**ANALISA PENERAPAN METODE MACHINE LEARNING PADA DIAGNOSA PENYAKIT JANTUNG**

**Lalu Yasril Imam 1, Billy Aprillio 2 & Yufis Azhar 3.**

1,2,3 Program Studi Teknik Informatika Universitas Muhammadiyah Malang

email: yasrilimam98@gmail.com 1, email: aprillionilly2@gmail.com2, email: yufis@umm.ac.id 3

(Received: x x x/ Accepted: x x x / Published Online: x x x)

**Abstrak**

Jantung merupakan organ paling penting dalam tubuh manusia yang memiliki fungsi utama untuk memompa darah ke seluruh bagian tubuh melalui pembuluh darah. World Health Organization (WHO) menyebutkan bahwa penyakit jantung koroner (PJK) menjadi salah satu masalah kesehatan dalam system kardiovaskular yang jumlahnya meningkat cepat dengan angka kematian 6,7 juta kasus pada tahun 2017. Tujuan paper ini adalah membandingkan tingkat akurasi dalam mendeteksi penyakit jantung akibat gangguan kardiovaskular. Paper ini menggunakan metode logistic regression, naive bayes dan k-nearest neighbour (KNN) untuk pendekatan sistematis. Pada penggunaan dataset penyakit jantung didapatkan klasifikasi k-nearest neighbour memiliki tingkat akurasi paling tinggi dibandingkan algoritma klasifikasi naive bayes dan regresi logistik yaitu mencapai tingkat akurasi sebesar 91% sedangkan algoritma naive bayes memperoleh tingkat akurasi 83% dan algoritma regresi logistik tingkat akurasi 85%.

**Kata kunci:** Penyakit Jatung, Regresi Logistik, Naive Bayes, k-nearest neighbor

***Abstract***

*The heart is the most important organ in the human body that has the main function to pump blood to all parts of the body through blood vessels. The World Health Organization (WHO) states that coronary heart disease (CHD) is one of the health problems in the cardiovascular system which is increasing rapidly with a mortality rate of 6.7 million cases in 2017. The purpose of this paper is to compare the level of accuracy in detecting heart disease due to cardiovascular disorders. This paper uses logistic regression, naive bayes and k-nearest neighbor (KNN) methods for a sistematic approach. In the use of the heart disease dataset, the k-nearest neighbor classification has the highest level of accuracy compared to the Naive Bayes classification algorithm and logistic regression, which reaches an accuracy level of 91% while the Naive Bayes algorithm obtains an 83% accuracy rate and a logistic regression algorithm with an 85% accuracy rate.*

***Keywords:*** *Heart Disease, Logistic Regression, Naive Bayes, k-nearest neighbor*

**PENDAHULUAN**

Jantung merupakan salah satu organ penting dalam tubuh manusia yang memiliki fungsi utama untuk memompa darah ke seluruh bagian tubuh melalui pembuluh darah. Apabila pembuluh darah mengalami suatu penyempitan, maka akan terjadi gangguan pada fungsi jantung sehingga menyebabkan penyakit jantung. Penyakit ini merupakan penyakit yang tidak menular dan dapat menyebabkan kematian. Kematian dini akibat penyakit jantung sekitar 4% terjadi di Negara berpenghasilan tinggi dan 42% terjadi di Negara berpenghasilan rendah. Dapat diperkirakan pada tahun 2030, angka kematian akibat penyakit jantung akan terus meningkat setiap tahunnya dan mencapai 23,3 juta penduduk. (Kementrian Kesehatan RI, 2014).

Menurut *World Health Organization* (WHO) setiap tahunnya penyakit jantung koroner (PJK) menjadi salah satu masalah kesehatan dalam *system kardiovaskular* yang jumlahnya meningkat cepat dengan angka kematian 6,7 juta kasus pada tahun 2017. Berdasarkan hasil laporan *World Health Organization* (WHO) angka kematian penyakit kardiovaskuler akan mengalami kenaikan sekitar 25% angka kematian pada tahun 2020 dan mengalami peningkatan khususnya di negara-negara berkembang, salah satu diantaranya berada di Asia Tenggara. Angka kematian pada tahun 2014 mencapai 1,8 juta kasus yang disebabkan oleh PJK, yang artinya PJK menjadi penyakit yang mematikan di kawasan Asia Tenggara salah satu negaranya adalah Indonesia.

