BAB I
PENDAHULUAN

1.1 Latar Belakang

Beberapa dekade kebelakang, perkembangan ilmu pengetahuan telah memasuki eranya mesin yang dapat berpikir dan mengambil keputusan sendiri atau lebih dikenal dengan pengertian *artificial intelligence* (AI). Didalam pengambilan keputusan tersebut, sebuah AI haruslah menjalani tahap pengujian terlebih dahulu dengan menggunakan algoritma – algoritma yang sering disebut *machine learning.* Kelebihan yang ditawarkan oleh AI ini tidak hanya dapat menyelesaikan permasalahan dalam bidang sains, bahkan mampu untuk menyelesaikan permasalahan kesehatan, manufaktur hingga pertanian. Pada penelitian ini penggunaan AI lebih ditekankan untuk menyelesaikan permasalan pada bidang Fisika.

Many-body problems yang erat kaitannya dengan sistem mikroskopis yang terdiri dari banyak partikel berinteraksi menyiratkan bahwa sistem mekanika kuantum bisa dijelaskan dengan fungsi gelombang, mengingat bahwa fungsi gelombang dari sistem *many-body problems* menyimpan berbagai informasi penting terkait sistem, salah satunya ialah energi sehingga solusi eksaknya sangat sulit untuk diperoleh dari fungsi gelombang tersebut bahkan hampir mustahil diselesaikan secara analitik, terkecuali untuk sistem yang sangat sederhana seperti atom Hidrogen. Salah satu langkah Peneliti untuk menyelesaikan permasalahan yang berkaitan dengan persamaan differensial adalah dengan melakukan pemodelan dan simulasi(Abimanyu, 2020).

Simulasi merupakan suatu proses untuk mendeskripsikan suatu masalah dalam rentang waktu tertentu. Simulasi ini memerlukan data untuk menarik kesimpulan karakteristik masalah yang diselesaikan. Pemodelan dan simulasi ini mendeskripsikan masalah dari suatu sistem yang dibuat secara matematika sehingga dapat diselesaikan secara numerik. Simulasi dilakukan dengan mengubah variabel dari model yang dibuat sehingga permasalahan dapat diprediksi dan divalidasi berdasarkan skenario validasi yang digunakan(Banks, 2010). Oleh karenanya, penyelesaian dari *many-body problems* bisa dilakukan dengan

pendekatan komputasi. Diantara pendekatan yang umumnya digunakan ialah Metode Hatree-Fock, *variational Monte Carlo*, dan *Density Functional Theory* (DFT). Dengan adanya pendekatan secara komputasi tersebut, informasi yang ingin ditemukan mengenai masalah bisa diperoleh tanpa menyelesaikan Persamaan Schrödinger secara analitis.

Pekembangan teori mekanika kuantum untuk memprediksi sifat dari suatu senyawa yang hasilnya akurat menjadi mudah bagi peneliti. Metode yang sering digunakan secara komputasi ialah *Density Functional Theory* (DFT). Dengan menggunakan metode ini penyelesaian Persamaan Schrödinger pada masalah banyak-partikel menjadi solusi yang cerdas dilakukan(Pongajow dkk., 2013). Metode *Density Functional Theory* (DFT) sering digunakan untuk menyelesaikan masalah geometri dan struktur elektron kompleks dari logam transisi. Ide dari metode ini adalah energi dari suatu molekul dapat ditentukan dari kerapatan elektron dari molekul tersebut. Pada logam transisi, metode DFT untuk menentukan struktur dan vibrasi energi lebih akurat daripada Metode Hatree-Fock(Pongajow dkk., 2017). Sehingga metode ini mampu mendekati hasil eksperimen tanpa membutuhkan waktu yang lama. Penggunaan metode ini memberikan kemudahan bagi peneliti yang melakukan kajian tentang geometri dan karakteristik agar informasi tentang geometri dan karakteristik ikatan tersebut dapat menententukan struktur dan stabilitas dari suatu senyawa kompleks tersebut(Pongajow dkk., 2017) Selanjutnya upaya untuk mendapatkan hasil yang lebih akurat tanpa menggunakan numerik telah dikemukakan ide untuk menyelesaikannya dengan algoritma *Machine Learning* dengan Matriks Coulomb sebagai deskriptornya.

