Perbandingan Tingkat Akurasi Support Vector Machine dengan Naive Bayes pada Studi Kasus Okupansi Lahan Berdasarkan Kondisi Cuaca

Comparison of Accuracy Level of Support Vector Machine with Naive Bayes on Land Occupancy Case Study Based on Weather Conditions

Yosua Marchel [1], Jondri Nasri [2]

[1,2]Prodi S1 Teknik Informatika, Fakultas Teknik, Universitas Telkom
[1]marchelyozz@student.telkomuniversity.ac.id, [2]jondri@telkomuniversity.ac.id

Abstrak

Kondisi sebuah ruangan pada suatu lahan dapat mempengaruhi tingkat huni ruangan tersebut. Kondisi ini dipengaruhi oleh beberapa faktor, antara lain temperatur, cahaya matahari, serta kelembapannya. Kita tidak akan mengerti dengan pasti faktor apa yang paling mempengaruhi kondisi sebuah lahan. Salah satu contoh metode yang sering digunakan untuk klasifikasi data yaitu Support Vector Machine (SVM) dan Naive Bayes. Dengan memproses data melalui atribut yang ada, akan dihasilkan keputusan di kelas mana data tersebut berada. Dengan begitu kita dapat menghitung tingkat keakurasian kedua metode tersebut terhadap suatu studi kasus. Melalui penelitian yang telah dilakukan, menghasilkan akurasi dengan menggunakan SVM lebih besar dibanding dengan Naive Bayes berdasarkan data okupansi lahan yang dipengaruhi oleh cuaca.

Kata kunci: Support Vector Machine, Naive Bayes, Okupansi

Abstract

The condition of a room on a land can affect the occupancy rate of the room. This condition is influenced by several factors, including temperature, sunlight, and humidity. We will not understand exactly what factors most affect the condition of a land. One example of a method often used for data classification is Support Vector Machine (SVM) and Naive Bayes. By processing the data through an existing attribute, a decision will be made in the class where the data is located. That way we can calculate the level of accuracy of both methods against a case study. Through research that has been done, it produces accuracy by using a larger SVM compared to Naive Bayes based on weather-affected land occupancy data.