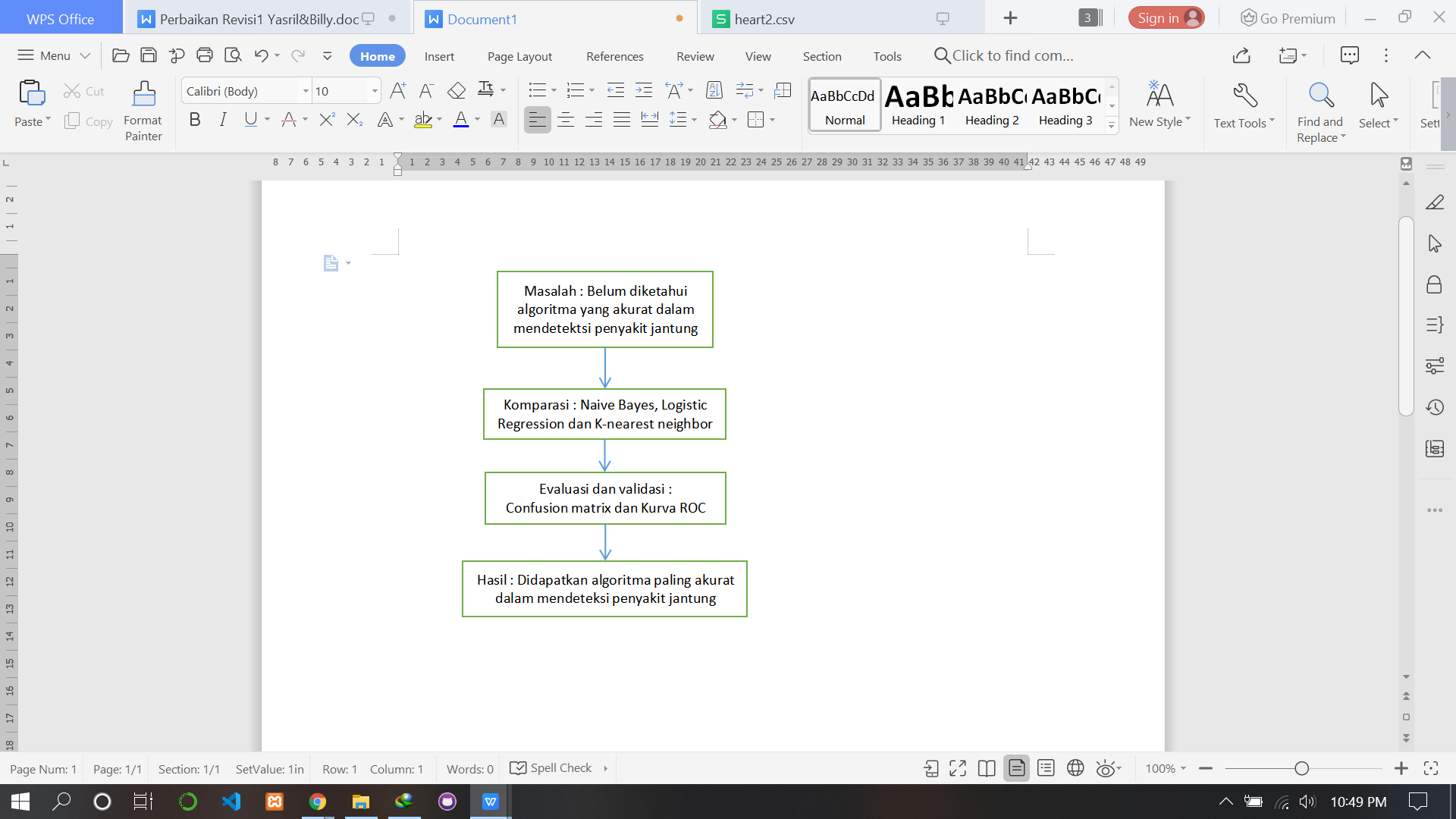
Data mining adalah proses penggalian sebuah data yang memanfaatkan suatu metode untuk memperoleh pola dari suatu data. Pendekatan yang dapat digunakan untuk menganalisis atau mengamati sekumpulan data adalah klasifikasi. Klasifikasi merupakan salah satu teknik data mining yang digunakan untuk mengelompokan sebuah data untuk membangun suatu model dari sampel data yang belum terklasifikasi untuk digunakan mengklasifikasi sampel data baru ke dalam kelas-kelas yang sejenis. Klasifikasi termasuk ke dalam supervised learning seolah-olah sekumpulan data akan dianalisis terlebih dahulu, kemudian pola dari hasil analisis tersebut digunakan untuk pengklasifikasian data uji. Terdapat banyak algoritma yang sering digunakan, namun yang populer dan sering digunakan diantaranya yaitu Naive Bayes, K-nearest neighbour dan Logistic Regression.

