1. PENDAHULUAN

Pada wanita, kanker payudara merupakan salah satu kanker yang paling banyak ditemukan dan terus meningkat setiap tahunnya. Terbukti pada penelitian sebelumnya yang mengulas masalah ini (Dewi, 2017). Kanker payudara merupakan penyakit yang tidak menular. Penyakit kanker payudara sampai saat masih belum diketahui apa yang menjadi penyebabnya, namun dapat digaris bawahi bahwa penyebab penyakit ini bersifat multifakturial yang saling mempengaruhi. Ada penelitian yang membahas tentang faktor risiko gaya hidup yang berpengaruh terhadap penyakit kanker payudara pada wanita di Rumah Sakit Kota Makassar. Dari penelitian tersebut dihasilkan bahwa mengkonsumsi lemak, obesitas, merokok, dan stress merupakan faktor risiko yang berpengaruh terhadap penyakit kanker payudara. Stres merupakan faktor risiko yang paling berpengaruh, dimana mereka yang stress akan terkena kanker payudara 2,698 kali dari yang tidak stress (Maria et al., 2017).

World Health Organization (WHO) dan *American Cancer Society* memberikan informasi tentang data kanker pada tahun 2018. Sebanyak 18 juta kasus baru terdiagnosis pada tahun 2018, yang paling banyak adalah kanker paru-paru (2,09 juta kasus), payudara (2,09 juta kasus), dan prostat (1,28 juta kasus). Kematian, kanker merupakan penyebab kematian kedua di seluruh dunia (8,97 juta kematian) setelah penyakit jantung iskemik, tetapi kemungkinan besar akan menjadi yang pertama pada tahun 2060 (~ 18,63 juta kematian) (Mattiuzzi & Lippi, 2019). Pada tahun 2011, survei terhadap 187 negara dengan tingkat kematian dan insiden kanker payudara dari tahun 1980 hingga tahun 2010 menunjukkan bahwa insiden kanker payudara global meningkat dari 641.000 kasus pada tahun 1980 menjadi 1.643.000 kasus pada tahun 2010, dengan tingkat peningkatan tahunan rata-rata 3,1% (Wang et al., 2018).

Kanker payudara merupakan penyakit yang tidak menular yang paling banyak menyerang wanita dengan angka kematian tinggi dan terus meningkat setiap tahun. Diagnosis penyakit memainkan peran penting dalam menentukan strategi pengobatan, yang sangat terkait dengan keselamatan pasien. Untuk itu, diperlukkan suatu *machine learning* untuk memprediksi suatu penyakit. *Machine learning* digunakan untuk menangani data dengan lebih efisien, keuntungan utama menggunakan *machine learning* adalah suatu algoritma dapat mempelajari suatu data dan algoritma tersebut dapat dengan otomatis melakukan tuganya (Dey, 2016). Analisi data kanker payudara ini peneliti menggunakan algoritma klasifikasi *Random Forest* (RF).

Pemilihan algoritma klasifikasi ini karena teknik klasifikasi digunakan untuk memprediksi nilai keputusan dari variabel kelas untuk jenis variabel kualitatif atau kategori dengan perhitungan dengan satu atau lebih variabel independen atau prediktor. Ini diterapkan secara luas di bidang-bidang seperti ilmu komputer (struktur data), kedokteran (diagnosis), botani (klasifikasi), dan psikologi (teori keputusan) (Jaiswal & Samikannu, 2017)

Random Forest (RF) dipilih dalam mengolah dataset kanker payudara karena memiliki banyak keuntungan yaitu (1) dapat menangani secara efektif database besar, (2) dapat menangani ribuan variabel input tanpa menghapus variabel, (3) menciptakan estimasi internal yang tidak bias dari kesalahan umum, (4) dapat memperkirakan pentingnya setiap variabel untuk klasifikasi, (5) mendemonstrasikan kinerja yang kuat dan akurat pada kumpulan data kompleks (Dou et al., 2019).

Terdapat penelitian sebelumnya tentang klasifikasi dan prediksi diagnostik kanker payudara menggunakan algoritma *Support Vector Machines* (SVM), K-NN, *Naive Bayes*, J48, *Random Forest* (RF) and *multilayer perceptron methods*. Penelitian ini menghasilkan bahwa *Random Forest* (RF) merupakan metode yang paling berhasil dengan nilai akurasi 98,77%. Metode tersukses kedua adalah *multilayer perceptron methods* dengan nilai akurasi 98,41% (Saygili, 2018).

