.14 %	87.10 %	97.30 %	90.00 %	100 %
	k = 5	92.37 %	82.86 %	82,86 %	97.30 %	90.00 %	95.45 %

4.2. Pembahasan

Pada sub bab ini akan dibahas hasil perbandingan kinerja metode yang digunakan dan hasil peningkatan akurasi pada metode *k-NN* dan *k-NN* + *boostrapping*.

4.2.1 Perbandingan Kinerja Metode

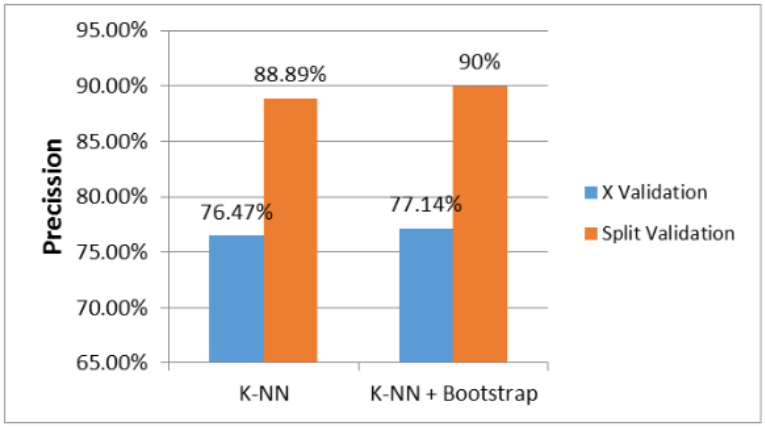
Pada bagian ini akan dibahas mengenai hasil pengukuran kinerja metode. Metode yang dibandingkan yaitu metode *k*-NN standar dengan *k*-NN berbasis *Bootstrap* dengan *k* =3. Gambar 2 menunjukkan perbandingan *Accuracy*, Gambar 3 menunjukkan perbandingan *Precission*,



Gambar 2. Diagram Perbandingan *Accuracy*

Gambar 2 menunjukkan diagram perbandingan *accuracy* pada masing-masing metode yang digunakan. Dapat diambil kesimpulan bahwa *accuracy* tertinggi didapatkan pada metode k-NN dengan Bootstrap yaitu sebesar 97.30%. Hal ini menunjukkan bahwa penerapan metode bootstrap mempengaruhi kinerja dari pengklasifikasi dan terjadi peningkatan accuracy yang cukup signifikan.

Gambar 3 menunjukkan diagram perbandingan precission pada masing-masing metode yang digunakan. Dapat diambil kesimpulan bahwa precission tertinggi didapatkan pada metode *k*-NN + *Bootstrap* sebesar 90%.



Gambar 3. Diagram Perbandingan *Precission*

4.2.2 Peningkatan Kinerja Metode

Tabel 5 menunjukkan hasil kinerja masing-masing pengklasifikasi terhadap dataset IPD Jepara 2014 yang mana dapat disimpulkan pada penggunaan *k*-NN dan *k*-NN + Bootstrap dengan metode *split validation* didapatkan hasil yang lebih baik sehingga diperoleh peningkatan hasil kinerja pengklasifikasi *k*-NN yang ditunjukkan pada Tabel 5 berikut:

Tabel 5. Peningkatan kinerja pengklasifikasi *K*-NN + *Bootstrap*

Metode	Peningkatan *Accuracy*
10-Fold Cross validation	0.12 % (92.34 % - 92.22 %)
Split Validation	5.41 % (97.30 % - 91.89 %)

Berdasarkan Tabel 5 dapat diambil kesimpulan bahwa penggunaan metode *k*-NN berbasis *Bootstrap* semakin dapat meningkatkan *accuracy* dari pengklasifikasi sebesar 5.41 %. Hal ini membuktikan bahwa penggunaan metode *Bootstrap* dapat meningkatkan kinerja dari pengklasifikasi *k*-NN.

5. PENUTUP

5.1. Kesimpulan

Berdasarkan masalah dan tujuan dari penelitian ini yaitu peningkatan akurasi pada algoritma *k*-NN pada saat diterapkan metode *sample bootstrapping* pada status desa maka dapat diambil

kesimpulan bahwa penggunaan metode *Bootstrap* terbukti dapat meningkatkan *accuracy* dari pengklasifikasi sebesar 5.41 %, dari pengujian *k-NN* mendapatkan accuracy 91,89% dan mengunakan Bootstap dengan *k*-NN 97,30%, sehingga dapat ditarik kesimpulahan untuk penentuan 5 status Desa, yang meliputi Desa Sangat Tertinggal, Desa Tertinggal, Desa Maju dan Desa Mandiri dapat menggunakan metode *k*-NN dengan Bootsrapping yang memberikan nilai *accuracy* yang tinggi.

