BAB II
TINJAUAN PUSTAKA

II.1. Tinjauan Pustaka

Data mining telah mendapatkan begitu besar perhatian pada dekade terakhir sehubungan dengan perkembangan *hardware* yang menyediakan kemampuan komputasi luar biasa yang memungkinkan pengolahan data besar (Ronsen, 2012). Mempersiapkan data adalah tahapan *preprocessing* yang sangat penting pada data mining, alasan utamanya adalah karena kualitas dari *input* data sangat mempengaruhi kualitas *output* analisis yang dihasilkan. Ada 7 (tujuh) tahapan proses data mining, dimana 4 (empat) tahap pertama disebut juga dengan data preprocessing (terdiri dari *data cleaning*, *data integration*, *data selection*, dan data transformation), yang dalam implementasinya membutuhkan waktu sekitar 60% dari keseluruhan proses (Junaedi et al., 2011).

Preprocessing data merupakan langkah penting dalam proses penemuan pengetahuan, karena keputusan-keputusan yang berkualitas harus didasarkan pada data yang berkualitas (Kumar & Chadha, 2012). *Preprocessing* data sering kali digunakan untuk mengurangi kesalahan data dan sistematis bias dalam data mentah sebelum analisis apapun terjadi (Tong et al., 2011). Pada penelitian tentang Data Preprocessing for Supervised Leaning (Kotsiantis et al., 2006) banyak faktor yang mempengaruhi keberhasilan Machine Learning (ML), yang pertama dan utama adalah representasi dan kualitas data, jika banyak data yang tidak relevan, terdapat *noise*, redudansi data dan data yang tidak handal maka

penemuan pengetahuan selama fase pelatihan lebih sulit. Dengan demikian, *preprocessing data* merupakan langkah yang penting dalam proses Machine Learning. Dengan algoritma *Future Subset Selection* maka dapat mengidentifikasi dan menghapus fitur yang tidak relevan dan *redundant*, hal ini akan mengurangi dimensi dari data dan memungkinkan algoritma pembelajaran untuk beroperasi lebih cepat dan efektif. Lalu pada penelitian tentang *Preprocessing of various Data Using Different Classification Algorihms for Evolutionary Programming* (Karthick & Malathi, 2015), dengan *preprocessing* maka dapat membantu untuk meningkatkan efisiensi data dan menghapus *noise* data yang membantu untuk mengidentifikasi *survival of the fittest.*

Kemudian pada penelitian yang dilakukan oleh Nawi et al. (2013) tentang menganalisis manfaat pada *preprocessing* data menggunakan teknik yang berbeda-beda dalam rangka meningkatkan konvergensi ANN yaitu *preprocessing* merupakan langkah penting dalam proses data mining, kualitas, keandalan, dan ketersediaan adalah beberapa faktor yang dapat menyebabkan kesuksesan interpretasi data di *neural network.* Dengan mengolah dataset dari UCI repository yaitu Wine, Iris dan Haberman dengan teknik preprocesing Min-Max Normalization, Z-Score Normalization dan Decimal Scaling Normalization maka hasil penelitian menunjukkan bahwa menggunakan teknik preprocessing dapat meningkatkan keakuratan klasifikasi ANN setidaknya lebih dari 95% sehingga dapat meningkatkan kekuatan dari algoritma yang diusulkan.

Penelitian selanjutnya adalah tentang *Accuracy of Pima Indian Diabetes Data Set Using Higher Order Neural Network and Principal Component Analysis(PCA)* (Anand et al., 2013), yaitu data preprocessing diperlukan untuk meningkatkan akurasi prediksi. Masalah *missing data* menjadi sulit dalam menganalisa dan dalam proses pengambilan keputusan sehingga *missing data* diganti sebelum menerapkan model Neural Network. Tanpa preprocessing ini proses pelatihan dalam neural network menjadi sangat lambat. Preprocessing juga berskala dalam rentang nilai data yang sama untuk setiap fitur input meminimalkan bias dalam jaringan saraf dalam satu fitur dan fitur yang lainnya. Dengan menggunakan *Principal Component Analysis(PCA)* maka *mean square error* menjadi menurun, dan *convergence* menjadi lebih cepat.

Iterative Partitioning Filter adalah salah satu algoritma *noisy data filtering* pada *preprocesing* data, tujuan dari algoritma ini adalah untuk menghilangkan *noise instance* dari data pelatihan. Algoritma ini ditemukan oleh T. M. Khoshgoftaar yaitu seorang profesor dari department Computer Science and Engineering Florida Atlantic University dan P. Rebours Magister di computer engineering Florida Atlantic University, Boca Raton, FL, USA pada tahun 2007. Algoritma ini memiliki empat buah parameter yaitu, *numPartitions* adalah jumlah partisi dari *training set, filterType* adalah skema penyaringan data penelitian, terdapat dua buah skema yaitu *consensus* dan *majority, confidence* adalah nilai *confidence* dari algoritma C4.5, dan itemsetsPerLeaf adalah jumlah minimum

itemsetperleaf dari algoritma C4.5. Ketika iterative partitioning filter menggunakan skema majority maka sebuah *instance* diberi label sebagai *noise* jika kesalahan klasifikasi lebih dari 50% dari model, dan bila menggunakan skema consensus maka menghapus sebuah *instance* jika prediksi dari semua n model berbeda dari label kelas yang sebenarnya dari instance. Dalam rangka untuk mengidentifikasi sebuah instance sebagai noise, sebuah instance harus terjadi kesalahan klasifikasi setidaknya oleh model yang mana terimbas pada model yang berisi *instance* itu (Khoshgoftaar & Rebours, 2007).