Implementation of Data Mining in Breast Cancer Diagnosis Classification using Logistic Regression Algorithm
[Implementasi Data Mining dalam Klasifikasi Diagnosa Kanker Payudara menggunakan Algoritma Logistic Regression]

Yahya Anugerah Dwi Khurrota A'yunan*[1)], Uce Indahyanti[2)]

[1,2)]Program Studi Teknik Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia
*Email : uceindahyanti@umsida.ac.id

Abstract. *Breast cancer is a very dangerous disease. It is considered as one of the most serious threats to women's health. To treat breast cancer, surgery and chemotherapy are two common approaches. It is important to diagnose breast cancer early to minimize the severity and increase the chance of cure. This study aims to classify breast cancer diagnoses using Logistic Regression. The data used is secondary data downloaded from Kaggle.com totaling 569 records. After going through pre-processing, the data that is ready to be processed is then divided into training and testing data with a ratio of 70%: 30%. This study resulted in an accuracy rate of 98% for predicting breast cancer patients after classification modeling and model testing using the confusion matrix method.*

Keywords *– Logistic Regression, Breast Cancer Diagnosis Classification, Confusion Matrix*

Abstrak. *Kanker payudara merupakan salah satu penyakit yang sangat berbahaya. Penyakit ini dianggap sebagai salah satu ancaman serius terhadap kesehatan perempuan. Untuk mengobati kanker payudara, pembedahan dan kemoterapi merupakan dua pendekatan yang umum dilakukan. Penting untuk mendiagnosis kanker payudara sejak dini guna meminimalisir tingkat keparahan dan meningkatkan peluang kesembuhan. Penelitian ini bertujuan untuk melakukan klasifikasi diagnosa kanker payudara menggunakan Logistic Regression. Data yang digunakan merupakan data sekunder yang diunduh dari Kaggle.com sebanyak 569 record. Setelah melalui pre processing, data yang telah siap diolah kemudian dibagi menjadi data training dan testing dengan perbandingan 70% : 30%. Penelitian ini menghasilkan tingkat akurasi sebesar 98% terhadap prediksi pasien kanker payudara setelah dilakukan pemodelan klasifikasi dan pengujian model menggunakan metode confusion matrix.*

Kata Kunci *– Regresi Logistik; Kalsifikasi Diagnosa Kanker Payudara; Confusion Matrix*

I. PENDAHULUAN

Kanker payudara adalah perkembangan sel yang tidak normal yang dimulai dari lapisan epitel saluran dan lobus. Selain itu, sel kanker payudara dapat menyerang lemak, pembuluh darah, dan saraf payudara, sehingga dapat dengan mudah menyebar ke beberapa bagian anggota tubuh lain apabila tidak cepat ditangani[1]. Prosedur dimulai ketika sel-sel di jaringan payudara saling tumbuh terlalu cepat. Kanker payudara adalah suatu bentuk kanker yang terjadi pada sel di payudara, yang dapat berasal dari komponen kelenjar seperti duktus dan epitel lobular, serta komponen non kelenjar seperti lemak, pembuluh darah, dan persarafan. Sel-sel ini biasanya membentuk tumor yang bisa dilihat melalui sinar-X atau terasa seperti benjolan di area payudara. Tanda-tanda yang sering muncul pada penderita adalah munculnya benjolan atau pengerasan pada payudara yang berbeda dengan jaringan sekitarnya dan terkadang juga dapat disertai dengan keluarnya darah dari puting payudara.Kanker payudara merupakan kanker yang sangat umum di kalangan wanita di dunia. Itu menyumbang 25% dari semua kasus kanker, dan mempengaruhi lebih dari 2,1 Juta orang pada tahun 2015 saja. Berdasarkan data GLOBOCAN[2] pada tahun 2012, International Agency for Research on Cancer (IARC) melaporkan terdapat 14.067.894 kasus kanker baru dan 8.201.575 kematian diakibatkan oleh penyakit kanker di seluruh dunia. Dilansir dari data tersebut dapat disimpulkan bahwa kanker payudara merupakan penyebab terbesar dari kasus baru kanker, yaitu mencapai 43,3% setelah dikontrol dengan umur, dan juga berkontribusi sebesar 12,9% pada kematian akibat kanker. Menurut data GLOBOCAN, persentase kematian akibat kanker payudara jauh lebih rendah dibandingkan dengan persentase kasus baru. Oleh karena itu, penemuan dan pencegahan kanker payudara pada tahap awal sangat penting untuk meningkatkan kemungkinan kesembuhan[3]. Menurut informasi yang disajikan oleh WHO (Organisasi Kesehatan Dunia), pada tahun 2018, kanker payudara menyumbang sebesar 42,5% dari total kasus kanker yang mencakup seluruh dunia dan sekitar 9,3 wanita meninggal setiap tahunnya karena hal ini[4].