Algoritma *Machine Learning* bergantung pada data yang diberikan dan deskriptor atau fungsi yang mewakili oleh simulasi yang dibuat(DQLab, 2021). Sebagai salah satu fungsi secara unik dapat mewakili sistem kuantum berdasarkan posisi dan muatan atom penyusunnya. Fitur atau deskriptor yang digunakan ialah Matriks Coulomb. Coulomb Matrix (CM) adalah deskriptor sederhana yang meniru interaksi elektrostatik antara inti(Tchagang & Valdés, 2019). Sejak Matriks Coulomb (CM) dipopulerkan oleh Rupp dkk(2012), deskriptor yang lebih canggih telah dikembangkan lagi, tetapi CM masih menjadi referensi peneliti ketika Peneliti membandingkan molekul satu dengan lainnya. Dengan menggabungkan *Machine*

Learning sebagai algoritmanya dan Matriks Coulomb sebagai deskriptor atau fitur penelitian ini menjadi mudah dan relatif cepat daripada menyelesaikannya secara manual. *Machine Learning* dapat digunakan dengan menggunakan fungsi dan data DFT untuk energi semprotan molekul. Namun sejak ditemukannya konsep Matriks Coulomb, penelitian tentang *Machine Learning* dari sistem kuantum dilakukan menggunakan algoritma *Machine Learning* sederhana hingga Jaringan Syaraf Tiruan (Rupp dkk., 2012).

Dengan algoritma *Machine Learning,* Peneliti dapat menyederhanakan desain bahan kimia dan padatan serta melihat sifat elektronik molekul tanpa menjalankan simulasi molekul satu persatu dari awal. Untuk penyelesaian data molekul yang diberikan pada struktur elektroniknya, *Machine Learning* memainkan peranan yang sangat penting dalam pengembangan sains dan teknologi. Kemajuan teknologi tidak lepas dari berkembangnya metode – metode yang digunakan dalam *Machine Learning* itu sendiri yang dimulai pada tingkat sederhana hingga dengan kompleksitas yang tinggi. Pada penelitian terdahulu, proses untuk memodelkan distribusi energi atomisasi molekul ini menggunakan metode lain yakni dengan *XGBoost* dan *Convulutioal Neural Network* yang mana dilakukan oleh (Abimanyu, 2020) yang mendapatkan hasil bahwa algoritma dengan metode regresi *XGBoost* memiliki nilai *Mean Absolute Error* (MAE) dan waktu yang lebih cepat dari algoritma *Convulutioal Neural Network*. Begitu juga dengan penelitian yang dilakukan oleh (Sumanto dkk., 2021) menghasilkan bahwa algoritma XGBoost lebih baik dan akurat memprediksi energi atomisasi molekul daripada algoritma *Convulutional Neural Network*. Berdasarkan dua penelitian sebelumnya, didapat bahwa algoritma XGBoost lebih bisa diandalkan pada studi kasus ini. Penelitian ini melanjutkan membandingkan dengan algoritma regresi yang lainnya seperti , *K-Neighbor Regressor* dan *Random Forest Regressor* dan mencoba mengoptimasi lagi algoritma XGBoost dengan menghitung koefisien determinasi(R^2), *root mean square error*(RMSE), sumber daya dan lama waktu yang dibutuhkan masing – masing algoritma menghasilkan distribusi energi atomisasi molekul tersebut. Algoritma *XGBoost* merupakan algoritma yang sering dipakai karena termasuk algoritma yang baik, pada dasarnya algoritma ini merupakan pengembangan dari algoritma *gradient boost* dengan penambahan beberapa proses sehingga lebih baik. Proses yang dimaksud ialah pemangkasan, *newton boosting* dan adanya parameter

pengacakan ekstra, karena penambahan fitur tersebut algoritma ini diberi nama *Extreme Gradient Boosting(XGBoost)*. Algortima kedua yang digunakan ialah *K-Neighbor Regressor*, algoritma ini termasuk algoritma sederhana dengan langkah pertama yang dilakukan ialah mengklasifikasikan berdasarkan kelompok tertentu dengan menghitung jarak antara euclidieanny(Wanto dkk., 2020). Algoritma ketiga ialah *Random Forest*, algoritma ini berdasarkan dari pohon keputusan yang membuat beberapa keputusan menggunakan data booststrap dan secara acak memilih subset variabel yang berada pada pohon keputusan. Hasil akurasi dari algoritma ini akan dibandingkan berdasarkan nilai koefisien determinasi (R^2), *root mean square error* (RMSE) dan lama waktu yang dibutuhkan masing – masing algoritma menyelesaikan pemodelannya. Dengan ini penelitiannya diberi judul dengan "**Evaluasi Algoritma *K-Nearest Neighbors, Random Forest* Dan *Xgboost* Pada Prediksi Energi Atomisasi Molekul**".