Algoritma K-Nearest Neighbor (KNN) adalah salah satu algoritma yang berfungsi untuk melakukan klasifikasi berdasarkan data pembelajarannya (train data sets)[1]. Salah satu masalah dari algoritma KNN adalah jarak antara record baru dan record yang tersedia yang memiliki semua atribut dalam sebuah dataset training menyebabkan proses klasifikasi yang tidak baik dan menurunkan akurasi algoritma klasifikasi. Hasil dari pendekatan ini adalah menggunakan atribut-atribut yang berbeda ketika menghitung jarak antara dua record data dan dibandingkan dengan menggunakan algoritma klasifikasi Naïve Bayes yang meghasilkan akurasi sebesar 83.707% dengan menggunakan dataset heart disease. Algoritma yang dihasil mengunakan teorema Bayes mengasumsikan semua atribut independen atau tidak saling ketergantungan pada setiap nilai variable kelas[2]. Algoritma Naïve Bayes merupakan sebuah metode klasifikasi dengan menggunakan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes dengan memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya[3]. Regresi logistik adalah model prediksi yang menghubungkan antara satu atau beberapa variabel independen (variabel bebas) dengan variabel dependen yang berupa kategori; biasanya 0 dan 1[4].

Dalam penelitian ini, diusulkan menggunakan 3 algoritma klasifikasi yaitu K-nearest neighbor, Naive Bayes dan Logistic Regression. Ketiga metode tersebut akan dibandingkan berdasarkan tingkat akurasinya dalam pengambilan keputusan yang paling optimal untuk klarifikasi penyakit jantung.

**METODE**

Penelitian ini merupakan penelitian eksperimen yang bertujuan untuk melakukan perbandingan dan evaluasi pada algoritma-algoritma klasifikasi data mining pada data penyakit jantung.



Gambar 1. Kerangka pemikiran

Pada gambar 1 diatas menunjukkan kerangka pemikiran dari penelitian ini. Analisa permasalahan (problem) terkait algoritma dalam mendeteksi penyakit jantung. Dilanjutkan dibuat menggunakan model algoritma Logistic Regression, Naive Bayes dan KNN Pengujian evaluasi dan validasi untuk mengukur akurasi menggunakan Confucion matrix dan Kurva ROC, serta hasil dari penelitian didapat di antara ketiga algoritma, didapat algoritma yang terbaik dalam prediksi penyakit jantung.

Data penelitian diambil dari, <http://archive.ics.uci.edu/ml/> dan meggunakan sebanyak 1025 record. Dataset ini sebenarnya memiliki 76 atribut atau kolom, tetapi pada percobaan yang telah di terbitkan hanya merujuk pada penggunaan 14 atribut[5]. Tabel dibawah ini menunjukkan potongan dataset.

Tabel 1. Dataset heart disease

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| age | sex | cp | tresbps | chol | Fbs | restecg | thalach | target |
| 52 | 1 | 0 | 125 | 212 | 0 | 1 | 168 | 0 |
| 53 | 1 | 0 | 140 | 203 | 1 | 0 | 155 | 0 |
| 34 | 0 | 1 | 118 | 210 | 0 | 1 | 192 | 1 |
| 51 | 0 | 2 | 140 | 308 | 0 | 0 | 142 | 1 |
| 58 | 1 | 2 | 140 | 211 | 1 | 0 | 165 | 1 |
| 60 | 1 | 2 | 140 | 185 | 0 | 0 | 155 | 0 |
| 67 | 0 | 0 | 106 | 223 | 0 | 1 | 142 | 1 |

Klasifikasi menentukan suatu class atau kelompok untuk tiap contoh data. Input dari suatu model klasifikasi adalah atribut dari data sample dan outputnya adalah class dari data samples itu sendiri, dalam machine learning untuk membangun model klasifikasi digunakan metode supervised learning[6]. Supervised learning adalah metode yang digunakan dalam menemukan hubungan antara atribut masukan dan atribut target, hubungan yang ditemukan diwakili suatu struktur yang disebut model.

