JNATIA Volume 1, Nomor 1, November 2022
Jurnal Nasional Teknologi Informasi dan Aplikasinya

Hyperparameter Tuning Algoritma KNN Untuk Klasifikasi Kanker Payudara Dengan Grid Search CV

Nyoman Hendradinata Dharma[a1] dan I Gede Santi Astawa[a2].

[a]Program Studi Informatika,
Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Udayana
Bali, Indonesia
[1]nyomanhendradinata20@gmail.com@email.com
[2]santi.astawa@unud.ac.id

Abstract

One of the deadliest diseases in the world is Breast cancer. Breast cancer is a disease caused by abnormal cells that grow and develop rapidly and malignantly in the human breast and spread quickly to the tissues or organs around the breast. Data from Riskesdas in 2019 stated that in Indonesia, the prevalence of breast cancer was 41.2 per 100,000 Indonesians with an average death rate of 17 per 100,000 Indonesians. Technology nowadays is increasingly advanced and developed which can help people to find out the disease they are suffering from early before carrying out further examinations with the doctor. Breast cancer can be detected early by classifying it with machine learning algorithm. In this research, Breast cancer will be classified using K-Nearest Neighbor algorithm with Grid Search to classify whether a person has breast cancer or not. K-Nearest Neighbor (KNN) is one of the classification algorithms, where classification is carried out on data objects based on learning data whose neighbors are closest to the data object. The performance results of the classification model using K-Nearest Neighbor are 83% accuracy, 73% precision, and 89% recall.

Keywords: *KNN, Kanker Payudara, Klasifikasi, Grid Search, Hyperparameter Tuning*

1. Introduction

Salah satu penyakit yang mematikan di dunia adalah Kanker Payudara. Kanker Payudara merupakan penyakit yang disebabkan oleh sel-sel abnormal yang tumbuh dan berkembang secara cepat dan ganas di payudara manusia dan cepat menyebar ke bagian jaringan atau organ sekitar payudara. Berdasarkan data dari Riskesdas pada tahun 2019 disebutkan di Indonesia, prevalensi kejadian untuk Kanker Payudara sebanyak 41,2 per 100.000 masyarakat Idonesia dengan rata-rata kematian 17 per 100.000 masyarakat Indonesia [1].

Teknologi dewasa ini sudah semakin maju dan berkembang dapat membantu masyarakat untuk mengetahui penyakit yang dideritanya lebih awal sebelum melakukan pemeriksaan lebih lanjut ke dokter. Kanker Payudara dapat diketahui lebih awal dengan melakukan klasifikasi menggunakan algoritma *machine learning*. Terdapat berbagai macam algoritma *machine learning* yang dapat digunakan untuk melakukan klasifikasi, salah satunya adalah KNN atau K-Nearest Neighbor. KNN merupakan algoritma *supervised* yang melakukan proses klasifikasi objek dari data pembelajaran yang memiliki ketetanggaan objek data paling dekat.

Penelitian sebelumnya dengan judul Hyperparameter Tuning pada Algoritma Klasifikasi dengan Grid Search oleh Wahyu Nugraha dan Agung Sasongko [5]. Pada penelitian tersebut dilakukan optimasi hyperparameter yaitu Grid Search terhadap 7 algoritma klasifikasi *machine learning* yaitu *XGBoost, Support Vector Machine (SVM), Random Forest, Logistic Regression, Gaussian Naive Bayes,* K-*Nearest Neighbors (k-NN), Decision Tree.* Pada penelitian tersebut menunjukan hasil eksperimen model *XGBoost* memperoleh nilai terbaik yaitu sebesar 0,772 sedangkan untuk *Decision tree* memiliki nilai terendah yaitu 0,701.

berdasarkan pemaparan diatas, penulis akan mengimplementasikan *hyperparameter tuning* algoritma KNN untuk klasifikasi kanker payudara dengan *Grid Search* untuk melakukan klasifikasi apakah seseorang terkena kanker payudara atau tidak.