Diagnosis dini penyakit kanker payudara dapat dilakukan melalui metode data mining. Proses ini bertujuan untuk menguraikan informasi baru dari dataset dan memanfaatkan teknik seperti statistik, matematika, artificial intelligence, dan machine learning untuk menemukan informasi dan pengetahuan yang berguna dalam database[5]. Dalam hal ini, klasifikasi kanker payudara memainkan peran penting dalam menentukan tingkat keparahan dan merumuskan rencana

pengobatan yang sesuai. Pada tahap awal, klasifikasi kanker payudara dapat dilakukan melalui pemeriksaan fisik, mammografi, atau biopsi, namun metode ini seringkali memiliki tingkat kesalahan dan kurang efektif. Oleh karena itu, penggunaan algoritma pembelajaran mesin seperti logistic regression sangat penting untuk meningkatkan akurasi dalam klasifikasi kanker payudara. Algoritma ini menggunakan rumus matematis untuk memprediksi probabilitas kanker payudara dan mengklasifikasikan kanker payudara ke dalam tingkatan yang berbeda[6].

Algoritma Regresi Logistik adalah salah satu konsep yang krusial dalam dunia statistik, analisis data, dan pembelajaran mesin[7]. Dalam konteks analisis data, regresi logistik digunakan untuk memprediksi probabilitas terjadinya suatu peristiwa atau kejadian berdasarkan variabel-variabel yang relevan[8]. Algoritma ini mendapatkan namanya dari bentuk fungsinya yang mirip dengan regresi linear, namun digunakan untuk masalah klasifikasi yang menghasilkan keluaran dalam bentuk probabilitas atau kelas-kelas diskret[9]. Kata binomial memiliki arti bahwa hanya terdapat dua kemungkinan hasil yang mungkin terjadi, tidak lebih dari itu[10]. Contohnya, 1 atau 0, ya atau tidak, hitam atau putih, dan lain sebagainya[11]. Oleh karena itu, logistic regression bisa diaplikasikan untuk menyelesaikan berbagai masalah yang berkaitan dengan klasifikasi (classification)[12].