Dalam klasifikasi kita dapat mengelompokkan sesuatu kedalam suatu kategori tertentu, contoh untuk masalah klasifikasi adalah mengelompokkan apakah seseorang pasien “mengidap” atau “tidak mengidap” penyakit tertentu. Informasi tentang pasien sebelumnya digunakan sebagai data latih algoritma untuk mendapatkan aturan. Salah satu tujuan klasifikasi adalah untuk meningkatkan kehandalan hasil yang diperoleh dari data[7].

**Logistic Regression**

Regresi logistik merupakan model regresi yang menunjukkan pengaruh dari variabel prediktor, baik berupa kontinu maupun kategorik, terhadap variabel respon berupa data kategorik. Pada regresi logistik biner, variabel respon terdiri dari dua kategori (biner) yaitu 0 dan 1. Dimana variabel respon untuk setiap observasi mengikuti distribusi Bernoulli. Model Regresi Logistik sebagai berikut:

(1)

Keterangan :

: persamaan yang biasa dikenal dalam OLS (Ordinary Least Squares)

: banyaknya variabel predicator

**Naive Bayes**

Naive Bayes merupakan metode pengelompokkan berdasarkan probabilitas sederhana yang dirancang untuk digunakan dengan asumsi antara suatu kelas dengan kelas lainnya tidak saling tergantung (independen). Pada naïve bayes, proses pembelajaran algoritma lebih ditekankan pada estimasi probabilitas. Keuntungan pada pendekatan ini yaitu pengelompokkan akan mendapatkan nilai error yang lebih kecil ketika menggunakan dataset berukuran besar[8]. Selain itu klasifikasi Naïve Bayes terbukti memiliki akurasi dan kecepatan yang tinggi saat digunakan ke dalam basis data dengan jumlah yang besar[9]. Formulasi Naïve Bayes untuk klasifikasi adalah sebagai berikut[10]:

(2)

Keterangan :

: probabilitas data dengan vektor pada kelas

: probabilitas awal pada kelas (prior probability)

: probabilitas independen kelas dari semua fitur dalam vector

: probabilitas dari

**K-Nearest Neighbor**

Algoritma K-Nearest Neighbor (K-NN) adalah metode pengelompokan terhadap objek berdasarkan data pembelajaran yang memiliki jarak paling dekat dengan objek tersebut. K-Nearest Neighbor (K-NN) merupakan salah satu metode pengklasifikasian data berdasarkan similaritas dengan label data[11]. Similaritas menggunakan matriks jarak dimana satuan jaraknya menggunakan Euclidean:

(3)

Keterangan :

