Jurnal Teknologi Informatika dan Terapan Vol. 05, No 02, Juli – Desember 2018 ISSN: 2354-838X
DOI: https://doi.org/10.25047/jtit.v5i2.84

PERBANDINGAN METODE K-NN DAN BAYES PADA MISSING IMPUTATION

Taufiq Rizaldi[#1], Fendik Eko Purnomo[*2], Aji Seto arifianto[#3]

[#] Jurusan Teknologi Informasi, Politeknik Negeri Jember*
Jalan Mastrip POBOX 164

[1]taufiq_r@polije.ac.id

[2]fendik_eko@polije.ac.id

[3]ajiseto@gmail.com

Abstract

The problem of data loss in a dataset is experienced in surveys for data collection which are usually caused by no response from units or items during the survey data collection process. The loss of a data can significantly influence the results of a study. The inaccuracy in choosing a solution to overcome these problems can result in a less than optimal outcome that tends to be biased. Some methods that are widely used to solve these problems are using the K Nearest Neighbor (K-NN) and Naïve Bayes methods, the main purpose of this study is to compare the performance of the two methods. From the results of the K-NN, the results were better, where the Mean Square Error (MSE) is bigger than 1 and MAPE around 10-16%, while Naïve Bayes got MSE values bigger than 1 and MAPE around 26%.

Keywords— *Missing Imputation, Naïve Bayes, Nearest Neighbor.*

PENDAHULUAN

Kasus hilangnya nilai pada dataset atau ketiadaan nilai pada data untuk atribut tertentu sangat banyak terjadi dan sering disebut dengan *missing imputation*. Salah satu penyebab terjadinya hal tersebut adalah tidak adanya respon dari unit atau item pada sebuah survey pengumpulan data yang mengakibatkan kesimpulan atau hasil penelitian dari survey tersebut menjadi kurang baik. Membuang data yang hilang pada survey dan melakukan tahap selanjutnya pada pemrosesan data sebagai solusi permasalahan tersebut akan mengakibatkan data yang kurang valid dan cenderung bias.

Salah satu solusi yang dapat dilakukan pada kasus diatas adalah dengan menggunakan metode yang tepat untuk melengkapi data berdasarkan nilai lain pada kelompok data tersebut. Beberapa metode pernah digunakan dalam penelitian untuk melengkapi data yang hilang. Salah satu metode yang banyak digunakan untuk mengisi nilai dari data yang hilang adalah metode K Nearest Neighbor (K-NN)[1][2][3].

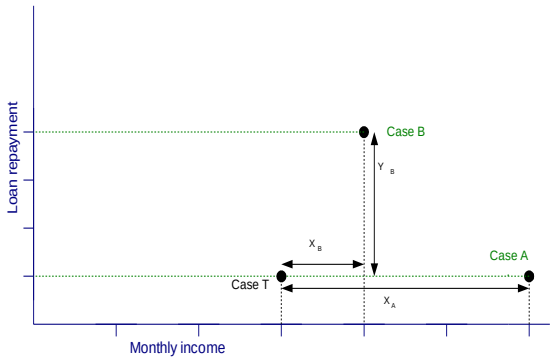
Penerapan metode K-NN menggunakan data bunga iris menunjukan tingkat kesalahan saat proses imputasi berada pada kisaran 4,02% hingga 5,48% [1]. Sedangkan penerapan imputasi missing dat amenggunakan data produktivitas tanaman pangan di provinsi lampung menunjukan tingkat kesalahan saat proses imputasi berada pada kisaran 4,73% hingga 13,57% [2].

Metode lain yang dapat digunakan untuk melakukan imputasi pada missing data adalah Naïve Bayes. Penggunaan Naïve Bayes untuk melakukan imputasi diterapkan pada imputasi untuk melengkapi data persebaran Demam Berdarah Dengue di kota Jember [4]. Pada penelitian ini dilakukan percobaan untuk mendapatkan perbandingan performa antara metode K-NN dan metode Naïve Bayes untuk melakukan imputasi atau pengisian pada sebuah data yang hilang menggunakan data untuk memprediksi persebaran demam berdarah.