Keyword: Support Vector Machine, Naive Bayes, Occupancy

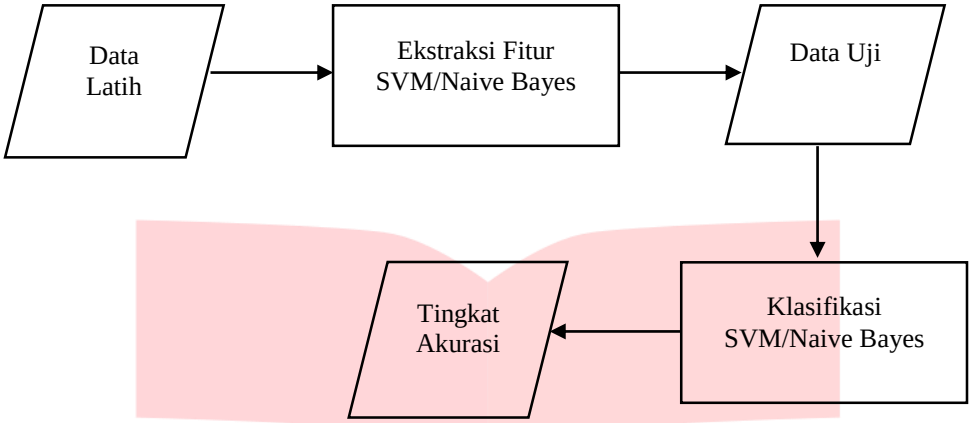
1. Pendahuluan

Kondisi sebuah lahan bangunan atau ruangan untuk dapat dihuni dipengaruhi oleh beberapa faktor-faktor alam yang secara natural mempengaruhi kondisi lahan bangunan atau suatu ruangan. Dengan mempelajari kumpulan data dari faktor-faktor tersebut, kita dapat mengklasifikasikan data-data tersebut untuk mengambil keputusan. Hasil pengklasifikasian data tersebut menghasilkan pengkelompokan data kedalam sebuah kelas. Dari studi kasus tersebut, kita akan mencoba mengklasifikasikan data-data tersebut dengan menggunakan SVM dan Naive Bayes. Bagaimanakah performansi SVM dan Naive Bayes dalam mengklasifikasikan data okupansi lahan tersebut. Metode manakah yang memiliki tingkat akurasi yang lebih baik. Banyak penelitian yang sebelumnya telah menerapkan metode Naive Bayes dan SVM dalam pengklasifikasian data [1] [2]. SVM adalah sebuah metode yang menggunakan persamaan *hyperplane* dalam memisahkan dua kelas data dengan menentukan margin terbesar antara data-data terluar [3]. Naive Bayes merupakan metode klasifikasi dengan menghitung probabilitas kemunculan data dalam suatu kelas [4].

2. Perancangan

2.1 Blok Proses

Alur sistem yang digunakan dalam penelitin ini adalah:



Gambar 2.1 Diagram Blok Proses

2.2 Data Set

Data set yang digunakan berupa enam atribut dan dua kelas (satu atribut sebagai kelas). Jumlah data latih yang digunakan sebanyak 8143 baris data dan data uji sebanyak 2665 baris data. Data set yang digunakan dapat diunduh pada [5]. Contoh data yang digunakan:

	A	B	C	D	E	F
1	Temperature	Humidity	Light	CO2	HumidityRatio	Occupancy
2	21.76	31.13333333	437.3333333	1029.666667	0.005021011	1
3	21.79	31	437.3333333	1000	0.005008581	1
4	21.7675	31.1225	434	1003.75	0.005021569	1
5	21.7675	31.1225	439	1009.5	0.005021569	1
6	21.79	31.13333333	437.3333333	1005.666667	0.005030298	1
7	21.76	31.26	437.3333333	1014.333333	0.005041605	1
8	21.79	31.1975	434	1018.5	0.005040749	1
9	21.79	31.39333333	437.3333333	1018.666667	0.005072649	1
10	21.79	31.3175	434	1022	0.005060296	1
11	21.79	31.46333333	437.3333333	1027.333333	0.005084053	1
12	21.79	31.525	437.75	1047.75	0.005094099	1
13	21.79	31.575	441.75	1049	0.005102244	1
14	21.79	31.395	442	1061.5	0.005072921	1
15	21.79	31.3925	441.75	1049	0.005072514	1
16	21.79	31.5	441.5	1048	0.005090026	1
17	21.815	31.5	438	1049.25	0.005097869	1
18	21.815	31.4725	449.5	1051.25	0.005093382	1
19	21.89	31.6	449.5	1060.5	0.005137857	1
20	21.79	31.55	449.5	1059.5	0.005098172	1
21	21.82333333	31.73	447.6666667	1072	0.005138035	1
22	21.865	31.7225	449.5	1072.333333	0.005150007	1
23	21.89	31.745	449.5	1075.333333	0.005161628	1

Sheet1

Gambar 2.2 Data set

2.3 Confusion Matriks

Confusion matriks akan digunakan dalam mengkelompokan data klasifikasi kedalam empat bagian, dan akan digunakan untuk menghitung besar akurasi pengujian.

