BAB I

PENDAHULUAN

1.1. Latar Belakang

Kanker payudara merupakan tipe kanker yang umumya terbentuk di sel-sel payudara dan sel-sel kanker tersebut tumbuh diluar kendali. Kanker payudara dapat terjadi pada semua gender, tetapi kanker ini umumnya lebih sering terjadi pada wanita. Di Australia kanker payudara merupakan tipe kanker yang sangat umum terjadi pada wanita dan kanker paling umum kedua yang menjadi penyebab kematian pada wanita setelah kanker paru-paru. Sedangkan di Indonesia sendiri, jumlah kanker payudara ini sampai menempati urutan pertama dibandingkan jenis kanker yang lain dan menjadi salah satu penyumbang kematian pertama. Lebih dari 2,3 juta wanita didiagnosis kanker payudara dengan tingkat kematian sekitar 670.000 pada tahun 2022 (*World Health Organization*, 2022).

Penderita kanker payudara sering mengalami tantangan psikologis seperti kecemasan, ketakutan, depresi, dan panik. Meskipun penyebab kanker payudara belum sepenuhnya dipahami, kemungkinannya bersifat multifaktorial. Upaya-upaya pencegahan telah dilakukan seperti pendidikan masyarakat bersama dengan penanganan yang disesuaikan dengan kebutuhan individu untuk mengatasi masalah kanker payudara. Pentingnya skrining kanker payudara sejak dini juga harus ditekankan. Kanker payudara juga dapat terbagi menjadi jenis ganas dan jinak. Jika diketahui jenisnya, maka pencegahan dan pengobatan akan segera dilakukan sesuai dengan jenis kanker payudaranya, sehingga dapat menghindari efek samping pada pasien dan bahkan kematian. Deteksi dini dan pengobatan dini sangat penting. Namun jika sudah didiagnosa terkena kangker payudara maka perlu dilakukan pemeriksaan lebih lanjut ke dokter.

Machine learning adalah cabang dari kecerdasan buatan dan juga komputer sains yang memiliki fokus dalam pemakaian data dan juga algoritma-algoritma agar bisa meniru cara manusia belajar untuk meningkatkan accuracy machine learning. Pemakaian machine ini juga dapat digunakan dalam bidang kesehatan dengan menggunakan berbagai algoritma untuk pengklasifikasian suatu penyakit.

Machine learning ini memerlukan algoritma agar bisa dipakai, dan untuk penelitian ini, peneliti akan memakai algoritma dengan supervised learning. Supervised learning cocok untuk dipakai karena masalah yang diatasi oleh supervised learning adalah klasifikasi dan regresi. KNN atau K-Nearest Neighbor adalah contoh algoritma supervised learning. KNN merupakan algoritma dasar untuk machine learning yang sering dipakai untuk klasifikasi.

Oleh karena itu, Dalam penelitian ini, peneliti menggunakan algoritma KNN untuk memprediksi penyakit kanker payudara. Algoritma KNN, atau K-Nearest Neighbors, bekerja dengan mengklasifikasikan data berdasarkan kesamaan dengan titik data terdekat dalam ruang variabel. Selain menggunakan KNN kita juga menggunakan random forest bagging, dan ANN untuk melakukan perbandingan antara ketiga algoritma tersebut dengan melihat tingkat akurasi algoritma tersebut untuk kasus diagnosis kanker payudara. Dengan menggunakan dataset kanker payudara yang tersedia dari Kaggle, peneliti akan melakukan langkah-langkah seperti pengolahan data, pembagian data menjadi set pelatihan dan set pengujian, dan evaluasi kinerja model berdasarkan tingkat akurasi, dan lainnya. Melalui penelitian ini, diharapkan dapat dieksplorasi kemampuan algoritma KNN dalam memprediksi apakah seorang pasien mengidap kanker payudara atau tidak berdasarkan variabel-variabel yang tersedia dalam dataset. Hasil dari penelitian ini akan memberikan wawasan yang berharga tentang potensi penggunaan algoritma KNN dalam konteks diagnosa kanker payudara.

1.2. Rumusan Masalah

Berdasarkan latar belakang masalah yang telah dijelaskan diatas maka rumusan masalah dalam penelitian ini adalah:

1. Bagaimana cara mengklasifikasikan kanker payudara menjadi kategori ganas atau jinak?
2. Bagaimana perbandingan algoritma KNN, Random Forest, Bagging, dan ANN dalam mengklasifikasikan kanker payudara untuk mengetahui tingkat keakuratan masing-masing algoritma?

1.3. Tujuan

Sejalan dengan rumusan masalah yang telah dipaparkan, maka tujuan yang diharapkan dari penelitian ini adalah sebagai berikut :

1. Mengetahui cara mengklasifikasikan kanker payudara menjadi kategori ganas atau jinak
2. Mengetahui perbandingan algoritma KNN, Random Forest, Bagging, dan ANN dalam mengklasifikasikan kanker payudara berdasarkan tingkat keakuratan masing-masing algoritma.