: Jarak antara data uji dan data latih

: data uji

: data latih

: jumlah data latih

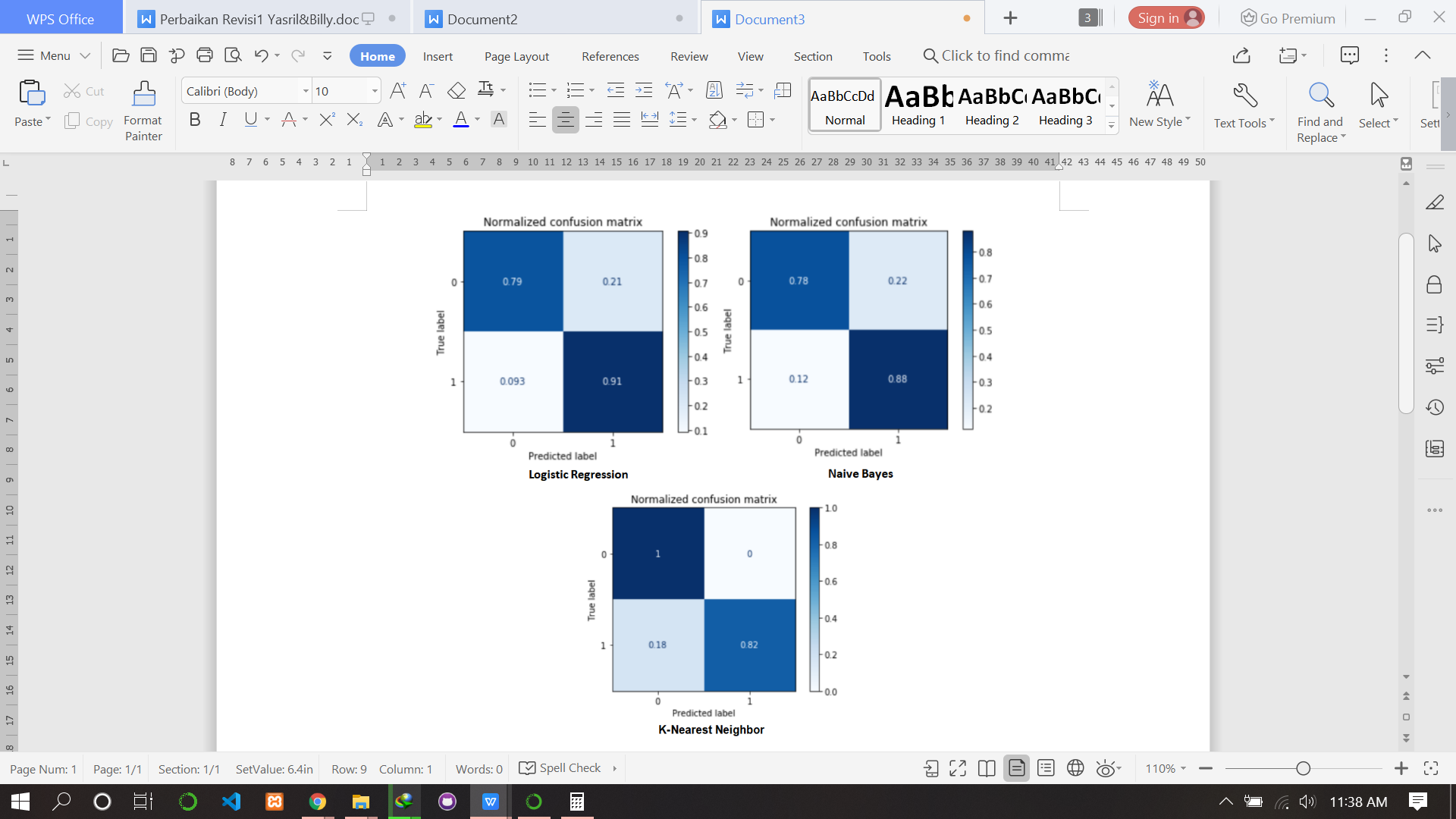
**HASIL DAN PEMBAHASAN**

Pada penelitian ini digunakan 1025 record data orang yang terjangkit. 80% atau 820 record data digunakan sebagai data latih (training data) dan 20% atau 205 record data digunakan sebagai data uji (testing data). Hasil pengujian dan evaluasi didapatkan perbandingan hasil peforma dari masing-masing algoritma yang dapat dijabarkan sebagai berikut.