2. Reseach Methods

Pada penelitian ini menggunakan *Breast Cancer Coimbra Dataset* yang merupakan data sekunder yang diperoleh dari https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra. Pada dataset ini memiliki 9 atribut dan 1 label dengan jumlah data sebanyak 166 baris data. Klasifikasi Kanker Payudara memiliki alur penelitian yang dimulai dengan (1) Menginputkan dataset dari *Breast Cancer Coimbra Dataset,* (2) Normalisasi Data, (3) Membagi data menjadi data training dan data testing, (4) *Hyperparameter Tuning* dengan *Grid Search Cross Validation*, (5) Klasifikasi KKN menggunakan parameter K optimal, (6) Mendapatkan hasil *accuracy, precision, dan recall*.

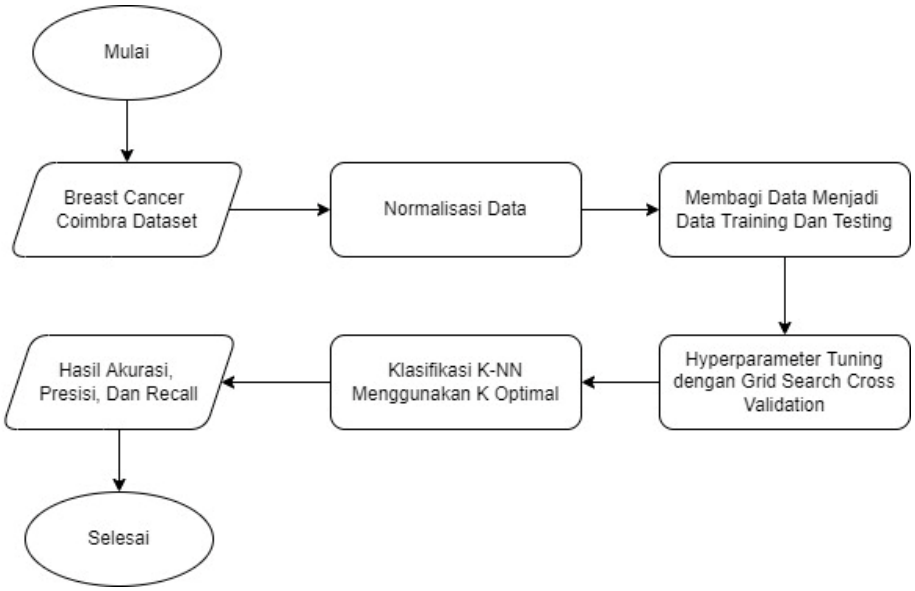

Figure 1. Alur Penelitian

2.1. Kanker Payudara

Payudara merupakan salah satu organ penting yang berfungsi untuk memproduksi ASI (Air Susu Ibu) dan Menyusui Bayi. Payudara terdiri dari jaringan, saraf, pembuluh darah, dan saraf yang berkumpul menjadi satu kesatuan. Kanker merupakan pertumbuhan sel-sel abnormal yang berkembang dengan cepat dan tidak dapat dikendalikan. Kanker Payudara merupakan penyakit tumor ganas yang berasal dari sel-sel abnormal yang tanpa kendali memperbanyak diri sehingga menyebar dan berpindah ke jaringan maupun organ di sekitar payudara [4]. Kanker Payudara tidak hanya menyerang perempuan namun laki-laki juga dapat terkena penyakit Kanker Payudara. Gejala-gejala dari seseorang yang mengidap Kanker Payudara adalah benjolan pada daerah payudara yang tidak sembuh-sembuh, keluar cairan kuning pada puting, warna kulit di area payudara berubah, dan puting menjadi masuk ke dalam.

2.2. K-Nearest Neighbor

K-Nearest Neighbor (KNN) merupakan salah satu algoritma klasifikasi, dimana klasifikasi dilakukan terhadap objek data berdasarkan data pembelajaran yang jarak ketetanggaannya paling dekat dengan objek data tersebut. K-Nearest Neighbor atau KNN merupakan salah satu algoritma *supervised learning,* algoritma *supervised learning* merupakan algoritma yang dilatih dengan diberikan data yang sudah diberikan label sehingga didapatkan suatu hasil tertentu. Dalam menghitung jarak ketetanggan paling dekat dapat digunakan dua cara yaitu *Euclidean*

distance dan *Manhattan distance,* namun *Euclidean distance* secara umum lebih sering digunakan dalam menghitung jarak ketetanggan. Rumus dari *Euclidean distance* dapat ditunjukkan dari persamaan dibawah

$$euc\ dist = \sqrt{\Sigma_{i=1}^{n} (ai - bi)^2} \quad (1)$$

Keterangan:
euc dist = jarak ketetanggan
ai = data training
bi = data testing
i = variabel atau record data
n = dimensi data