1.4. Manfaat

Adapun manfaat yang ingin dicapai dari hasil penelitian ini adalah sebagai berikut.

1. Memberikan pemahaman yang lebih dalam tentang proses prediksi penyakit kanker payudara menggunakan algoritma KNN.
2. Meningkatkan kesadaran akan pentingnya deteksi dini dan pengobatan yang tepat pada kasus kanker payudara
3. Menyediakan panduan praktis bagi peneliti, profesional medis, dan masyarakat umum dalam upaya pencegahan dan pengobatan kanker payudara.

1.5. Batasan Masalah

Dari identifikasi masalah yang ditetapkan dalam penelitian ini, maka dirasa perlu dilakukan pembatasan masalah sebagai berikut.

1. Penelitian ini berfokus pada penggunaan algoritma KNN (K-Nearest Neighbors) untuk melakukan prediksi penyakit kanker payudara.
2. Penggunaan dataset kanker payudara yang tersedia dari Kaggle sebagai basis penelitian.
3. Evaluasi kinerja model akan dilakukan berdasarkan tingkat akurasi yang didapatkan.

BAB II

TINJAUAN PUSTAKA

2.1 Literature Sebelumnya

Penelitian yang dilakukan oleh Shler Farhad Khorshid dan Adnan Mohsin Abdulazeez (2021) dengan judul "*Breast Cancer Diagnosis Based On K-Nearest Neighbors*" menyajikan tinjauan tentang algoritma klasifikasi yang digunakan untuk kanker payudara dengan mengimplementasikan algoritma K-NN. Setiap algoritma menggunakan teknik klasifikasi yang berbeda, dan dataset yang paling sering digunakan bervariasi di berbagai penelitian Secara keseluruhan, implementasi menggunakan K-NN relatif mudah. Dalam hal akurasi, KNN memberikan prediksi paling akurat (99,12%). Untuk diagnosis kanker payudara dengan menggunakan berbagai algoritma di masa depan, metode yang lebih efektif dan dapat diandalkan akan ditawarkan. Sedangkan pada penelitian lain dengan menggunakan K-Nearest Neighbors yang dilakukan oleh Tsehay Admassu Assegie (2020) didapat kesimpulan bahwa model KNN yang telah dioptimalkan untuk memprediksi kanker payudara dengan menggunakan pendekatan pencarian grid untuk menemukan hiperparameter terbaik. Perbandingan dilakukan antara penggunaan hiperparameter default dan hiperparameter yang telah disesuaikan. Hasilnya menunjukkan bahwa kinerja model meningkat secara signifikan ketika nilai parameter terbaik atau nilai K digunakan untuk melatih KNN. Dengan hiperparameter default, kinerja KNN mencapai akurasi 90,10%. Namun, akurasi deteksi kanker payudara yang lebih baik dicapai oleh KNN ketika menggunakan hiperparameter terbaik yang dipilih melalui pendekatan pencarian grid, dengan kinerja tertinggi mencapai 94,35%.

Sedangkan penelitian lainnya yang dilakukan oleh Rudi Hartono et al (2023) dengan judul "Analisa Perbandingan Kinerja Algoritma Klasifikasi Untuk Prediksi Penyakit Kanker Payudara", dengan algoritma yang digunakan adalah Random forest, Decision Tree, K-Nearest, , Support Vector Machine, dan Naïve Bayes memiliki hasil bahwa dari perbandingan kelima algoritma tersebut baik menggunakan teknik Ensemble Learning Bagging atau tidak, algoritma dengan nilai kinerja akurasi terbaik yaitu algoritma Random Forest. Semua kinerja akurasi pada setiap algoritma bertambah dengan rata-rata kenaikan 1% dengan menggunakan teknik Ensemble Learning Bagging, menunjukan bahwa teknik ini dapat digunakan untuk meningkatkan kinerja

akurasi dibandingkan dengan teknik default pada setiap algoritma ataupun dengan perhitungan manual, selain itu jumlah record juga dapat mempengaruhi kinerja algoritma.

2.2 Kanker Payudara

Kanker payudara merupakan salah satu jenis kanker yang paling umum pada wanita di seluruh dunia. Lebih dari 2,3 juta wanita didiagnosis kanker payudara dengan tingkat kematian sekitar 670.000 pada tahun 2022 (*World Health Organization*, 2022). Kanker payudara terjadi di semua negara di dunia dan dapat menyerang perempuan sejak masa pubertas, namun prevalensinya semakin tinggi seiring bertambahnya usia (*World Health Organization*, 2022). Sel-sel kanker payudara muncul di dalam saluran susu dan / atau lobulus penghasil susu payudara. Bentuk paling awal yang muncul (in situ) tidak mengancam jiwa dan dapat dideteksi pada tahap awal. Sel-sel kanker dapat menyebar ke jaringan payudara terdekat (invasi). Hal ini menciptakan tumor yang menyebabkan benjolan atau penebalan. Jenis tumor ini termasuk dalam kategori tumor ganas dengan kejadian yang meningkat di seluruh dunia dengan tingkat tertinggi berada di negara industri (Beata Smolarz et al, 2022). Kanker payudara memiliki kemampuan untuk bermetastasis dan sering kali menyebar ke organ-organ jauh seperti tulang, hati, paru-paru, dan otak, yang pada umumnya menyebabkan kondisi yang sulit untuk disembuhkan. Deteksi awal penyakit ini dapat menghasilkan prognosis yang positif dan meningkatkan tingkat kelangsungan hidup.

