JIKO (Jurnal Informatika dan Komputer)
Vol. 6, no. 1, Febuari 2022, hal. 1–5
e-ISSN : 2477-3964 — p-ISSN : 2477-4413

KOMPARASI KINERJA ALGORITMA XGBOOST DAN ALGORITMA SUPPORT VECTOR MACHINE (SVM) UNTUK DIAGNOSA PENYAKIT KANKER PAYUDARA

Muhammad Ravly Andryan[1],**Muhamad Fajri**[2] , dan **Nina Sulistyowati**[3]
[1,2,3]Teknik Informatika, Universitas Singaperbangsa Karawang
Jl. HS. Ronggo Waluyo, Telukjambe Timur, Karawang, Jawa Barat, Indonesia - 41361
Email: ravly.andryan18100@student.unsika.ac.id[1], muhammad.fajri18218@student.unsika.ac.id[2], nina.sulistio@unsika.ac.id[3]

Abstrak

Kanker payudara merupakan kanker yang sering ditemukan pada wanita di indonesia. kanker ini terjadi karena pertumbuhan neoplasma yang tidak normal yang berasal dari parenchyma. penyakit ini banyak ditemukan di negara-negara maju. Di Indonesia sendiri kanker payudara menempati peringkat kedua setelah kanker serviks. Kanker payudara dapat dideteksi secara dini melalui tumor pada payudara, umumnya dapat terbagi menjadi dua yaitu benign dan malignant. metode yang digunakan pada penelitian ini adalah Knowledge Data Discovery (KDD) dengan menggunakan algoritma XGBoost dan SVM, kemudian dilakukan klasifikasi untuk menentukan apakah kanker yang dianalisa itu bernilai benign atau malignant. Data yang digunakan dalam penelitian ini adalah data publik yang dirilis oleh UCI Machine Learning berjudul Wisconsin Breast Cancer Diagnostic. Hasil kinerja yang didapat setelah melakukan penelitian menggunakan kedua algoritma adalah Xgboost yang memiliki akurasi terbaik sebesar 95.12% dan nilai ROC_AUC sebesar 0.99 dan algoritma SVM memiliki akurasi terendah sebesar 90.24% dan nilai ROC_AUC sebesar 0.98.
Kata Kunci: *Ensamble learning, GridsearchCV, Kanker Payudara, KDD, Support Vector Machine, XGBoost*

Abstract

Breast cancer is a common cancer found by women in Indonesia. This cancer occurs due to the growth of abnormal neoplasms originating from the parenchyma. This disease is often found in developed countries. In Indonesia itself ranks second after cervical cancer. Breast cancer can be detected early through tumors in the breast, generally can be divided into two, namely benign and malignant. The method used in this study is Knowledge Data Discovery (KDD) using the XGBoost and SVM algorithms, then classification is carried out to determine whether the analyzed cancer is suitable for benign or malignant cancer. The data used in this study is public data released by UCI Machine Learning with the title Wisconsin Breast Cancer Diagnostic. The results obtained after conducting research using the second algorithm are Xgboost which has the best accuracy of 95.12% and the ROC_AUC value of 0.99 and the SVM algorithm has the lowest accuracy of 90.24% and the ROC_AUC value of 0.98.
KeyWords : *Breast Cancer, Ensemble learning, GridsearchCV, KDD, Support Vector Machine, XGBoost*

I. PENDAHULUAN

Kanker payudara merupakan kanker yang paling sering ditemukan pada wanita di indonesia. kanker ini terjadi karena pertumbuhan neoplasma yang tidak normal yang berasal dari parenchyma. penyakit ini banyak ditemukan di negara-negara maju [1]. Pada tahun 2020, terdapat 2,3 juta wanita yang terdiagnosis kanker payudara dan 685,000 kematian di seluruh dunia. Hingga akhir tahun 2020, ada 7,8 juta wanita hidup yang didiagnosis menderita kanker payudara dalam 5 tahun terakhir, menjadikan kanker payudara sebagai kanker paling umum di dunia. Tercatat angka kanker payudara yang dialami oleh wanita di Indonesia mencapai 42,1 per 100.000 penduduk dengan rata-rata kematian 17 per 100.000 penduduk. Melihat bahaya yang diakibatkan oleh kanker payudara maka sangat penting untuk mendiagnosa secara dini tumor yang terdapat di payudara. Oleh karena itu, tindakan analisa yang akan dilakukan menggunakan teknik data mining.Objek penelitian menggunakan data bersifat publik yang dikeluarkan oleh *UCI Machine Learning* dengan 569 *record* yang berisikan 30 atribut dari kanker payudara seperti *ID, Diagnosis*, dll. Hasil dari penilitian ini diharapkan agar dapat menjadi rujukan bagi peneliti selanjutnya dalam penggunaan metode *data mining*.