2.3. Normalisasi Data

Normalisasi merupakan proses yang digunakan untuk menskalakan nilai atribut data pada rentang tertentu sehingga nilai atribut tidak memiliki selisih yang cukup jauh. Dengan adanya normalisasi data dapat membantu model klasifikasi untuk mempelajari data lebih mudah. Salah satu cara untuk menormalisasi data adalah *Z-score normalization,* pada *Z-score normalization* data akan dinormalisasi berdasarkan dari nilai rata-rata atau *mean* dan *standar deviation* [2]. Persamaan dari *Z-score normalization* adalah *sebagai berikut.*

$$z = \frac{x - \mu}{\sigma} \quad (2)$$

Keterangan:
z = nilai baru
x = nilai lama
μ = mean
σ = standar deviasi

2.4. Grid Search Cross Validation

Grid Search Cross Validation merupakan salah satu metode yang digunakan untuk mencari suatu parameter yang optimal dalam suatu model. *Grid Search* adalah algoritma yang bekerja dengan cara menjadikan titik-titik *grid* memiliki jarak yang mirip, lalu mengkalkulasikan kesalahan-kesalahan untuk setiap titik-titik parameter tersebut, dimana dalam hal ini titik yang mempunyai nilai kesalahan terendah merupakan parameter yang paling optimal [3]. *Grid Search Cross Validation* akan melakukan pencocokan atau validasi dari semua model yang dikombinasikan dan hyperparameter secara otomatis [5].

2.5. *Accuracy, Precision, dan Recall*

Accuracy, Precision, dan *Recall* merupakan salah satu cara untuk memperhitungkan tingkat kebenaran dari pengklasifikasian suatu data. *Accuracy* digunakan untuk mengukur tingkat kemiripan nilai antara nilai prediksi dengan nilai sebenarnya, *precision* digunakan untuk mengukur tingkat kecocokan antara informasi yang diperlukan dengan bagian data yang diambil, *recall* digunakan untuk mengukur suatu tingkat keberhasilan sistem untuk menemukan suatu informasi kembali.

$$Accuracy = \frac{TP+FN}{TP+FN+FP+FN} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

Keterangan:
TP = Jumlah data yang memiliki nilai sebenarnya positif dan nilai prediksi positif
FP = Jumlah data yang memiliki nilai sebenarnya negatif dan nilai prediksi positif
FN = Jumlah data yang memiliki nilai sebenarnya positif dan nilai prediksi negatif
TN = Jumlah data yang memiliki nilai sebenarnya negatif dan nilai prediksi negatif

3. Result and Discussion

Pada penelitian ini dataset yang digunakan adalah *Breast Cancer Coimbra Dataset.* Pada dataset tersebut berisi 166 baris dengan 9 atribut dan 1 label, dimana dalam label tersebut terdapat dua nilai yaitu 1 untuk *Healthy controls* dan nilai 2 untuk *Patients.*

c

	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Classification
0	48	23.500000	70	2.707	0.467409	8.8071	9.702400	7.99585	417.114	1
1	83	20.690495	92	3.115	0.706897	8.8438	5.429285	4.06405	468.786	1
2	82	23.124670	91	4.498	1.009651	17.9393	22.432040	9.27715	554.697	1
3	68	21.367521	77	3.226	0.612725	9.8827	7.169560	12.76600	928.220	1
4	86	21.111111	92	3.549	0.805386	6.6994	4.819240	10.57635	773.920	1
5	49	22.854458	92	3.226	0.732087	6.8317	13.679750	10.31760	530.410	1
6	89	22.700000	77	4.690	0.890787	6.9640	5.589865	12.93610	1256.083	1
7	76	23.800000	118	6.470	1.883201	4.3110	13.251320	5.10420	280.694	1
8	73	22.000000	97	3.350	0.801543	4.4700	10.358725	6.28445	136.855	1
9	75	23.000000	83	4.952	1.013839	17.1270	11.578990	7.09130	318.302	1

Figure 2. *Breast Cancer Coimbra Dataset*

Selanjutnya dilakukan korelasi atribut untuk mencari keberpengaruhan atribut-atribut yang ada dalam *Breast Cancer Coimbra Dataset* dengan label kalsifikasi.

