Jurnal Nasional Komputasi dan Teknologi Informasi
Vol. 3 No. 1, April 2020
P-ISSN 2620-8342
E-ISSN 2621-3052

Prediksi Nilai Calon Mahasiswa dengan Algoritma *Backpropagation* (Studi Kasus: Data Kaggle)

Rizki Ardhian Ahmad[1*], Nur Nafi'iyah[1], Siti Mujilahwati[3]
[1,2,3] Prodi Teknik Informatika Fakultas Teknik Universitas Islam Lamongan
Jalan Veteran Nomor 53A Lamongan 62211
e-mail: [1]rizkiardhianahmad@gmail.com, [2]mynaff26@gmail.com

Abstrak— Mahasiswa yang akan melakukan pendaftaran ke perguruan tinggi, baik ke jenjang sarjana atau pascasarjana pasti harus diseleksi. Proses seleksi dengan tes dan serangkaian kegiatan lainnya. Nilai-nilai tes tersebut kemudian dianalisa untuk mengetahui apakah mahasiswa layak diterima atau tidak. Beberapa perguruan tinggi di Amerika Serikat atau Inggris melakukan serangkaian tes mulai tes akademik, tes bahasa Inggris dan kemampuan meneliti. Dari beberapa data hasil seleksi atau ujian dapat digunakan untuk memprediksi calon mahasiswa baru yang akan masuk perguruan tinggi. Tujuan penelitian ini adalah memprediksi nilai calan mahasiswa yang akan masuk di perguruan tinggi. Studi kasus ini mengambil dari data kaggle, yang akan diprediksi dengan menggunakan algoritma backpropagation. Variabel yang menjadi input adalah GRE score, TOEFL score, Universiy rating, SOP, LOR, GPA, Research. Output dari prediksi nilai calon mahasiswa dalam angka. Proses training backpropagation menggunakan toole Matlab dengan arsitektur jaringan 2 model. Model ke-1 menggunakan 7-5-1 dengan hasil MSE 0,00272. Model ke-2 menggunakan 7-4-1 dengan hasil MSE 0,0029.
Kata kunci: *prediksi nilai, backpropagation, kaggle, Matlab.*

1. Pendahuluan

Prediksi merupakan suatu peramalan atau forecasting satu data pada waktu yang akan datang. Proses prediksi bisa dilakukan jika terdapat dataset latih yang dapat dikenali polanya. Banyak algoritma yang digunakan untuk memprediksi suatu data atau kejadian yang akan datang. Melakukan peramalan dapat dilakukan pada data musiman, atau data yang setiap waktu berubah. Prediksi merupakan sebagian ilmu dari data mining, di mana data mining adalah menggali informasi atau pengetahuan dari kumpulan dataset beberapa kurun waktu yang lampau. Dari beberapa penelitian terkait prediksi dijabarkan sebagai berikut:

Saat algoritma Backpropagation dan Naïve Bayes dibandingkan dalam mengklasifikasi jenis kelamin manusia berdasarkan citra panoramik gigi, hasilnya menunjukkan bahwa algoritma backpropagation nilai akurasinya 85%, sedangkan Naïve Bayes hanya 80%[1]. Prediksi mengenai harga emas dengan menerapkan algoritma Fuzzy Mamdani, Regresi Linear, dan Backpropagation. Hasilnya dalam memprediksi harga emas, algoritma yang baik dengan nilai akurasinya tinggi adalah Regresi Linear, dan Backpropagation[2]. Di mana kedua penelitian [1][2] membandingkan algoritma untuk klasifikasi dan prediksi hasilnya backpropagation mempunyai nilai akurasi tinggi. Selain algoritma backpropagation dan regresi linear, naïve bayes juga memberikan nilai akurasi yang tinggi.

