

BAB I PENDAHULUAN

1.1 Latar Belakang

Teknologi pada saat ini berkembang dengan pesat seiring dengan perkembangan umat manusia. Penerapan teknologi informasi juga telah memasuki berbagai bidang khususnya bidang kesehatan. Data mining dapat digunakan pada bidang kesehatan untuk mendiagnosis penyakit berdasarkan dataset medis pasien terdahulu. Klasifikasi merupakan salah satu bagian penting dari *data mining*. Pada bidang klasifikasi terdapat banyak cabang salah satunya adalah pohon keputusan (*decision tree*). Salah satu penyakit yang banyak menyerang manusia jaman sekarang yaitu Diabetes. Diabetes merupakan penyakit yang mematikan. Faktor yang mempengaruhi penyakit diabetes antara lain dikarenakan gaya hidup kurang sehat, jarang olahraga, dan makan melebihi porsi dari apa yang dibutuhkan tubuh tiap hari [1].

Penelitian yang telah dilakukan oleh Intan Kemala dkk [2]. dengan tujuan untuk membandingkan kinerja metode *bagging* dan *non-ensemble* pada klasifikasi wilayah di Indonesia. Penelitian tersebut menghasilkan akurasi tertinggi pada algoritma *Random Forest* untuk mewakili *bagging* sebesar 95,14%. Penelitian yang dilakukan oleh Andi Kurniawan dan Agung Prihandono [3] untuk meningkatkan akurasi klasifikasi dalam menentukan blogger profesional. Penelitian tersebut menggunakan metode *Naïve bayes* dan *Bagging*. Hasil dari penelitian tersebut menunjukkan bahwa ada peningkatan akurasi sebesar 3% pada algoritma setelah diterapkannya metode *Bagging* terhadap *Naïve Bayes*. Penelitian yang telah dilakukan sebelumnya oleh Mirqotussa'adah, dkk [4] menggunakan metode *ensemble bagging* dan algoritma *C4.5* pada data diabetes menghasilkan sistem dengan akurasi yang lebih tinggi dibandingkan sistem tanpa menggunakan metode *ensemble bagging*. Penelitian tersebut menunjukkan hasil akurasi 68,61% untuk sistem tanpa menggunakan metode *ensemble*. Sedangkan untuk sistem dengan menggunakan metode *ensemble* menghasilkan akurasi 74,87%. Dari hasil tersebut dapat dilihat bahwa akurasi meningkat sebanyak 6,26%. Penelitian lainnya yang dilakukan oleh Desi Ayu Wulandari, dkk [5] menggunakan metode *ensemble bagging* dan metode *classification and regression tree (CART)* telah menghasilkan

sebuah sistem dengan akurasi 57,5%. Lebih tinggi dari sistem yang hanya menggunakan metode *CART* yaitu 50%. Dari nilai tersebut didapatkan peningkatan akurasi sebesar 7,5%. Penelitian yang dilakukan oleh Noviandi [6] menggunakan metode *C4.5* terhadap data yang sama yaitu Pima Indian Diabetes menghasilkan akurasi sebesar 70.32%. Perbedaan penelitian ini dengan penelitian yang telah dilakukan Noviandi adalah pada *pre-processing* data dan metode untuk membentuk model klasifikasinya. Penelitian yang dilakukan Noviandi dilakukan tanpa melalui tahapan *pre-processing* terlebih dahulu.

Penelitian kali ini menggunakan data dari PIDD (*Pima Indian Diabetes*) yang didapatkan dari *kaggle UCI Repository*. Data ini sudah banyak digunakan pada penelitian sebelumnya untuk tujuan menguji sebuah algoritma klasifikasi [4][1]. Dengan alasan tersebut maka saya menggunakan dataset ini dalam penelitian sekarang. Dataset tersebut bertipe data kontinu dan juga terdapat berapa *missing value*. Oleh karena itu sebelum membentuk sistem data harus dilakukan *pre-procesing* terlebih dahulu. Untuk mengatasi *missing value* tersebut menggunakan imputasi dan dilakukan diskritisasi untuk mengatasi data kontinu. *Missing value* adalah sebuah keadaan dimana hilangnya data pada suatu atribut tertentu. Keadaan tersebut biasanya disebabkan oleh kesalahan saat mengumpulkan data atau ketidakmampuan narasumber untuk memberikan jawaban dari pertanyaan yang diberikan. *Missing value* berpengaruh kepada performa data karena *missing value* dapat menurunkan kinerja dari data tersebut saat dilakukan proses klasifikasi.[7] Untuk mengatasi *missing value* terdapat 3 metode yaitu *Litwise deletion*, *Pairwise deletion*, dan imputasi. *Litwise deletion* mengatasi *missing value* dengan cara menghilangkan observasi yang mengandung *missing value*. *Pairwise deletion* mengatasi *missing value* dengan cara menghapus data yang mengandung *missing value*. Sedangkan imputasi mengatasi *missing value* dengan cara mengisi data kosong berdasarkan data yang telah diketahui. Penggunaan *Litwise deletion* dan *Pairwise deletion* kurang efektif untuk digunakan karena dapat menghilangkan informasi penting dari data sehingga dalam penelitian ini digunakan metode imputasi untuk mengatasi *missing value*. Imputasi terbagi menjadi beberapa jenis misalkan *mean*, *median*, dan *KNN*. Dalam penelitian ini menggunakan imputasi *KNN* sebagai metode untuk mengatasi *missing value* [8]. Diskritisasi merupakan

sebuah metode untuk mengkonversi data kontinu menjadi data diskrit dengan membaginya kedalam kategori tertentu. Diskritisasi mampu menurunkan penggunaan memori sistem dan meningkatkan efisiensi algoritma. Selain itu metode diskritisasi juga dapat membuat data lebih ringkas dan mudah dipahami [4][1].