2.3. Metode Pendekatan

Pengumpulan data adalah komponen yang penting dalam penelitian. Data adalah sekumpulan fakta yang memberikan gambaran dan mendukung pemahaman tentang fenomena atau peristiwa yang diteliti. Data memungkinkan peneliti untuk menganalisis dan menjelaskan situasi tertentu dengan lebih akurat dan dapat diandalkan. Metode pengumpulan data yang tepat sangat penting untuk memastikan bahwa hasil penelitian representatif dan kesimpulan yang dihasilkan valid.

Penelitian ini menerapkan dataset "*Breast Cancer Prediction*" dari *Kaggle* yang berisi data untuk digunakan dalam analisis diagnosis kanker payudara. Dataset ini mencakup informasi berikut:

1. Radius: Rata-rata jarak dari pusat massa tumor ke titik-titik di sepanjang keliling tumor.

2. *Texture*: *Texture* mengukur variasi intensitas cahaya yang ada pada gambar, sehingga dapat menunjukkan heterogenitas atau variasi dalam komposisi sel tumor.
3. Perimeter: Panjang total dari keliling tumor. Perimeter memberikan ukuran linear dari ukuran tumor.
4. Area: Area adalah luas permukaan dari tumor, dihitung berdasarkan jumlah piksel yang membentuk tumor dalam gambar.
5. *Smoothness*: *Smoothness* mengukur seberapa halus tepi dari tumor. Semakin kecil variasinya, semakin halus tepi tumor.
6. *Compactness*: *Compactness* adalah ukuran yang dihitung dengan rumus $perimeter^2/area - 1.0$. Hal ini memberikan indikasi seberapa padat atau *compact* tumor tersebut. Tumor yang lebih bulat cenderung memiliki nilai *compactness* yang lebih rendah.
7. *Concavity*: Mengukur tingkat keparahan dari bagian cekung pada kontur tumor. *Concavity* menunjukkan seberapa dalam bagian cekung atau lekukan pada tepi tumor.
8. *Concave Point*: Mengacu pada jumlah bagian cekung pada kontur tumor. *Concave Point* menghitung berapa banyak lekukan atau cekungan yang ada di sepanjang tepi tumor.
9. *Symmetry*: Mengukur simetri dari bentuk tumor untuk membandingkan bentuk dari satu sisi tumor dengan sisi lainnya untuk melihat seberapa simetris bentuk tumor tersebut.
10. *Fractal dimension*: Mengacu pada dimensi fraktal dari tepi tumor, yang dihitung dengan pendekatan garis pantai (*coastline approximation*) untuk mengukur kompleksitas dari tepi tumor. Semakin tinggi dimensi fraktalnya, semakin kompleks atau tidak teratur tepi tumor tersebut.

Dataset ini sangat bermanfaat untuk pengembangan model yang dapat digunakan untuk analisis diagnosis kanker payudara untuk mengetahui apakah tingkat kanker tersebut jinak atau ganas.

2.4 K-Nearest Neighbor untuk Analisis Kanker Payudara

Teknik K-Nearest Neighbor merupakan salah satu algoritma klasifikasi *Machine Learning* yang paling awal dan sederhana. K-NN *Classifier* dikenal sebagai metode yang efektif untuk membedakan antara kasus sehat dan sakit setelah pemilihan fitur dilakukan. Meskipun sederhana, K-NN dapat menghasilkan hasil yang kompetitif dan, dalam beberapa kasus, kombinasi secara cerdas dengan dapat mencapai kinerja terbaik. Algoritma K-NN

mengkategorikan setiap contoh yang tidak diketahui dalam data pelatihan berdasarkan mayoritas tanda di antara tetangga terdekatnya. Kinerjanya sangat bergantung pada matrik jarak yang digunakan untuk menentukan tetangga terdekat. Sebagian besar pengklasifikasi K-NN menggunakan matrik Euclidean untuk mengukur perbedaan antara contoh yang direpresentasikan sebagai vektor input tanpa adanya informasi tambahan. Selain matrik jarak konvensional seperti Minkowski dan Chebyshev, metode pengukuran jarak lainnya yang disarankan termasuk perhitungan jarak Xing. Jarak Euclidean dihitung seperti yang ditunjukkan dalam rumus berikut.