Tabel 2.1 Confusion Matriks

	0	1
0	TN	FP
1	FN	TP

Keterangan:
TN = True Negative
FP = False Positive
FN= False Negative
TP = True Positive

Tingkat akurasi pengujian akan dihitung dengan rumus:

$$Akurasi = \frac{TN + TP}{P + N} = \frac{TN + TP}{TP + FP + TN + FN} \quad (2.1)$$

3. Pembahasan

3.1 Hasil klasifikasi Naive Bayes

Dari hasil pengujian dengan menggunakan metode Naive Bayes, didapatkan hasil:

Tabel 3.1 Hasil klasifikasi Naive Bayes

	0	1
0	1638	55
1	6	966

Dari hasil tersebut dihasilkan tingkat akurasi Naive Bayes sebesar **0.9771**

3.2 Hasil klasifikasi SVM

Dari hasil pengujian dengan menggunakan metode SVM, didapatkan hasil:

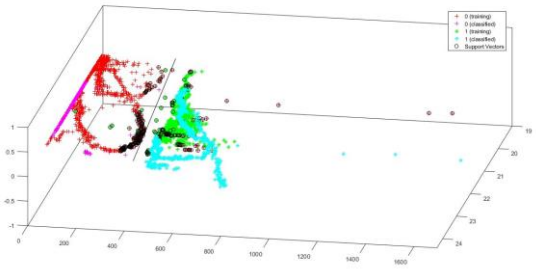
Tabel 3.2 Hasil klasifikasi SVM

	0	1
0	1638	55
1	2	970

Dari hasil tersebut dihasilkan tingkat akurasi Naive Bayes sebesar **0.9786**

3.3 Visualisasi Data

Dari hasil pengujian diatas dapat terlihat bahwa perbedaan kedua metode tersebut terletak pada hasil FN dan TP. Metode *Naïve Bayes* gagal mengklasifikasikan empat data dari data uji. Untuk dapat mengambil kesimpulan, penulis menggunakan visualisasi dua dimensi dengan menggunakan program matlab. Alasan mengapa digunakan visualisasi dua dimensi agar data yang ditampilkan dapat terlihat jelas.



Gambar 3.1 Visualisasi dengan matlab

Dari gambar diatas, diperlihatkan posisi data ketika dilakukan pengujian dengan menggunakan metode SVM. Data berwarna biru yang berada jauh dengan kumpulan data lain adalah data yang gagal diklasifikasikan dengan metode *Naïve Bayes*. Data tersebut yang seharusnya berada pada kelas 1, namun dihasilkan sebagai kelas 0 oleh *Naïve Bayes*. Hal ini dikarenakan posisi data yang sangat jauh dengan posisi kelas 1 dari data training yang menghasilkan nilai rata-rata serta standar deviasi untuk kelas 1. Serta adanya beberapa data latih outlier yang mempengaruhi hasil pengklasifikasian, sehingga data tersebut diklasifikasikan sebagai kelas 0 oleh *Naïve Bayes.*

4. Kesimpulan

Dari hasil pengujian yang dilakukan dengan menggunakan dua metode yaitu SVM dan Naive Bayes, terlihat bahwa metode SVM memiliki tingkat performansi yang lebih besar daripada Naive Bayes dengan perbedaan tingkat akurasi sebesar 0.0015. Hal ini dikarenakan ketidakmampuan Naive Bayes dalam mengatasi masalah data outlier. Data outlier sangat berpengaruh dalam perhitungan parameter yang akhirnya menyebabkan turunnya performansi Naive Bayes.

5. References

[1] M. K. S. Varma, D. N. K. K. Rao, K. K. Raju and D. G. P. S. Varma, "Pixel-based Classification Using Support Vector Machine Classifier," *Advanced Computing (IACC), 2016 IEEE 6th International Conference on*, 2016.

[2] J. D. S. Mora, D. A. Hurtado and O. L. R. Sandoval, "Prediction of Endocrine System Affectation in Fisher 344 Rats by Food Intake Exposed with Malathion, Applying Naïve Bayes Classifer and Genetic Algorithms," *International Journal of Preventive Medicine*, vol. 7, p. 111, 2016.

[3] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Leaming*, vol. 20, pp. 273-297, 1995.

[4] K. P. Murphy, Naive Bayes classifiers, University of British Columbia, 2006.

[5] "UC Irvine Machine Learning Repository," [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+.