BAB I

PENDAHULUAN

1.1 LATAR BELAKANG

Dalam era modern ini, Kemajuan dalam kimia dan fisika *molekuler* memberikan pemahaman tentang energi *atomisasi*, yaitu jumlah energi yang diperlukan untuk memisahkan semua *atom* dalam suatu *molekul*[1] yang berperan dalam berbagai bidang seperti farmasi dan ilmu material. Namun, penentuan energi ini secara eksperimen atau komputasi kuantum membutuhkan waktu, biaya, dan fasilitas kompleks. Dengan adanya permasalahan tersebut sebagai solusi, penelitian ini menggunakan *machine learning*, khususnya algoritma *Random Forest* dan *Gradient Boosting*, untuk memprediksi energi *atomisasi molekul* dengan efisien dan akurat, menggunakan data fitur *molekuler* dari *Kaggle*. Fokus penelitian ini adalah mengevaluasi performa kedua algoritma dalam menghasilkan prediksi yang akurat dan efisien.

Berdasarkan beberapa penelitian yang dilakukan, penerapan *machine learning* dalam bidang kimia dan material menunjukkan potensi yang signifikan. Penelitian sistem persediaan bahan kimia di laboratorium *forensik* menggunakan algoritma *Random Forest* untuk memprediksi stok bahan kimia dengan hasil yang efektif[2]. Penelitian lain menggunakan *Extreme Learning Machine (ELM)* untuk klasifikasi senyawa kimia, dengan akurasi terbaik pada dua kelas, namun menurun pada tiga kelas[3]. Di sisi lain, penelitian tentang *inhibisi korosi* senyawa obat menunjukkan bahwa *XGBoost*, setelah penyetelan *hyperparameter*, mengungguli algoritma

lainnya[4]. Terakhir, penelitian tentang modeling energi *atomisasi* molekul dengan *machine learning* menunjukkan akurasi tinggi dan dapat memperluas aplikasi dalam desain senyawa serta reaksi kimia. Semua penelitian ini menekankan pentingnya pemilihan dan optimasi algoritma yang tepat untuk aplikasi di bidang kimia dan material[5].

Penelitian tentang prediksi energi molekul menunjukkan bahwa algoritma *machine learning* seperti *Random Forest* dan *Gradient Boosting* efektif dan efisien, seperti *Random Forest* unggul dalam mengatasi *overfitting* dan *stabilitas*, Algoritma ini juga menunjukkan *stabilitas* yang tinggi dan dapat menangani *dataset* besar dengan baik, sehingga cocok untuk aplikasi di bidang kimia yang melibatkan data molekuler yang beragam[6], namun memiliki kekurangan dalam interpretabilitas model dan waktu pemrosesan yang lama jika jumlah pohon terlalu besar[7]. Sebaliknya, *Gradient Boosting* menghasilkan model dengan akurasi tinggi, tetapi lebih rentan terhadap *overfitting* dan membutuhkan waktu pelatihan lebih lama[8]. Solusi untuk mengatasi kekurangan ini adalah dengan melakukan rekayasa fitur, yang terbukti dapat meningkatkan hasil prediksi energi molekul. Beberapa penelitian menunjukkan bahwa penerapan rekayasa fitur ini mampu memberikan hasil yang lebih baik dibandingkan metode konvensional dalam konteks pemodelan energi[9]. Berdasarkan penelitian tersebut, kombinasi antara kedua algoritma dengan rekayasa fitur berbasis multipol bentuk molekul diharapkan dapat menghasilkan prediksi yang lebih akurat dan efisien.

Melalui penelitian ini, penulis berupaya memberikan solusi terhadap permasalahan prediksi energi molekul dengan mengangkat judul "**PREDIKSI**

ENERGI MOLEKUL MENGGUNAKAN RANDOM FOREST DAN GRADIENT BOOSTING BERBASIS MACHINE LEARNING”.

1.2 RUMUSAN MASALAH

Berdasarkan latar belakang diatas, maka rumusan masalah pada penelitian ini adalah :

1. Bagaimana menerapkan algoritma *Random Forest* dan *Gradient Boosting* untuk memprediksi energi molekul dengan akurasi tinggi menggunakan pendekatan *machine learning*?
2. Bagaimana mengevaluasi performa algoritma *Random Forest* dan *Gradient Boosting* dalam memprediksi energi molekul berdasarkan akurasi prediksi yang sesuai?

