OPTIMASI ALGORITMA K-NEAREST NEIGHBOR DENGAN SELEKSI FITUR MENGGUNAKAN XGBOOST

Muflih Ihza Rifatama[1], Mohammad Reza Faisal[2], Rudy Herteno[3], Irwan Budiman[4], Muhammad Itqan Mazdadi[5]

[12,]Program Studi Ilmu Komputer, Universitas Lambung Mangkurat
Jl. Brigjen Jalan Hasan Basri, Pangeran, Kec. Banjarmasin Utara, Kota Banjarmasin, Kalimantan Selatan 70123
[1] 1711016210011@mhs.ulm.ac.id, [2] reza.faisal@ulm.ac.id, [3] rudy.herteno@ulm.ac.id,
[4] irwan.budiman@ulm.ac.id, [5] mazdadi@ulm.ac.id

Abstract

Cancer is a general term for a large group of diseases that can affect any part of the body. One of the most dangerous cancers is breast cancer. Breast cancer prevention can be done in one way, namely screening or early diagnosis. Diagnostics can use machine learning with several algorithms, for example K-Nearest Neighbor. The KNearest Neighbor (K-NN) classification algorithm is a fairly well-known and often used algorithm, but there is a weakness in the KNN algorithm, namely that this algorithm is very influential in the presence of noise or irrelevant data if the feature scale is inconsistent with its importance. One way to overcome this is by selecting features. The feature selection used is using Extreme Gradient Boosting (XGBoost) based on the importance of the features obtained. The results show that KNN with XGBoost feature selection outperforms the KNN model without feature selection, for the KNN value with XGBoost feature selection gets an accuracy of 0.977 while KNN without feature selection gets an accuracy of 0.974.

Keywords : *KNN, K Nearest Neighbour, Feature Selection, XGBoost.*

Abstrak

Kanker merupakan istilah umum untuk sekelompok besar penyakit yang dapat menyerang bagian tubuh mana pun. Salah satu kanker yang berbahaya adalah Kanker payudara. Pencegahan kanker payudara dapat dilakukan dengan salah satu cara yaitu skrining atau diagnosa dini. Pendiagnosaan dapat menggunakan Machine learning dengan beberapa algoritma contohnya K-Nearest Neighbor. Algortima klasifikasi K-Nearest Neighbor (K-NN) merupakan algortima yang cukup terkenal dan sering digunakan, tetapi terdapat kelemahan pada algoritma KNN yaitu algoritma ini sangat berpengaruh dengan adanya data yang noise atau tidak relevan jika skala fitur tidak konsisten dengan kepentingannya. Salah satu cara mengatasinya adalah dengan cara menyeleksi fitur. Seleksi fitur yang digunakan yaitu menggunakan Extreme Gradient Boosting (XGBoost) berdasarkan kepentingan fitur yang didapatkan. Hasilnya menunjukkan bahwa KNN dengan seleksi fitur XGBoost menggungguli model KNN tanpa seleksi fitur, untuk nilai KNN dengan seleksi fitur XGBoost mendapatkan akurasi sebesar 0,977 sedangkan KNN tanpa seleksi fitur mendapatkan akurasi sebesar 0,974.

Kata kunci : *KNN, K Nearest Neighbour, Seleksi Fitur, XGBoost.*

1. PENDAHULUAN

Machine learning adalah metode atau teknik yang menggunakan algoritma pembelajaran pada sejumlah besar data untuk membantu menangani dan membuat prediksi [1]. Machine learning dapat diartikan sebagai metode komputasi yang mengambil data sebelumnya dan belajar sehingga menaikkan kinerja atau membuat prediksi yang benar.

Algoritma K-Nearest Neighbor termasuk dalam keluarga algoritma pembelajaran yang diawasi. KNN adalah salah satu metode lebih baik

ketika tidak ada pengetahuan yang tepat tentang distribusi data yang masuk. Algoritma ini menghitung jarak dari data ke seluruh data pelatihan dan menetapkannya sebagai kelas yang diwakili oleh label mayoritasnya k-terdekat [2].

Salah satu tahapan preprocessing adalah seleksi fitur. Seleksi fitur secara langsung mempengaruhi hasil klasifikasi. Dalam pengenalan pola seleksi fitur sangat penting untuk analisis data. Proses ini bertujuan untuk memilih fitur terbaik dari fitur asli dan dapat mengurangi dimensi data dalam jumlah besar dan keluar dari masalah curse of dimensionality, sehingga meningkatkan kinerja metode klasifikasi. [3].

Seleksi fitur bertujuan untuk memilih variable penting dalam melakukan klasifikasi menggunakan K-Nearest Neighbor (KNN). Tujuan utama seleksi fitur adalah untuk fokus menemukan data relevan. Seleksi fitur harus dapat membedakan data, karena fitur yang tidak relevan dan berlebihan akan mempengaruhi hasil [4]. Salah satu teknik efektif untuk meningkatkan algortima klasifikasi yaitu dengan yeknik seleksi fitur yang dapat menghasilkan fitur-fitur relevan.