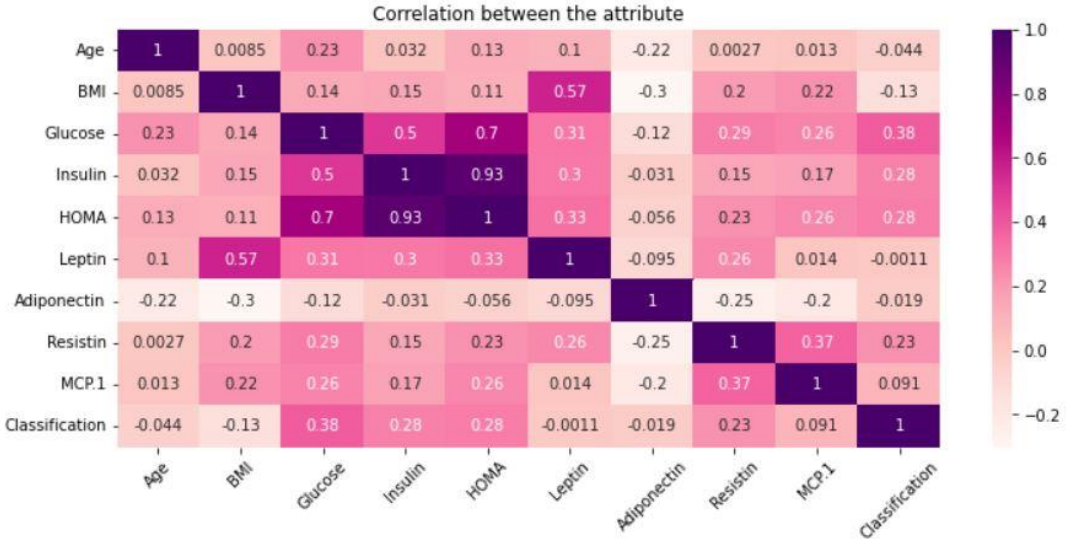

Figure 3. Korelasi Atribut

3.1. Normalisasi Data

Selanjutnya akan dilakukan normalisasi pada dataset. Hal ini dilakukan agar rentang nilai data dari 9 atribut yang ada tidak terlalu jauh. Dengan rentang nilai data yang berdekatan akan memudahkan model untuk melakukan klasifikasi.

```
array([[-0.57979363, -0.81667527, -1.23922225, -0.72873938, -0.61428241,
        -0.93233407, -0.07022151, -0.54551749, -0.34125061],
       [ 1.60182096, -1.37875056, -0.25829943, -0.68803819, -0.54824045,
        -0.93041264, -0.69734988, -0.86421418, -0.1912238 ],
       [ 1.53948912, -0.89176446, -0.30288683, -0.55007314, -0.46475236,
        -0.45421914,  1.79799836, -0.4416602 ,  0.05821407],
       [ 0.66684328, -1.24330321, -0.92711044, -0.67696507, -0.57420965,
        -0.87602119, -0.44194467, -0.15886735,  1.14271758],
       [ 1.7888165 , -1.29460116, -0.25829943, -0.6447433 , -0.52108087,
        -1.04268238, -0.78688094, -0.33635201,  0.69471601]])
```

Figure 4. Hasil Korelasi Atribut 5 Data Teratas

3.2. Membagi Data menjadi *Data Training* Dan *Data Testing*

Setelah dataset di normalisasi, akan dilakukan pemisahan dataset menjadi dua bagian yaitu menjadi data training dan data testing. Masing-masing pembagian dari dataset tersebut adalah 92 data atau 80% untuk data training dan 24 data atau 20% data testing.

