Implementasi Naive Bayes Pada Studi QSAR untuk Identifikasi Inhibitor CDK2 sebagai Anti-Kanker

Rizki Achmad Riyanto, Hendro Pratama Saragih

Abstract—Pada Studi QSAR (Quantitative structure–activity relationship) dengan menggunakan metode Naïve Bayes dilakukan untuk melakukan identifikasi terhadap inhibitor dari CDK2 (Cyclin Dependent Kinase 2). Terdapat 2 bentuk representative dari metode naïve bayes yang digunakan yaitu Gaussian Naïve Bayes yang memperoleh akurasi sebesar 94,69% dan Bernoulli Naïve Bayes sebesar 93,11%

1 Introduction

Karena semakin berkembangnya dunia Kimia-informatika banyak peneliti melakukan penelitian terhadap bidang tersebut. Oleh karena itu penelitian ini melakukan sebuah penerapan atau implementasi sebuah metode untuk melakukan identifikasi inhibitor CDK2 sebagai anti-kanker . CDK2 (Cyclin Dependent Kinase 2) adalah gen pengkodean Protein. Penyakit yang terkait dengan CDK2 termasuk Kanker Payudara dan Glioblastoma Multiforme. Di antara jalur yang terkait adalah meiosis Oosit dan Signaling ERK. Anotasi Gen Ontologi (GO) yang terkait dengan gen ini mencakup aktivitas transferase, mentransfer gugus yang mengandung fosfor, dan aktivitas protein tirosin kinase. Paralog penting dari gen ini adalah CDK3. Pada penelitian terfokus untuk melakukan identifikasi terhadap zat yang dapat menghambat CDK2 (Cyclin Dependent Kinase 2) yang sebagai anti kanker

2 Metode

2.1 Seleksi Fitur

Pada penelitian ini kami menggunakan metode *Random Forest* untuk melakukan seleksi fitur berdasarkan dataset yang tersedia. Random forest merupakan pengembangan dari Decision Tree dengan menggunakan beberapa Decision Tree, dimana setiap DecisionTree telah dilakukan training menggunakan sampel individu dan setiap atribut dipecah pada tree yang dipilih antara atribut subset yang bersifat acak. Dan pada proses klasifikasi, individunya didasarkan pada vote dari suara terbanyak pada kumpulan populasi tree.

Random Forest yang dihasilkan memiliki banyak tree, dan setiap tree ditanam dengan cara yang sama. Tree dengan variabel x akan ditanam sejauh mungkin dengan tree dengan variabel y. Dan dalam perkembangannya, sejalan dengan bertambahnya data set, maka tree pun ikut berkembang. Penempatan tree yang saling berjauhan membuat apabila terdapat tree disekitar tree x berarti pohon tersebut merupakan perkembangan dari tree x. Beberapa fungsi learning yang dihasilkan random forest digunakan strategi ensemble ”bagging” untuk mengatasi masalah overfitting apabila dihadapkan data set yang kecil.

Dalam penelitian ini, model Random Forest menggunakan seluruh fitur dan entri dalam dataset yaitu terdiri dari 34 atribut yang mana memiliki 1626 data entri. Berikut merupakan hasil dari pe-rankingan 34 fitur yang ada, sebagai berikut :

Feature	Ranking	Feature	Ranking
1. MLFER_E	0.160285	18. AATS3i	0.0158
2. maxaaN	0.129107	19. MAXDP	0.013725
3. AATS8v	0.086772	20. minHBa	0.011011
4. ATS0m	0.084293	21. AATSC0v	0.010643
5. SssNH	0.079935	22. ATSC1m	0.008704
6. AATS6m	0.051273	23. AATS7s	0.007863
7. AATS1i	0.040376	24. SaasC	0.006182
8. AATS8e	0.037748	25. AATS2s	0.006158
9. MDEC33	0.033887	26. SRW5	0.006117
10. MIC2	0.031459	27. MAXDN	0.005073
11. CrippenLogP	0.025894	28. ALogP	0.004233
12. SdO	0.024618	29. ATSC1i	0.003943
13. nHeteroRing	0.024031	30. ATSC8i	0.003939
14. C1SP2	0.023443	31. ETA_dBeta	0.003822
15. nHBDon	0.020056	32. ATSC3i	0.003332