Terdapat masalah pada K-Nearest Neighbor atau KNN yaitu algoritma ini dapat sangat turun dengan adanya data yang noise atau tidak relevan jika skala fitur tidak konsisten dengan kepentingannya [5]. Dengan masalah tersebut KNN diperlukan proses seleksi fitur agar dapat mengurangi data yang noise dan fitur yang kurang relevan sehingga dapat meningkatkan akurasi pada metode tersebut [6].

Extreme gradient boosting (XGBoost) adalah sistem penambah pohon yang dapat diskalakan yang menggabungkan secara berurutan untuk membentuk model akhir hingga kesalahan diminimalkan. Pemilihan fitur ansambel berdasarkan pendekatan pohon dan model klasifikasi ensemble menggunakan XGboost dapat meningkatkan kinerja klasifikasi serta pengurangan fitur selain itu juga mengurangi waktu komputasi [7]. XGBoost berfungsi untuk pemilihan fitur terbaik berdasarkan feature rangking diharapkan agar meningkatkan akurasi pada K-Nearest Neighbor yang memiliki masalah pada akurasi yang dapat sangat terpengaruhi dengan adanya fitur yang noise atau tidak relevan.

Bedasarakan latar belakang diatas maka penelitian yang akan di angkat adalah seleksi fitur pada klasifikasi Algoritma K-Nearest Neighbor menggunakan XGBoost pada dataset Breast Cancer Wisconsin (Diagnostic) Dataset dari UCI Machine Learning.

Penelitian terdahulu menjadi acuan penulis dalam melakukan penelitian yang memiliki keterkaitan dengan penelitian ini dan bertujuan untuk menentukan posisi dan perbedaan penelitian yang akan dilakukan. Pada penelitian Ahmet Saygili tahun 2018 melakukan penelitian menggunakan metode seleksi fitur Gain Ratio dan Random Sampling pada KNN dengan K yaitu 1, 3 dan 5 mendapatkan hasil akurasi (97.36%), Sensitivity (0.974), Specificity (0.973) dan AUC (0.991)[8]. Pada penelitian lain N. Manju, B S Harish dan V Prajwal tahun 2019 menggunakan XGBoost serta membandingkan dengan metode Decision Tree, Random Forest dan Adaboost pada 248 fitur dan 8 fitur menggunakan dataset adalah subset turunan dari dataset Cambridge yang sudah di normalisasi dan dataset standar Cambridge yang imbalance. Mendapatkan hasil pada dataset yang dinormalisasi 248 fitur mendapatkan 96,97% dan 8 fitur mendapatkan 98.51% sedangkan dataset imbalance 248 fitur mendapatkan 87,48% dan 8 fitur mendapatkan 93.54%[7]. Penelitian Rizki Tri Prasetio tahun 2020 melakukan perbandingan metode KNN basik dan KNN dengan optimasi Algoritma Genetika untuk pemilihan fitur dan parameter k dengan 5 dataset yang berbeda mendapatkan hasil peningkatan akurasi pada metode KNN yang di optimasi Algoritma Genetika terhadap 5 dataset dibanding menggunakan KNN. Pada KNN mendapat akurasi 94.15% (breast-cancer (D)), 78.75% (breast-cancer (P)), 61.16% (diabetic-retinopathy), 77.5% (heart (SPECTF)) dan 90.91% (cardiotocography) sedangkan KNN yang dioptimasi mendapatkan akurasi 99.2.% (breast-cancer (D)), 86.44% (breast-cancer (P)), 71.69% (diabetic-retinopathy), 87.5% (heart (SPECTF)) dan 98.59% (cardiotocography)[5]. Penelitian Ichwanul Muslim Karo Karo tahun 2020 menggunakan XGBoost untuk klasifikasi pada kebakaran hutan dan lahan. Dari 12 fitur dipilih 6 atau 7 fitur yang berpengaruh besar untuk klasifikasi mendapatkan akurasi sebesar 89.52%[9].

Berdasarkan penelitian sebelumnya pada penelitian ini melakukan perbandingan KNN dengan seleksi fitur menggunakan XGBoost dan KNN tanpa seleksi fitur pada dataset Breast Cancer Wisconsin (Diagnostic).

1.1. Data Mining

Data mining didefinisikan sebagai cara menemukan pola dalam data. Berdasarkan

tugasnya, data mining dibagi menjadi deskripsi, estimasi, prediksi, klasifikasi, clustering, dan asosiasi. Proses fase data mining terdiri dari tiga langkah utama. Lakukan perbaikan data terlebih dahulu. Pada langkah ini data dipilih, dibersihkan dan diproses sesuai dengan pedoman dan pengetahuan ahli domain. Selanjutnya, penggunaan algoritma data mining dilakukan pada langkah ini untuk mengeksplorasi data terintegrasi untuk memfasilitasi identifikasi informasi yang berharga. Tiga tahap analisis data mengevaluasi hasil data mining untuk menemukan domain pengetahuan berupa aturan yang diekstraksi dari jaringan [20].

