PENERAPAN DATA MINING CLASSIFICATION UNTUK DATA BLOGGER MENGGUNAKAN KNN DAN EVALUASI MODEL DENGAN HOLD OUT ESTIMATION

Hanan Nadia - 1810511098
Irza Ramira Putra - 18105111100
Deo Haganta Depari - 1810511104
Nadhifa Zhafira - 1810511111
Quina Alifa – 1810511115

Abstrak

Situs *blogger* merupakan salah satu media penyebaran informasi dan data yang masih banyak diakses di internet. *Blogger* adalah layanan penerbitan blog yang menerima blog multi-pengguna dengan entri bertanda waktu. Saat ini, jumlah pengguna situs *blogger* lebih menurun dibandingkan saat-saat trendnya, tetapi masih berjumlah jutaan. Oleh karena itu, penelitian ini dilakukan untuk mengklasifikasikan data pengguna situs blogger untuk mengetahui apakah pengguna tersebut termasuk *Blogger Professional* atau *Blogger Musiman*. Penelitian ini menggunakan dataset dari penelitian sebelumnya sebagai referensi. Metode yang digunakan untuk pengklasifikasian data adalah metode *KNN* (*K-Nearest Neighbor*) dan untuk evaluasi model menggunakan metode *Hold Out Estimation*. Data yang digunakan berupa dataset *blogger* yang didapatkan dari penelitian sebelumnya yang diubah dari tipe *excel* menjadi tipe *csv* agar dapat diolah. Dari hasil penelitian pertama dengan menggunakan *weights distance* dan *test size* bernilai 0.1(Hanya menggunakan 10 persen data dari *dataset* sebagai data testing) mendapatkan nilai akurasi sebesar 100%. Sedangkan penelitian kedua dengan *weights distance* dan *test size* bernilai 0.2(Hanya menggunakan 20 persen data dari *dataset* sebagai data testing) mendapatkan nilai akurasi sebesar 90 %.

I. Pendahuluan

1.1 Latar Belakang

Seiring dengan perkembangan teknologi dalam bidang informasi, informasi dan data yang terdapat dalam internet juga semakin banyak. Salah satu media penyebaran informasi ini adalah situs *Blogger*. *Blogger* adalah layanan penerbitan blog yang menerima blog multi-pengguna dengan entri bertanda waktu. Jumlah pengguna situs *blogger* walaupun saat ini menurun dibandingkan saat-saat trendnya, tetap memiliki jumlah pengguna yang melebihi jutaan. Sampai saat ini pun berdasarkan *hostingtribunal.com* tercatat bahwa terdapat lebih dari 500 juta blog dari total 1.7 miliar *website* di dunia dengan lebih dari 2 juta penulis blog yang rutin mengunggah postingan. Oleh karena itu, pengklasifikasian pengguna *blogger* sebagai Blogger professional maupun musiman dapat mempermudah pembaca blog.

Klasifikasi merupakan penempatan objek-objek ke salah satu dari beberapa kategori yang telah ditetapkan sebelumnya. KNN (K-Nearest Neighbor) merupakan salah satu metode yang digunakan untuk mengklasifikasi. KNN melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang berjarak paling dekat dengan objek tersebut.

Berdasarkan kasus Blogger di atas, maka dilakukannya klasifikasi dengan menggunakan algoritma KNN (K-Nearest Neighbor) dan melakukan evaluasi model dengan menggunakan metode Hold Out Estimation. Penelitian referensi yang digunakan adalah penelitian yang dilakukan oleh Recha Abriana Anggraini, Galih Widagdo, Arief Setya Budi, dan M. Qomaruddin yang berjudul Penerapan Data Mining Classification untuk Data Blogger Menggunakan Metode Naïve Bayes (Recha Abriana Anggraini, Galih Widagdo, Arief Setya Budi, M. Qomaruddin, 2019)

1.2 Penelitian Terdahulu

Penelitian terdahulu menjadi salah satu acuan bagi penulis dan sebagai landasan teori untuk melakukan penelitian yang baru. Penulis menggunakan dataset yang sama yaitu dataset *blogger* yang akan diklasifikasikan menjadi 2 kelompok yaitu *Blogger Professional* (BP) dan *Blogger Musiman* (BM). Berikut merupakan penelitian terdahulu berupa jurnal yang berkaitan dengan topik penulis.