Kanker payudara dapat dideteksi secara dini melalui tumor pada payudara, umumnya dapat terbagi menjadi dua yaitu *benign* dan *malignant*. Dengan metode *data mining* menggunakan algoritma XGBoost dan SVM, akan dilakukan klasifikasi dari beberapa parameter apakah tumor yang dianalisa itu bernilai *benign* atau *malignant* dan akan dinilai dari segi *accuracy, recall, precision* dan *roc_auc*, lalu akan dibandingkan kinerja dari 2 algoritma tersebut. Objek penelitian menggunakan data bersifat publik yang dikeluarkan oleh *UCI Machine Learning* dengan 569 *record* yang berisikan 30 atribut dari kanker payudara seperti *ID, Diagnosis*, dll. Hasil dari penelitian ini diharapkan agar dapat menjadi rujukan bagi peneliti selanjutnya dalam penggunaan metode data mining klasifikasi.

Berdasarkan penelitian sebelumnya yang dilakukan oleh Ma'arif, F., & Arifin, T. (2017) bahwa Metode Algoritma *Support Vector Machine* (SVM) adalah Algoritma yang baik untuk klasifikasi Kanker Payudara menggunakan Dataset WBC (*Wisconsin Breast Cancer*), dimana nilai klasifikasi performansi Akurasi dan AUC nya adalah yang tertinggi diantara algoritma yang telah penulis uji, sedangkan untuk penggabungan algoritma seleksi fitur *Backward Elimination* dan *Support Vector Machine* (SVM)

mendapatkan peningkatan Akurasi sebesar 14% sehingga nilai tingkat akurasi akhirnya sebesar 97.14% dan nilai AUC mencapai 0.995 [2]. Penelitian yang dilakukan oleh Sinha, et al (2020) didapatkan akurasi klasifikasi terbaik untuk memprediksi kanker payudara, di mana pengklasifikasi Xgboost memberikan akurasi sebesar 98% [3]. Kemudian pada penelitian yang dilakukan oleh Prihartiwi et.al. (2021) bahwa Penyakit kanker payudara dapat diprediksi dengan menerapkan pengetahuan data mining. Dari hasil eksperimen diperoleh Algoritma *Support Vector Machine* menghasilkan tingkat *Accuracy* tertinggi yaitu sebesar 74,29% [4]. Hasil uji *Accuracy*, *Sensitivity*, PPV dan NPV pada Algoritma *Decision Tree* menunjukan bahwa Algoritma *Decision Tree* memiliki performa terburuk.

II. METODE

Pada penelitian ini, metode yang akan dipergunakan adalah metode *Knowledge Discovery in Databases* (KDD). *Knowledge Discovery in Database* (KDD) adalah penerapan metode saintifik pada *data mining*. *Data Mining* (DM) adalah inti dari proses KDD, melibatkan kesimpulan dari algoritma yang mengeksplorasi data, mengembangkan model dan menemukan pola yang sebelumnya tidak diketahui [5].Tahapan-tahapan dari KDD adalah sebagai berikut:

Data Selection

Pada tahap *Data Selection* data yang telah dikumpulkan kemudian dipilih berdasarkan atribut-atribut yang relevan untuk penelitian ini dengan melakukan penghapusan pada atribut-atribut yang dianggap tidak relevan. Data diperoleh dari data public yang berasal dari *kaggle dataset Wisconsin Breast Cancer Diagnostic* yang berisikan rata-rata, standard error dan ”worst” atau terbesar (rata-rata dari tiga nilai terbesar) dari fitur-fitur ini dihitung untuk setiap gambar, menghasilkan 30 atribut. Atribut.dapat dilihat pada Tabel I.