1.2. Machine learning

Machine learning adalah susunan teknik yang membantu memproses dan memprediksi data dalam jumlah yang sangat besar dengan menyajikan data menggunakan algoritma pembelajaran. Machine learning juga dapat didefinisikan sebagai metode komputasi eksperimental untuk meningkatkan kinerja atau membuat prediksi yang akurat. Yang dimaksud dengan pengalaman di sini adalah informasi awal yang sudah tersedia dan dapat digunakan sebagai data pelatihan. [18].

Ada dua jenis konsep pembelajaran dalam Machine learning. Pertama, pembelajaran terawasi adalah teknik pembelajaran mesin yang memproses kumpulan data. Perbedaannya adalah pada Unsupervised Learning tidak adanya pengklasifikasian dari dataset yang dipakai sedangkan Supervised Learning dataset telah terklasifikasi sebelum melakukan pemrosesan model [19].

1.3. K-Nearest Neighbour

K-fold Cross Validation merupakan teknik validasi yang berguna untuk membagi data *training* dan data *testing*. Dimana *K-fold Cross Validation* ini akan membagi keseluruhan dataset sebanyak k. Kemudian *K-fold Cross validation* akan melakukan proses klasifikasi sebanyak k kali dengan tiap iterasinya salah satu subset menjadi data *test* dan subset lainnya akan menjadi data training [4]. K-fold Cross Validation digunakan untuk menemukan kombinasi data yang terbaik.

1.4. K-Nearest Neighbour

Algoritma K-Nearest Neighbor (KNN) merupakan metode yang menggunakan algoritma supervised, dan algoritma dapat dibagi menjadi dua jenis yaitu supervised learning dan unsupervised learning. Algoritme pembelajaran yang diawasi bertujuan untuk mempertahankan pola baru, sementara pembelajaran tanpa pengawasan mempertahankan pola dalam data. Keakuratan algoritma KNN tergantung pada ada tidaknya data yang tidak relevan atau apakah bobot fitur sesuai dengan relevansinya dengan klasifikasi. [4].

K-Nearest Neighbor (KNN) berupaya untuk mendapatkan pola data baru dengan cara menghubungkan pola data sebelumnya dengan pola data baru guna mengklasifikasikan kemiripan data ke dalam beberapa kelas berdasarkan atribut yang ada. Kemiripan data bisa lebih besar dari 1, sehingga KNN dapat mengambil k data yang paling mirip dan mengklasifikasikan berdasarkan data yang paling mirip. [10]. Jarak yang digunakan adalah jarak Euclidean Distance.

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} + \cdots + (x_n - y_n)^2 \quad [1]$$

Keterangan :
d = jarak
x = data latih
y = data uji
n = jumlah data

1.5. XGBoost

XGBoost adalah implementasi dari gradient boosted decision tree yang dirancang untuk kecepatan ,kinerja dan skalabilitas [11]. Konsep dasar dari algoritma ini adalah menyesuaikan parameter pembelajaran secara iteratif untuk mengurangi cost function. XGBoost membangun struktur pohon menggunakan model yang lebih terorganisir, yang dapat meningkatkan kinerja, mengurangi kompleksitas model, dan menghindari overfitting. [12].

Pemilihan fitur ansambel berdasarkan pendekatan pohon dan model klasifikasi ensemble menggunakan XGBoost dapat meningkatkan kinerja kalsifikasi. Pengurangan fitur juga dapat mengurangi waktu komputasi. [7].

1.6. Seleksi Fitur

Seleksi fitur merupakan suatu teknik preprocessing yang penting dalam melakukan klasfikasi. Dalam pemilihan fitur dapat menyebabkan penurunan kinerja dari klasifikasi [15]. Berdasarkan karakteristiknya, fitur dapat di

bagi menjadi tiga yaitu fitur yang tidak relevan, fitur yang relevan dan fitur yang redundan [16].

Seleksi fitur merupakan pemilihan subset dari fitur asli sesuai dengan kriteria seleksi fitur tertentu. Hasil dari pemilihan fitur yaitu menghapus data yang tidak relevan dan berlebihan, dapat mengurangi waktu komputasi, meningkatkan akurasi pembelajaran, dan menyederhanakan hasil pembelajaran [17].

Fitur yang relevan merupakan fitur yang mempengaruhi output, dan peran fitur tersebut tidak dapat digantiakn oleh fitur lainnya. Selanjutnya ada fitur yang tidak relevan yaitu fitur yang tidak memilki pengaruh terhadap output. Sedangkan fitur redundan merupakan fitur yang dapat menggantikan peran dari fitur yang lain. Tujuan dari seleksi fitur ini yaitu untuk mendapatkan fitur yang relevan dan menghilangkan fitur tidak relevan dan fitur redundan.

Metode yang digunakan untuk seleksi fitur yaitu embedded, filter dan wrapper. Pendekatan embedded merupakan algoritma yang digunakan dalam satu kesatuan atau sebagian dari algoritma data mining. Pendekatan filter merupakan seleksi fitur yang tidak dilakukan bersamaan dengan permodelan yang dilakukan. Pendekatan yang digunakan pada penelitian ini adalah wrapper. Pendekatan wrapper yaitu seleksi fitur yang dilakukan bersamaan saat permodelan. Seleksi ini memanfaatkan tingkat klasifikasi dari metode pengklasifikasian atau pemodelan yang digunakan.

