Laporan Analisis Sistem Klasifikasi menggunakan Metode Bagging berbasis Naïve Bayes.

Fauzan Firdaus – 1301164317 – IF-40-04

A. Abstrak

Diketahui sebuah dataset yang terdiri menjadi 2 bagian yaitu data train dan data test. Pada data train, terdapat 3 kolom dan 298 baris data. 2 kolom pertama berisi atribut, sedangkan kolom terakhir menunjukkan klasifikasi / label dari data tersebut. Pada data test sendiri terdapat 75 baris dan 2 kolom saja yang menunjukkan atribut dari data tersebut. Maka dari itu, dengan metode bagging akan dicari label/kelas untuk setiap data test. Bahasa yang digunakan untuk program ini adalah bahasa **Matlab**, dengan versi tools adalah **R2018A**.

B. Metode Bagging

Sedikit bahasan mengenai metode bagging, bagging adalah salah satu metode / algoritma dalam ensemble learning yang merupakan gabungan algoritma pembelajaran mesin (machine learning) yang dirancang untuk meningkatkan stabilitas dan akurasi dari algoritma machine learning yang digunakan dalam klasifikasi statistik dan regresi. Bagging juga mengurangi varians dan membantu untuk menghindari terjadinya overfitting.

Dalam program ini, output utama dari bagging adalah menentukan label/kelas pada data test. Selain itu, program ini juga dapat menghasilkan akurasi yang didapatkan dari proses training berdasarkan setiap bag/bootstrap terhadap data train.

C. Strategi Penyelesaian Masalah

Dalam penyelesaian permasalahan kali ini, bagging digunakan untuk mendapatkan sebuah kualitas klasifikasi label/klas yang akurat. Dengan metode bagging tersebut, muncul hipotesa awal bahwa penggabungan model klasifikasi akan mempertinggi akurasi yang menunjukkan jumlah persentase ketepatan label yang sedang diklasifikasi.

Pertama-tama, hal yang harus ditentukan adalah menentukan kebutuhan yang akan digunakan pada program. Kebutuhan tersebut adalah menentukan jumlah bootstrap/bag yang akan menjadi model klasifikasi, dan menentukan jenis klasifikasi yang akan digunakan pada metode bagging. Karena pada kasus ini diharuskan untuk menggunakan metode naïve bayes, maka metode naïve bayes akan digunakan sebagai basis klasifikasi untuk metode bagging tersebut.

Setelah menetapkan 2 kebutuhan tersebut, hal yang harus dilakukan selanjutnya adalah membuat bootsrap. Bootstrap adalah sejumlah *n* data yang diambil secara random dari dataset, dengan pengambilan data yang boleh berulang. Jumlah bootstrap yang akan dibuat diusahakan berjumlah ganjil. Hal ini dikarenakan akan memudahkan pada proses *majority vote.* Proses tersebut akan melakukan voting pada label yang telah didapatkan. Jika jumlah bootstrapnya genap, dikhawatirkan terdapat kasus dimana 2 label memiliki jumlah vote yang sama.

Setelah membuat bootstrap dengan sejumlah data *n* yang baru, hal yang harus dilakukan selanjutnya adalah proses training setiap data dengan label yang ada pada bootstrap tersebut. Keluaran dari proses ini adalah label baru yang akan diisi pada sebuah model di langkah selanjutnya.

Kemudian, jika label baru sudah didapatkan berdasarkan proses training, selanjutnya perlu membuat model. Model adalah data yang jumlah dan isi atributnya sama dengan data train, namun sebagian labelnya akan diganti dengan label baru pada proses training bootstrap sesuai dengan indeks data yang sama. Misalnya, bootstrap A berisi indeks 1,2,3,3,3, dan 6. Maka, label yang baru dengan indeks ke 1,2,3, dan 6 akan menggantikan label yang lama, dan disimpan pada sebuah model.

Setelah ke-5 model didapatkan, terdapat proses opsional yaitu menghitung akurasi dari setiap model terhadap data asli pada data train.

Lalu, proses selanjutnya adalah melakukan *majority vote*. Proses ini bertujuan untuk melakukan voting pada setiap label dalam ke 5 model untuk menghasilkan label yang baru. Label baru tersebut akan digunakan untuk proses terakhir yaitu memprediksi label/kelas pada data testing.

D. Eksekusi dan Output Program

Sesuai mekanisme metode bagging pada poin sebelumnya, penulis mengimplementasikan metode bagging berbasis naïve bayes untuk menentukan label/kelas pada data testing.

Ketentuan programnya (asumsi/batasan masalah) adalah sebagai berikut:

- Jumlah / ukuran bootstrap / bag adalah 100 data yang merupakan 30% kurang dari jumlah data train yaitu 298 data.
- Karena terdapat beberapa data bootstrap yang tidak terambil pada data train, maka label pada data yang tidak terambil tersebut dikembalikan pada label/data train yang asli.

Setelah diimplementasikan, output program adalah menghasilkan akurasi ke 5 model yang telah dibuat, dan akurasi total setelah 5 model tersebut dilakukan proses voting. Outputnya adalah sebagai berikut.

```
Akurasi Bag/Bootstrap 1 : 95.6376 %
Akurasi Bag/Bootstrap 2 : 94.6309 %
Akurasi Bag/Bootstrap 3 : 94.2953 %
Akurasi Bag/Bootstrap 4 : 95.302 %
Akurasi Bag/Bootstrap 5 : 95.6376 %
Akurasi Total Setelah Majority Vote : 100 %
File 'TebakanTugas4ML.csv' telah berhasil dibuat
```

Berikut adalah file tebakan / hasil prediksi label/kelas pada data test.

1	2
2	2
3	2
4	2
5	2
6	2
7	2
8	2
9	2
10	2
11	2
12	1
13	2
14	1
15	2
16	1
17	1
18	1
19	1
20	1
21	2
22	1
23	2
24	1
25	1
26	1
27	1
28	2
29	1
30	2
31	2
32	1
33	2
34	1
35	1
36	1
37	1
38	1
39	1
40	1

41	1
42	2
43	1
44	1
45	1
46	1
47	1
48	1
49	1
50	1
51	1
52	1
53	1
54	1
55	1
56	1
57	1
58	1
59	1
60	1
61	1
62	1
63	1
64	1
65	1
66	1
67	1
68	1
69	1
70	1
71	1
72	1
73	1
74	1
75	1