Tabel I: Daftar Atribut

Informasi Atribut	***ID Number***
Sepuluh fitur bernilai nyata dihitung untuk setiap inti sel	*Diagnosis (M= Malignant, B=Benign)*
	radius (rata-rata jarak dari pusat ke titik-titik pada keliling)
	texture (standar deviasi nilai skala abu-abu)
	perimeter
	area
	smoothnes (variasi lokal dalam panjang radius)
	compactness (perimeter2 / area - 1.0)
	concavity (tingkat keparahan bagian cekung dari kontur)
	concave points(jumlah bagian cekung dari kontur)
	symmetry
	fractal dimension

Data Preprocessing

Data preprocessing memainkan peran penting dalam *data mining* dan *machine learning* [6]. Pada tahap ini dilakukan perubahan *value* pada kolom “Diagnosis” dengan menggunakan metode *hot encoding* yaitu pemetaan dari value bernilai “M” menjadi 1 dan “B” menjadi 0 dan juga pengecekan *missing value*. Untuk proses *cleaning data* yang didapatkan sudah merupakan data yang bersih sehingga untuk proses untuk ini hanya dilakukan juga dilakukan pemilihan atribut yang penting-penting saja seperti membuang data *“ID” record*. Pembersihan data, juga disebut pembersihan atau penggosokan data, berkaitan dengan pendeteksian dan penghapusan kesalahan dan inkonsistensi dari data untuk meningkatkan kualitas data. Masalah kualitas data hadir dalam koleksi data tunggal, seperti *file* dan *database*, misalnya, karena salah ejaan selama entri data, informasi yang hilang atau data tidak *valid* lainnya [7].

Data Transformation

Pada tahap *data transformation* dilakukan *Principal Component Analysis* (PCA) untuk mereduksi jumlah atribut pada data, selanjutnya dilakukan *tuning* yaitu proses memaksimalkan kinerja algoritma tanpa *overfitting, underfitting*, atau menciptakan *varians* yang tinggi [8]. Proses *tuning* dilakukan dengan menggunakan *GridSearchCV*. Lalu dilakukan normalisasi data menggunakan *StandardScaler*. Metode normalisasi *StandardScaler* berkontribusi untuk menyesuaikan semua nilai atribut ke skala tertentu [9].

Data Mining

Proses *data mining* dilakukan dengan menggunakan 2 algoritma yaitu *Xgboost* (XGB) dan *Support Vector Machine Learning* (SVM). Selanjutnya kedua algoritma tersebut akan dilakukan komparasi menggunakan *ROC_AUC* untuk menemukan *insight* yang tersembunyi. *Receiver operating characteristics* (ROC) berguna untuk mengatur pengklasifikasi dan memvisualisasikan kinerjanya [10]. Algoritma *Support Vector Machine* (SVM) adalah alat prediksi klasifikasi dan regresi yang menggunakan teori pembelajaran mesin untuk memaksimalkan akurasi prediksi sambil secara otomatis menghindari kecocokan data yang berlebihan

[11]. Karena kesederhanaan dan fleksibilitasnya yang relatif untuk mengatasi berbagai masalah klasifikasi, SVM secara khusus memberikan kinerja prediksi yang seimbang, bahkan dalam studi di mana ukuran sampel mungkin terbatas [12]. Oleh karena itu algoritma SVM dipilih pada penelitian ini dikarenakan kesederhanaan dan fleksibilitasnya. *Xgboost* adalah kependekan dari *eXtreme Gradient Boosting package*. Algoritma *XGBoost* dapat mengatasi data medis yang kompleks dan beragam, dan dapat memenuhi persyaratan ketepatan waktu dan akurasi diagnosis tambahan dengan lebih baik [13]. Melihat sifat *XGBoost* yang dapat diskalakan, algoritma tersebut memiliki potensi besar untuk diterapkan secara luas pada tugas klasifikasi biner [14]. Berdasarkan alasan tersebut algoritma Xgboost dipilih pada penelitian ini dikarenakan potensinya yang besar untuk klasifikasi biner.

Interpretation/Evaluation

Berdasarkan hasil dari proses data mining yang telah dilakukan, didapat hasil dari kedua algoritma yang telah dievaluasi menggunakan ROC_AUC. Lalu akan ditemukan algoritma mana yang memiliki kinerja lebih baik.

III. HASIL

A. Data Selection

Data yang digunakan dalam penelitian ini adalah data publik yang dirilis oleh UCI *Machine Learning* berjudul *Wisconsin Breast Cancer Diagnostic*.Pada tahap ini dilakukan penghilangan kolom yang tidak diperlukan seperti "ID" dan "Unnamed 32".