1.7. Confusion Matrix

Confusion Matrix adalah alat evaluasi yang digunakan dalam pembelajaran mesin. Dalam Confusion Matrix, kolom mewakili skor kelas yang diprediksi dan baris mewakili skor kelas yang sebenarnya.. Confusion Matrix dapat menampilkan semua kemungkinan kasus masalah pada klasifikasi [13].

Pada Confusion Matrix terdapat empat buah nilai matrix yaitu, True Positive (TP), False Positive (FP), False Negative (FN), dan True Negative (TN). True Positive adalah jumlah record yang diprediksi class-nya dengan benar, False Positive adalah jumlah record pada kolom yang bersangkutan kecuali TP, False Negative adalah jumlah record pada baris yang bersangkutan kecuali TP, True Negative untuk suatu class adalah jumlah semua record pada baris dan kolom kecuali baris dan kolom untuk class tersebut.

TABEL 1. CONFUSION MATRIX

		Kelas Prediksi	
		True	False
Kelas Sebenarnya	True	TP	FN
	False	FP	TN

Pengukuran Confusion Matrix terdapat accuracy, precision, dan recall. Accuracy merupakan rasio perdiksi benar dari positif dan negatif dengan keseluruhan data. Pengukuran accuracy dapat menggunakan rumus :

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad [2]$$

Precision merupakan rasio prediksi benar positif dibanding seluruh hasil prediksi positif. Pengukuran precision menggunakan rumus :

$$Precison = \frac{TP}{TP+FP} \quad [3]$$

Recall merupakan rasio prediksi benar positif di banding seluruh data yang benar positif. Pengukuran recall menggunakan rumus :

$$Recall = \frac{TP}{TP+FN} \quad [4]$$

1.8. Min-Max Scalling

Min-max scaling atau disebut juga dengan Min-max normalization adalah suatu metode normalisasi data dengan melakukan transformasi linier terhadap data asli sehingga menghasilkan keseimbangan nilai perbandingan antara data sebelum dan sesudah proses. Mix-max scaling digunakan untuk melakukan normalisasi data dengan rentang 0 sampai 1 bertujuan agar tidak adanya fitur yang memiliki nilai yang tidak seimbang [14].

$$Normalized(x)\ \frac{minRange+(x-minValue)(maxRange-minRange)}{maxValue-minValue} \quad [5]$$

Keterangan :

x = data

minRange = rentang nilai terendah setelah di normalisasi

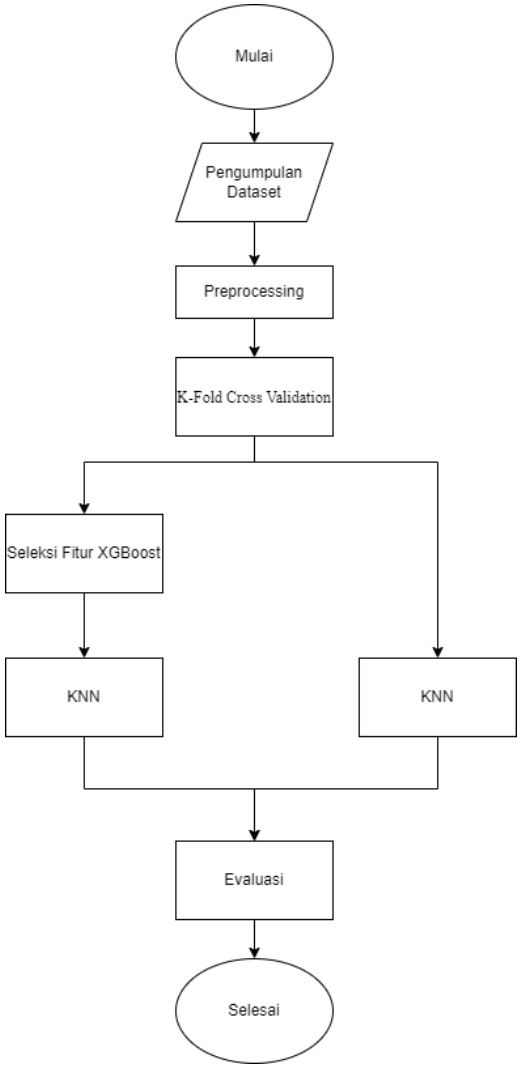
maxRange = rentang nilai tertinggi setelah di normalisasi

minValue = nilai terendah dari rentang nilai
maxValue = nilai tertinggi dari rentang nilai

2. METODOLOGI PENELITIAN

2.1. Skema Alur Penelitian

Beberapa Proses tahapan yang dilakukan pada penelitian ini yaitu pengumpulan, data yang dipakai pada penelitian ini adalah Breast Cancer Wisconsin (Diagnostic) berasal dari UCI Machine Learning. Setelah dataset didapat lalu dilakukan normalisasi. Data yang telah di normalisasi dilakukan pembagian data menggunakan K-Fold Cross Validation. Setelah data di bagi dilakukan perbandingan klasifikasi KNN dengan seleksi fitur dan tanpa seleksi fitur sehingga di evaluasi menggukanan Confusion Matrix untuk mendapatkan akurasi kinerja dari metode tersebut. Untuk alur diagram penelitian dapat dilihat pada gambar 1.