16. AATS8s	0.017665	33. n6Ring	0.001404
17. ZMIC5	0.016068	34. n6HeteroRing	0.00114

2.2 Naive Bayes

Naïve Bayes merupakan klasifikasi sederhana yang menerapkan teorema bayes dengan menganggap semua fitur saling tidak berhubungan. Pengguna algoritma bayes ini menggunakan keseluruhan probabilitas, yaitu probabilitas dokumen terhadap kategori (prior). Kemudian teks akan terkategori berdasarkan probabilitas maksimum (posterior). Dengan kata lain metode ini mengasumsikan bahwa ada atau tidaknya fitur tertentu dari kelas tidak berhubungan dengan ada atau tidaknya fitur yang lain (Yuan, 2010). Berikut merupakan salah satu dari persamaan dari metode Naïve Bayes :

$$p(c_j \mid w_i) = \frac{p(w_i \mid c_j)\, p(c_j)}{p(w_i)}$$

$p(c_j \mid w_i)$	Peluang kategori j ketika terdapat kemunculan kata i
$p(w_i \mid c_j)$	Peluang sebuah kata i masuk ke dalam kategori j
$p(c_j)$	Peluang kemunculan sebuah kategori j
$p(w_i)$	Peluang kemunculan sebuah kata

Berikut merupakan 2 bentuk representasi dari metode naïve bayes yang digunakan dalam penelitian ini :

1. Gaussian Naive Bayes

Gaussian Bayes biasanya digunakan untuk merepresentasikan probabilitas bersyarat dari fitur continue pada sebuah kelas $P(Xi \mid Y)$, dan dikarakteristikkan dengan dua parameter : mean dan varian.

2. Bernoulli Naive Bayes

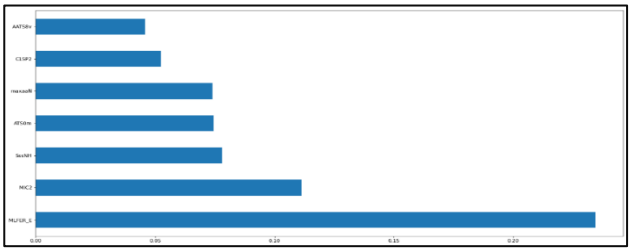
Pada Bernaulli Naïve Bayes, pembobotan dilakukan dengan menggunakan binary (0 dan 1) dalam pembobotan tiap term, hal ini berbeda dengan perhitungan term frekuensi yang melakukan pembobotan pada setiap term.

2.3 Validasi Model

Pada penelitian di ambil 7 parameter terbaik berdasarkan model Random Forest. Berikut table dibawah ini merupakan fitur yang diambil :

Feature	Ranking
1. MLFER_E	0.160285
2. maxaaN	0.129107
3. AATS8v	0.086772
4. ATS0m	0.084293
5. SssNH	0.079935
6. AATS6m	0.051273
7. AATS1i	0.040376

Berikut merupakan visualisasi dari 7 fitur terbaik yang diambil berdasarkan model Random Forest :



3 Hasil dan Pembahasan

3.1 Seleksi Fitur

Fitur yang diambil adalah *AATS8v, C1SP2, MAXAAn, ATS0m, SssNH, MIC2* dan *MLFER_E* karena berdasarkan peringkat yang lebih tinggi dan dianggap lebih baik.

3.2 Model Prediksi

Berdasarkan Model Naïve Bayes yang digunakan penelitian ini mengambil 2 bentuk representasi dari naïve bayes untuk memprediksi hasil dari data yang tersedia. Berikut merupakan hasil prediski dari kedua bentuk representasi yang di uji :

1. **Gaussian Naive Bayes**
   Berdasarkan model dari Gaussian naïve bayes yang telah di coba, diperoleh akurasi sebesar 94,69%.

```
Accuracy Gaussian Naive Bayes: 0.9469153515064562
```

2. **Bernoulli Naive Bayes**
   Berdasarkan model dari Bernoulli naïve bayes yang telah di coba, diperoleh akurasi sebesar 93,11%.

```
Accuracy Bernoulli Naive Bayes: 0.9311334289813487
```

4 Kesimpulan

Berdasarkan pada pengujian yang telah dilakukan pada studi QSAR untuk mengidentifikasi inhibitor CDK2 sebagai anti-kankerc dengan menggunakan 2 bentuk repsresentatif metode naïve bayes dapat disimpulkan. Fitur *AATS8v, C1SP2, MAXAAn, ATS0m, SssNH, MIC2* dan *MLFER_E* dianggap sebagai fitur yang terbaik didapatkan menggunakan model Random Forest yang mana sebelumnya telah dilakukan normalisasi data. Sehingga saat melakukan prediksi menggunakan model Naïve Bayes diperoleh nilai akurasi 94,69% untuk model Gaussian Naïve Bayes dan 93,11% untuk model Bernoulli Naïve Bayes.