5.2. Saran

Penelitian ini telah memberikan kontribusi dalam peningkatan kinerja pengklasifikasi klasik *k*-NN. Hasil dari penelitian ini menunjukkan bahwa metode *Bootstrap* dapat mempengaruhi dan meningkatkan kinerja dari pengklasifikasi *k*-NN. Disarankan untuk penelitian selanjutnya dapat menggunakan metode *Feature Selection* dan *Hybrid* K-Means dan *k*-NN untuk mendapatkan kinerja pengklasifikasi yang lebih baik lagi.

PERNYATAAN ORIGINALITAS

"Saya menyatakan dan bertanggung jawab dengan sebenarnya bahwa artikel ini adalah hasil karya saya sendiri kecuali cuplikan dan ringkasan yang masing-masing telah saya jelaskan sumbernya".
[Eko Siswanto – P31.2014.01593]

DAFTAR PUSTAKA

[1] Undang-Undang Nomor 6, "Undang-Undang Republik Indonesia Nomor 6 Tahun 2014 Tentang Desa," pp. 1–71, 2014.

[2] . Pendidikan, D. A. N. Kebudayaan, and R. Indonesia, "Salinan salinan," pp. 6–8, 2014.

[3] Lei Xu *et al.*, "Information Security in Big Data: Privacy and Data Mining," *IEEE Access*, vol. 2, 1149–1176, 2014.

[4] V. K. Deepa and J. R. R. Geetha, "Rapid development of applications in data mining," *2013 Int. Conf. Green High Perform. Comp*, pp. 1–4, 2013.

[5] H. Prasetyo and A. Pu arianti, "International Conference on Information Technology Systems and Innovation (ICITSI)," *Comparison of Distance Measures for Clustering Data with Mix Attribute pes for Indonesian Potential-based Regional Grouping*, 2014.

[6] I. H. Witten, E. Frank, and M. a. Hall, *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition*, vol. 54, no. 2. 2011.

[7] X. Wu *et al.*, *Top 10 algorithms in data mining*, vol. 14, no. 1. 2008.

[8] F. D. Anton, "Data *centroid k-mean* dan *gap indicator* untuk pengambilan keputusan prioritas pembangunan Desa, 2016.

[9] *Change the world with data . We ' ll show you how . .*

[10] Re i yoga, "Pengelompokan status desa menggunakan algoritma *k-means*,", 2016.

[11] S. A. Dudani, "The Distance-Weighted k-Nearest-Neighbor Rule," *IEEE Trans. Syst. Man Cybern.*, vol. SMC-6, no. 4, pp. 325–327, 1976.

[12] Y. C. Liaw, C. M. Wu, and M. L. Leou, "Fast k-nearest neighbors search using modified principal axis search tree," *Digit. Signal Process. A Rev. J.*, vol. 20, no. 5, pp. 1494–1501, 2010.

[13] Teguh Tamrin, "Klasifikasi Status Perkembangan Desa Menggunakan Algoritma K-Nearest eighbor," 2016.

[14] V. Krishnaiah, G. Narsimha, and N. S. Chandra, "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques," *Int. J. Comput. Sci. Inf. Technol.*, vol. 4, no. 1, pp. 39–45, 2013.

[15] M. V. D, "Data Mining Classification Techniques Applied to Analyze the Impact of Ambient Conditions on Aero Engine Performance - A Case Study Using Xlminer."

[16] T. Pang-Ning, M. Steinbach, and V. Kumar, "Introduction to data mining," *Libr. Congr.*, p. 796, 2006.

[17] M. (Morga. K. (Publishers. . Han, J & Kamber, "Data Mining Concept and Techniques," no. 0, 2012.

[18] D. T. Larose, *Data Ming Methods and Models*. Canada: John Wiley & Sons, Inc., Hoboken, New Jersey, 2005.

[19] and M. A. H. I. H. Witten, E. Frank, *Data Mining Practical Machine Learning Tools and hnique*. Burlington: Morgan Kaufmann Publishe, 2011.

[20] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," *Int. Jt. Conf. Artif. Intell.*, vol. 14, no. 12, pp. 1137–1143, 1995.

[21] Y. Zhang and S. Wang, "Detection of Alzheimer's disease by displacement field and machine rning.," *PeerJ*, vol. 3, p. e1251, 2015.

[22] C. Vercellis, *Business Intelligence: Data Mining and Optimization for Decision Making*, Edition Fi. Politecnico di Milano, Italy, 2009.