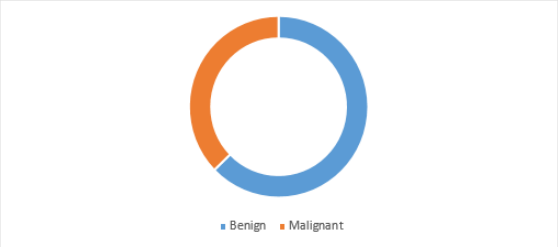


Gambar 1. Alur Penelitian

2.2. Pengumpulan Data

Data yang digunakan pada penelitian ini menggunakan Breast Cancer Wisconsin (Diagnostic) yang memiliki 32 fitur dengan 569 record. Terdapat 2 kelas yaitu 212 bersifat maligant dan 357 bersifat benign. pada fitur id karena hanya berisi nomor urut id dan tidak berpengagruh terhadap hasil maka dilakukan pengahpuas fitur id.

Pada dataset Breast Cancer Wisconsin (Diagnostic) memiliki perbandingan data antar kelas yaitu sebesar 62,74 % kelas Benign dan 37,26 % kelas Malignant seperti pada gambar 2.



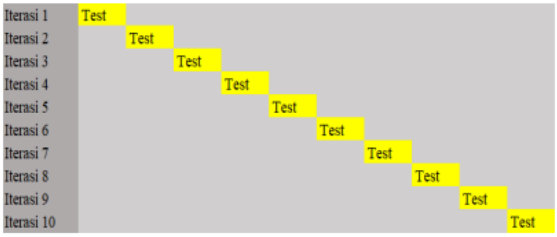
Gambar 2. Perbandingan Kelas

2.3. Preprocessing

Sebelum data digunakan, diperlukan preprocessing agar data memiliki nilai yang seimbang. Data yang memiliki nilai yang tidak seimbang dapat mempengaruhi kinerja sebuah model. Preprocessing yang digunakan yaitu MinMax Scaling. MinMax Scaling adalah metode dimana mengubah data menjadi rentang dari 0 dan 1 dengan cara mengurangan dari nilai sekarang dan nilai minimum lalu dibagi nilai maximum dikurang nilai minimum.

2.4. K-Fold Cross Validation

Pada tahap ini dilakukan pembagian data bertujuan agar menemukan kombinasi data yang terbaik dalam melatih model dengan membagi data latih dan data uji. Pembagian data menggunakan 10-Fold Cross Validation. Dengan nilai k tersebut maka rasio pembagian data latih 90% dan data uji 10%.



Gambar 3. Pembagian Data

Seperti pada gambar 3 diatas, cara kerja K-Fold Cross Validation adalah menentukan nilai k, maka dataset akan dibagi menjadi nilai k subset data. Jika nilai k = 10 maka terdapat 10 subset data. Lalu subset dibagi menjadi data latih dan data uji. Pada nilai k = 10 maka terdapat 9 subset data yang menjadi data latih dan 1 subset data yang menjadi data uji. Proses dilakukan secara berulang berdasar nilai k yang telah di tentukan dengan memilih data uji yang belum pernah di uji sebelumnya dari subset data hingga semua dari subset data menjadi data latih.

2.5. Seleksi Fitur Menggunakan XGBoost

Setelah dilakukan pembagian data latih dan data uji selanjutnya menyeleksi fitur menggunakan XGBoost. Penyeleksian fitur dengan metode XGBoost berdasarkan feature importance atau fitur yang berpengaruh bedasarkan metode XGBoost. Proses seleksi fitur dilakukan di setiap fold dalam 10-Fold Cross Validation pada data latih, hal ini dilakukan untuk menghindari adanya kebocoran data dari data uji (data leakage).

Pada tahap ini dilakukan seleksi fitur menggunakan metode wrapper. Seleksi fitur dilakukan dalam K-Fold Cross Validation karena metode seleksi fitur bergantung pada label atau kelas. Proses seleksi fitur dilakukakan menggunakan XGBoost pada sehingga mendapatkan subset yang memiliki ranking terbaik. Hasil yang didapat terdapat 10 subset fitur, 10 subset fitur tersebut akan dipilih pada data latih dan data uji dan fitur lainnya akan dihapus.

Parameter yang digunakan pada klasifikasi XGBoost yaitu menggunakan parameter default seperti pada tabel 2.