Selain algoritma Backpropagation, dan Naïve Bayes, algoritma C4.5 dan KNN juga digunakan sebagai prediksi atau klasifikasi kelulusan mahasiswa. Algoritma C4.5[3] juga digunakan untuk memprediksi tingkat kelulusan mahasiswa dengan output sistem adalah lulus cepat, lulus tepat, lulus terlambat, dan drop out. Hasil klasifikasi tingkat kelulusan mahasiswa dengan algoritma C4.5 adalah 87,5%[3]. Misalnya dalam penelitian memprediksi kelulusan mahasiswa [4][5][6][7]. Penelitian Jananto melakukan prediksi kelulusan mahasiswa dengan output sistem tepat waktu dan tidak tepat waktu, dengan dataset acuan sebanyak 254, dan hasil akurasinya 76%[6]. Melakukan prediksi dengan menggunakan metode Naïve Bayes membutuhkan dataset acuan. Sama halnya dengan melakukan prediksi dengan metode backpropagation. Karena algoritma backpropagation akan mengenali dataset training sebagai acuan menghitung bobot atau bias. Dari hasil bobot dan bias training digunakan sebagai perhitungan data ujicoba/testing. Algoritma naïve bayes juga sama, membutuhkan dataset acuan untuk menghitung nilai probabilitas dataset ujicoba/testing.

Algoritma C4.5 juga membutuhkan dataset training untuk mengenali pola, hasil training algoritma C4.5 adalah sebuah pohon/tree dan rule. Penelitian David Hartanto menggunakan dataset training 60 baris, dan 40 baris dataset testing. Contoh lagi mengenai prediksi tingkat kelulusan mahasiswa dengan metode KNN[8], di mana menggunakan dataset sebanyak 1718 baris untuk training dan testing. Proses prediksi kelulusan mahasiswa dengan metode KNN menghasilkan output tepat dan terlambat. Di mana setiap data input akan dihitung jaraknya terhadap dataset acuan. Sedangkan nilai k adalah nilai ketetanggan yang paling dekat dengan hasil output. Penelitian Abdul Rohman dalam menentukan k untuk uji coba memprediksi kelulusan mahasiswa dengan k=1 nilai akurasinya 82,25%, k=2 nilai akurasinya 79,45%, k=3 nilai akurasinya 83,95%, k=4 nilai akurasinya 82,62%, dan k=5 adalah k ketetanggaan yang akurasinya tertinggi, yakni 85,15%[8].

Dari beberapa penelitian yang diulas sebelumnya, penulis ingin melakukan prediksi nilai mahasiswa

berdasarkan dataset Kaggle, dengan variabel GRE score, TOEFL score, University Rating, SOP, LOR, GPA, Research, dan outputnya adalah Chance of Admit. Prediksi nilai mahasiswa dari data kaggle ini output berupa nilai angka, dengan algoritma backpropagation digunakan untuk memprediksi.