Tujuan dari penelitian ini adalah

2. METODE

2.1 *Data Collection*

Pengumpulan data merupakan tahap awal dalam sebuah penelitian. Dalam proses data mining terdapat variabel dan atribut yang memudahkan dalam proses penelitian. Data yang didapat dari penyakit kanker payudara ini ada 569 data dan 32 atribut didalamnya. Terdapat ID number, Diagnosis, dan 10 atribut yang dibagi lagi menjadi 3 : *mean*, *standard error*, dan *worst* atau *largest.* Misalnya pada atribut radius menjadi Mean Radius, Radius SE, dan Worst Radius. Berikut adalah deskripsi atribut dan variable yang ditunjukkan pada Tabel 1.

Tabel 1. Variabel dan Atribut Data Penyakit Kanker Payudara

Variabel	Atribut	Deskripsi
X1	*ID number*	ID number dari data
X2	*Radius_mean*	Rata-rata jarak dari pusat ke titik-titik di sekeliling
X3	*Texture_mean*	Standar deviasi dari nilai skala abu-abu
X4	*Perimeter_mean*	Ukuran rata-rata tumor inti

X5	*Area_mean*	
X6	*Smoothness_mean*	Variasi lokal dalam Panjang radius
X7	*Compactness_mean*	(keliling^ 2 / luas - 1,0)
X8	*Concavity_mean*	Tingkat keparahan bagian cekung dari kontur
X9	*Concave points_mean*	Jumlah bagian cekung dari kontur
X10	*Symmetry_mean*	
X11	*Fractal dimension_mean*	
X12	*Radius_se*	
X13	*Texture_se*	
X14	*Perimeter_se*	
X15	*Area_se*	
X16	*Smoothness_se*	
X17	*Compactness_se*	
X18	*Concavity_se*	
X19	*Concave point_se*	
X20	*Symmetry_se*	
X21	*Fractal dimension_se*	
X22	*Radius_worst*	
X23	*Texture_worst*	
X24	*Perimeter_worst*	
X25	*Area_worst*	
X26	*Smoothness_worst*	
X27	*Compactness_worst*	
X28	*Concavity_worst*	
X29	*Concave point_worst*	
X30	*Symmetry_worst*	
X31	*Fractal dimension_worst*	
Y	*Diagnosis*	M = Ganas, B = Jinak

Data ini merupakan data valid yang didapat dari UCI dengan judul "Breast Cancer Wisconsin (Diagnostic)" mengenai penyakit kanker payudara. Data kanker payudara ini dibuat oleh Dr. William H. Wolberg dari Departemen Bedah Umum Universitas Wisconsin, W. Nick Street dari Departemen Ilmu Komputer Universitas Wisconsin, dan Olvi L. Mangasarian, dari Jurusan Ilmu Komputer Universitas Wisconsin.

2.2 *Data Preprocessing*

A. *Data Splitting*

Setelah mendapatkan dataset, tahap selanjutnya melakukan proses *splitting* pada dataset penyakit kanker payudara. *Data splitting* adalah desain studi yang banyak digunakan dalam dataset berdimensi tinggi dan dimungkinkan untuk membagi dataset asli yang tersedia menjadi data *training* dan data *testing.* Data *training* adalah subset dari kumpulan data asli yang digunakan untuk memperkirakan dan mempelajari parameter algoritma *machine learning* yang diperlukan. Data *testing* adalah subset dari kumpulan data asli yang digunakan untuk memperkirakan performa model pembelajaran yang diperlukan (Tabassum & Iqbal, 2020). Pada penelitian kanker payudara, dataset dibagi menjadi 70% data *training* dan 30% data *testing.*

B. *Data Normalisasi*

Normalisasi adalah teknik penskalaan nilai atribut dari data sehingga bisa terletak pada rentang tertentu (Anggoro & Novitaningrum, 2021). Normalisasi data dapat bermanfaat dalam tahap pemrosesan data tanpa peningkatan besar dalam memori dan kebutuhan pemrosesan daya (Patro & sahu, 2015). Terdapat empat teknik normalisasi, dan dalam penelitian kanker payudara ini peneliti menggunakan metode *min-max.* Penggunaan metode *min-max* ini karena telah diuji pada penelitian penyakit diabetes menggunakan algoritma *Random Forest* (RF) dimana menguji 2 metode normalisasi yaitu *min-max* dan *z-score,* akurasi terbaik dihasilkan dengan normalisasi *min-max* sebesar 95,45% dibanding dengan *z-score* 95% (Agung BS, Gde, Adiwijaya, Dwifebri P, 2021). *Min-max* menskalakan dan menerjemahkan setiap fitur secara individual dengan rentang yang diberikan pada data *training* . Metode *min-max* ditunjukkan dalam persamaan 1.

$$\mathbf{V' = \frac{V - min}{max - min} (newMax - newMin) + newMin} \qquad (1)$$

Keterangan :

