Solusi Untuk Masalah "Titanic: Machine Learning From Disaster"

Julisa Bana Abraham
Department of Electrical Engineering and Information Technology
Universitas Gadjah Mada
Daerah Istimewa Yogyakarta
julisa.bana.a@mail.ugm.ac.id

***Abstrak* - Pada penelitian kali ini didesain sebuah sistem cerdas untuk mengklasifikasi apakah korban dari kecelakaan kapal RMS Titanic meninggal atau bertahan hidup menggunakan data aktual dari para korban kapal RMS Titanic sebelum mengalami kecelakaan. Penelitian ini berusaha untuk menyelesaikan masalah "Titanic: Machine Learning from Disaster" yang terdapat pada situs kaggle.com menggunakan model klasifikasi dengan metode Gaussian Naïve Bayes yang memiliki akurasi 75,119%, metode MLP yang memiliki akurasi 72,48%, dan metode kNN yang memiliki akurasi 61,72%.**

Index Terms – Titanic, Machine Learning, MLP, KNN, GNB

I. PENDAHULUAN

Saat ini sistem cerdas banyak diaplikasikan pada berbagai bidang, seperti kesehatan, sains, ekonomi, biologi, dan lain-lain. Penggunaan sistem cerdas yang sangat fleksibel sehingga dapat diaplikasikan ke semua bidang. Pada penelitian kali ini didesain sebuah sistem cerdas untuk mengklasifikasi apakah korban dari kecelakaan kapal Titanic meninggal atau bertahan hidup menggunakan data aktual dari para korban kapal titanic sebelum mengalami kecelakaan. Tenggelamnya kapal Titanic merupakan salah satu kecelakaan kapal paling terkenal sepanjang sejarah yang terjadi pada 15 April 1912 setelah menabrak gunung es sehingga membunuh 1502 dari 2224 penumpang dan awak kapal. Pada penelitian ini dipilih Gaussian Naïve Bayes, K-Nearest Neighbors, dan Multilayer Perceptron sebagai metode-metode pengklasifikasi.

II. METODE PENELITIAN

Tujuan dari penelitian ini adalah untuk membuat model classifier yang cocok yang dapat mengklasifikasi selamat atau tidaknya penumpang RMS Titanic berdasarkan atribut-atribut penumpang sebelum mengalami kecelakaan.

A. Dataset

Dataset training dan testing didapatkan dari kaggle.com sebuah situs untuk mengasah kemampuan machine learning dan data mining dalam kasus "Titanic: Machine Learning From Disaster". Pengetesan hasil klasifikasi dilakukan langsung pada situs tersebut. *Dataset* untuk training dan testing memiliki 10 atribut. Disediakan 891 data penumpang untuk training dan 418 data penumpang untuk testing. Dari 891 data penumpang untuk training digunakan 66,7% untuk melatih ketiga model klasifikasi dan sisanya 33,3% untuk validasi model, hal ini dilakukan untuk menghindari terjadinya *overfitting* pada model selain itu juga digunakan untuk mengetahui akurasi model sebelum di lakukan *submit* untuk validasi secara *online* di kaggle.com. Dari 10 atribut penumpang hanya diambil 7 atribut yaitu parch, umur, *fare*, *embarked*, Pclass, SibSip, dan jenis kelamin. Atribut nama, nomor kabin, dan tiket tidak dimasukan karena dinilai tidak relevan untuk pemodelan. Definisi dari atribut-atribut pada data ialah sebagai berikut *embarked* adalah pelabuhan mana penumpang memulai perjalanan, Pclass ialah kelas tiket penumpang, parch adalah jumlah orang tua yang ikut dalam perjalanan, SibSp adalah jumlah saudara atau pasangan yang ikut dalam perjalanan, *fare* adalah tarif tiket, *ticket* adalah nomor tiket penumpang. Variable embarked, pclass, dan *sex* merupakan atribut-atribut diskrit sedangkan parch, sibsp, fare dan age merupakan atribut yang bersifat kontinu. Untuk atribut *age* terdapat banyak data yang kosong sehingga data yang kosong tersebut diisi dengan rerata umur penumpang yaitu 35 tahun.