TINJAUAN PUSTAKA

K Nearest Neighbor (K-NN)

Metode K Nearest Neighbor (K NN) adalah salah satu metode yang menerapkan algoritma supervised learning yang memiliki tujuan untuk menghubungkan pola dari data yang sudah ada untuk menemukan pola baru sehingga nantinya dapat diteukan polanya. K-NN menghitung jarak antara data baru atau data testing dengan data lama atau data training sehingga didapatkan nilai similaritnya[5] seperti yang digambarkan pada gambar 1.



Gambar 1.Pengelompokan pada *Nearest Neighbor (NN).*

Alur dari metode K-NN adalah sebagai berikut :

Memilih jumlah kasus optimal dari K-NN

Hitung jarak / nilai kesamaan (similarity) antara gen target dan gen lainnya tanpa nilai yang hilang. Ada berbagai ukuran untuk penentuan similarityantar gen seperti koefisien korelasi Pearson, Manhattan Distance, Euclidean Distance dan variance minimization[11]. Pada umumnya jarak antara gen target dengan gen lainnya dihitung denganmenggunakanEuclidean Distance, dimana jika terdapat Xi = {xi1, xi2, ..., xin} dan Xj = {xj1, xj2, ..., xjn}, jarak antara Xi dan Xj dihitung dengan menggunakan rumus

$$dist(x_i, x_j) = \sqrt{\sum_{l=1}^{n} (x_{il} - x_{jl})^2} \qquad (1)$$

Dimana n adalah jumlah fitur dalam sebuah kasus.

Urutkan hasil perhitungan gen target terhadap gen lainya dari kecil ke besar. Maka didapatkanlah kelompok yang mendekati inputan.

Naïve Bayes

Naïve Bayes adalah sebuah model yang memiliki konsep bahwa sebuah fitur pada sebuah data berkaitan tidak dengan fitur lain dalam data yang sama[4]. Teorema bayes yang digunakan utuk melakukan prediksi menggunakan persamaan seperti berikut ini :

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{q} P(X_i|Y)}{P(X)} \qquad (2)$$

Dimana :

1) P(H|E) adalah Probolitas akhir bersyarat (conditional probability) suatu hipotesis H terjadi jika (evidence) E terjadi.

P(E|H) adalah Probabilitas sebuah kejadian E terjadi tanpa memandang evidence apapun.

P(E) adalah Probabilitas awal (priori) kejadian E terjadi tanpa memandang hipotesis/evidence lain.

Ada beberapa hal yang penting dari aturan Bayes tersebut, yaitu:

1) sebuah probabilitas awal/prior H atau P(H) adalah probabilitas dari suatu hipotesis sebelum diamati.

2) sebuah probabilitas akhir H atau P(H|E) adalah probabilitas dari suatu hipotesis setelah diamati.

METODE PENELITIAN

Pada bagian ini akan dibahas tentang metodologi yang digunakan pada kegiatan penelitian.

Alur Metode *K-NN*

Alur dari metode K Nearest Neigbor (K-NN) untuk mendapatkan nilai imputasi pada data yang hilang adalah sebagai berikut :

1) Menentukan parameter K yang akan digunakan. Parameter K adalah jumlah observasi terdekat atau tetangga terdekat. Pada penelitian ini jumlah K yang akan digunakan adalah K dengan nilai 1(NN), 3, 5, 7, 9.

Melakukan perhitungan jarak antara kelompok data yang mengandung missing data dengan kelompok data yang lengkap menggunakan persamaan (1).

Hasil perhitungan jarak diurutkan berdasarkan dengan nilai jarak terkecil. Kemudian dipilih berdarakan nilai K yang telah ditentukan.

DOI: https://doi.org/10.25047/jtit.v5i2.84

Melakukan proses imputasi dengan menghitung nilai weight mean estimation (WME) pada K observasi terdekat yang tidak mengandung nilai missing data dengan persamaan :

$$x_j = \frac{\sum_{k=1}^{K} w_k v_k}{\sum_{k=1}^{K} w_k} \quad (3)$$

dimana adalah x_j estimasi rata-rata berbobot, adalah v_k nilai pada data lengkap pada variabel yang mengandung missing data berdasarkan observasi dari k, K adalah jumlah observasi terdekat yang digunakan, k adalah observasi dari K, w_k adalah bobot observasi tetangga terdekat ke-K dengan Persamaan :

$$w_k = \frac{1}{d\,(x_{ak} x_{bk})^2} \quad (4)$$

dimana d adalah jarak observasi K.