Nas_Jurnal #05 METODE SAMPLE BOOSTRAPING

ORIGINALITY REPORT

23%	16%	11%	19%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Universitas Sam Ratulangi Student Paper	2%
2	es.scribd.com Internet Source	1%
3	Submitted to Universitas Kristen Satya Wacana Student Paper	1%
4	digilib.uin-suka.ac.id Internet Source	1%
5	Submitted to Universiti Putra Malaysia Student Paper	1%
6	Herlambang Brawijaya, Samudi Samudi, Slamet Widodo. "Komparasi Algoritma K-Nearest Neighbor dan Naiive Bayes Pada Pengobatan Penyakit Kutil Menggunakan Cryotheraphy", JUITA : Jurnal Informatika, 2019 Publication	1%
7	Submitted to Universitas Islam Indonesia Student Paper	1%

Ji Zhang, Raid Luaibi Lafta, Xiaohui Tao, Yan Li,

8	Fulong Chen, Yonglong Luo, Xiaodong Zhu. "Coupling a Fast Fourier Transformation With a Machine Learning Ensemble Model to Support Recommendations for Heart Disease Patients in a Telehealth Environment", IEEE Access, 2017 Publication	1%
9	gaib.itb.ac.id Internet Source	1%
10	media.neliti.com Internet Source	1%
11	Submitted to Sriwijaya University Student Paper	1%
12	Submitted to Syiah Kuala University Student Paper	1%
13	elib.uni-stuttgart.de Internet Source	1%
14	www.selasar.com Internet Source	1%
15	tutcris.tut.fi Internet Source	1%
16	ieeexplore.ieee.org Internet Source	<1%
17	Submitted to Universitas Airlangga Student Paper	<1%

18	Submitted to iGroup Student Paper	<1%
19	www.dicsr-qnt.com Internet Source	<1%
20	Bing-Fei Wu, Shih-Jhe Yao, Li-Wei Hou, Po-Ju Chang, Wan Ju Tseng, Ching-Wei Huang, Yung-Shin Chen, Po-Yu Yang. "Intelligent shopping assistant system", 2016 International Automatic Control Conference (CACS), 2016 Publication	<1%
21	core.ac.uk Internet Source	<1%
22	Submitted to Forum Komunikasi Perpustakaan Perguruan Tinggi Kristen Indonesia (FKPPTKI) Student Paper	<1%
23	desacidewa.blogspot.com Internet Source	<1%
24	link.springer.com Internet Source	<1%
25	150.214.191.180 Internet Source	<1%
26	Submitted to Universitas Muria Kudus Student Paper	<1%
27	ejournal-umht.org Internet Source	<1%

28	Submitted to Universitas Brawijaya Student Paper	<1%
29	journal.budiluhur.ac.id Internet Source	<1%
30	zombiedoc.com Internet Source	<1%
31	www.coursehero.com Internet Source	<1%
32	Submitted to UIN Sultan Syarif Kasim Riau Student Paper	<1%
33	Submitted to Universitas Negeri Surabaya The State University of Surabaya Student Paper	<1%
34	publish.kne-publishing.com Internet Source	<1%
35	"Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020)", Springer Science and Business Media LLC, 2020 Publication	<1%
36	repository.bsi.ac.id Internet Source	<1%
37	Chih-Chiang Wei. "Comparing lazy and eager learning models for water level forecasting in	<1%

river-reservoir basins of inundation regions", Environmental Modelling & Software, 2015
Publication

38	eprints.ums.ac.id Internet Source	<1%
39	Submitted to Universiti Teknologi Malaysia Student Paper	<1%
40	eprints.binus.ac.id Internet Source	<1%
41	pt.scribd.com Internet Source	<1%
42	catatanyusufjabung.blogspot.com Internet Source	<1%
43	Submitted to Universitas Sebelas Maret Student Paper	<1%
44	ejournal.unsrat.ac.id Internet Source	<1%
45	ejournal.uika-bogor.ac.id Internet Source	<1%
46	Submitted to Sultan Agung Islamic University Student Paper	<1%
47	id.123dok.com Internet Source	<1%

www.scribd.com

48	Internet Source	<1%
49	repository.its.ac.id Internet Source	<1%
50	Rangga Sanjaya, Fitriyani Fitriyani. "Prediksi Bedah Toraks Menggunakan Seleksi Fitur Forward Selection dan K-Nearest Neighbor", Jurnal Edukasi dan Penelitian Informatika (JEPIN), 2019 Publication	<1%
51	Submitted to Cedar Valley College Student Paper	<1%
52	Submitted to Universitas Diponegoro Student Paper	<1%
53	Mustika Mentari, Yuita Arum Sari, Ratih Kartika Dewi. "Deteksi Kanker Kulit Melanoma dengan Linear Discriminant Analysis-Fuzzy k-Nearest Neigbhour Lp-Norm", Register: Jurnal Ilmiah Teknologi Sistem Informasi, 2016 Publication	<1%
54	Ali Alamsyah Kusumadinata. "PEMANFAATAN MEDIA INFORMASI DALAM PROGRAM RUMAH TIDAK LAYAK HUNI (RTLH)", QARDHUL HASAN: MEDIA PENGABDIAN KEPADA MASYARAKAT, 2019 Publication	<1%

55	"Information and Communication Technology for Intelligent Systems", Springer Science and Business Media LLC, 2019 Publication	<1%
56	Submitted to Universitas Jenderal Soedirman Student Paper	<1%

Exclude quotes	Off	Exclude matches	< 5 words
Exclude bibliography	On		

Nas_Jurnal #05 METODE SAMPLE BOOSTRAPING

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11