Tabel 2. Hasil Komparasi Kinerja Algoritma

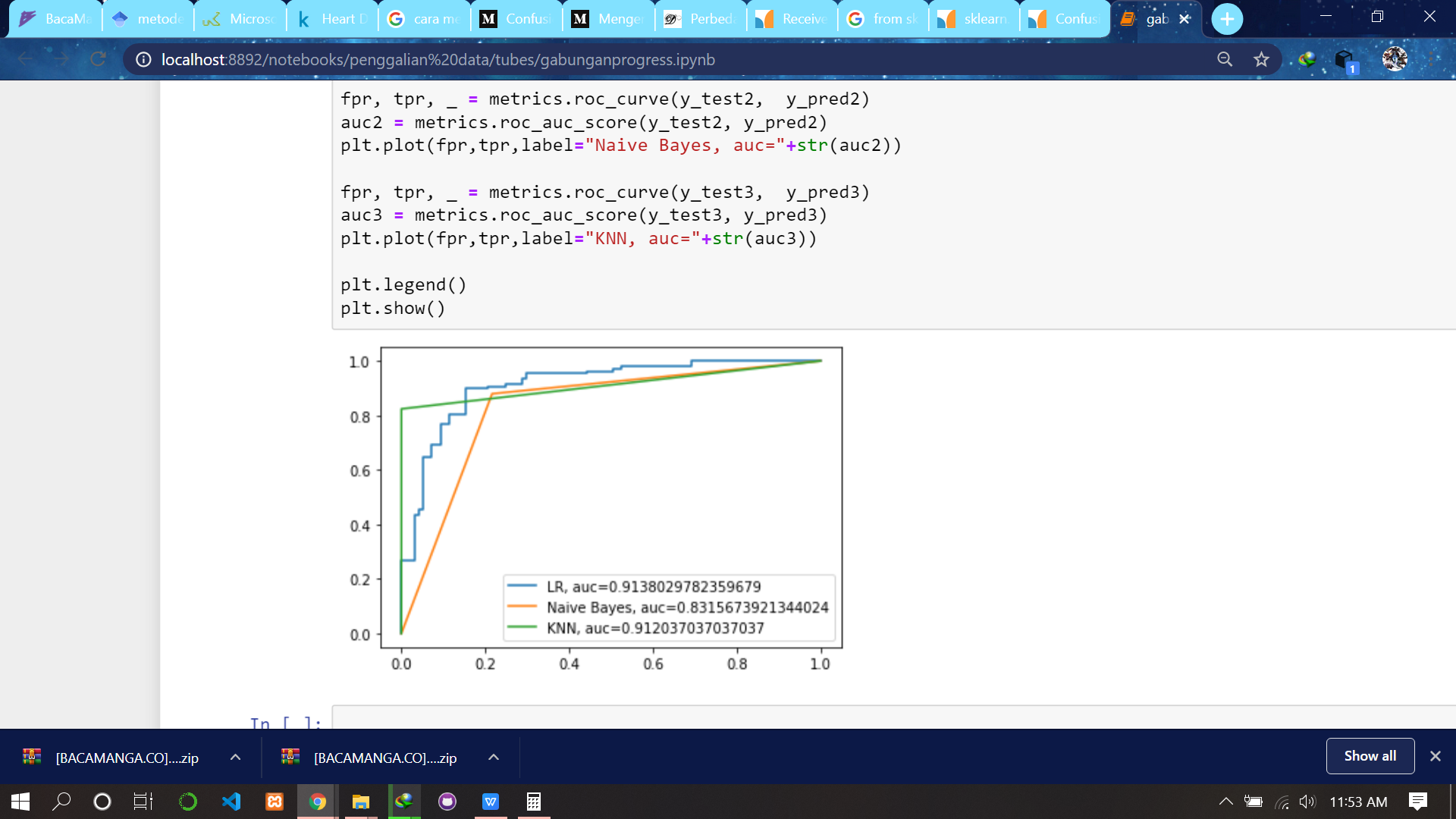
|  |  |  |  |
| --- | --- | --- | --- |
| Parameter Kinerja | Logistic Regression | K-Nearest Neighbor | Naïve Bayes |
| *Accuracy* | 0.85 | 0.91 | 0.83 |
| *Precision* | 0.83 | 1 | 0.82 |
| *Recall* | 0.91 | 0.82 | 0.88 |
| *AUC* | 0.91 | 0.91 | 0.83 |

Dari tabel confusion matrix yang dihasikan, terbukti bahwa hasil akurasi tertinggi pada perbandingan hasil kinerja klasifikasi didapatkan oleh algoritma KNNsebesar 0.90731 dengan nilai K sebanyak 2 sesuai dengan jumlah nilai target. Kemudian algoritma Logistic Regression sebesar 0.84878 dan yang terakhir adalah naïve bayes dengan tingkat akurasi terendah sebesar 0.83414.



Gambar 2. Normalized confusion matrix

Gambar 2 menunjukkan hasil dari confusion matrix yang telah di normalisasi pada masing-masing algoritma. Proses pengujian metode dimulai dari evaluasi menggunakan *Area Under Curve* (AUC) untuk mengukur hasil akurasi dari ketiga algoritma yang memiliki tingkat yang berbeda antara *Logistic Regression*, KNN dan *Naïve bayes.* Terjadi peningkatan pada algoritma LogisticRegression dan KNN namun tidak pada algoritma *Naïve Bayes.* Gambar 3 menunjukkan nilai AUC pada ketiga algoritma.