Penelitian yang saya jelaskan sebelumnya menggunakan metode *C4.5* dan *Ensemble Bagging* pada data pima Indian diabetes menghasilkan akurasi yang cukup tinggi yaitu 74,87%. Penelitian kali ini akan mencoba menggunakan metode *ID3* dan *Bagging* untuk membangun sistem pada penyakit diabetes. Penggunaan *ID3* didasari pada penelitian sebelumnya yang menggunakan metode *C4.5 Bagging* sehingga timbul rasa keingintahuan apakah bagging juga akan berpengaruh kepada algoritma *ID3*. Perbedaan metode *C4.5* dan *ID3* terdapat pada saat menentukan *root node*. Metode *ID3* menggunakan nilai *Information Gain* untuk menentukan *root node*, sedangkan *C4.5* menggunakan nilai *Gain Ratio* untuk menentukan *root node* [9]. Metode ensemble merupakan metode yang dipakai untuk meningkatkan performa algoritma klasifikasi dengan cara membangun beberapa model dari data train. Kemudian pada saat menentukan hasil prediksi metode ini menggunakan teknik majority vote dari classifier-classifier tersebut. Proses penggabungan hasil prediksi ini biasanya disebut dengan *aggregating* [4]. Contoh dari metode ensemble adalah metode *Bootstrap Aggregating* atau juga bisa disebut *Bagging*. Berdasarkan hasil dari penelitian sebelumnya, metode *ensemble bagging* terbukti mampu untuk meningkatkan performa akurasi klasifikasi.

Dengan adanya penelitian ini diharapkan dapat terbentuk sistem yang memiliki akurasi tinggi.

1.2 Perumusan Masalah

Berdasarkan pada uraian latar belakang diatas maka dapat dipaparkan rumusan masalah sebagai berikut :

1.2.1 Permasalahan

Dataset Pima Indian Diabetes terdapat beberapa nilai yang kosong. Jika dibiarkan maka akan mempengaruhi kinerja dari sistem. Dataset

tersebut masih bertipe data kontinu sehingga harus dilakukan diskritisasi agar data dapat digunakan dengan metode yang diusulkan. Akurasi sangat penting dalam sebuah sistem, oleh karena itu dibutuhkan sebuah metode tambahan untuk menambah akurasi dari sistem yang akan dibuat.

1.2.2 Usulan Solusi

Solusi yang diusulkan dalam penelitian ini adalah dengan menggunakan Imputasi KNN untuk mengatasi missing value dan menggunakan diskritisasi fuzzifikasi untuk mengatasi atribut kontinu serta mengkombinasikan teknik *ensemble bagging* dan algoritma *ID3* untuk menambah akurasi sistem.

1.2.3 Pertanyaan Penelitian

1. Bagaimana cara mengatasi *missing value* dan data kontinu pada data Pima Indian Diabetes ?
2. Berapakah pertambahan akurasi pada metode *ID3* setelah dikombinasikan dengan *Ensemble Bagging* terhadap studi kasus Pima Indian Diabetes ?

1.3 Tujuan & Manfaat

1.3.1 Tujuan Penelitian

Tujuan dari penelitian ini adalah :

1. Mengatasi *missing value* dan data kontinu pada dataset Pima Indian Diabetes.
2. Menambah akurasi dari algoritma klasifikasi *ID3*.

1.3.2 Manfaat Penelitian

Hasil dari penelitian ini adalah dapat menjadi metode usulan untuk penelitian mengenai klasifikasi yang akan datang serta menjadi rujukan untuk mengembangkan sistem dengan akurasi yang lebih tinggi pada penelitian dengan menggunakan metode *Ensemble Bagging*.

1.4 Batasan-batasan

Batasan masalah di penelitian ini adalah :

1. Dataset menggunakan data Pima Indian Diabetes dari *UCI Repository* dengan jumlah data sebanyak 768 data. (<https://www.kaggle.com/uciml/pima-indians-diabetes-database>)
2. Dataset Pima Indian Diabetes memiliki 9 atribut dengan ketentuan 8 atribut (Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPredigreeFunction) sebagai atribut prediktor dan 1 atribut (Outcome) sebagai kelas.
3. Bahasa pemrograman yang digunakan adalah bahasa Python 3.10.5.
4. Menggunakan teknik *bagging* dengan algoritma *ID3* sebagai pembentuk tree.
5. Pengujian menggunakan metode *K-Fold* dan pengukuran performa menggunakan *Confusion Matrix*.

1.5 Sistematika Proposal

BAB I PENDAHULUAN

Bab ini membahas tentang latar belakang, perumusan masalah, tujuan, manfaat, batasan – batasan, dan sistematika proposal.

BAB II TINJAUAN PUSTAKA

Bab ini menjelaskan teori – teori yang mendukung penelitian, serta menjadi dasar dalam pemecahan masalah.

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan apa saja langkah-langkah penelitian yang akan digunakan.

BAB IV IMPLEMENTASI METODE

Bab ini membahas tentang implementasi metode-metode yang digunakan dan juga tahapan-tahapan yang dibutuhkan untuk melakukan implementasi tersebut, serta hasil dari pengujian.

BAB V PENUTUP

Bab terakhir membahas mengenai kesimpulan yang didapat dalam penelitian dan juga saran untuk penelitian kedepannya.