Tabel 1. Penelitian Terdahulu

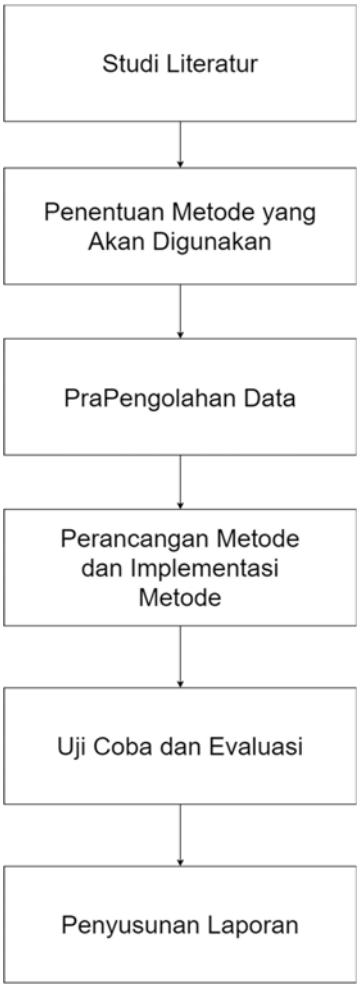
Nama Peneliti	Judul Penelitian	Hasil Penelitian
Recha Abriana, Galih Widagdo, Arief Setya, & M. Qomaruddin, 2019.	Penerapan Data Mining Classification untuk Data Blogger Menggunakan Metode Naïve Bayes	Pengolahan data untuk mengklasifikasikan dataset blogger menggunakan algoritma naïve bayes memiliki hasil akurasi yang baik dan dapat dijadikan sebagai acuan bagi orang-orang yang ingin mengetahui klasifikasi blogger musiman atau blogger professional.
Perbedaan : Pada penelitian sebelumnya Recha, Galih, Arief, dan M. Qomaruddin menggunakan algoritma naïve bayes dan evaluasi model dengan performance vector lalu class precision dan class recall, sedangkan penulis menggunakan algoritma KNN dan evaluasi model dengan *Hold Out Estimation*.		

1.3 Usulan Metode

Dalam penelitian ini, penulis menggunakan algoritma KNN (*K-Nearest Neighbor*) untuk mengklasifikasikan dataset *blogger* menjadi dua kelompok. Kemudian untuk evaluasi model, penulis menggunakan metode *Hold Out Estimation*.

II. Metodologi Penelitian

Secara umum, penelitian ini dilakukan dalam beberapa tahapan yang diawali dari studi literatur, penentuan metode yang akan digunakan, pra pengolahan data, perancangan metode dan implementasi metode, uji coba dan evaluasi dan penyusunan laporan.



Gambar 1. Rancangan Penelitian

2.1 Studi Literatur

Mempelajari Jurnal Penelitian terdahulu dan Jurnal Penelitian yang menghasilkan *dataset* yang digunakan.

2.2 Penentuan Metode yang Akan Digunakan

Dari informasi yang telah didapatkan dari studi literatur, penulis akan menentukan metode apa yang akan digunakan selain dari metode yang telah digunakan pada jurnal pada penelitian terdahulu, metode yang digunakan adalah algoritma KNN (*K-Nearest Neighbor*) untuk mengklasifikasikan dataset *blogger* menjadi dua kelompok. Kemudian untuk evaluasi model, penulis menggunakan metode *Hold Out Estimation*.

