ABSTRAKSI

Data *mining* merupakan teknik pengambilan informasi dari data yang besar berupa transaksi bisnis, data ilmiah, gambar, video dan data lainnya untuk mendapatkan informasi baru. Salah satu kegunaan dari data *mining* adalah klasifikasi. Klasifikasi bertujuan untuk pengelompokan data berdasarkan ikatan antara *variable* dalam data. Tujuan dari penelitian ini melakukan perbandingan penggunaan metode klasifikasi *Naïve Bayes*, *Extreme Gradient Boosting*, dan *Backpropagation Neural Network* pada beberapa jenis data. Penelitian ini memiliki 4 tahapan yaitu *preprocessing*, perancangan model, implementasi model, dan analisis hasil klasifikasi. Dalam tahapan *preprocessing* terbagi menjadi 4 tahap yaitu *encoding*, penanganan *missing value*, *split* data *train* dengan *test*, dan *scalling*. Encoding dilakukan pada atribut data berjenis kategorik. Dalam penanganan *missing value* terdapat 5 metode terdiri dari pengisian dengan nilai *mean*, median, modus, penghapusan, dan prediksi dengan regresi linear. Lalu pada tahap *split* data *train* dan *test* dengan bobot 80:20. Kemudian *scalling* terdapat 4 skenario diantaranya tanpa, *standard*, *robust*, dan *minmax scalling*. Pada perancangan model klasifikasi menggunakan library python seperti sklearn untuk klasifikasi *Naïve Bayes* dan *Extreme Gradient Boost*. Kemudian Tensorflow untuk klasifikasi *Backpropagation Neural Network*. Implementasi model klasifikasi merupakan kombinasi dari penanganan *missing value*, metode *scalling*, dan metode klasifikasi yang menghasilkan 72 skenario. Tahapan Analisis dan Kesimpulan merupakan perbandingan hasil skenario dari 5 hasil karakteristik klasifikasi yaitu *Precision*, *Recall*, *Accuracy*, *F1-Score*, dan *Error Rate* pada dua jenis dataset berbeda. Kemudian dari hasil pengujian skenario didapatkan *Backpropagation Neural Network* menjadi metode klasifikasi terbaik pada dataset *univariate* dan *variate*. Klasifikasi pada dataset NBA berguna untuk pemain meningkatkan kemampuan agar dapat bermain mencapai 5 tahun, dan tim saat rekrutmen pemain. Kemudian untuk dataset *Car Insurance* berguna bagi pengemudi menjaga perilaku berkendara dan kondisi kendaraanya. Terakhir untuk perusahaan asuransi mengurangi kerugian.

Kata kunci : Data *mining*, Klasifikasi, *Naïve Bayes*, *Extreme Gradient Boosting*, *Backpropagation Neural Network*, Jenis data, *Scalling*, *Missing Value*, Python.

ABSTRACT

Data mining is a technique of retrieving information from big data in the form of business transactions, scientific data, images, videos, and other data to obtain new information. One of the uses of data mining is classification. Classification aims to group data based on the ties between variables in the data. This study aims to compare the use of Naïve Bayes classification methods, Extreme Gradient Boosting, and Backpropagation Neural Networks on several types of data. This study has four stages: preprocessing, model design, model implementation, and analysis of classification results. The preprocessing stage comprises four stages: encoding, handling missing values, splitting train data with tests, and scaling. First, encoding is done on data attributes of the categorical type. Then, in handling missing values, there are five methods consisting of filling with mean, median, mode, deletion, and predictions with linear regression. Then, at the stage of split data, train and test with a weight of 80:20. Then there are four scaling scenarios: without, standard, robust, and minmax scaling in designing the classification model using python libraries such as sklearn for Naïve Bayes classification and Extreme Gradient Boost. Then Tensorflow for Backpropagation Neural Network classification. The implementation of the classification model is a combination of handling missing values, scaling methods, and classification methods resulting in 72 scenarios. The analysis and conclusion stages compare the scenario results from 5 classification characteristics, namely Precision, Recall, Accuracy, F1-Score, and Error Rate, on two different types of datasets. Then from the results of scenario testing, it was found that Backpropagation Neural Network became the best classification method on univariate and variate datasets. Classification in the NBA dataset is useful for players upgrading skills to be able to play for up to 5 years and for teams during player recruitment. Then the Car Insurance dataset is useful for drivers to maintain their driving behavior and vehicle condition. Lastly, insurance companies reduce losses.

Keywords : Data mining, Classification, Naïve Bayes, Extreme Gradient Boosting, Backpropagation Neural Network, Types of data, Scaling, Missing Value, Python