Naïve Bayes

Alur dari metode Naïve Bayes untuk mendapatkan nilai imputasi pada data yang hilang adalah sebagai berikut :

Mencari nilai mean menggunakan persamaan :

$$\text{Mean } (\mu) = \frac{\sum_{i=1}^{n} x_i}{n} \quad (5)$$

Mencari nilai variansi didapat dari hasil nilai varian yang sudah diproses dan hasil dijadikan bilangan bulat menggunakan persamaan variansi :

$$\sigma^2 = \frac{1}{n-1} \sum_{i=n}^{n} (x_i - \mu)^2 \quad (6)$$

Dari nilai tiap variabel diproses kedalam rumus Naïve Bayes, kecuali data kosong dengan menggunakan persamaan Naïve Bayes seperti berikut :

$$\varphi_{\mu,\sigma}{}^{(X)} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \quad (7)$$

μ dan σ dapat diestimasi dari data, untuk setiap kelas. Gunakan $\varphi_{\mu,\sigma}{}^{(X)}$ untuk P (X|C).

Data

Data yang digunakan sebagai uji coba pada penelitian ini adalah data angka bebas jentik yang digunakan untuk memprediksi penyebaran penyakit pada kabupaten jember yang telah lengkap, kemudian salah satu nilai pada data tersebut dihilangkan yang nantinya akan dibandikan hasilnya dengan nilai aslinya. Contoh data lengkap dan data yang telah dihilangkan seperti pada tabel 1.

TABEL IV
DATA

Kode	Data	
	Lengkap	Tidak Lengkap
B1	100	100
B2	100	100
B3	100	100
B4	100	0
B5	95	95
B6	100	100
B7	100	100
B8	100	100
B9	100	100
B10	100	100
B11	100	100
B12	100	100
B13	100	100
B14	100	100
B15	100	100
B16	100	100
B17	100	100
B18	100	100
B19	100	100
B20	100	100
B21	100	100
B22	100	100
B23	100	100
B24	100	100
B25	100	100
B26	100	100
B27	100	100
B28	98	98

HASIL DAN PEMBAHASAN

Pada bagian ini akan dibahas tentang hasil yang didapat dari kegiatan penelitian. Langkah awal yang dilakukan untuk mendapatkan imputasi dengan *K-NN* adalah memisahkan antara data lengkap dengan data yang tidak lengkap, kemudian dilakukan perhitungan untuk mendapatkan nilai similarity yang terdekat dengan menggunakan persamaan (1) seperti contoh yang ditampilkan pada tabel 2.

TABEL VI
HASIL PERHITUNGAN *ECLUDIAN DISTANCE*

No Urut	Kode	Euclidean
1	B10	9.486832981
2	B9	11.44552314
3	B29	13.78404875
4	B11	18.41195264
5	B15	19.67231557
6	B5	19.74841766
7	B1	27.31300057
8	B24	30.09983389
9	B17	34.94281042
10	B2	38.02630668
11	B7	39.57271787
12	B13	40.32369031
13	B22	42.21374184
14	B19	49.82971001
15	B16	63.51377803
16	B12	66.49060084
17	B25	72.01388755
18	B3	73.01369735
19	B21	89.11228871
20	B6	91.20307012
21	B8	95.13674369
22	B14	105.4798559
23	B20	105.579354
24	B23	121.3383699
25	B27	217.1888579
26	B28	219.1141255
27	B26	409.0452298

Dari hasil perhitungan diatas dilakukan pemrosesan dengan nilai K yang sudah ditentukan, kemudian dihitung nilai *weight mean estimation* (WME) menggunakan persamaan (3) dan (4) dengan hasil seperti yang ditunjukan pada tabel 3.