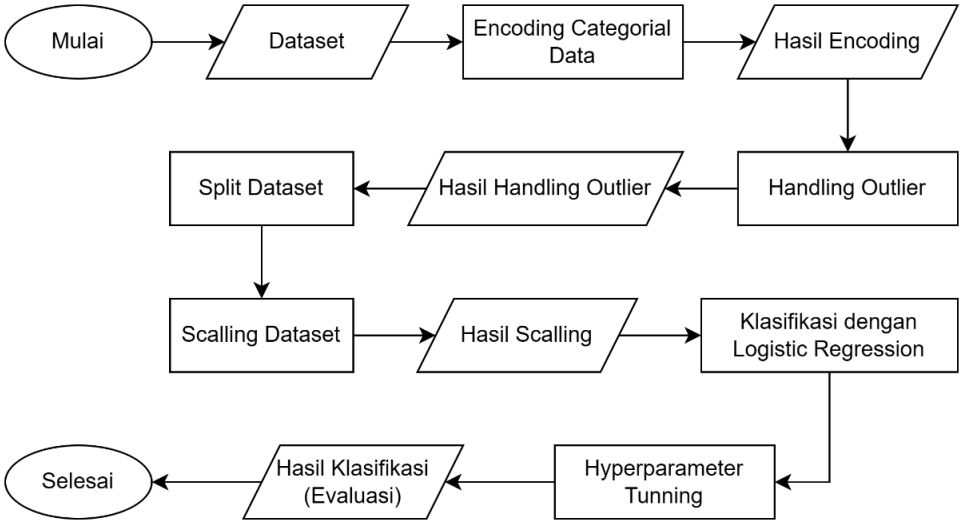
Beberapa penelitian sejenis sebelumnya telah dilakukan oleh Achmad 2022 dengan judul "Klasifikasi Breast Cancer Menggunakan Logistic Regression" yang mengklasifikasi breast cancer hanya menggunakan 120 data dari 569 data dan 2 atribut. Pada penelitian ini dilakukan menggunakan metode logistic regression sehingga didapatkan presentase hasil akurasi sebesar 76.04% untuk data latih dan sebesar 83.33% untuk data uji[13].Penelitian kanker payudara menggunakan metode K Nearest Neighboor dilakukan Atthalla, Jovandy, dan Habibie yang melakukan klasifikasi untuk menghitung jarak kemiripan. Hasil akhir penelitian yang dilakukan pada tahun 2018 menghasilkan nilai akurasi sebesar 93%[14].

Penelitian tentang analisis sentimen pada kasus Covid – 19 yang dilakukan oleh Santoso, Aloysius K.S, Astrid N, Aliyah K, Bagus D.W, Ahmad N pada tahun 2021. Data yang diambil merupakan hasil *scrapping* Twitter dengan mengelompokkan label negative, positif, dan netral. Dari hasil tersebut penelitian ini memakai metode Logistic Regression dan variasi *hyperparameter* L2 dan hyperparameter none. Sehingga hasil yang didapatkan untuk model terbaik adalah dengan menerapkan variasi *hyperparameter* L2[15].

Berdasarkan beberapa penelitian terakhir yang menggunakan algoritma logistic regression. Metode ini sangat efektif dan efisien jika ditambah dengan beberapa pengukuran kinerja dan beberapa evaluasi tambahan sehingga mendapat hasil yang lebih informatif untuk digunakan oleh penelitian lain lebih lanjut.Sehingga mencapai tingkat akurasi yang tinggi dapat diterapkan pada studi kasus data yang diambil oleh peneliti dan menggunakan bebrapa tambahan evaluasi kinerja yang membuat hasil lebih lebih sempurna. Dengan latar belakang tersebut, maka penelitian ini diusulkan guna dapat mengetahui kinerja dari metode logistic regression dalam mengklasifikasian kanker payudara menggunakan dataset Breast Cancer Wisconsin (Diagnostic), dan menggunakan teknik pengukuran kinerja Confusion Matrix.

II. METODE

Terhadap beberapa tahapan yang digunakan dalam penelitian ini. Adapun tahapan – tahapan tersebut adalah sebagai berikut:



Gambar 1. Flowchart Penelitian

Pada Gambar 1 merupakan langkah langkah penelitian akan dilakukan dalam klasifikasi diagnosa kanker payudara.

1. Pengambilan Dataset

   Dataset yang dipakai di penelitian ini merupakan data yang bersumber pada kaggle dengan nama *Breast Cancer Wisconsin (diagnostic)* (https://www.kaggle.com/data sets/yasserh/breast-cancer-dataset/code) pada tahun 2015. Dataset ini memiliki 569 isi data dan 31 kolom.

