**Laporan NLP Klasifikasi Teks**

**Nama : Fikri Rozan Imadudin/1301150768,Riko Bintang Purnomo Putra/1301154714**

**Kelas : IF-40-10**

**Analisis dan strategi penyelesaian masalah.**

**Tujuan dan Masalah** :

1. Diberikan data komentar kaskus gender.csv dan komentar kaskus gender test.csv yang memiliki label yaitu Komentar dan Gender. komentar kaskus gender akan digunakan sebagai data training dan komentar kaskus gender test akan digunakan sebagai data testing.
2. Pada tugas klasifikasi teks kali ini akan dibuat sebuah sistem klasifikasi menggunakan metode metode *Bagging*, salah satu teknik *Ensemble Learning*, berbasis *Multinomial Naïve Bayes* untuk menentukan kelas/label data uji dalam kaskus gender test.

**Teori dan Kajian :**

Gender Klasifikasi

Gender klasifikasi adalah secara otomatis dengan metode *machine learning* membedakan suatu gender **Pria** atau **Wanita** berdasarkan isi dari sebuah data corpus yang telah dilabelkan.

Preprocessing

MNNB dan *Bagging*MNNB adalah salah satu metode dari *Naïve Bayes* dengan model multinomial.

Bagging merupakan salah satu dari *Ensemble Learning* yang digunakan untuk mengurangi variansi, meningkatkan akurasi algoritma klasifikasi dan dapat menghindari *overfitting*. Idenya yaitu membagi data kedalam beberapa *subspace* lalu mengkombinasikan data secara acak dengan ukuran yang sama dengan data asli dalam masing-masing subspace sehingga didapatkan dari masing-masing *subspace* rata-rata untuk prediksi atau klasifikasi yang akan digunakan nanti.

**Metode Penelitian :**

Metodologi yang digunakan terdiri dari beberapa tahap yaitu memasukan dataset, *Preprocessing*, memisahkan dataset menjadi dua bagian yaitu (*x\_train, y\_train, x\_test, y\_test*), perancangan sistem, pengujian model dan analisa hasil. Akan digunakan sebuah library sklearn BaggingClassifier yang merupakan metode bagging dan MultinomialNB sebagai naive bayes.

**Anlisis dan Hasil :**

Experiment dilakukan terhadap *Naïve Bayes* tanpa menggunakan *Bagging* dan *Naïve Bayes* dengan *Bagging.* Pada tabel berikut berupa hasil yang telah diterapkan dengan random\_state = 7 dan n\_estimator pada bagging = 100.

|  |  |  |
| --- | --- | --- |
| Algoritma | Akurasi | KFold |
| *Naïve Bayes* tanpa menggunakan *Bagging* | 93.9597 % | 87.66 % |
| *Naïve Bayes* dengan *Bagging* | 94.6308 % | 86.99 % |

Hasil experiment yang ada pada tabel berikut menyatakan bahwa penerapan teknik bagging pada *Naïve Bayes* meningkatkan nilai akurasi sebesar 0.6711 % dan hasil KFoldnya *Naïve Bayes* lebih unggul. Pada percobaan ke dua kali ini random\_state = 10 dan n\_estimator pada bagging = 100.

|  |  |  |
| --- | --- | --- |
| Algoritma | Akurasi | KFold |
| *Naïve Bayes* tanpa menggunakan *Bagging* | 93.9597 % | 87.6 % |
| *Naïve Bayes* dengan *Bagging* | 93.9597 % | 86.9 % |

Hasil experiment yang ada pada tabel berikut menyatakan bahwa penerapan teknik *Naïve Bayes* tanpa menggunakan *Bagging* hasil Kfoldnya lebih tinggi dan akurasinya sama. Pada percobaan ke tiga kali ini random\_state = 7 dan n\_estimator pada bagging = 9.

|  |  |  |
| --- | --- | --- |
| Algoritma | Akurasi | KFold |
| *Naïve Bayes* tanpa menggunakan *Bagging* | 93.9597 % | 87.6 % |
| *Naïve Bayes* dengan *Bagging* | 93.9597 % | 88 % |

Hasil experiment yang ada pada tabel berikut menyatakan bahwa penerapan teknik *Naïve Bayes* menggunakan *Bagging* hasil Kfoldnya lebih tinggi dan akurasinya sama. Pada percobaan ke empat kali ini random\_state = 10 dan n\_estimator pada bagging = 9.

|  |  |  |
| --- | --- | --- |
| Algoritma | Akurasi | KFold |
| *Naïve Bayes* tanpa menggunakan *Bagging* | 93.9597 % | 87.66 % |
| *Naïve Bayes* dengan *Bagging* | 94.6308 % | 87.66 % |

Hasil experiment yang ada pada tabel berikut menyatakan bahwa penerapan teknik *Naïve Bayes* menggunakan *Bagging* hasil lebih tinggi akurasinya dan KFoldnya sama.

**Kesimpulan :**

Pada percobaan satu dengan teknik bagging Akurasinya unggul tetapi KFold lebih rendah, lalu pada percobaan ke empat akurasi bagging lebih unggul dan KFoldnya sama. Teknik Bagging terbukti meningkatkan akurasi walaupun tidak besar karena data yang digunakan jumlahnya sedikit dan dari experiment tersebut bahwa subspace/n\_estimatornya dan random\_state mempengaruhi nilai dari akurasi dan KFold.