2. Tinjauan Pustaka

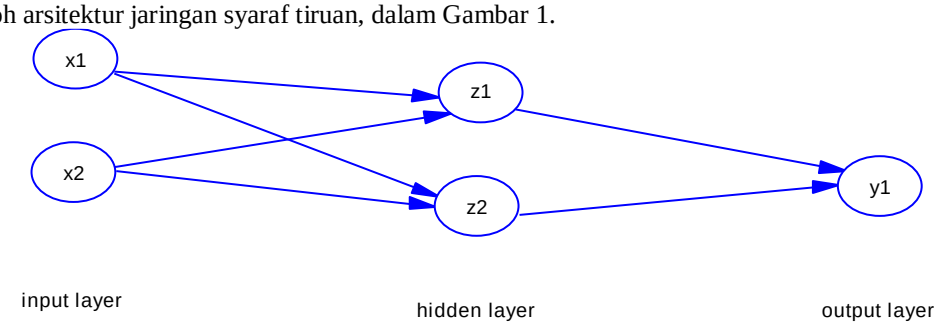
Jaringan syaraf tiruan suatu sistem yang mengolah suatu data dengan meniru jalan kerja otak manusia[11]. Di mana otak manusia terdapat aliran yang saling berhubungan yang akan memberikan informasi atau neuron. Jaringan syaraf tiruan sering digunakan untuk pengenalan pola atau klasifikasi data yang di mana membutuhkan proses learning atau training. Jaringan syaraf tiruan mempunyai kemampuan belajar yang diberikan dari data dengan data yang kompleks atau rumit. Hasil dari learning tersebut akan menghasilkan model atau bobot yang digunakan untuk uji coba pengenalan suatu data baru. Learning tersebut dilakukan agar bisa melakukan perubahan bobot dan menghasilkan model. Jaringan syaraf tiruan mempunyai beberapa layer yang di mana layar tersebut akan membaca data input serta menyusun hidden layer dari layar tersebut akan menghasilkan output. Jaringan syaraf tiruan mempunyai kelebihan diantaranya mampu mempelajari bagaimana belajar dari data yang diberikan atau melatih diridari data awal. Jaringan syaraf tiruan membuat suatu organisasi sendiri atau merepresentasikan diri sendiri dari informasi yang diterima saat belajar[9][10]. Berikut beberapa layer yang ada dalam jaringan syaraf tiruan:

a. Input layer: unit-unit dalam lapisan input disebut unit-unit input yang bertugas menerima pola inputan dari luar yang menggambarkan suatu permasalahan.
b. Hidden layer: unit-unit dalam lapisan tersembunyi disebut unit-unit tersembunyi, yang mana nilai outputnya tidak dapat diamati secara langsung.
c. Output layer: unit-unit dalam lapisan output disebut unit-unit output, yang merupakan solusi jaringan syaraf tiruan terhadap suatu permasalahan.

Berdasarkan model matematis, model jaringan syarf tiruan ditentukan oleh beberapa hal, diantaranya:

a. Arsitektur jaringan, yaitu sebuah arsitektur yang menentukan pola atau neuron
b. Model pembelajaran (*learning method*), yaitu metode yang digunakan untuk menentukan dan mengubah bobot
c. Fungsi aktivasi

Contoh arsitektur jaringan syaraf tiruan, dalam Gambar 1.



Gambar 1. Arsitektur Jaringan Syaraf Tiruan

Di mana input layer terdapat 2 node, hidden layer terdapat 2 node dan output layer 1 node. Adapun algoritma backpropagation adalah[9][10]:

a. Inisialisasi bobot
b. Menentukan nilai kondisi berhenti
c. Melakukan proses perhitungan dari input layer ke hidden layer, seperti Persamaan 1

$$z\_input_j = v_{0j} + \sum_{i=1}^{n} x_i v_{ij} \quad (1)$$

d. Melakukan aktivasi nilai dari node hidden layer, seperti Persamaan 2

$$z_j = f(z\_input_j) \quad (2)$$

e. Menghitung dari hidden layer ke output layer, seperti Persamaan 3

$$y\_input_k = w_{0k} + \sum_{j=1}^{p} z_j w_{jk} \quad (3)$$

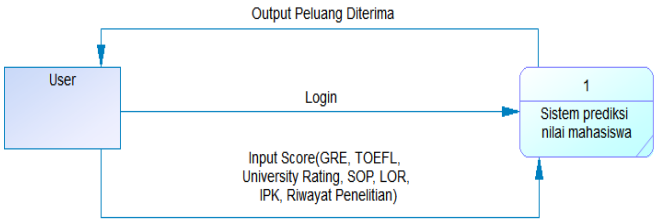
f. Melakukan aktivasi nilai dari node output layer, seperti Persamaan 4

$$y_k = f(y\_input_k) \quad (4)$$

g. Melakukan perhitungan nilai error antara hasil aktivasi dari node output layer terhadapat target, serta melakukan perambatan mundur
h. Melakukan koreksi nilai bobot di node output layer dan node hidden layer
i. Perbaikan bobot