$$d(xi, xj) = \sqrt{\sum_{r=1}^{n} wr(ar(xi) - ar(xj))^2}$$

Di mana sebuah contoh direpresentasikan sebagai vektor $x = (a1,\ a2,\ a3,\ .\ .\ .,\ an)$, dengan n merupakan dimensi vektor input atau jumlah atribut dari contoh tersebut. ar adalah atribut ke-r dari contoh tersebut, dengan r berkisar dari 1 hingga n. Semakin kecil nilai $d(xi,\ xj)$, semakin mirip kedua contoh tersebut. Pada algoritma KNN, label kelas dari contoh uji ditentukan berdasarkan suara mayoritas dari K-Nearest Neighbor.

$$y(d_i) = arg\ max \sum_{x_j \in kNN} y(x_j, c_k)$$

Dalam konteks ini, di adalah contoh uji, xj adalah salah satu *nearest neighbor* dari *set training*, dan $Y(xj,\ ck)$ menunjukkan apakah xj termasuk dalam kelas ck. Persamaan tersebut menyatakan bahwa kelas dengan mayoritas anggota di antara K-Nearest Neighbor akan menjadi prediktor. Sebagai contoh, jika 5 algoritma K-Nearest Neighbor digunakan sebagai pengklasifikasi, dan tiga dari lima nearest neighbor termasuk dalam satu kategori sedangkan dua lainnya termasuk kategori yang berbeda, maka kita dapat mengasumsikan bahwa contoh uji tersebut termasuk dalam kategori pertama (Shler Farhad Khorshid, 2021).

Teknik ini adalah dasar dari algoritma nearest neighbor, di mana label kelas sampel ditentukan hanya dengan mengidentifikasi nearest neighbor (NN). K-NN mengasumsikan bahwa probabilitas kondisional kelas stabil secara lokal dan bahwa dimensi yang besar diperoleh dari bias. K-NN adalah sistem klasifikasi yang sangat serbaguna dan tidak memerlukan data pelatihan

untuk diproses sebelumnya. Oleh karena itu, algoritma K-NN yang disempurnakan harus fokus pada menemukan jumlah k yang tepat, dimana k adalah nearest neighbor itu sendiri, untuk mendapatkan label kelas yang paling mungkin bagi setiap contoh yang diuji (Shler Farhad Khorshid, 2021).

Langkah-langkah algoritma K-NN adalah sebagai berikut:

1. Masukkan kumpulan data dan bagi menjadi satu *set training* dan satu *set testing*.

2. Pilih sebuah *instance* dari *set training* dan ukur jaraknya terhadap semua *instance* dalam *set training*.

3. Urutkan daftar jarak tersebut dalam urutan menaik.

4. Tentukan kelas dari *instance* yang dipilih berdasarkan kelas mayoritas dari tiga *instance* terdekat dalam set pelatihan (dengan k = 3).

2.5 Random Forest untuk Analisis Kanker Payudara

Random Forest memiliki beberapa pohon, dan attribute dipilih secara random dengan memakai bagging. Random Forest memakai sekumpulan pohon di mana setiap pohon tersebut memiliki nilai vekor *arbitrary* yang di sampel secara terpisah untuk semua pohon di Random Forest. Fase algoritma Random Forest dibagi menjadi 3 fase, yang pertama adalah pohon sebanyak i dibuat. Fase kedua menggabungkan semua pohon menjadi Random Forest. Yang terakhir adalah memberikan output. Output dari Random Forest dihasilkan melalui hasil voting dari setiap pohon (Vincent Angkasa et al, 2022).

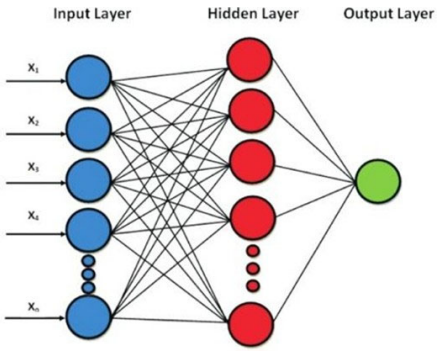
Random Forest juga dapat menggunakan metode bagging dalam membangun setiap pohon. Untuk node tiap tree diciptakan dengan memilih random column dari bagging. Setelah column dipilih, column tersebut dilakukan *ascending sorting*. Random row dipilih dan row dibawahnya dipilih untuk diambil *value* dari row tersebut dan dibagi 2. Jika random row yang dipilih merupakan row terakhir maka yang dipilih row kedua adalah row diatasnya. Berdasarkan kedua *value* tersebut maka node dibangun dan split bisa terjadi. Setelah node tersebut dibangun maka *gini index* dapat dicari. Jika *gini index* sudah 0 maka split tersebut akan stop. Tetapi jika

belum 0 maka split akan terus terjadi hingga menghasilkan *pure leaf*. Setelah itu jika sudah selesai looping, untuk setiap tree, maka algoritma Random Forest menjadi berhenti dan selesai.