TABLE 2. PARAMETER XGBOOST

No	Parameter	Nilai
1	learning_rate	0,3
2	max_depth	6
3	base_score	0,5
4	n_estimator	100
5	gamma	0

Sebelum menemukan fitur importance dilakukan klasifiaksi menggunakan XGBoost. Klasifikasi dilakukan dengan cara mengacak nilai disalah satu fitur seperti pada gambar 3, lalu buat klasifikasi dengan model yang sama dan menggunakan kumpulan data yang dihasilkan. Gunakan akurasi klasifikasi ini dan nilai akurasi sebenarnya untuk menghitung berapa banyak fungsi kerugian yang didapat dari pengacakan. Penurunan kinerja itu mengukur fitur importance yang baru saja diacak. Kembalikan data ke urutan awal membatalkan shuffle dari langkah awal. Sekarang ulangi mengacak dengan nilai fitur berikutnya di dataset, sampai menghitung fitur importance setiap fitur.

radius_mean_minmax	texture_mean_minmax	perimeter_mean_minmax	area_mean_minmax
0,521	0,023	0,546	0,364
0,643	0,273	0,616	0,502
0,601	0,39	0,596	0,449
0,21	0,361	0,234	0,103
0,63	0,157	0,631	0,489
0,259	0,203	0,268	0,142
0,533	0,347	0,524	0,38
0,318	0,376	0,321	0,184
0,285	0,41	0,302	0,16
0,259	0,485	0,278	0,141
0,428	0,458	0,407	0,278
0,416	0,277	0,413	0,27
0,577	0,51	0,612	0,415

Gambar 4. Mengacak Nilai Fitur

2.6. Klasifikasi dan Evaluasi

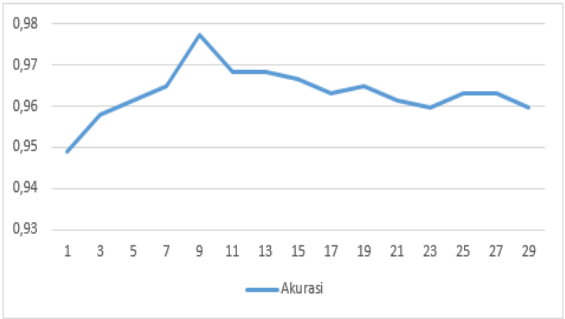
Dataset yang telah selesai dilakukan pembagian dan seleksi fitur menggunakan XGBoost dilanjutkan dengan proses klasifikasi pada K-Nearest Neighbor menggunakan fitur-fitur yang didapat pada metode XGBoost. Parameter yang digunakan pada klasifikasi KNN yaitu nilai ganjil k = 1 sampai k = 30.

Pada seleksi fitur dilakukan di dalam K-Fold Cross Validation dengan nilai K = 10 yang menggunakan XGBoost sehingga setiap iterasi fold mendapatkan 1 subset fitur. Dari 1 subset fitur tersebut terdiri dari beberapa kumpulan fitur lalu diklasifikasi menggunakan KNN.

Untuk mengetahui kinerja model maka dilakukan evaluasi menggunakan Confusion Matrix untuk mendapatkan akurasi dari kinerja metode.

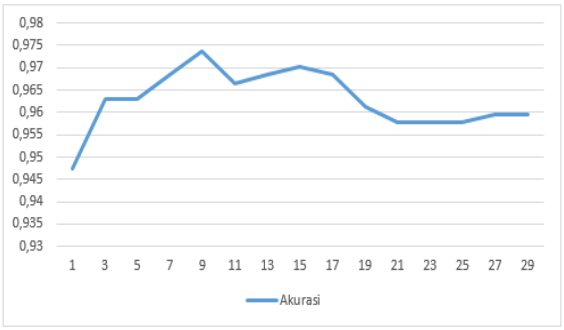
3. HASIL DAN PEMBAHASAN

Hasil dari seleksi fitur pada dataset Breast Cancer Wisconsin (Diagnostic) yaitu 10 subset fitur dari 1 subset fitur terdapat beberapa fitur. Sepuluh subset seleksi fitur tersebut didapat dari seleksi fitur XGBoost berdasarkan feature importance. Lalu dilakukan klasifikasi menggunakan KNN dengan nilai ganjil k=1 sampai k=30. Pada KNN dengan menggunakan seleksi fitur mendapatkan hasil akurasi rata-rata dari 10 fold yaitu 97,7 % seperti pada gambar 5.



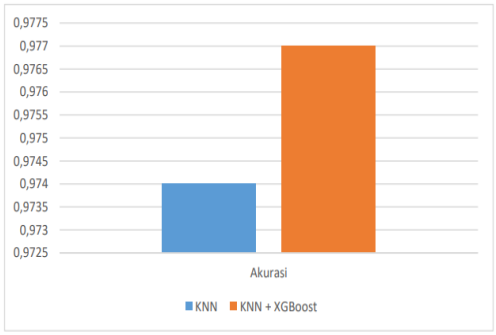
Gambar 5. Akurasi KNN Dengan Seleksi Fitur

Sedangkan pada KNN tanpa menggunakan seleksi fitur mendapatkan hasil rata-rata dari 10 fold yaitu 97,4% seperti pada gambar 6.