3. Metode Penelitian

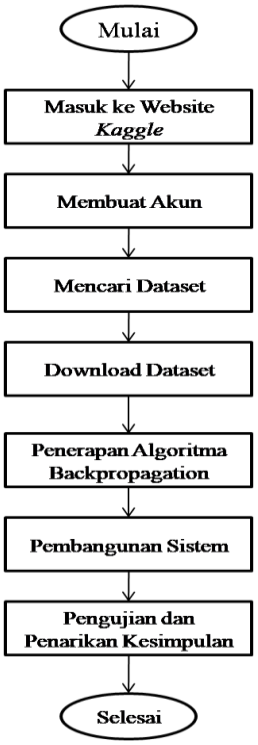
Alur dari sistem ini seperti dalam Gambar 2. Di mana pengguna memasukkan dataset training agar mendapatkan nilai bobot dan bias. Selanjutnya nilai bobot dan bias disimpan dalam tabel yang akan digunakan untuk testing data baru.



Gambar 2. Alur Sistem

Penelitian ini menggunakan dataset Kaggle untuk memprediksi nilai mahasiswa. Di mana data diambil dan diolah untuk mengambil variabel-variabel yang berkorelasi dengan nilai mahasiswa. Output dari sistem prediksi ini adalah nilai mahasiswa, dengan variabel input, yaitu: GRE score, TOEFL score, University Rating, SOP, LOR, GPA, dan Research. Alur memproses dataset seperti dalam Gambar 3. Penjelasan dari masing-masing variabel, adalah:

a. GRE score adalah Graduate Record Examination. Hasil nilai tes standar agar dapat masuk perguruan tinggi di Amerika Serikat atau Universitas di Negeri yang menggunakan bahasa Inggris.
b. TOEFL score (Test of English as a Foreign Language) adalah hasil ujian kemampuan bahasa Inggris.
c. University rating adalah peringkat Universitas (baik dari publikasi, akademik dan prestasi lainnya)
d. SOP (Statement of Purpose) adalah suatu pernyataan mahasiswa dan tujuan mahasiswa mendaftar di perguruan tinggi tersebut.
e. LOR (Letter of recommendation) adalah hasil penilaian dari dosen pembimbing atau dosen promotor terhadap mahasiswa. Baik dari segi kemampuan, karakteristik kepribadian dan akademiknya.
f. GPA (Grade Point Average) nilai IPK mahasiswa Indeks Prestasi Komulatif
g. Research adalah pengalaman dari penelitian atau hasil belajar meneliti dengan dosen mahasiswa.
h. Chance of Admit adalah output dari prediksi yang berupa hasil angka nilai mahasiswa



Gambar 3. Alur mengolah Dataset

Penelitian ini mempunyai 7 variabel input, dan dataset yang ada sebanyak 400, yang akan digunakan untuk training sebanyak 350 baris dan testing sebanyak 50. Contoh data training seperti Tabel 1. Dataset yang menjadi variabel input adalah angka, output dari prediksi juga angka. Algoritma yang digunakan untuk prediksi adalah algoritma backpropagation. Di mana angka terlebih dahulu dinormalisasi menjadi 0-1 dengan Persamaan 5. Variabel yang akan dilakukan normalisasi adalah GRE score, TOEFL score, University rating, SOP, LOR, GPA. Sedangkan variabel research tidak dilakukan normalisasi karena value sudah dalam rentang 0-1. Hasil normalisasi seperti Tabel 2.

$$N\_baru = \frac{(N\_lama - N\_Min\_lama)}{(N\_Maks\_lama - N\_Min\_lama)} * (N\_Maks\_baru - N\_Min\_baru) \quad (5)$$

Sedangkan arsitektur jaringan backpropagation yang digunakan menggunakan 2 model. Model ke-1: node input sebanyak 7, node hidden sebanyak 5, dan node output sebanyak 1. Model ke-2 node input sebanyak 7, node hidden sebanyak 4, dan node output sebanyak 1.

Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	GPA	Research	Chance of Admit
1	337	118	4	4.5	4.5	9.65	1	0.92
2	324	107	4	4	4.5	8.87	1	0.76
3	316	104	3	3	3.5	8	1	0.72
4	322	110	3	3.5	2.5	8.67	1	0.8
5	314	103	2	2	3	8.21	0	0.65
6	330	115	5	4.5	3	9.34	1	0.9
7	321	109	3	3	4	8.2	1	0.75
8	308	101	2	3	4	7.9	0	0.68
9	302	102	1	2	1.5	8	0	0.5
10	323	108	3	3.5	3	8.6	0	0.45

Tabel 1. Contoh Dataset Training

GRE Score	TOEFL Score	University Rating	SOP	LOR	GPA	Research	Chance of Admit
0.94	0.93	0.75	0.88	0.88	0.91	1.00	0.92

0.68	0.54	0.75	0.75	0.88	0.66	1.00	0.76
0.52	0.43	0.50	0.50	0.63	0.38	1.00	0.72
0.64	0.64	0.50	0.63	0.38	0.60	1.00	0.80
0.48	0.39	0.25	0.25	0.50	0.45	0.00	0.65
0.80	0.82	1.00	0.88	0.50	0.81	1.00	0.90
0.62	0.61	0.50	0.50	0.75	0.45	1.00	0.75
0.36	0.32	0.25	0.50	0.75	0.35	0.00	0.68
0.24	0.36	0.00	0.25	0.13	0.38	0.00	0.50
0.66	0.57	0.50	0.63	0.50	0.58	0.00	0.45

Tabel 2. Hasil Normalisasi Data

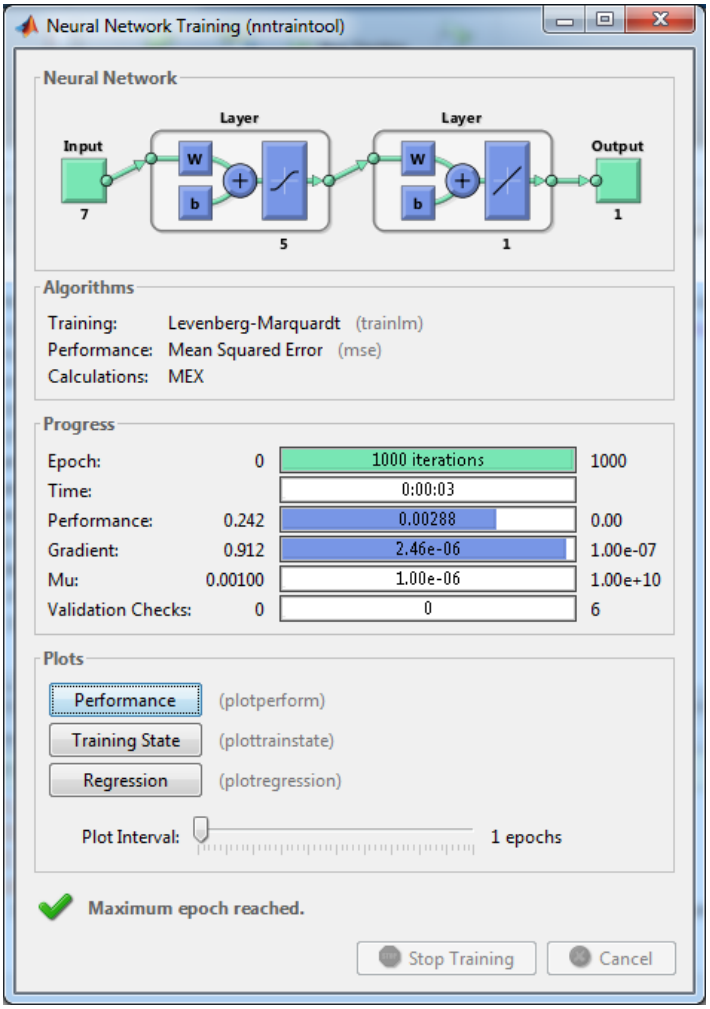
4. Hasil dan Pembahasan

Model ke-1 proses melakukan training menggunakan tool Matlab, Gambar 4 proses training. Model ke-1 menggunakan node input sebanyak 7, node hidden sebanyak 5, dan node output sebanyak 1.