V’ = hasil normalisasi

V = nilai yang akan dinormalisasi

min = nilai minimum dari dataset

max = nilai maximum dari dataset

newMax = batas maximal normalisasi

newMin = batas minimal normalisasi

2.3 *Data Processing*

A. *Random Forest* (RF)

Dalam algoritma ini, sejumlah besar *decision tree* dibangun saat mereka beroperasi bersama. *Decision tree* bertindak sebagai pilar dalam algoritma ini. *Random Forest* adalah kelompok *decision tree* yang simpulnya ditentukan pada langkah pra-pemrosesan. Setelah membangun beberapa pohon, fitur terbaik dipilih dari subset fitur acak. Terdapat beberapa fitur *random forest*, ini diantaranya :

i. Menangani sejumlah besar variabel input tanpa menghapus variabel.
ii. Menunjukkan variabel-variabel yang penting dalam klasifikasi.
iii. Basis data yang besar juga berjalan secara efisien.
iv. Pohon atau hutan yang dihasilkan dapat disimpan untuk penggunaan di masa mendatang juga.

Langkah-langkah dari algoritma *Random Forest* (RF) :

Langkah 1: Dari data *training*, pilih K titik dari data acak.

Langkah 2: Bangun pohon keputusan dengan titik data K ini.

Langkah 3: Sebelum mengulangi langkah 1 dan 2, pilih nomor *NTree* dari pohon yang ingin dibuat.

Langkah 4: Prediksikan nilai y dengan membuat masing-masing *NTree* pohon untuk titik data baru dan tetapkan rata-rata titik data baru untuk semua nilai y yang diprediksi.

Rumus matematika untuk pengklasifikasi *random forest*

2.4 *Evaluation Model*

Akurasi

Keterangan :

TP = *True Positive*

TN = *True Negative*

FP = *False Positive*

FN = *False Negative*

3. HASIL DAN PEMBAHASAN

4. PENUTUP

DAFTAR PUSTAKA

Agung BS, Gde, Adiwijaya, Dwifebri P, M. (2021). *Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk Deteksi Pasien Penderita Diabetes Menggunakan Metode Normalisasi*. *1*(10), 114–122.

Anggoro, D. A., & Novitaningrum, D. (2021). Comparison of accuracy level of support vector machine (SVM) and artificial neural network (ANN) algorithms in predicting diabetes mellitus disease. *ICIC Express Letters*, *15*(1), 9–18. https://doi.org/10.24507/icicel.15.01.9

Dewi, M. (2017). Sebaran Kanker di Indonesia, Riset Kesehatan Dasar 2007. *Indonesian Journal of Cancer*, *11*(1), 1–8. https://doi.org/10.33371/ijoc.v11i1.494

Dey, A. (2016). Machine Learning Algorithms: A Review. *International Journal of Computer Science and Information Technologies*, *7*(3), 1174–1179. www.ijcsit.com

Dou, J., Yunus, A. P., Tien Bui, D., Merghadi, A., Sahana, M., Zhu, Z., Chen, C. W., Khosravi, K., Yang, Y., & Pham, B. T. (2019). Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Island, Japan. *Science of the Total Environment*, *662*(January), 332–346. https://doi.org/10.1016/j.scitotenv.2019.01.221

Jaiswal, J. K., & Samikannu, R. (2017). Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression. *Proceedings - 2nd World Congress on Computing and Communication Technologies, WCCCT 2017*, 65–68. https://doi.org/10.1109/WCCCT.2016.25

Maria, I. L., Sainal, A. A., & Nyorong, M. (2017). Risiko Gaya Hidup Terhadap Kejadian Kanker Payudara Pada Wanita. *Media Kesehatan Masyarakat Indonesia*, *13*(2), 157. https://doi.org/10.30597/mkmi.v13i2.1988

Mattiuzzi, C., & Lippi, G. (2019). Current Cancer Epidemiology glossary. *Journal of Epidemiology and Global Health*, *9*(4), 217–222.

Patro, S. G. K., & sahu, K. K. (2015). Normalization: A Preprocessing Stage. *Iarjset*, *2*(3), 20–22. https://doi.org/10.17148/iarjset.2015.2305

Saygili, A. (2018). Classification and Diagnostic Prediction of Breast Cancers via Different Classifiers. *International Scientific and Vocational Journal*, *2*(December), 48–56.

Tabassum, H., & Iqbal, M. M. (2020). Enactment Ranking of Supervised Algorithms Dependence of Data Splitting Algorithms : A Case Study of Real Datasets. *International Journal of Computer Science & Information Technology (IJCSIT)*, *12*(2), 15–22. https://doi.org/10.5121/ijcsit.2020.12202

Wang, H., Zheng, B., Yoon, S. W., & Ko, H. S. (2018). A support vector machine-based ensemble algorithm for breast cancer diagnosis. *European Journal of Operational Research*, *267*(2), 687–699. https://doi.org/10.1016/j.ejor.2017.12.001