Random Forest menggunakan pohon $gk(A,\theta k)$ Dimana k^{th} sebagai *base learners* adalah tumpukan variabel acak yang bersifat independen untuk k=1,..., K. Untuk *training data* D = (a_1, b1), ..., (aN, yN), di mana aI = (a_I, 1, . . , a_I,p)^T mewakili *m predictor* dan b1 sebagai respons dan sebuah realisasi spesifik θk dari Θk. Pohon yang sudah *fit* diwakili sebagai $\hat{g}k(a,\theta k, D)$. Rumus tersebut dipakai untuk menghasilkan acak pada 2 fase. Pada fase pertama termasuk bagging, sebuah pola *bootstrap* independen yang diambil dari data *original* dan di-*fit* ke masing-masing pohon. Sampling *bootstrap* memanfaatkan fungsi acak yang memberikan satu bagian dari θk. Pada tahap kedua, ketika pemilihan acak, *splitter* baik ditemukan dari *subset* variabel r predictor secara terpisah pada tiap-tiap node ketika melakukan *splitting. Predictor* pengambilan sampel memberikan bagian θk yang tersisa dengan pengacakan (Vincent Angkasa et al, 2022).

2.6 Artificial Neural Network untuk Analisis Kanker Payudara

Pada tahun 1943, Mc.Culloch dan Pitts memperkenalkan model matematika yang merupakan penyederhanaan dari struktur sel saraf yang sebenarnya.



Gambar 2.1 McCulloch & Pitts neuron model

Gambar diatas memperlihatkan bahwa sebuah neuron memiliki tiga komponen :

1. Synapse (w1, w2,…, wn)T

2. Alat penambah (adder)

3. Fungsi aktifasi (f)

Korelasi antara ketiga komponen ini dirumuskan pada persamaan $y = f \sum_{i=1}^{n} xi \times wi$. Signal x berupa vektor berdimensi $n\,(x1,\ x2,\ \ldots xn)T$ akan mengalami penguatan oleh synapse $w\ (w1,\ w2,\ \ldots,\ wn)T$. Selanjutnya akumulasi dari penguatan tersebut akan mengalami transformasi oleh fungsi aktifasi f. Fungsi f ini akan memonitor, apabila akumulasi penguatan *signal* itu telah melebihi batas tertentu, maka sel neuron yang semula berada dalam kondisi "0", akan mengeluarkan signal "1". Berdasarkan nilai output tersebut (=y), sebuah neuron dapat berada dalam dua status: "0" atau "1". Neuron disebut dalam kondisi *firing* bila menghasilkan output bernilai "1" (Fitra Septia Nugraha, 2019).

Pada algoritma ini diawali dengan persiapan data yang telah didapat agar dapat digunakan pada saat melakukan pemodelan dan menjalankan tahap evaluasi. Tahap *preprocessing* atau pemrosesan data mencakup kegiatan membangun data serta juga membersihkan data agar siap untuk digunakan ke tahap *modelling* atau pemodelan data. Berikut ini tahap pemrosesan data antara lain mengecek *Missing Values*, mengecek distrusi data, mengecek tipe fitur dan encoding data. Pada tahap ini melakukan analisis data berdasarkan algoritma yang telah ditentukan yakni Artificial Neural Network (ANN). Pada tahapan ini dilakukan *Split Data* yakni proses membagi dataset menjadi 2 bagian yang mana menjadi data *testing* dan menjadi data *training* (Tommy Dwi Putra, 2022). Hasil pengujian model yang dilakukan adalah mengklasifikasikan kanker payudara jinak dan ganas dengan Artificial Neural Network (ANN) untuk mendapatkan nilai akurasi terbaik.

BAB III

PEMBAHASAN DAN ANALISIS DATA

3.1 Metode Analisis

3.1.1 Bootstrap Aggregating (Bagging)

Bagging atau Bootstrap Aggregating adalah teknik ensemble learning yang efisien dalam meningkatkan klasifikasi. Bagging bermanfaat untuk meningkatkan kinerja algoritma klasifikasi dalam machine learning dan meningkatkan akurasi dengan menggabungkan beberapa klasifikasi tunggal. Bagging menggunakan teknik Bootstrap atau pengambilan sampel berulang dari data asli sebanyak n kali dengan penggantian, untuk membuat set pelatihan. Proses Bagging meliputi tiga tahap utama: Bootstrap Sampling, Pelatihan Model, dan Penggabungan.

3.1.2 Random Forest

Random Forest adalah metode statistik nonparametrik yang diperkenalkan oleh Breiman untuk meningkatkan akurasi dan mengurangi risiko overfitting dengan membangkitkan atribut secara acak untuk setiap node. Metode ini terdiri dari sekumpulan decision tree yang digunakan untuk mengklasifikasikan data ke dalam kelas tertentu dan menangani masalah regresi. Misalkan kita memiliki variabel penjelas X dan variabel respon Y. Set pembelajaran terdiri dari observasi independen dari vektor X dan Y dengan asumsi bahwa nilai Y dipengaruhi oleh X ditambah gangguan acak. Dalam konteks ini, fungsi regresi diestimasi oleh model Random Forest. Algoritma Random Forest juga efektif dalam mengklasifikasikan data yang tidak seimbang dalam jumlah besar, menghasilkan performa yang baik dan waktu eksekusi yang cepat.

