ABSTRAK

Diabetes atau sering disebut sebagai penyakit kencing manis merupakan suatu penyakit akibat kelainan metabolik yang diakibatkan oleh tingginya kadar glukosa darah di tubuh dalam waktu yang lama. Jika tidak ditangani lebih awal, diabetes bisa menyebabkan terjadinya komplikasi terhadap penyakit lain. *International Diabetes Federation* (IDF) memperkirakan sedikitnya terdapat 463 juta jiwa di seluruh dunia pada usia rentang 20-79 tahun menderita penyakit diabetes pada tahun 2019. Angka tersebut diprediksi akan terus mengalami peningkatan hingga mencapai 578 juta jiwa pada tahun 2030 dan 700 juta jiwa pada tahun 2045. Negara Indonesia sendiri berada di urutan ke-7 dari total 10 negara dengan jumlah penderita diabetes terbanyak, yaitu sebesar 10,7 juta. Jumlah tersebut diprediksi akan terus mengalami kenaikan dengan jumlah 13,7 juta jiwa pada tahun 2030 dan 16,6 juta jiwa pada tahun 2045. Banyak orang terdiagnosis penyakit diabetes setelah mengalami komplikasi. Oleh karena itu, para praktisi dan peneliti memusatkan perhatiannya dalam mendeteksi penyakit diabetes dengan penggunaan teknik *data mining* dalam menggali informasi berharga dari kumpulan data penyakit diabetes. *Dataset* yang digunakan pada penelitian ini adalah *dataset* yang berasal dari *National Institute of Diabetes and Digestive and Kidney Diseases* dan dapat diakses di UCI *Machine Learning Repository: Pima Indians Diabetes Database*. *Dataset* ini berisikan 768 pasien wanita dengan 8 atribut diagnosis kondisi medis yang berbeda dan 1 atribut tujuan atau atribut label. Penelitian ini membandingkan algoritma *K-Nearest Neighbor* dan *Logistic Regression* untuk klasifikasi data *Pima Indians Diabetes Database*. Pada penelitian ini, penulis melakukan penanganan *missing value* terhadap data dan menggunakan metode *Grid Search* untuk menemukan model dengan hasil akurasi yang optimal. Hasil akurasi dievaluasi dengan menggunakan *confusion matrix* dan menghitung nilai AUC. Dari hasil klasifikasi yang dilakukan, diperoleh hasil algoritma *K-Nearest Neighbor* dengan nilai akurasi sebesar 85,06% dan algoritma *Logistic Regression* dengan akurasi sebesar 77,92%.

Kata kunci—***diabetes, data mining, klasifikasi, k-nearest neighbor, logistic regression***