2. Preprocessing Data

   Dataset Breast Cancer Wisconsin (*diagnostic*) akan dianalisa terlebih dahulu bagaimana isi dan atribut – atributnya. Analisa ini dilakukan selain untuk mengetahui missing value dan tipe data juga untuk mendapatkan kolom yang merupakan target atau label[16]. Sebuah mesin tidak dapat memproses data yang bersifat kategorik. Sehingga pada tahap awal *preprocessing*, peneliti melakukan encoding terhadap data yang bersifat kategorik. Pada penelitian ini, kolom yang bersifat kategorik adalah kolom "diagnosis". Karena kolom "diagnosis" merupakan kolom target, maka peneliti melakukan Teknik *label encoder* untuk melakukan encoding. Pada tahap ini indikator diagnosis berisi B atau M maka akan diubah menjadi angka 0 dan 1. Setelah melalui tahapan encoding, selanjutnya yaitu tahap *handling outlier* dimana tahapan ini digunakan untuk menghilangkan data yang bernilai ekstrem atau *outlier*[17]. Data *outlier* merupakan data yang memiliki nilai jauh dari rata rata. Pada penelitian ini, peneliti menggunakan Teknik Z-Score yang memiliki aturan umum nilai yang kurang dari -3 atau lebih dari +3 menunjukkan bahwa nilai data adalah ekstrem[18]. Sehingga data yang melebihi batas bawah dan bataas atas akan dihapus. Dalam distribusi normal standar, Z-score menggambarkan seberapa jauh nilai pengamatan berjarak dari rata-rata dalam satuan deviasi standar. Sebagian besar (sekitar 99.7%) nilai dalam distribusi normal standar berada dalam rentang -3 hingga +3.

3. Processing Data

   Tabel 1 merupakan penentuan parameter yang akan digunakan untuk memanggil fungsi Grid Search agar dapat melakukan pencarian parameter yang optimal untuk model Logistic Regression. Pada parameter penalty peneliti menggunakan 2 regularisasi yaitu L1 dan L2. L1 akan memberikan hasil model dengan sparse (jarang), sedangkan L2 akan memberikan hasil model dengan bobot yang lebih merata[7]. Parameter C berguna untuk mengatur kekuatan regularisasi yang sudah disesuaikan nilainya dengan karakteristik data yang digunakan seperti jumlah sampel dan jumlah fitur pada data. Parameter solver yang dipilih `meupakan solver paling umum digunakan dalam regresi logistik[19].

Tabel 1. Hyperparameter tunning

Parameter	Value
Penalty	["l1","l2"]
C	[0.001, 0.009, 0.01, 0.09, 1, 5, 10, 25]
Solver	['lbfgs','newton-cg', 'liblinear', 'sag', 'saga']

4. Evaluasi

Setelah sukses menerapkan algoritma, model yang tercipta dievaluasi untuk menguji kinerjanya. Pada tahap penilaian digunakan sebuah *confusion matrix*, hal ini berusaha untuk mengungkapkan perbandingan hasil klasifikasi oleh model dengan hasil klasifikasi yang sebenarnya. *confusion matrix* diwakili oleh tabel dengan empat nilai berbeda yang berisi kombinasi nilai prediksi dan aktual pada tabel 2[20].

Tabel 2. Confusion Matrix

	Positif	**Negatif**
Positif	True Positif (TP)	False Positif (FP)
Negatif	False Negatif (FN)	False Negatif (FN)

III. HASIL DAN PEMBAHASAN

A. Pengambilan Dataset

Naskah Kolom diagnosis pada dataset merupakan target atau label dari dataset. Kolom diagnosis hanya memiliki 2 macam label, yaitu M (Malignant) untuk menyatakan diagnosa pada kanker ganas, dan B (Benign) untuk melabeli kanker jinak seperti pada Tabel 3. Namun kolom diagnosis akan dihilangkan sebelum melakukan pembagian data (*splitting data*), sehingga jumlah kolom atau atribut yang digunakan pada metode Logistic Regression adalah 30 kolom

saja. Pengambilan data dari kaggle dengan nama *Breast Cancer Wisconsin (diagnostic)* (https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset/code) pada tahun 2015. Dataset ini memiliki 569 isi data dan 31 kolom. Data disimpan berbentuk csv. Dataset awal seperti pada Tabel 3 masih berupa data awal sebelum dilakukan encoding categorial data.