B. Data Preprocessing

Data *value* pada kolom "Diagnosis" akan diubah dari "M" yang berarti *malignant* dan "B" *benign* menjadi angka 1 dan 0 untuk kemudahan memproses data. Lalu dilakukan pengecekan *missing value* dan ditemukan bahwa data tidak memilki *missing value*.

C. Data Transformation

Pada tahap ini dilakukan reduksi dari jumlah atribut dengan menggunakan metode *Principal Component Analysis* (PCA). Lalu dilakukan proses *tuning* menggunakan *GridsearchCV*. *Grid Search Cross Validation* (CV) telah digunakan untuk melatih dan mengoptimalkan model untuk memberikan hasil terbaik [15]. Pertama data dibagi menjadi dua yaitu data latih (80%) dan data uji (20%), lalu dilakukan pencarian secara menyeluruh menggunakan *GridsearchCV* terhadap parameter spesifik. Pada Tabel II dapat dilihat opsi-opsi argumen untuk tuning *hyperparameter* dengan tujuan mendapatkan model terbaik dengan mencoba semua kombinasi kemungkinan dari *tuning hyperparameter* yang telah dibuat. Kemudian untuk Tabel III adalah hasil untuk parameter dari atribut *best_params_* yang ada pada *GridSearchCV best_params_* ini didapat dari nilai *mean_validation_score* yaitu nilai rata-rata terbaik dari hasil uji validasi pada tahap *GridSearchCV*.

Tabel II: Daftar Hyperparameters

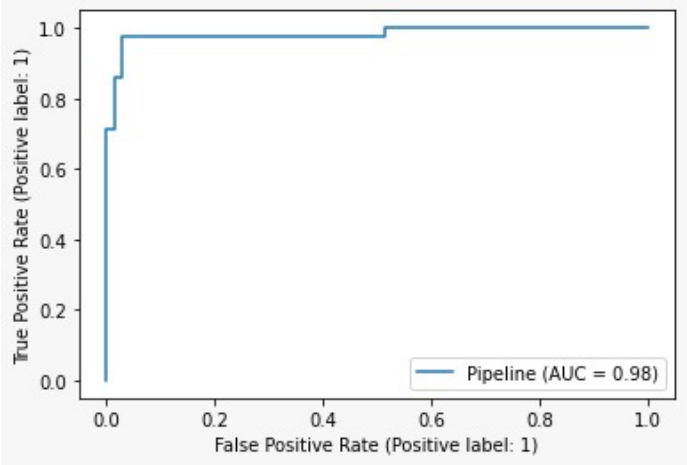
classifier	***hyper-parameters***	**Nilai**
SVM	*C*	[0.1,1, 10, 100]
	gamma	[1,0.1,0.01,0.001]
	kernel	[rbf, poly, sigmoid]
XGB	*max_dept*	[1, 2, 3]
	n estimator	[100, 150, 200, 250]
	learning_rate	[0.1, 0.2,... ,0.9]

Tabel III: Hasil dari GridsearchCV

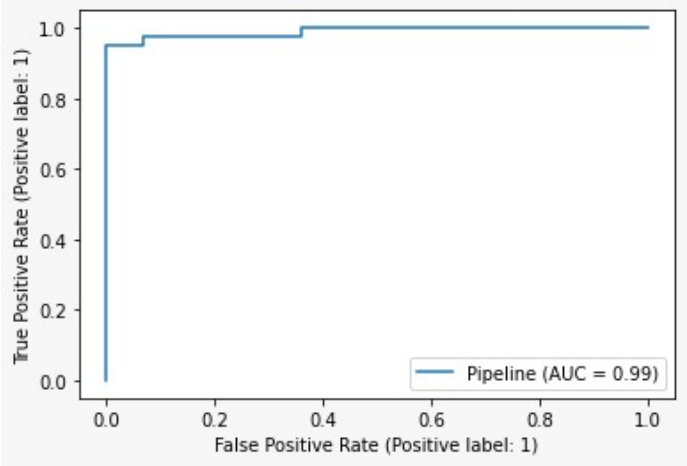
classifier	***hyper-parameters***	**Hasil**
SVM	*C*	10
	gamma	0,01
	kernel	rbf
XGB	*max_dept*	1
	n estimator	100
	learning_rate	0,5

D. Data Mining

Dilakukan proses data mining menggunakan algoritma *Xgboost* dan *Support Vector Machine*. Salah satu penilaian yang dilakukan adalah dengan menggunakan Kurva ROC_AUC. Kurva ini menunjukan bagan yang memvisualisasikan *tradeoff* antara *true positive rate* (TPR) dan *false positive rate* (FPR). Hasil dari kurva ROC_AUC SVM dapat dilihat pada Gambar 1 dan hasil dari kurva ROC_AUC *Xgboost* dapat dilihat pada Gambar 2.