TABEL VIII
HASIL IMPUTASI

K	**WME**
NN	100
k-NN 3	73.97090926
k-NN 5	88.21870703
k-NN 7	89.21498831
k-NN 9	97.50108726

Untuk proses imputasi dengan naïve bayes dimulai dengan mencari nilai mean tanpa mengikut sertakan baris data yang hilang menggunakan persamaan (5) sehingga didapatkan nilai mean :

$$\text{Mean } (\mu) = \frac{2693}{27} = 99.74074074$$

Kemudian dicari nilai variasi dengan menggunakan persamaan (6) sehingga didapatkan hasil :

$$\sigma^2 = \frac{27.18519}{702} = 0.03875$$

Dari nilai variasi yang sudah didapat kemudian dilakukan perhitungan dengan persamaan naïve bayes(2) untuk mendapatkan nilai imputasi.

$$\varphi_{\mu,\sigma}{}^{(X)} = 0.465239$$

Untuk mengevaluasi maka dilakukan perhitungan nilai rata – rata MSE (Mean Square Error) dan MAPE (Mean Absolute Percentage Error) dengan hasil seperti pada tabel 4.

TABEL VIIX
RATA – RATA NILAI MSE DAN MAPE

	MSE	MAPE
NN	0.0745	10.7040816 %
NN 3	0.2587	16.9346939 %
NN 5	0.1151	10.7275863 %
NN 7	0.1471	12.3040816 %
NN 9	0.1057	13.1265306 %
Bayes	2.0876	26.3040816 %

DOI: https://doi.org/10.25047/jtit.v5i2.84

KESIMPULAN

Performa dari metode k-Nearest Neighbor (k-NN) untuk proses imputasi pada data yang hilang di sebuah kelompok data mempunyai performa yang cukup memuasakan dibandingkan dengan Naïve Bayes, berdasarkan perhitungan MSE (Mean Square Error) dimana semakin kecil nilai MSE semakin baik, rata-rata nilai MSE untuk K-NN bernilai dibawah 1 yaitu untuk NN = 0.0745, k-3 = 0.2587, k-5 = 0.1151, k-7 = 0.1471, dan k-9 = 0.1057, Sedangkan untuk Naïve Bayes mendapatkan rata-rata nilai MSE sebesar 2.0876.

Sedangkan untuk MAPE dimana jika nilai MAPE < 10% berarti sangat baik dan jika diantara 10% - 20% berarti baik, mendapatkan hasil untuk K-NN adalah NN = 5.7040816 %, k-3 = 11.9346939 % , k-5 = 5.7275863 %, k-7 = 7.3040816 %, dan k-9 = 8.1265306 %, untuk Naïve Bayes nilai MAPE sebesar 26.3040816 %.

UCAPAN TERIMA KASIH

Ucapan terima kasih Penulis berikan pada Kementerian Riset, Teknologi, dan Pendidikan Tinggi, P3M Politeknik Negeri Jember, dan Jurusan Teknologi Informasi Politeknik Negeri Jember atas dukungannya pada kegiatan penelitian ini.

DAFTAR PUSTAKA

Susanti, Shantika Martha, Evy Sulistianingsih. 2018. K Nearest Neighbor Dalam Imputasi Missing Data. Buletin Ilmiah Math. Stat. dan Terapannya (Bimaster) Volume 07, No. 1 (2018), hal 9 -14.

Siregar. S.Y, Toni Toharudin, Bertho Tantular. 2014. Performa Metode K Nearest Neighbor Imputation (KNNI) Untuk Menangani Multivariate Missing Data. Prosiding Seminar Nasional Statistika Departmen Statistika FMIPA Unpad Vol 4, No 1 (2014).

Irawan. N.D., Wijono, Onny Setyawati. 2017. Perbaikan Missing Value Menggunakan Pendekatan Korelasi Pada Metode K-Nearest Neighbor. Jurnal Infotel Vol.9 No.3 Agustus 2017.

Arifianto. A.S., Didit Rahmat Hartadi, Anik Nur Novitasari E. S. 2016. Prediksi Missing Imputation Untuk Data Penyebaran Demam Berdarah Degue Menggunakan Naïve Bayes. Jurnal Teknologi Informasi dan Terapan, Vol. 03, No. 01, Juli-Desember 2016.

Rizaldi. T, M.A. Muslim, E. Yudaningtyas. 2014. Knowledge Management System untuk Diagnosis Infeksi Nosokomial. Jurnal EECCIS Vol.8 (2), 105-110. Universitas Brawijaya:Malang.