3.3. *Hyperparameter Tuning* dengan *Grid Search Cross Validation*

Selanjutnya dilakukan hyperparameter tuning, dimana dalam hal ini akan dicari *hyperparameter k* (jumlah tetangga terdekat) terbaik dalam pengklasifikasian menggunakan *Grid Search Cross Validation*. Dimana digunakan 10 *Cross Validation* dan range dari parameter k adalah 1-10. Setelah dilakukan *hyperparameter tuning dengan* data training didapatkan jumlah k yang paling baik atau optimal adalah k=7.

```
print(grid_search.best_params_)
```

```
{'n_neighbors': 7}
```

Figure 5. Hasil Hyperparameter Tuning

3.4. *Klasifikasi KNN Dengan Parameter k Optimal*

Selanjutnya dibangun model klasifikasi K-Nearest Neighbor (KNN) pada *Breast Cancer Coimbra Dataset* dengan parameter k (jumlah tetangga terdekat) terbaik yang sudah didapatkan di *Grid Search Cross Validation* yaitu k=7. Dimana hasil yang didapatkan dari model klasifikasi menggunakan data testing adalah tingkat dari akurasi 83%, tingkat presisi 73%, dan recall 89%.:

Table 1. Perbandingan Hasil Performa KNN

	Precision	Recall	Accuracy
KNN dengan Grid Search CV, k=7	0.73%	0.89%	0.83%

4. Conclusion

Pada penelitian ini, klasifikasi kanker payudara dibantu dengan algoritma K-Nearest Neighbor (KNN) dengan parameter *k* (jumlah tetangga terdekat) yaitu 7. Untuk meningkatkan performa algoritma KNN dilakukan *Hyperparameter Tuning* dengan *Grid Search Cross Validation.* Dimana performa yang dimaksud disini adalah *Accuracy, Precision,* dan *Recall.* Hasil performa dari model klasifikasi menggunakan K-Nearest Neighbor adalah tingkat dari akurasi 83%, tingkat presisi 73%, dan recall 89%.

Saran yang diharapkan untuk penelitian selanjutnya adalah dapat membandingkan algoritma-algoritma lain untuk dilakukan *Hyperparameter Tuning* dengan *Grid Search Cross Validation*

dalam mendeteksi penyakit kanker payudara, agar didapatkan algoritma mana yang memiliki performa lebih akurat dan efisien. Sehingga dapat ditemukan algoritma yang lebih baik dalam pengklasifikasian penyakit kanker payudara.

References

[1] Ariq. Naupal. Azmi, Bambang. Kurniawan, Andi. Lukman and Ade. Utia. Detty, "Hubungan Faktor Keturunan Dengan Kanker Payudara DI RSUD Abdoel Moeloek", *Jurnal Ilmiah Kesehatan Sandi Husada,* vol. 9, no. 2, pp.702-707, 2020.
[2] Darnisa. Azzahra. Nasution, Hidayah. Husnul. Khotimah, and Nurul. Chamidah, "PERBANDINGAN NORMALISASI DATA UNTUK KLASIFIKASI WINE MENGGUNAKAN ALGORITMA K-NN", *CESS (Journal of Computer Engineering System and Science),* vol. 4, no. 1, pp.78-82, 2021.
[3] Dewi. Satriani, Latifah. Uswatun. Khasanah and Nanda. Arista. Rizki, "PENERAPAN METODE GRID-SEARCH DALAM MENENTUKAN PARAMETER MODEL PERTUMBUHAN PENDUDUK DI KOTA SAMARINDA", *Prosiding Seminar Nasional Matematika, Statistika, dan Aplikasinya,* vol. 1, no. 5, pp.65-74, 2019.
[4] Laili. Rahayuwati, Iqbal. Abdul. Rizal, Tuti. Pahria, Makmat. Lukman and Neti.Juniarti, "Pendidikan Kesehatan tentang Pencegahan Penyakit Kanker dan Menjaga Kualitas Kesehatan", *Media Karya Kesehatan,* vol. 3, no. 1, pp.59-69, 2020.
[5] Wahyu. Nugraha and Agung. Sasongko, "Hyperparameter Tuning pada Algoritma Klasifikasi dengan Grid Search", *SISTEMASI: Jurnal Sistem Informasi,* vol. 11, no.21, pp.391-401, 2020.