Tabel 3. Dataset breast cancer

Atribut \ Responden	1	2	3	4
Diagnosis	M	M	M	B
radius_mean	17. 99	20.57	19.69	13.540
texture_mean	10.38	17.77	21.25	14.36
perimeter_meanM	122.8	132.9	130	87.46
area_mean	1001	1326	1203	566.3
smoothness_mean	0.1184	0.08474	0.1096	0.09779
compactness_mean	0.2776	0.07864	0.1559	0.08129
concavity_mean	0.3001	0.0869	0.1974	0.06664
concave points_mean	0.1471	0.07017	0.1279	0.04781
symmetry_mean	0.2419	0.1812	0.2069	0.1885
fractal_dimension_mean	0.07871	0.05667	0.05999	0.05766
radius_se	1.095	1.095	0.7456	0.2699
texture_se	0.9053	0.9053	0.9053	0.7886
perimeter_se	8.589	3.398	4.585	2.058
area_se	153.4	74.08	74.08	23.56
smoothness_se	0.006399	0.006399	0.00615	0.008462
compactness_se	0.00615	0.01308	0.04006	0.0146
concavity_se	0.05373	0.0186	0.03832	0.02387
concave points_se	0.01587	0.01587	0.02058	0.01315
symmetri_se	0.03003	0.01389	0.0225	0,0198
fractal_dimension_se	0.006193	0.003532	0.004571	0,0023
radius_worst	25.38	24.99	23.57	15,11
texture_ worst	17.33	23.41	25.53	19,26
perimeter_ worst	184.6	158.8	152.5	99,7
area_ worst	2019	1956	1709	711,2
smoothness_ worst	0.1622	0.1238	0.1444	0,144
compactness_ worst	0.6656	0.1866	0.4245	0,1773
concavity_ worst	0.7119	0.2416	0.4504	0,239
concave points_ worst	0.2654	0.186	0.186	0,1288
symmetri_ worst	0.4601	0.275	0.3613	0,2977
fractal_dimension_ worst	0.1189	0.08902	0.08758	0,07259

B. Preprocessing

Dataset tidak langsung digunakan oleh sistem. Oleh karena itu, beberapa preprocessing harus dilakukan untuk sedikit memodifikasi data guna meningkatkan kualitas data yang digunakan.

1) Encoding Categorial Data

Data awal pada kolom diagnosis isinya masih bersifat kategorial yaitu B dan M kemudian diubah menjadi numerik 0 dan 1 menggunakan *library scikit-learn.* LabelEncoder diinstantiate menjadi objek bernama "labelencoder". Kemudian, kolom "diagnosis" dalam sebuah dataframe (df) akan diberikan nilai numerik dengan memanggil metode "fit_transform" dari objek "labelencoder". Setelah itu, nilai numerik baru tersebut akan disimpan ke dalam kolom "diagnosis" yang sama dalam dataframe (df). Untuk hasil encoding categorial data dapat dilihat pada tabel 4.

Tabel 4. Dataset setelah encoding

Atribut \ Responden	1	2	3	4
Diagnosis	1	1	1	0
radius_mean	17. 99	20.57	19.69	13.540
texture_mean	10.38	17.77	21.25	14.36
perimeter_meanM	122.8	132.9	130	87.46
area_mean	1001	1326	1203	566.3
smoothness_mean	0.1184	0.08474	0.1096	0.09779
compactness_mean	0.2776	0.07864	0.1559	0.08129
concavity_mean	0.3001	0.0869	0.1974	0.06664
concave points_mean	0.1471	0.07017	0.1279	0.04781
symmetry_mean	0.2419	0.1812	0.2069	0.1885
fractal_dimension_mean	0.07871	0.05667	0.05999	0.05766
radius_se	1.095	1.095	0.7456	0.2699
texture_se	0.9053	0.9053	0.9053	0.7886
perimeter_se	8.589	3.398	4.585	2.058
area_se	153.4	74.08	74.08	23.56
smoothness_se	0.006399	0.006399	0.00615	0.008462
compactness_se	0.00615	0.01308	0.04006	0.0146
concavity_se	0.05373	0.0186	0.03832	0.02387
concave points_se	0.01587	0.01587	0.02058	0.01315
symmetri_se	0.03003	0.01389	0.0225	0,0198
fractal_dimension_se	0.006193	0.003532	0.004571	0,0023
radius_worst	25.38	24.99	23.57	15,11
texture_ worst	17.33	23.41	25.53	19,26
perimeter_ worst	184.6	158.8	152.5	99,7
area_ worst	2019	1956	1709	711,2
smoothness_ worst	0.1622	0.1238	0.1444	0,144
compactness_ worst	0.6656	0.1866	0.4245	0,1773
concavity_ worst	0.7119	0.2416	0.4504	0,239
concave points_ worst	0.2654	0.186	0.186	0,1288
symmetri_ worst	0.4601	0.275	0.3613	0,2977
fractal_dimension_ worst	0.1189	0.08902	0.08758	0,07259