Gambar 1: Kurva ROC_AUC SVM



Gambar 2: Kurva ROC_AUC Xgboost

Tabel IV: Hasil Komparasi Kinerja Algoritma

	accuracy	***precision***	***recall***	***roc_auc***
SVM	0,942857143	0,936738737	0,913445378	0,975800215
XGB	0,962637	0,969548	0,935798	0,986499

Hasil dari *accuracy, precision, recall, roc_auc* dapat dilihat pada Tabel IV. Terlihat dari hasil tersebut kinerja dari kedua algoritma. *Accuracy* adalah persentase total prediksi yang benar dibagi dengan jumlah total kejadian, *precision* (juga disebut nilai prediktif positif) adalah fraksi dari *instance* yang relevan di antara *instance* yang diambil, sedangkan *recall* (juga dikenal sebagai sensitivitas) adalah fraksi dari instance relevan yang diambil lalu yang terakhir terdapat roc_auc yang berarti nilai evaluasi *classifier*. Angka tersebut menunjukan nilai pada area di bawah kurva ROC_AUC.

E. Interpretation/Evaluation

Dari hasil pengujian pada tahap *data mining* terlihat bahwa kedua algoritma mempunyai nilai kinerja yang berbeda. Terdapat beberapa kategori penilaian seperti *accuracy, precision, recall* dan *roc_auc*. *Accuracy* adalah rasio antara sampel yang

terklasifikasi dengan baik (*true positives and true negatives*) dan jumlah total sampel yang terdapat pada data. Algoritma SVM memiliki nilai *accuracy* sebesar 94.28% dan *Xgboost* memiliki nilai *accuracy* 96.26%. Precision adalah rasio tp / (tp + fp) di mana tp adalah jumlah *true positive* dan fp adalah jumlah *false positive*.

Pada algoritma SVM memiliki nilai *precision* 0.93 dan algoritma *Xgboost* memiliki nilai *precision* sebesar 0.96. *Recall* adalah rasio tp / (tp + fn) di mana tp adalah jumlah *true positive* dan fn jumlah *false negative*. Recall secara intuitif adalah kemampuan untuk mengklasifikasikan dengan tujuan untuk menemukan semua sampel positif. Algoritma SVM memiliki nilai *recall* sebesar 0.91 dan untuk algoritma *Xgboost* memiliki nilai *recall* sebesar 0.93. Nilai roc_auc memiliki rentang nilai 0 hingga 1, dan mengurutkan kemungkinan prediksi. Nilai roc_auc untuk algoritma SVM adalah 0.975 dan untuk algoritma Xgboost bernilai 0.986. Data training yang menggunakan SVM mempunyai akurasi sebesar 94.28% sedangkan untuk algoritma *Xgboost* mempunyai akurasi sebesar 96.26% . Berdasarkan hasil kurva ROC pada Gambar 1 dan 2 kedua algoritma mempunyai nilai ROC_AUC diatas 0.98.

Setelah dilakukan uji menggunakan data testing ditemukan bahwa akurasi untuk algoritma SVM bernilai 90.24% dan untuk algoritma Xgboost memiliki akurasi bernilai 95.12%.

IV. Pembahasan

Berdasarkan hasil penelitian, terdapat perbedaan dengan rentang 0.01-0.03 dari kinerja kedua algoritma baik dari segi *accuracy, precision, recall* atau roc_auc. Kedua algoritma mempunya rentang akurasi sebesar 0.90 – 1.00 sehingga masuk kategori Excellent classification. Gambar 1 dan 2 menunjukan bahwa kurva ROC dari kedua algoritma mendekati sumbu Y(1,00) dengan demikian kinerja dari kedua algoritma untuk classifier dapat dibilang sangat baik [16] untuk membedakan jenis tumor M (*malignant*) dan B (*benign*).