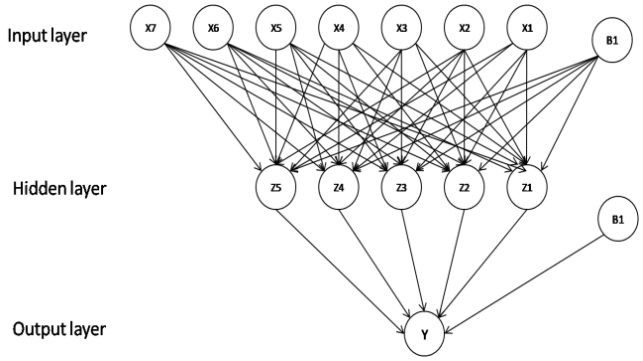
Hasil setiap training akan menghasilkan nilai bobot dan bias. Arsitektur jaringan untuk memprediksi nilai mahasiswa model ke-1 seperti Gambar 5. Model ke-2 proses melakukan training menggunakan tool Matlab, Gambar 6 proses training. Model ke-2 menggunakan node input sebanyak 7, node hidden sebanyak 4, dan node output sebanyak 1.

Training dilakukan dan testing dilakukan sebanyak 3 kali, tujuannya untuk mendapatkan nilai akurasi yang tinggi. Sistem ini digunakan untuk memprediksi nilai mahasiswa, dengan interface ujicoba seperti Gambar 7. Di mana terdapat 7 buah textbox untuk menginputkan nilai GRE score, TOEFL score, University rating, SOP, LOR, GPA, dan Research.

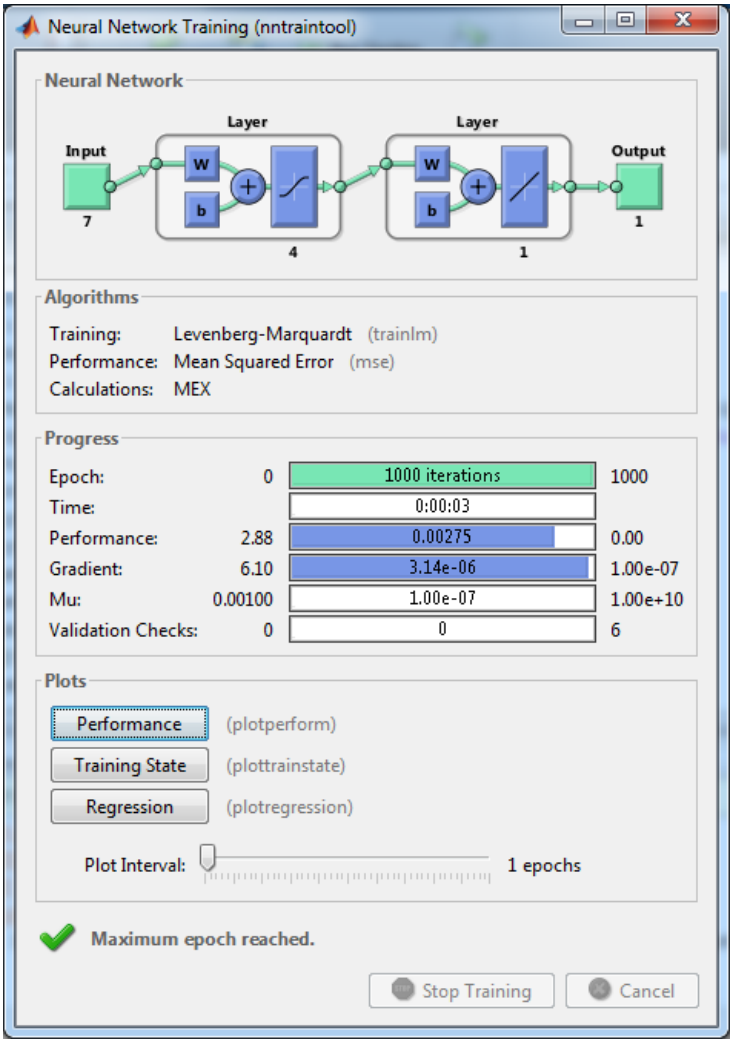
Di mana hasil training dengan model ke-1 mendapatkan hasil error 0,00272, seperti Gambar 8. Dan Gambar 9 hasil training model ke-2 dengan nilai error 0,0029.