3.1.3 K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) adalah metode klasifikasi non-parametrik dan lazy learning yang mengklasifikasikan objek berdasarkan kedekatannya dengan data pelatihan. KNN menggunakan data latih yang ada untuk menentukan jarak dan

mengklasifikasikan data baru, sering kali dengan menggunakan jarak Euclidean. Metode ini mengurangi masalah overfitting dan dalam jumlah sampel besar, tingkat kesalahannya mendekati optimal Bayes. KNN banyak digunakan dalam analisis data dan pengembangan AI karena efisiensinya dalam mengidentifikasi pola. Rumus jarak Euclidean adalah:

$$d(x.y) = \sqrt{\sum_{i=1}^{n} (x_i + y_i)}$$

3.1.4 Artificial Neural Network

Artificial Neural Network (ANN) adalah teknik komputasi yang meniru jaringan biologis melalui perhitungan non-linear. ANN terdiri dari beberapa lapisan neuron yang saling terhubung dan mampu mempelajari hubungan non-linear dalam data. Struktur dasar ANN meliputi lapisan input yang menerima data, lapisan tersembunyi yang melakukan proses komputasi, dan lapisan output yang menghasilkan prediksi. Proses dalam ANN melibatkan beberapa tahap, yaitu forward propagation di mana data input melewati jaringan, perhitungan loss untuk mengukur seberapa baik model memprediksi output target menggunakan fungsi loss, backpropagation untuk memperbarui bobot berdasarkan kesalahan, dan gradient descent untuk mengoptimalkan model.

Cross-Entropy Loss:

$$L(x, y) = -\sum_{i=1}^{c} x_i \log_2 y_i$$

3.2 Deskripsi Data

	id	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean
count	5.690000e+02	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000
mean	3.037183e+07	14.127292	19.289649	91.969033	654.889104	0.096360	0.104341	0.088799	0.048919	0.181162
std	1.250206e+08	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813	0.079720	0.038803	0.027414
min	8.670000e+03	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380	0.000000	0.000000	0.106000
25%	8.692180e+05	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920	0.029560	0.020310	0.161900
50%	9.060240e+05	13.370000	18.840000	86.240000	551.100000	0.095870	0.092630	0.061540	0.033500	0.179200
75%	8.813129e+06	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400	0.130700	0.074000	0.195700
max	9.113205e+08	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.426800	0.201200	0.304000

texture_worst	perimeter_worst	area_worst	smoothness_worst	compactness_worst	concavity_worst	concave points_worst	symmetry_worst	fractal_dimension_worst	Unnamed: 32
569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	0.0
25.677223	107.261213	880.583128	0.132369	0.254265	0.272188	0.114606	0.290076	0.083946	NaN
6.146258	33.602542	569.356993	0.022832	0.157336	0.208624	0.065732	0.061867	0.018061	NaN
12.020000	50.410000	185.200000	0.071170	0.027290	0.000000	0.000000	0.156500	0.055040	NaN
21.080000	84.110000	515.300000	0.116600	0.147200	0.114500	0.064930	0.250400	0.071460	NaN
25.410000	97.660000	686.500000	0.131300	0.211900	0.226700	0.099930	0.282200	0.080040	NaN
29.720000	125.400000	1084.000000	0.146000	0.339100	0.382900	0.161400	0.317900	0.092080	NaN
49.540000	251.200000	4254.000000	0.222600	1.058000	1.252000	0.291000	0.663800	0.207500	NaN

Gambar 3.1 Deskripsi Data

Berdasarkan data yang ada pada dataset dari situs Kaggle, total data yang diperoleh adalah 569 entri dengan 33 kolom variabel dengan rincian sebagai berikut:

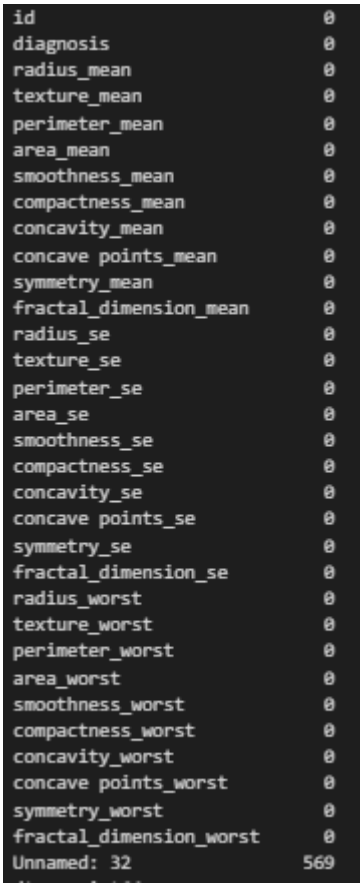
- id: Nomor identifikasi unik untuk setiap sampel.
- diagnosis: Diagnosis dari sampel, dimana 'M' menunjukkan malignan (kanker) dan 'B' menunjukkan benign (jinak).
- radius_mean: Rata-rata dari radius inti sel.
- texture_mean: Rata-rata dari variasi tekstur (skala abu-abu) inti sel.
- perimeter_mean: Rata-rata dari keliling inti sel.
- area_mean: Rata-rata dari area inti sel.
- smoothness_mean: Rata-rata dari kelancaran inti sel, dihitung sebagai variasi lokal dalam panjang radius.
- compactness_mean: Rata-rata dari kekompakan inti sel, dihitung sebagai $\frac{perimater^2}{area} - 1.0$
- concavity_mean: Rata-rata dari cekungan inti sel, yaitu tingkat cekungan kontur inti sel.
- concave points_mean: Rata-rata dari jumlah titik cekung pada kontur inti sel.
- symmetry_mean: Rata-rata dari simetri inti sel.
- fractal_dimension_mean: Rata-rata dari dimensi fraktal inti sel, dihitung sebagai "coastline approximation" - 1.
- radius_se: Kesalahan standar dari radius inti sel.
- texture_se: Kesalahan standar dari variasi tekstur inti sel.
- perimeter_se: Kesalahan standar dari keliling inti sel.
- area_se: Kesalahan standar dari area inti sel.
- smoothness_se: Kesalahan standar dari kelancaran inti sel.

- compactness_se: Kesalahan standar dari kekompakan inti sel.
- concavity_se: Kesalahan standar dari cekungan inti sel.
- concave points_se: Kesalahan standar dari jumlah titik cekung pada kontur inti sel.
- symmetry_se: Kesalahan standar dari simetri inti sel.
- fractal_dimension_se: Kesalahan standar dari dimensi fraktal inti sel.
- radius_worst: Nilai terburuk (terbesar) dari radius inti sel.
- texture_worst: Nilai terburuk dari variasi tekstur inti sel.
- perimeter_worst: Nilai terburuk dari keliling inti sel.
- area_worst: Nilai terburuk dari area inti sel.
- smoothness_worst: Nilai terburuk dari kelancaran inti sel.
- compactness_worst: Nilai terburuk dari kekompakan inti sel.
- concavity_worst: Nilai terburuk dari cekungan inti sel.
- concave points_worst: Nilai terburuk dari jumlah titik cekung pada kontur inti sel.
- symmetry_worst: Nilai terburuk dari simetri inti sel.
- fractal_dimension_worst: Nilai terburuk dari dimensi fraktal inti sel.
- Unnamed: 32

3.3 Data Preprocessing

3.3.1 Handling Missing Values

Missing value adalah data yang hilang dalam dataset yang dapat mempengaruhi analisis dan kinerja model. Penting untuk mendeteksi dan menangani missing value, misalnya dengan melakukan imputasi atau penghapusan data.



Gambar 3.2 Jumlah Missing Value pada Variabel

Semua atribut dalam dataset ini tidak memiliki nilai yang hilang. Setiap baris dan kolom memiliki entri yang berisi nilai "0" atau false, menunjukkan bahwa seluruh dataset terisi lengkap tanpa ada data yang kosong.

3.3.2 Feature Selection

Menghapus kolom 'id' dan 'Unnamed: 32' dari dataset dengan menggunakan fungsi drop() adalah langkah penting dalam memilih fitur untuk meningkatkan kualitas dan kinerja model. Tindakan ini bertujuan untuk menghilangkan fitur yang tidak relevan atau duplikat yang tidak memberikan kontribusi pada analisis atau prediksi yang akurat. Dengan fokus pada fitur yang lebih informatif, proses pembelajaran mesin menjadi lebih efisien dan hasilnya lebih dapat diandalkan, menjaga kebersihan dan kemudahan interpretasi dataset.

3.3.3 Feature Categorization

Pada tahap ini, perlu untuk memisahkan fitur-fitur dalam dataset ke dalam dua jenis utama sebelum analisis, yaitu kategorik dan numerik. Sebagai contoh, kolom 'diagnosis' dapat dianggap sebagai fitur kategorik yang memuat label 'M' (malignant) dan 'B' (benign), sementara kolom-kolom lainnya dianggap sebagai fitur numerik. Pendekatan ini memungkinkan penerapan teknik preprocessing yang tepat, seperti encoding untuk fitur kategoris dan normalisasi untuk fitur numerik. Dengan cara ini, dataset dapat disiapkan secara optimal untuk analisis data lebih lanjut.

3.3.4 Target Variable Encoding

Pada tahap encoding variabel target, kolom 'diagnosis' telah diubah ke dalam format numerik. Secara khusus, label 'M' (malignant) dipetakan ke angka 1, sedangkan label 'B' (benign) dipetakan ke angka 0. Pendekatan ini umumnya digunakan dalam masalah klasifikasi biner, di mana variabel target perlu dalam format numerik untuk proses pelatihan model.