2) Handling Outlier

Data hasil dari encoding data tadi selanjutnya akan di proses dalam handling outlier menggunakan metode Z-score. Script yang digunakan sebagai berikut :

```
for col in cols:
    zscore = abs(stats.zscore(df[col]))
    filtered_entries = (zscore < 3) &
filtered_entries

df = df[filtered_entries]
```

Dari script tersebut memproses data yang berjumlah 569 kemudian difilter menjadi 495 dikarenakan nilai yang diluar rentang dari -3 sampai 3 akan dibuang dan tidak akan digunakan.

3) Split Dataset

Data yang telah di outlier akan dipisah menjadi 70% data train dan 30% data latih. Pada bagian data X disini kolom diagnosis dihilangkan. Sedangkan pada data Y dibiarkan tetap seperti awal.

```
x_train, x_test, y_train, y_test =
train_test_split(X, y, test_size=.3,
random_state=42)
```

4) Scalling Dataset

Pada proses scalling dataset, dataframe dibagi menjadi dua yaitu X dan Y, kemudian melakukan transformasi pada data fitur (X) dengam memanggil fungsi "StandarScaller" dari *library sklearn.* Fungsi StandardScaler akan mengubah data fitur (X) dengan mengatasi perbedaan skala antar fitur dengan membawa setiap fitur ke skala yang sama. Ini bertujuan untuk menghindari perbedaan bobot pada algoritma machine learning yang dipengaruhi oleh skala fitur yang berbeda. Fungsi ini mengatasi perbedaan skala dengan membawa setiap fitur ke skala normal (mean = 0 dan standard deviation = 1).

```
    from sklearn.preprocessing import MinMaxScaler,StandardScaler
    # Transform features by scaling each feature to a given range
    X = StandardScaler().fit_transform(X)
    X = pd.DataFrame(X)
    X.head()
```

C. Klasifikasi dan Evaluasi

Langkah selanjutnya dilakukan import beberapa fungsi dari pustaka sklearn.metrics, yaitu confusion_matrix, roc_auc_score, classification_ report, dan accuracy_score. Kemudian, model lgr disesuaikan dengan parameter terbaik yang didapatkan dari grid_result_lgr menggunakan metode set_params(). Model lgr dilatih dengan menggunakan data latih (x_train, y_train), dan kemudian dilakukan prediksi pada data uji (x_test) menggunakan model yang telah dilatih. Hasil prediksi tersebut disimpan dalam variabel y_pred_lgr. Selanjutnya, dilakukan pencetakan hasil evaluasi model, dimulai dengan classification_report yang mencetak laporan klasifikasi yang mencakup presisi, recall, f1-score, dan support untuk setiap kelas target. Kemudian, confusion_matrix mencetak matriks kebingungan yang menunjukkan jumlah prediksi yang benar dan salah untuk setiap kelas target. Selanjutnya, mencetak skor ROC-AUC dengan menggunakan roc_auc_score yang menghitung area di bawah kurva ROC. Terakhir, mencetak skor akurasi dengan menggunakan accuracy_score yang menghitung persentase prediksi yang benar.

```
From sklearn.metrics import
confusion_matrix, roc_auc_score,
classification_report, accuracy_score
lgr =
lgr.set_params(**grid_result_lgr.best_params_)
lgr.fit(x_train, y_train)
y_pred_lgr = lgr.predict(x_test)

print(classification_report(y_test, y_pred_lgr))
print(confusion_matrix(y_test, y_pred_lgr))
print(f'ROC-AUC score : {roc_auc_score(y_test,
y_pred_lgr)}')
print(f'Accuracy score : {accuracy_score(y_test,
y_pred_lgr)}')
```

