

ABSTRAK

Penerapan teknologi informasi telah memasuki berbagai bidang khususnya bidang medis. Teknologi data mining bisa digunakan untuk memprediksi penyakit diantaranya adalah diabetes. Model dari data mining diantaranya adalah klasifikasi. Di bidang klasifikasi terdapat banyak cabang yang berkembang salah satunya adalah pohon keputusan. Algoritma tree yang digunakan pada penelitian ini adalah *ID3*. Dalam penelitian ini dataset yang digunakan yaitu Pima Indian Diabetes yang diambil dari Kaggle UCI repository. Dataset berjumlah 768 data dengan pasien keseluruhannya adalah perempuan dengan umur minimal 21 tahun. Data ini bertipe data *kontinyu* (numerik) dan masih mengandung *missing value* sehingga perlu dilakukan *pre-processing* terlebih dahulu sebelum data ini diolah ke metode selanjutnya. *Pre-processing* dari dataset ini terdapat 2 proses yaitu imputasi *missing value* kemudian diskritisasi. Setelah data di *pre-processing* maka dapat dilakukan proses pembentukan sistem menggunakan metode *ID3 ensemble bagging*. Metode *ensemble bagging* merupakan metode yang digunakan untuk meningkatkan performa akurasi algoritma klasifikasi. Di dalam metode *ensemble bagging* terdapat dua proses yaitu pembentukan data baru atau *Bootstrap data* dan penyatuan hasil prediksi atau *agregatting*. Pembuatan model pada penelitian ini adalah dengan menerapkan 10-fold cross validation. Setelah itu evaluasi hasil pengujian menggunakan data testing. Pengujian metode *ID3* memiliki rata-rata akurasi sebesar 74,89%. Sedangkan pengujian menggunakan metode *ID3* dengan bagging memiliki rata-rata akurasi tertinggi sebesar 76,83%. Hasil tersebut menunjukkan penambahan akurasi sebesar 1,94%.

Kata kunci : *Decision Tree, Ensemble, Bagging, ID3, Diabetes, PIDD*

ABSTRACT

The application of information technology has entered various fields, especially the medical field. Data mining technology can be used to predict diseases such as diabetes. One of the models of data mining is classification. In the field of classification, there are many branches that develop, one of which is the decision tree. The tree algorithm used in this study is ID3. In this study, the dataset used was Pima Indian Diabetes, which was taken from the Kaggle UCI repository. The dataset totals 768 data with all patients being women with a minimum age of 21 years. This data is of continuous (numeric) data type and still contains a missing value, so pre-processing is required before this data is processed to the next method. There are 2 processes for pre-processing this dataset, namely imputation of missing values and then discretization. After the data is pre-processed, the system formation process can be carried out using the ID3 ensemble bagging method. The ensemble bagging method is a method used to improve the performance of classification algorithm accuracy. In the ensemble bagging method, there are two processes, namely the formation of new data or bootstrap data and the unification of predicted results or aggregating. Making the model in this study is to apply 10-fold cross validation. After that, evaluate the test results using data testing. The ID3 method test has an average accuracy of 74.89%. While testing using the ID3 method with bagging has the highest average accuracy of 76.83%. These results indicate an additional accuracy of 1.94%.

Keywords : *Decision Tree, Ensemble, Bagging, ID3, Diabetes, PIDD*