1.3 BATASAN MASALAH

Agar pembahasan masalah tidak meluas dari penelitian ini, maka penulisan memberikan Batasan masalah sebagai berikut :

1. Bahasa pemrograman yang digunakan adalah *Pyhton* dengan menggunakan *Google Colab*.
2. Penelitian ini menggunakan algoritma *Random Forest* dan *Gradient Boosting*.
3. Penelitian ini menggunakan *dataset* "*Ground State Energies of Molecules*" yang terdiri dari 16.242 *entri* yang berisi energi keadaan dasar 16.242 molekul yang dihitung dengan simulasi mekanika kuantum.
4. *Dataset* yang digunakan berjumlah 16.242 baris.

5. Atribut dalam *dataset* meliputi 1275 kolom pertama *entri* dalam *matriks Coulomb* yang berfungsi sebagai fitur *molekuler*. Kolom ke-1276 adalah *Pubchem Id* tempat struktur *molekuler* diperoleh. Kolom ke-1277 adalah energi *atomisasi (Eat)* yang dihitung melalui simulasi menggunakan paket *Quantum Espresso*.
6. Target yang digunakan adalah kolom energi *atomisasi (eat)*.
7. Evaluasi model yang digunakan adalah *Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared (R2)*.

1.4 TUJUAN PENELITIAN

Adapun tujuan yang ingin dicapai dari penelitian ini adalah sebagai berikut:

1. Menerapkan algoritma *Random Forest* & *Gradient Boosting* untuk memprediksi energi molekul menggunakan *dataset 'Energy Molecule'* dari *Kaggle.*
2. Mengevaluasi akurasi model prediksi energi molekul yang dikembangkan menggunakan *Random Forest* & *Gradient Boosting*.

1.5 MANFAAT PENELITIAN

Manfaat yang akan didapatkan dari penelitian ini, yaitu:

1. Mendukung penerapan algoritma *Random Forest* dan *Gradient Boosting* dalam prediksi energi molekul menggunakan *dataset 'Ground State Energies of Molecules'* dari *Kaggle*.

2. Menghasilkan model prediksi energi molekul dengan akurasi yang lebih tinggi melalui evaluasi model yang dikembangkan menggunakan algoritma *Random Forest* dan *Gradient Boosting*.
3. Menjadi referensi bagi penelitian selanjutnya yang menggunakan algoritma *machine learning* untuk prediksi sifat molekul.

1.6 SISTEMATIKA PENULISAN

Sistematika ini menggambarkan tentang pembahasan yang penulis buat pada tugas akhir untuk memudahkan dalam memahami penulisan laporan ini. Adapun sistematika penulisan dalam penelitia ini disusun sebagai berikut:

BAB I : PENDAHULUAN

Dalam bab pendahuluan penulis membahas tentang latar belakang masalah, perumusan masalah, batasan masalah, tujuan dan manfaat penelitian serta sistematika penulisan.

BAB II : LANDASAN TEORI

Dalam bab ini penulis menguraikan teori-teori yang menjadi dasar dalam penelitian ini. Adapun teori-teori ini bersumber dari buku dan jurnal-jurnal untuk mendukung pemahaman.

BAB III : METODOLOGI PENELITIAN

Dalam bab ini penulis menjelaskan metodologi yang digunakan alam penelitian, termasuk desain penelitian, teknik pengumpulan data, dan merode analisis data.

BAB IV : ANALISIS DAN HASIL

Dalam bab ini penulis memparkan analisis data yang diperoleh selama penelitian dan hasil dari analisis tersebut. Bab ini juga mencakup interpretasi hasil serta diskusi mengenai temuan penelitian.

BAB V : PENUTUP

Dalam bab ini penulis memberikan kesimpulan dari hasil penelitian dan menyampaikan saran-saran yang relevan untuk penelitian lebih lanjut atau penerapan praktis dari temuan penilitian.

Dengan sistematika penulisan ini, diharapkan pembaca dapat dengan mudah mengikuti alur penelitian dan memahami setiap bagian dari laporan penelitian yang disusun.