Script diatas akan memunculkan output sebagai berikut:

Tabel 5. Output confusion matrix

	Precision	Recall	F1-score	support
0	0.98	0.99	0.98	93
1	0.98	0.96	0.97	56
Accuracy			0.98	149
Macro avg	0.98	0.98	0.98	149
Weighted avg	0.98	0.98	0.98	149

.

VII. SIMPULAN

Kesimpulan penelitian ini menggunakan algoritma regresi logistik untuk membangun model klasifikasi kanker payudara. Klasifikasi tersebut diharapkan dapat menjadi prediksi awal kondisi kanker payudara dan selanjutnya berkembang menjadi prediksi awal penyakit pernapasan. Dari hasil dan proses evaluasi yang menggunakan teknik pengukuran kinerja confusion matrix, model klasifikasi diperoleh tingkat akurasi yang sangat baik yaitu 98%. Penelitian ini dapat dikembangkan menggunakan data privat yang lebih banyak dan menambahkan algoritma klasifikasi lainnya untuk membandingkan tingkat akurasi yang diperoleh.

UCAPAN TERIMA KASIH

Terimakasih kepada Universitas Muhammadiyah Sidoarjo yang telah memberikan dukungan pada penelitian ini.

REFERENSI

[1] A. Suyanto, *Data Mining in Early Diagnosis of Breast Cancer*. Journal of Medical Systems, 2017.
[2] Kemenkes RI., "Infodatin. Bulan Peduli Kanker Payudara Jakarta Kemenkes RI.," *Jakarta Selatan, Indones. Kementeri. Kesehat. Republik Indones.*, pp. 1–17, 2016.
[3] E. Susilowati, A. T. Hapsari, M. Efendi, and P. Edi, "Diagnosa Penyakit Kanker Payudara Menggunakan Metode K - Means Clustering," *J. Sist. Informasi, Teknol. Inform. dan Komput.*, vol. 10, no. 1, pp. 27–32, 2019.
[4] I. Mubarog, A. Setyanto, and H. Sismoro, "Sistem Klasifikasi Pada Penyakit Breast Cancer Dengan Menggunakan Metode Naïve Bayes," *Creat. Inf. Technol. J.*, vol. 6, no. 2, p. 109, 2021, doi: 10.24076/citec.2019v6i2.246.
[5] Suyanto, *Data mining untuk klasifikasi dan klasterisasi data*. Bandung: Informatika Bandung, 2017.
[6] N. Meilani and O. Nurdiawan, "Data Mining untuk Klasifikasi Penderita Kanker Payudara Menggunakan Algoritma K-Nearest Neighbor," *J. Wahana Inform.*, vol. 2, no. 1, pp. 177–187, 2023, [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer.
[7] M. I. Gunawan, D. Sugiarto, and I. Mardianto, "Peningkatan Kinerja Akurasi Prediksi Penyakit Diabetes Mellitus Menggunakan Metode Grid Seacrh pada Algoritma Logistic Regression," *J. Edukasi dan Penelit. Inform.*, vol. 6, no. 3, p. 280, 2020, doi: 10.26418/jp.v6i3.40718.
[8] A. Bimantara and T. A. Dina, "Klasifikasi Web Berbahaya Menggunakan Metode Logistic Regression," *Annu. Res. Semin.*, vol. 4, no. 1, pp. 173–177, 2019, [Online]. Available: https://seminar.ilkom.unsri.ac.id/index.php/ars/article/view/1932.
[9] G. P. PB, "Klasifikasi Persetujuan Permohonan Pinjaman Pada Koperasi Simpan Pinjam Menggunakan Algoritma Logistic Regression," *J. Ilmu Data*, vol. 2, no. 12, pp. 1–12, 2022, [Online]. Available: http://ilmudata.org/index.php/ilmudata/article/view/281%0Ahttp://ilmudata.org/index.php/ilmudata/article/download/281/270.