Gambar 3. Nilai AUC ketiga algoritma

Secara keseluran dari ketiga algoritma tersebut didapat bahwa algoritma *Logistic Regression* dan KNN berdasarkan nilai AUC masuk pada predikat *Excelent Classification* sedangkan *Naïve bayes* masuk dalam predikat *Good Classification.*

Tabel 3. Nilai AUC dan Keterangan

|  |  |
| --- | --- |
| Nilai AUC | Klasifikasi |
| 0.90 – 1.00 | *Excellent classification* |
| 0.80 – 0.90 | *Good classification* |
| 0.70 – 0.80 | *Fair classification* |
| 0.60 – 0.70 | *Poor classification* |
| 0.50 – 0.60 | *failure* |

**SIMPULAN**

Kesimpulan yang diperoleh pada penelitian ini adalah hasil perbandingan alogoritma klasifikasi K-nearest neighbour, naive bayes dan regresi logistik yang menggunakan dataset *heart disease* menyatakan bahwa algoritma klasifikasi k-nearest neighbour merupakan algoritma klasifikasi yang memiliki tingkat akurasi paling tinggi dibandingkan algoritma klasifikasi naive bayes dan regresi logistik yaitu mencapai tingkat akurasi sebesar 91% sedangkan algoritma naive bayes memperoleh tingkat akurasi 83% dan algoritma regresi logistik tingkat akurasi 85%.

**REFERENSI**

[1] M. Moradian and A. Baraani, “Knnba : K-Nearest-Neighbor-Based-Association Algorithm,” *Theor. Appl. Inf. Technol.*, 2009.

[2] S. Bahri, D. Marisa Midyanti, R. Hidayati, J. Sistem Komputer, and F. Mipa, “Perbandingan Algoritma Naive Bayes dan C4.5 Untuk Klasifikasi Penyakit Anak,” *Semin. Nas. Apl. Teknol. Inf.*, pp. 24–31, 2018.

[3] D. Teknik, I. Universitas, N. Bayes, N. Bayes, and N. Bayes, “Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi,” vol. 8, no. 1, 2014, doi: 10.26555/jifo.v8i1.a2086.

[4] T. Purwa, “Perbandingan Metode Regresi Logistik dan Random Forest untuk Klasifikasi Data Imbalanced (Studi Kasus: Klasifikasi Rumah Tangga Miskin di Kabupaten Karangasem, Bali Tahun 2017),” *J. Mat. Stat. dan Komputasi*, vol. 16, no. 1, p. 58, 2019, doi: 10.20956/jmsk.v16i1.6494.

[5] K. Baati, T. M. Hamdani, and A. M. Alimi, “A modified hybrid naïve possibilistic Classifier for heart disease detection from heterogeneous medical data,” *6th Int. Conf. Soft Comput. Pattern Recognition, SoCPaR 2014*, pp. 353–358, 2014, doi: 10.1109/SOCPAR.2014.7008032.

[6] H. H. Hsu, “Advanced data mining technologies in bioinformatics,” *Adv. Data Min. Technol. Bioinforma.*, pp. 1–329, 2006, doi: 10.4018/978-1-59140-863-5.

[7] H. Kahramanli and N. Allahverdi, “Design of a hybrid system for the diabetes and heart diseases,” *Expert Syst. Appl.*, vol. 35, no. 1–2, pp. 82–89, 2008, doi: 10.1016/j.eswa.2007.06.004.

[8] M. A. jabbar, B. L. Deekshatulu, and P. Chandra, “Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm,” *Procedia Technol.*, vol. 10, pp. 85–94, 2013, doi: 10.1016/j.protcy.2013.12.340.

[9] J. Han, M. Kamber, and J. Pei, “Data Mining: Concepts and Techniques,” *Data Min. Concepts Tech.*, 2012, doi: 10.1016/C2009-0-61819-5.

[10] E. Prasetyowati and N. Ramadhani, “Sistem Evaluasi Dan Klasifikasi Kinerja Akademik Mahasiswa Universitas Madura Menggunakan Naive Bayes Dengan Dirichlet Smoothing,” *JUTI J. Ilm. Teknol. Inf.*, vol. 16, no. 2, p. 192, 2018, doi: 10.12962/j24068535.v16i2.a688.

[11] D. T. Larose, *Data Mining Methods and Models*. 2006.