1.2 Ruang Lingkup Masalah

Ruang lingkup permasalahan yang diteliti pada penelitian ini ialah bagaimana cara mendistribusikan energi atomisasi dengan algoritma *Xgboost Regressor, K-Neighbors Regressor* dan *Random Forest Regressor* dan divalidasi menggunakan nilai koefisien determinasi, *Root Mean Square Error*(RMSE) serta lama waktu masing – masing algoritma tersebut. Data acuan yang digunakan berasal dari database PubChem Chemical Compound sebanyak 16242 molekul yang terdiri dari 2 – 50 atom – atom penyusun molekul tersebut. Membangun simulasi distribusi energi atomisasi molekul tersebut menggunakan bahasa pemprograman Python dengan bantuan *library Machine Learning* seperti *library* Scikit-Learn dan pendukungnya matplotlib, seaborn, numpy.

1.3 Rumusan Masalah

1. Bagaimana mensimulasikan distribusi energi atomisasi molekul dengan algoritma *Xgboost Regressor, K-Neighbors Regressor* dan *Random Forest Regressor* ?
2. Bagaimana menentukan perbandingan keakuratan dari algoritma *Xgboost Regressor, K-Neighbors Regressor* dan *Random Forest* dengan mempenghitungkan koefisien determinasi (R^2) nya?

3. Bagaimana menentukan perbandingan algoritma *Xgboost Regressor*, *K-Neighbors Regressor* dan *Random Forest Regressor* yang paling optimal dengan memperhitungkan *Root Mean Square Error* (RMSE) nya ?
4. Bagaimana keefektifan waktu yang dibutuhkan algoritma *Xgboost Regressor, K-Neighbors Regressor* dan *Random Forest Regressor* untuk menyelesaikan simulasi distribusi energi atomisasi molekul tersebut ?
5. Bagaimana penggunaan sumber daya komputasi yang dibutuhkan algoritma *Xgboost Regressor, K-Neighbors Regressor* dan *Random Forest Regressor* untuk menyelesaikan simulasi distribusi energi atomisasi molekul tersebut ?

1.4 Batasan Masalah

1. Algoritma yang digunakan untuk mensimulasikan distribusi energi atomisasi molekul ialah *Random Forest Regressor, K-Neighbors Regressor,* dan *XGB Regressor.*
2. Menggunakan koefisien determinasi (R^2) untuk membandingkan keakuratan algoritma *Random Forest Regressor, K-Neighbors Regressor,* dan *XGB Regressor.*
3. Menggunakan *Root Mean Square Error* (RMSE) untuk membandingkan optimasi algoritma *Random Forest Regressor, K-Neighbors Regressor,* dan *XGB Regressor*.
4. Menggunakan lama waktu untuk mengukur efektifitas algoritma *Random Forest Regressor, K-Neighbors Regressor,* dan *XGB Regressor.*
5. Menggunakan sumber daya komputasi untuk mengukur efisiensi algoritma *Random Forest Regressor, K-Neighbors Regressor,* dan *XGB Regressor.*
6. Dataset bersumber dari website Pubchem Chemical and Chompound dengan spesifikasi molekul disusun oleh atom C, H, N, O, P dan S dengan tiap molekul mengandung 2-50 atom.

1.5 Tujuan Penelitian

1. Mensimulasikan distribusi energi atomisasi molekul dengan algoritma *Random Forest Regressor, K-Neighbors Regressor,* dan *XGBoost Regressor* untuk memprediksi energi atomisasi molekul berdasarkan elemen Matriks Coulumb nya.
2. Mendapatkan hasil yang paling akurat dari algoritma *Random Forest Regressor, K-Neighbors Regressor,* dan *XGB Regressor* untuk menentukan distribusi energi atomisasi molekul berdasarkan hasil koefisien determinasi (R^2).
3. Membandingkan algoritma yang optimal dengan menghasilkan *Root Mean Square Error* (RMSE) sekecil mungkin..
4. Membandingkan keefektifan waktu yang digunakan algoritma *Xgboost Regressor, K-Neighbors Regressor* dan *Random Forest Regressor* ketika menyelesaikan distribusi energi atomisasi molekul.
5. Membandingkan efisiensi sumber daya komputasi yang digunakan algoritma *Xgboost Regressor, K-Neighbors Regressor* dan *Random Forest Regressor* untuk mendistribusikan energi atomisasi molekul dari sisi penggunaan prosesor dan memory

1.6 Manfaat Penelitian

Penelitian yang dilakukan diharapkan dapat memberikan manfaat sebagai berikut:

1. Model yang dibuat pada penelitian ini dapat digunakan sebagai referensi untuk penelitian selanjutnya.
2. Model yang dibuat pada penelitian ini dapat dijadikan sebagai pembanding untuk penelitan serupa.
3. Model yang dibuat pada penelitian ini dapat membantu peneliti lain yang menggunakan aplikasi tertentu untuk mendesain bahan kimia/material.