2.3 Pra Pengolahan Data

Jumlah responden dalam penelitian ini sebesar 100 responden, dengan dataset blogger ini, peneliti akan mengklasifikasikan jenis blogger kedalam 2 kelompok yaitu *Blogger Professional(BP)* dan *Blogger Musiman(BM)*.

Berikut merupakan data training dari dataset blogger.

Tabel 2. DATASET BLOGGER

Degree	Caprice	Topics	LMT	LPSS	PB
high	left	impression	yes	yes	yes
high	left	political	yes	yes	yes
medium	middle	tourism	yes	yes	yes
. . .	. . .	. . .	. . .	. . .	. . .
medium	right	news	yes	yes	no
medium	left	impression	yes	yes	yes

Tabel 2 merupakan dataset blogger yang dipakai dalam penelitian ini. Data tersebut berjumlah 100 data yang direpresentasikan dalam bentuk tabel.

Dataset yang digunakan pada penelitian terdahulu, yang juga terdapat pada situs *"UCI Machine Learning Repository"* merupakan *spreadsheet* bertipe *excel,* maka dari itu *dataset* perlu diubah dulu menjadi tipe *csv*, sehingga bisa digunakan pada proses Implementasi Metode.

2.4 Perancangan Metode dan Implementasi Metode

Berdasarkan metode yang telah ditentukan, maka *library* yang digunakan adalah:

- *Pandas*
- *Numpy*
- *Preprocessing* dari *sklearn*
- *Train_test_split* dari *sklearn.model_selection*
- *StratifiedKFold* dari *sklearn.model_selection*
- *Itertools*
- *sys*

Langkah implementasi metodenya adalah sebagai berikut:

- Membaca *dataset csv* yang telah diolah sebelumnya sesuai dengan kolom yang tepat, lalu menyimpannya ke *variabel array*
- Konversi kategori dari tipe *string* menjadi *tipe numeric* untuk setiap kolom, dengan menggunakan *preprocessing label encoder*
- Pisahkan *dataset array* menjadi dua kategori, satu *array* fitur dan satu lagi *array* untuk kelas

- Pisahkan data fitur dan data kelas menjadi 2 kategori, satu data untuk *training* dan satu untuk *testing,* menggunakan *train test split* dengan parameter data dan ukuran data *testing*
- Menyimpan dan mencetak hasil prediksi dari data *testing*, lalu mencetak hasil kelas yang sebenarnya
- Membuat perhitungan besarnya nilai prediksi yang salah (error) dan mencetak nilai nya
- Membuat perhitungan nilai prediksi yang akurat dengan cara mengurangi 100% dengan nilai prediksi yang salah dan mencetak nilai nya
- Menerapkan algoritma evaluasi Hold Out Estimation pada model data yang digunakan dan mencetak hasil akurasi*, sensitivity dan specificity* dari model

2.5 Uji Coba dan Evaluasi

Pengujian dilakukan untuk mendapatkan pengaturan *n neighbors, weights(uniform/distance)*, dan *test size* yang terbaik berdasarkan nilai error prediksi yang paling kecil dan akurasi, nilai sensitivity dan specificity yang paling besar.

Untuk memudahkan dalam mendapatkan pengaturan yang terbaik, penulis menggunakan perulangan dan menyimpan setiap *output* ke dalam *file notepad*, dari evaluasi *output* tersebut didapatkan 2 pengaturan terbaik, keduanya memiliki *n neighbors* bernilai 7 dan menggunakan *weights distance,* pengaturan pertama menggunakan *test size* bernilai 0.1, sedangkan pengaturan kedua menggunakan *test size* bernilai 0.2.

2.6 Penyusunan laporan

Penyusunan laporan dilakukan mulai dari awal hingga akhir penelitian.