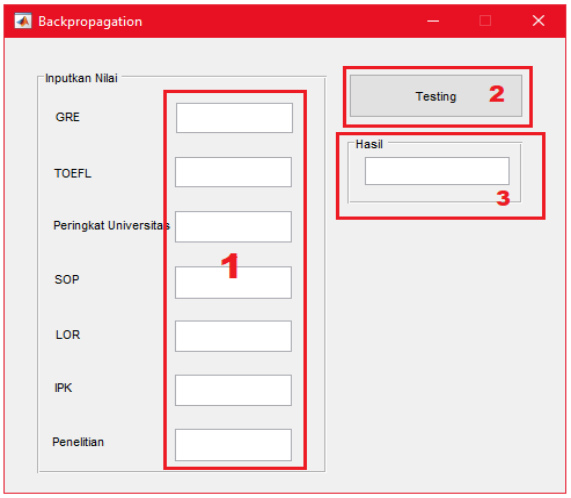
Gambar 4. Proses Training Model ke-1



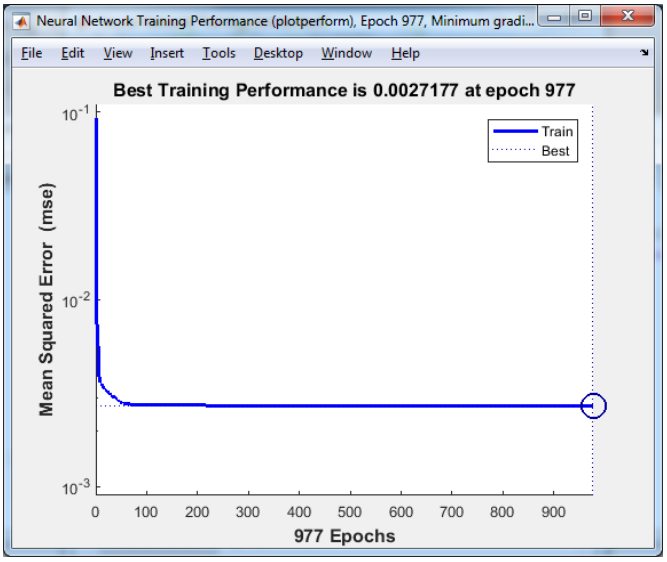
Gambar 5. Arsitektur Jaringan backpropagation Model ke-1