[10] F. M. Faruk, F. M. Faruk, F. S. Doven, and B. Budyanra, "Penerapan Metode Regresi Logistik Biner Untuk Mengetahui Determinan Kesiapsiagaan Rumah Tangga Dalam Menghadapi Bencana Alam," *Semin. Nas. Off. Stat.*, vol. 2019, no. 1, pp. 379–389, 2020, doi: 10.34123/semnasoffstat.v2019i1.146.
[11] N. G. Ramadhan, F. D. Adhinata, A. J. T. Segara, and D. P. Rakhmadani, "Deteksi Berita Palsu Menggunakan Metode Random Forest dan Logistic Regression," *JURIKOM (Jurnal Ris. Komputer)*, vol. 9, no. 2, p. 251, 2022, doi: 10.30865/jurikom.v9i2.3979.
[12] A. K. A. I, F. Nurhadi, I. K. O. Setiawan, I. A. Rizky, and R. B. Manurung, "Pengaruh Normalisasi Data pada Klasifikasi Harga Ponsel Berdasarkan Spesifikasi Menggunakan Klasifikasi Naive Bayes dan Multinomial Logistic Regression," *J. Rekayasa Elektro Sriwij.*, vol. 3, no. 1, pp. 8–16, 2022.
[13] A. D. Achmad, "KLASIFIKASI BREAST CANCER MENGGUNAKAN METODE LOGISTIC REGRESSION," vol. 9, no. 1, 2022.
[14] I. N. Atthalla, A. Jovandy, and H. Habibie, "Klasifikasi Penyakit Kanker Payudara Menggunakan Metode K Nearest Neighbor," *Pros. Annu. Res. Semin.*, vol. 4, no. 1, pp. 148–151, 2018.
[15] A. K. Santoso, A. Noviriandini, A. Kurniasih, B. D. Wicaksono, and A. Nuryanto, "Klasifikasi Persepsi Pengguna Twitter Terhadap Kasus Covid-19 Menggunakan Metode Logistic Regression," *JIK (Jurnal Inform. dan Komputer)*, vol. 5, no. 2, pp. 234–241, 2021.
[16] R. Nofitri and N. Irawati, "Analisis Data Hasil Keuntungan Menggunakan Software Rapidminer," *JURTEKSI (Jurnal Teknol. dan Sist. Informasi)*, vol. 5, no. 2, pp. 199–204, 2019, doi: 10.33330/jurteksi.v5i2.365.
[17] A. Saleh and F. Nasari, "Penerapan Equal-Width Interval Discretization Dalam Metode Naive Bayes Untuk

Meningkatkan Akurasi Prediksi Pemilihan Jurusan Siswa (Studi Kasus: Mas Pab 2 Helvetia,Medan)," *Masy. Telemat. Dan Inf. J. Penelit. Teknol. Inf. dan Komun.*, vol. 8, no. 1, p. 1, 2018, doi: 10.17933/mti.v8i1.98.
[18] N. Barkah, E. Sutinah, and N. Agustina, "Metode Asosiasi Data Mining Untuk Analisa Persediaan Fiber Optik Menggunakan Algoritma Apriori," *J. Kaji. Ilm.*, vol. 20, no. 3, pp. 237–248, 2020, doi: 10.31599/jki.v20i3.288.
[19] O. I. Desanti, I. Sunarsih, and Supriyati, "Persepsi Wanita Berisiko Kanker Payudara Tentang Pemeriksaan Payudara Sendiri Di Kota Semarang, Jawa Tengah," *Ber. Kedokt. Masy.*, vol. 26, no. 3, pp. 152–161, 2010.
[20] A. Alharthi, Abdulrahman ; Al-Mutairi, "Performance evaluation of classification models using confusion matrix," *Int. J. Adv. Comput. Sci. Appl.*, pp. 427–432, 2020.

Conflict of Interest Statement:
The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.