V. Simpulan Dan Saran

Berdasarkan hasil analisis data dan pembahasan pada bab sebelumnya dapat diperoleh kesimpulan :

1) 1. Dari hasil penelitian diperoleh persentase kinerja dari masing – masing algoritma yaitu algoritma *Support Vector Machine* dengan akurasi sebesar 94.28% dengan *precision* bernilai 93.67%, recall memiliki nilai 91.34% dan roc_auc dengan nilai 97.58% dan algoritma Xgboost dengan akurasi 96.26% dengan *precision* bernilai 96.95%, *recall* memiliki nilai 93.57% dan roc_auc dengan nilai 98.64%.
2) Hasil kinerja terbaik yang didapat setelah melakukan penelitian menggunakan kedua algoritma adalah *Xgboost* yang memiliki akurasi terbaik sebesar 95.12% dan nilai AUC sebesar 0.99 dan SVM memiliki akurasi dan nilai AUC terendah bernilai 90.24% dan 0.98.

Berdasarkan hasil penlitian diatas maka didapat beberapa saran untuk penelitian selanjutnya :

1) Sebaiknya menggunakan data yang sudah terjamin kelayakannya untuk memudahkan proses pengolahan data.
2) Dilakukan pengembangan lebih lanjut pada tahap preprocessing data.
3) Melakukan penelitian dengan menggunakan metode lain ataupun metode hybrid.

Pustaka

[1] Fauzi, A., Supriyadi, R., & Maulidah, N. (2020). Deteksi Penyakit Kanker Payudara dengan Seleksi Fitur berbasis Principal Component Analysis dan Random Forest. Jurnal Infortech, 2(1), 96-101.
[2] Ma'arif, F., & Arifin, T. (2017). Optimasi Fitur Menggunakan Backward Elimination Dan Algoritma SVM Untuk Klasifikasi Kanker Payudara. Jurnal Informatika, 4(1).
[3] Sinha, N. K., Khulal, M., Gurung, M., & Lal, A. (2020). Developing a web based system for breast cancer prediction using xgboost classifier. International Journal of Engineering Research Technology (IJERT), 9.
[4] Prahartiwi, L. I., & Dari, W. (2021). Komparasi Algoritma Naive Bayes, Decision Tree dan Support Vector Machine untuk Prediksi Penyakit Kanker Payudara.
[5] Maimon, O., & Rokach, L. (2005). Introduction to knowledge discovery in databases. In Data mining and knowledge discovery handbook (pp. 1-17). Springer, Boston, MA.
[6] Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. IEEE Data Eng. Bull., 23(4), 3-13.
[7] Jiang, Y., Cukic, B., & Menzies, T. (2008, July). Can data transformation help in the detection of fault-prone modules?. In Proceedings of the 2008 workshop on Defects in large software systems (pp. 16-20).
[8] Bhattacharya, S., Maddikunta, P. K. R., Meenakshisundaram, I., Gadekallu, T. R., Sharma, S., Alkahtani, M., & Abidi, M. H. (2021). Deep neural networks based approach for battery life prediction. Computers, Materials & Continua, 69(2), 2599-2615.
[9] Paper, D., & Paper, D. (2020). Scikit-Learn Classifier Tuning from Complex Training Sets. Hands-on Scikit-Learn for Machine Learning Applications: Data Science Fundamentals with Python, 165-188.
[10] Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874.
[11] Jakkula, V. (2006). Tutorial on support vector machine (svm). School of EECS, Washington State University, 37.
[12] Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In Machine Learning (pp. 101-121). Academic Press.
[13] Li, S., & Zhang, X. (2020). Research on orthopedic auxiliary classification and prediction model based on XGBoost algorithm. Neural Computing and Applications, 32(7), 1971-1979.
[14] Wang, C., Deng, C., & Wang, S. (2020). Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost. Pattern Recognition Letters, 136, 190-197.
[15] Ranjan, G. S. K., Verma, A. K., & Radhika, S. (2019, March). K-nearest neighbors and grid search cv based real time fault monitoring system for industries. In 2019 IEEE 5th international conference for convergence in technology (I2CT) (pp. 1-5). IEEE.
[16] Handayani, A., Jamal, A., & Septiandri, A. A. (2017). Evaluasi Tiga Jenis Algoritme Berbasis Pembelajaran Mesin untuk Klasifikasi Jenis Tumor Payudara. Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI), 6(4), 394-403.