III. Hasil dan Pembahasan

Dengan menggunakan 2 pengaturan terbaik yang didapat pada proses Uji Coba dan Evaluasi, didapatkan hasil klasifikasi seperti berikut :

Hasil Klasifikasi dengan pengaturan pertama dengan *n neighbors* bernilai 7, menggunakan *weights distance* dan *test size* bernilai 0.1(Hanya menggunakan 10 persen data dari *dataset* sebagai data testing)

Tabel 3. Tabel Hasil Klasifikasi Pengaturan Pertama

		Data									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
HASIL	**Klasifikasi**	1	1	1	1	1	0	1	0	1	1
	Sebenarnya	1	1	1	1	1	0	1	0	1	1

Hasil nilai evaluasi klasifikasi:

- Error Prediction = 0.00 %
- Accuracy = 100.00 %

Hasil nilai evaluasi model:

- Accuracy = 1.0
- Sensitivity = 1.0
- Specificity = 1.0

Hasil Klasifikasi dengan pengaturan kedua dengan *n neighbors* bernilai 7, menggunakan *weights distance* dan *test size* bernilai 0.2(Hanya menggunakan 20 persen data dari *dataset* sebagai data testing)

Tabel 4. Tabel Hasil Klasifikasi Pengaturan Kedua

		Data																			
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
HASIL	**KLASIFIKASI**	1	1	1	1	1	0	1	0	1	1	1	1	1	0	0	1	1	1	1	0
	SEBENARNYA	1	1	1	1	1	0	1	0	1	1	1	1	1	0	1	1	0	1	1	0

Hasil nilai evaluasi klasifikasi:

- Error Prediction = 10.00 %
- Accuracy = 90.00 %

Hasil nilai evaluasi model:

- Accuracy = 0.9
- Sensitivity = 0.9333333333333333
- Specificity = 0.8

IV. Kesimpulan dan Saran

Berdasarkan hasil dari kedua penelitian yang dilakukan, dapat dikatakan bahwa pengklasifikasian data menggunakan metode KNN memiliki hasil yang baik. Hasil klasifikasi dari penelitian pertama dengan pengaturan *n neighbors* bernilai 7, menggunakan *weights distance* dan *test size* bernilai 0.1 (Hanya menggunakan 10 persen data dari *dataset* sebagai data testing) mendapatkan nilai prediksi yang salah (Error) sebesar 0% dan nilai Accuracy sebesar 100%. Sedangkan untuk penelitian kedua dengan pengaturan *n neighbors* bernilai 7, menggunakan *weights distance* dan *test size* bernilai 0.2 (Hanya menggunakan 20 persen data dari *dataset* sebagai data testing) mendapatkan nilai prediksi yang salah (Error) sebesar 10% dan nilai Accuracy sebesar 90%.

Selain pengklasifikasian data, pada penelitian ini juga dilakukan evaluasi model menggunakan metode *Hold Out Estimation* dan juga memiliki hasil yang baik. Hasil dengan pengaturan pertama mendapatkan nilai Accuracy, Sensitivity dan Specificity sebesar 1.0. Sedangkan untuk hasil dengan pengaturan kedua mendapatkan nilai Accuracy sebesar 0.9, Sensitivity sebesar 0.933 dan nilai Specificity sebesar 0.8.

Pengolahan *dataset blogger* untuk mengklasifikasikan data *blogger profesional dan blogger* musiman bisa dicoba dengan metode lainnya, yang mungkin memiliki nilai akurasi yang tinggi dengan data testing yang lebih tinggi

V. Daftar Pustaka

Anggraini, Recha Abriana, Galih Widagdo, Arief Setya Budi, dan M. Qomaruddin. 2019. Penerapan Data Mining Classification untuk Data Blogger Menggunakan Metode Naïve Bayes. Jurnal Sistem dan Teknologi Informasi, 7(1), 47-51.

Archive.ics.uci.edu. 2012. BLOGGER Data Set (Data XLSX). Diunduh pada 1 Juni 2019 dari https://archive.ics.uci.edu/ml/datasets/BLOGGER

Gharehchopogh, Farhad Soleimanian, dan Seyyed Reza Khaze. 2012. Data Mining Application for Cyber Space Users Tendency in Blog Writing: A Case Study. International Journal of Computer Applications, 47(18), 40-46.