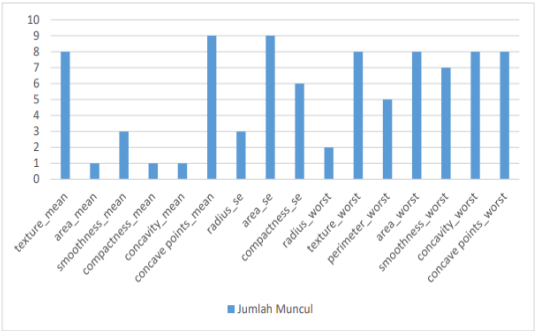
Gambar 6. Akurasi KNN Tanpa Seleksi Fitur

Sehingga mendapatkan perbandingan akurasi dari 2 model yaitu seperti pada gambar 7.



Gambar 7. Perbandingan Akurasi

Terdapat fitur-fitur yang sering muncul pada setiap fold karena fitur tersebut adalah fitur yang dianggap penting oleh metode XGBoost yaitu seperti pada gambar 8.

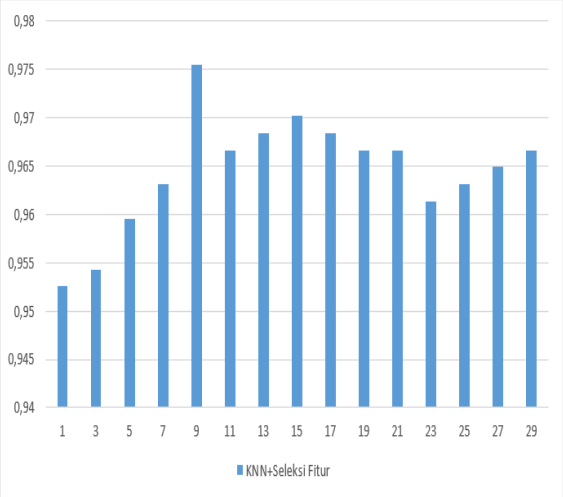


Gambar 8. Jumlah Kemunculan Fitur

Untuk mengetahui fitur yang di dapat berpengaruh dilakukan klasifikasi ulang menggunakan KNN dengan menggunakan beberapa fitur yang sering muncul pada seleksi

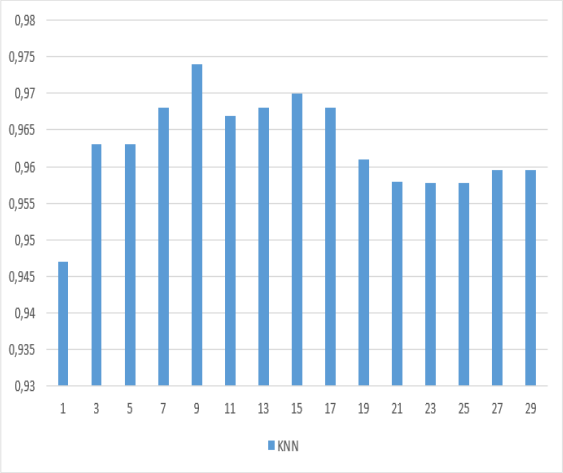
fitur yang digunakan pada XGBoost. Dari gambar 5 yang digunakan yaitu texture_mean, concave points_mean, area_se, compactness_se, texture_worst, perimeter_worst, area_worst, concave points_worst dan symmetry_worst sehingga mendapatkan hasil akurasi. Berikut perbandingan akurasi KNN dengan fitur yang telah dipilih dengan KNN dengan tanpa pemilihan fitur.

Pada gambar 9 KNN dengan seleksi fitur mendapatkan hasil akurasi 97,5%.



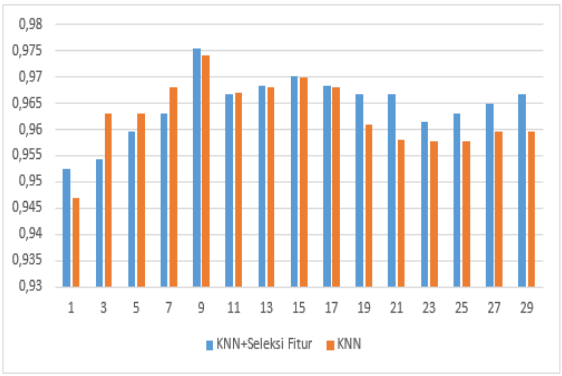
Gambar 9. Akurasi KNN Dengan 9 Fitur

Sedangkan KNN tanpa seleksi fitur mendapatkan 97,4% seperti pada gambar 10.



Gambar 10. Akurasi KNN tanpa seleksi fitur

Sehingga mendapatkan perbadingan akurasi seperti pada gambar 11. KNN dengan seleksi fitur lebih unggul dan memiliki fitur yang lebih sedikit.



Gambar 11. Perbandingan Akurasi

4. Kesimpulan dan Saran

Dari penelitian diatas dapat disimpulkan KNN dengang seleksi fitur XGBoost dapat mengungguli KNN tanpa seleksi fitur pada akurasi serta jumlah fitur yang digunakan. Akurasi yang didapat KNN dengan seleksi fitur XGBoost yang menggunakan 9 fitur mendapatkan 97,5% sedangkan KNN tanpa seleksi fitur mendapatkan 97,4%.