B. Pemodelan

Penelitian ini dilakukan menggunakan bahasa pemrograman python dengan bantuan library scikit learn untuk memodelkan *classifier* dengan parameter yang diatur berdasarkan akurasi yang didapatkan dari dataset untuk validasi. Dipilih tiga metode klasifikasi untuk penelitian ini yaitu Gaussian Naïve Bayes, k-Nearest Neighbor, dan Multilayered Perceptron. Ketiga metode ini dipilih karena penulis familiar dengan metode-metode tersebut.

1) Gaussian Naïve Bayes

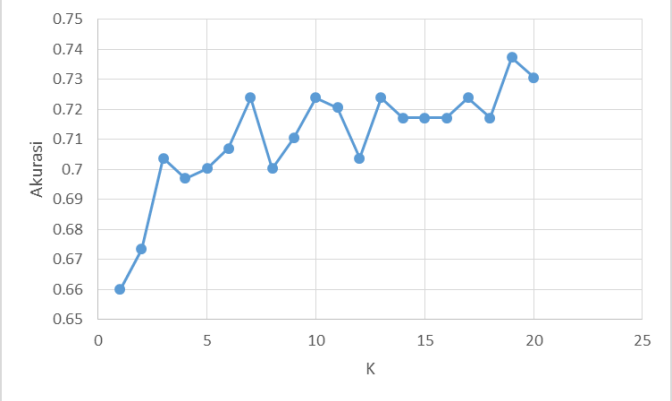
Metode ini menerapkan teorema bayes, yang dikemukakan oleh Thomas Bayes pada abat ke 18. Pada penelitian ini digunakan model Gaussian naïve bayes yang biasa digunakan untuk data yang bersifat kontinu. P(xi|Ci) didefinisikan sebagai berikut:

$$p(x = v \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}} \quad (1)$$

Dengan μ adalah rerata dan σ adalah standar deviasi [1]. Tidak ada parameter khusus untuk metode ini pada library scikit learn.

2) k-Nearest Neighbor

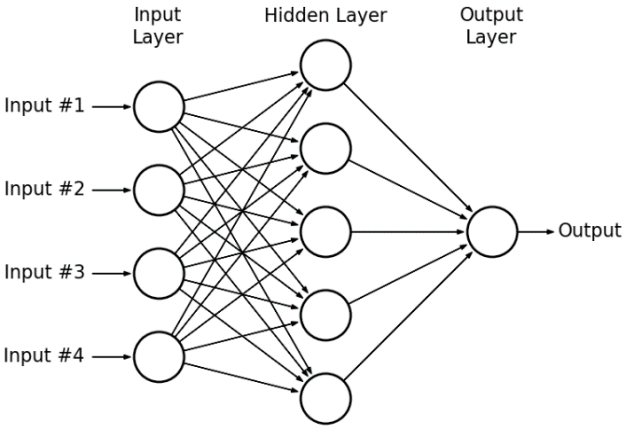
Metode k-Nearest Neighbor bekerja dengan mencari sejumlah k objek data atau pola yang paling dekat dengan pola masukan, kemudian memilih kelas dengan jumlah pola terbanyak di antara k pola tersebut [1]. Pada penelitian ini penentuan jumlah k dilakukukan dengan memeriksa akurasi pada setiap k pada nilai 1 sampai 20. Dipilih k=19 untuk pemodelan kNN dikarenakan memiliki akurasi terbaik pada validasi model. Berikut grafik k terhadap akurasi.



Gambar 1 Grafik K terhadap akurasi

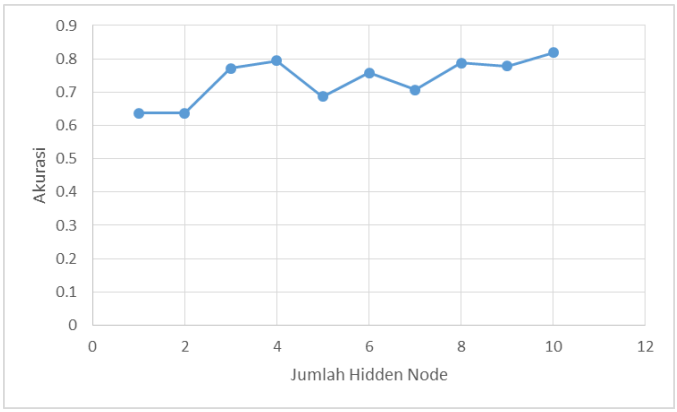
3) Multilayer Perceptron

Multilayer Perceptron (MLP) adalah arsitektur neural network yang memiliki banyak lapisan (layer). MLP memiliki satu atau lebih lapis tersembunyi atau *hidden layer,* dengan node yang berhubungan dengannya yang disebut *hidden node* [2]. MLP diilustrasikan pada gambar di bawah ini:



Gambar 2 Ilustrasi MLP [2]

Pada penelitian ini ditetapkan 3 buah *hidden layer* dan masing-masing *hidden layer* ditetapkan untuk memiliki 10 *hidden node*, yang ditetapkan berdasarkan evaluasi terhadap akurasi yang ditampilkan pada grafik dibawah.



Gambar 3 Grafik jumlah *hidden node* terhadap akurasi

III. HASIL DAN PEMBAHASAN

Penelitian ini dilakukan untuk mengetahui dan membuat model yang mampu untuk mengklasifikasi selamat atau tidaknya penumpang RMS Titanic. Telah dimodelkan 3 *classifier* dengan menggunakan metode Gaussian Naïve Bayes, k-Nearest Neighbor, dan Multilayer Pereceptron untuk menyelesaikan masalah ini. Hasil akurasi aktual dari penelitian ini didapatkan pada waktu *submit* hasil klasifikasi pada situs kaggle.com, sedangkan akurasi untuk validasi digunakan dari sebagian data latih yang sengaja tidak dimasukan sewaktu proses pelatihan sebanyak sekitar 33,3% dari jumlah data latih keseluruhan.

1) Gaussian Naïve Bayes

Pemodelan *classifier* menggunakan metode Gaussian Naïve Bayes menghasilkan *classifier* yang memiliki akurasi aktual yang cukup baik yaitu 75,119%, tidak jauh berbeda dari akurasi pada sewaktu validasi yaitu 80,47% yang berarti classifier ini tidak mengalami *overfitting.*

2) k-Nearest Neighbor

Pemodelan *classifier* menggunakan metode kNN dengan jumlah k=19 menghasilkan *classifier* yang memiliki akurasi aktual yaitu 61,72% agak jauh dari akurasi pada sewaktu validasi yaitu 73,47% yang dapat dimaklumi karena metode ini merupakan metode yang tidak memperhitungkan seluruh data dalam proses klasifikasinya atau bisa dikatakan termasuk algoritma *greedy* [1].

3) Multilayer Perceptron

Pemodelan *classifier* menggunakan metode MLP dengan jumlah hidden layer 3 dan setiap hidden layer memiliki 10 *hidden node* menghasilkan *classifier* yang memiliki akurasi aktual yaitu 72,48% tidak jauh dari akurasi pada sewaktu validasi yaitu 79,79% sehingga metode ini bisa dikatakan tidak mengalami *overfitting* sewaktu pelatihannya

IV. KESIMPULAN DAN SARAN

Penggunaan sistem cerdas pada masa ini tidak memiliki batasan pada jenis masalah yang dapat diselesaikan, baik itu masalah sains, sosial, maupun masalah yang cenderung aneh seperti klasifikasi keselamatan penumpang RMS Titanic ini. Model terbaik dalam menyelesaikan masala "Titanic: Machine Learning from Disaster" pada penelitian ini adalah model klasifikasi yang menggunakan metode Gaussian Naïve Bayes memiliki akurasi tertinggi yaitu 75,119% dengan metode MLP menempati peringkat kedua dengan akurasi 72,48%, sedangkan model klasifikasi menggunakan metode kNN memiliki akurasi terendah yaitu 61,72%.

REFERENSI

[1] Suyanto, "Data Mining Untuk Klasifikasi dan Klasterisasi data," *Penerbit Informatika*, February 2017.

[2] Hassan Mohammed et al.,"Assesment of artificial neural network for bathymetry estimation using high resolution Satelite Imagery in Shallow Lakes., International Water Technology Journal, 2013.