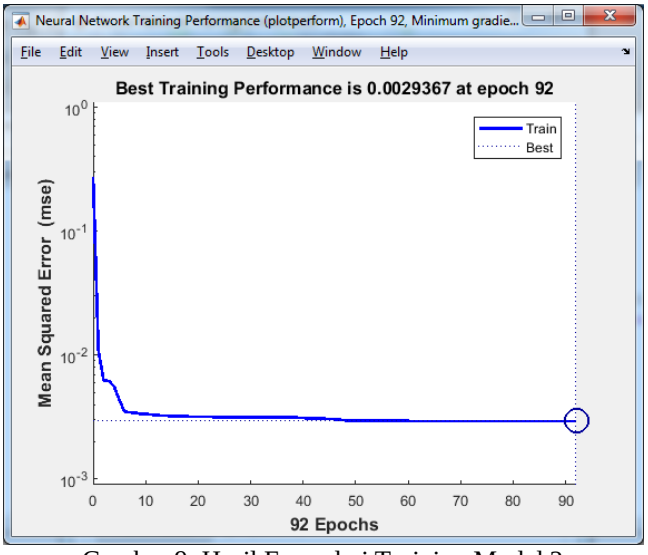
Gambar 6. Proses Training Model ke-2



Gambar 7. Interface Sistem



Gambar 8. Hasil Error dari Training Model 1



Gambar 9. Hasil Error dari Training Model 2

5. Kesimpulan

Training dilakukan sebanyak tiga kali agar mendapatkan nilai akurasinya tinggi. Nilai error terendah pada model ke-1 adalah 0,0016. Dan model ke-2 nilai error terendah adalah 0,0012. Sedangkan MSE secara berturut-turut model ke-1 dan ke-2 adalah 0,0027, dan 0,0029.

6. Daftar Pustaka

[1] Nur Nafi'iyah, Siti Mujilahwati, "Analisis Algoritma Backpropagation dan Naive Bayes," in *SENASIF*, Malang, 2018.

[2] N. Nafi'iyah, "Perbandingan Regresi Linear, Backpropagation dan Fuzzy Mamdani dalam Memprediksi Harga Emas," in *SENIATI*, Malang, 2016.

[3] David Hartanto Kamagi, Seng Hansun, "Implementasi Data Mining dengan Algoritma C4.5 untuk Memprediksi Tingkat Kelulusan Mahasiswa," *ULTIMATICS*, vol. 6, no. 1, pp. 15-20, 2014.

[4] Mujib Ridwan, Hadi Suyono, M. Sarosa, "Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier," *Jurnal EECCIS*, vol. 7, no. 1, pp. 59-64, 2013.

[5] Diana Laily Fithri, Eko Darmanto, "SISTEM PENDUKUNG KEPUTUSAN UNTUK MEMPREDIKSI KELULUSAN MAHASISWA MENGGUNAKAN METODE NAÏVE BAYES," in *SNATIF*, Universitas Muria Kudus, 2015.

[6] A. Jananto, "Algoritma Naive Bayes untuk Mencari Perkiraan Waktu Studi Mahasiswa," *Jurnal Teknologi Informasi DINAMIK*, vol. 18, no. 1, pp. 9-16, 2013.

[7] A. A. Murtopo, "Prediksi Kelulusan Tepat Waktu Mahasiswa STMIK YMI Tegal Menggunakan Algoritma Naïve Bayes," *CSRID Journal*, vol. 7, no. 3, pp. 145-154, 2015.

[8] A. Rohman, "MODEL ALGORITMA K-NEAREST NEIGHBOR (K-NN) UNTUK PREDIKSI KELULUSAN MAHASISWA," *Neo Teknika*, vol. 1, no. 1, pp. 1-9, 2015.

[9] Sutojo, Edy Mulyanto, Vincent Suhartono, Kecerdasan Buatan, Yogyakarta: Andi, 2011.

[10] Arief Hermawan, Jaringan Saraf Tiruan Teori dan Aplikasi, Yogyakarta: Andi, 2006.

[11] S. Susmanto, Z. Zulfan, and M. Munawir, "Sistem Penerapan Fuzzy Multi Attribute Decision Making (MADM) Dalam Mendukung Keputusan Untuk Menentukan Lulusan Terbaik Pada Sekolah Tinggi Teknik Poliprofesi Medan," *J. Nas. Komputasi dan Teknol. Inf.*, vol. 1, no. 1, 2018.