Adapun saran yang diberikan pada penelitian ini adalah pada penelitian selanjutnya menggunakan dataset dengan karakteristik yang berbeda atau menggunakan metode seleksi fitur lainnya.

5. UCAPAN TERIMA KASIH

Bagian ini menguraiakan ucapan terima kasih pada Dosen Ilmu Komputer FMIPA Universitas Lambung Mangkurat dan teman-teman yang mendukung penelitian ini.

Daftar Pustaka:

[1] Danukusumo, K. P., Implementasi Deep Learning Menggunakan Convolutional Neural Network Untuk Klasifikasi Citra Candi Berbasis Gpu. Universitas Atma Jaya Yogyakarta. 2017.

[2] Baharuddin, M. M., Tafsir H., and Huzain A., Analisis Performa Metode K-Nearest Neighbor Untuk Identifikasi Jenis Kaca. ILKOM Jurnal Ilmiah, Vol. 11, No. 3, pp 269-274.

[3] Pratama, Y.A., Analisis Metode Seleksi Fitur Untuk Meningkatkan Akurasi Pada Variant Metode Klasifikasi K-Nearest Neighbor (KNN). Universitas Sumatera Utara. 2019.

[4] Bode, A., K-Nearest Neighbor Dengan Feature Selection Menggunakan Backward Elimination Untuk Prediksi Harga Komoditi

Kopi Arabika. ILKOM Jurnal Ilmiah. Vol. 9, No. 2, pp. 188-195.
[5] Prasetio, R.T., Seleksi Fitur Dan Optimasi Parameter K-Nn Berbasis Algoritma Genetika Pada Dataset Medis. Jurnal Responsif, Vol. 2 No.2, pp. 213-221.
[6] Putri, L.A.A.R., Seleksi Fitur Dalam Klasifikasi Genre Musik. Jurnal Ilmiah ILMU KOMPUTER Universitas Udayana, Vol. 10, No.1, pp. 19-26.
[7] Manju, N., B S Harish and V Prajwal. Ensemble Feature Selection and Classification of Internet Traffic using XGBoost Classifier. I. J. Computer Network and Information Security, Vol. 11, No. 7, pp 37-44.
[8] Saygılı, A., Classification and Diagnostic Prediction of Breast Cancers via Different Classifiers. International Scientific And Vocational Journal, Vol. 2, No.2, pp 48-56.
[9] Karo, I. M. K., Implementasi Metode XGBoost dan Feature Importance untuk Klasifikasi pada Kebakaran Hutan dan Lahan. Journal of Software Engineering, Information and Communication Technology, Vol. 1, No. 1, pp 10-16.
[10] Mahardika, K. W., Yuita A. S. and Achmad A., Optimasi K-Nearest Neighbour Menggunakan Particle Swarm Optimization pada Sistem Pakar untuk Monitoring Pengendalian Hama pada Tanaman Jeruk. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, Vol. 2, No. 9, pp 3333-3344.
[11] Mardiansyah, H. Penanganan Masalah Data Kredit Untuk Kelas Tidakseimbang Menggunakan SmoteXGBoost. Universitas Sumatera Utara. 2019.
[12] Tama, B. A., Lewis N., S.M. Riazul Islam and Kwak K. S. An enhanced anomaly detection in web traffic using a stack of classifier ensemble. Vol. 8, pp 24120–24134.
[13] Xu, J., Yuanjian Z. and Duoqian M., Three-way confusion matrix for classification: A measure driven view. Information sciences, Vol. 507, pp 772-794.
[14] Nasution, D. A., H. H. Khotimah dan N. 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.
[15] Alkaromi, M. A., Information Gain Untuk Pemilihan Fitur Pada Klasifikasi Heregistrasi Calon Mahasiswa Dengan Menggunakan K-NN. Universitas Dian Nuswantoro. 2014.
[16] Ladha, L. and T. Deepa, Feature Selection Methods and Algorithms. International Journal on Computer Science and Engineering, Vol. 3, No. 5, pp. 1787-1797.
[17] Cai, J., Luo J., Wang S., Yang S., Feature selection in machine learning: a new perspective. Neurocomputing, Vol. 300, pp. 70-79.
[18] Fermansah, D., Penggunaan Metode Traditional Transformations Data Augmentation Untuk Peningkatan Hasil Akurasi Pada Model Algoritma Convolutional Neural Network (CNN) di Klasifikasi Gambar. Universitas Siliwangi. 2019.
[19] Khoiruddin, M., Klasifikasi Penyakit Daun Padi Menggunakan Convolutional Neural Network. Institut Teknologi Telkom Purwokerto. 2021.
[20] Leidiana, H., Penerapan algoritma K-Nearest Neighbor untuk penentuan resiko kredit kepemilikan kendaraan bemotor. PIKSEL: Penelitian Ilmu Komputer Sistem Embedded and Logic, Vol.1, No. 1, pp. 65-76.