Langkah ini penting karena banyak algoritma machine learning memerlukan variabel target yang bersifat numerik. Dengan memetakan 'M' ke 1 dan 'B' ke 0, model dapat memahami dan memprediksi kategori diagnosa dengan lebih baik selama proses pelatihan dan evaluasi. Ini juga mempermudah perhitungan metrik evaluasi seperti akurasi, presisi, recall, dan lainnya yang bergantung pada nilai numerik dari variabel target.

3.3.5 Exploratory Data Analysis

Fungsi `value_counts()` digunakan untuk menghitung jumlah setiap kelas (malignant dan benign) dalam kolom diagnosis. Ini memberikan informasi penting tentang bagaimana data terdistribusi antara kedua kelas, yang sangat relevan dalam pengembangan dan evaluasi model.

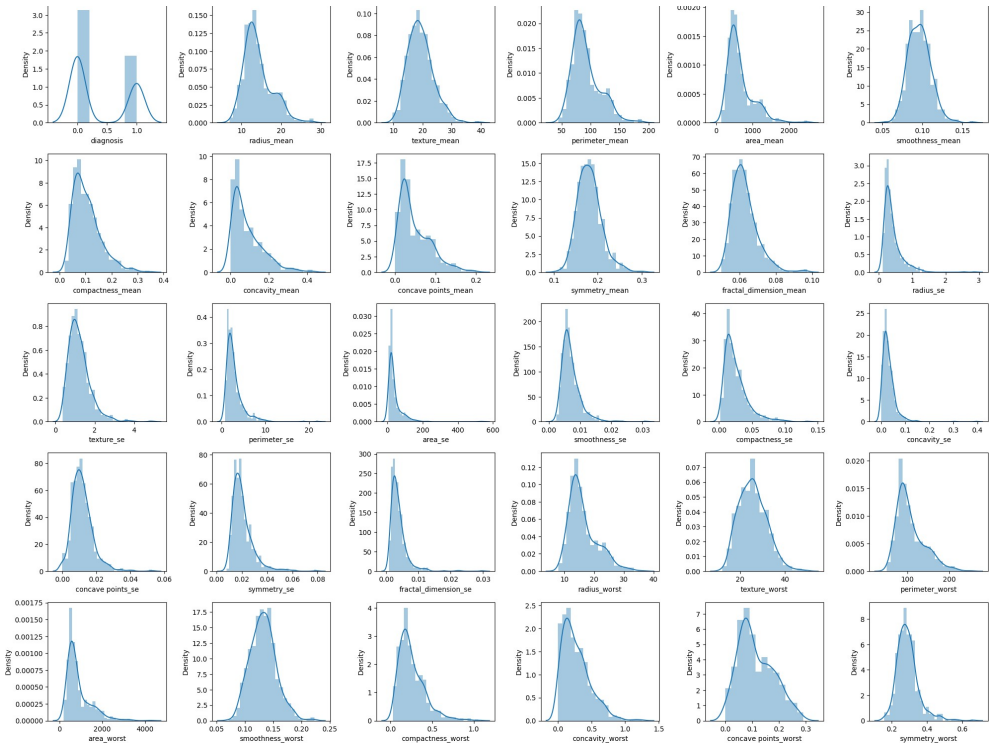
Dengan menggunakan `value_counts()`, kita dapat melihat seberapa banyak sampel yang termasuk dalam kategori 'malignant' dan 'benign'. Ini membantu

dalam memahami sebaran data, yang mempengaruhi cara kita menyesuaikan model. Misalnya, jika satu kelas memiliki jumlah sampel yang jauh lebih banyak daripada yang lain, hal ini dapat memengaruhi bagaimana model kita diterapkan dan dievaluasi. Analisis distribusi kelas ini digunakan untuk menyesuaikan strategi preprocessing dan penggunaan metrik evaluasi, agar model dapat memberikan prediksi yang akurat terhadap dataset yang sesungguhnya.

3.4 Visualisasi Data

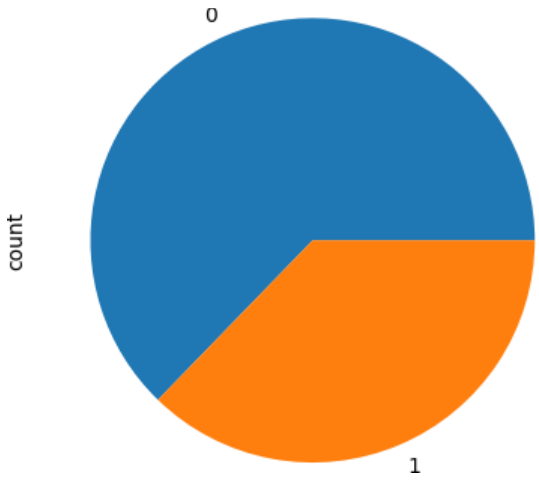
3.4.1 Plot Histogram

Grafik ini memberikan gambaran tentang bagaimana nilai-nilai dari setiap variabel dalam dataset yang terkait dengan analisis diagnosis kanker payudara tersebar dan didistribusikan.



Gambar 3.3 Plot Histogram Distribusi Variabel

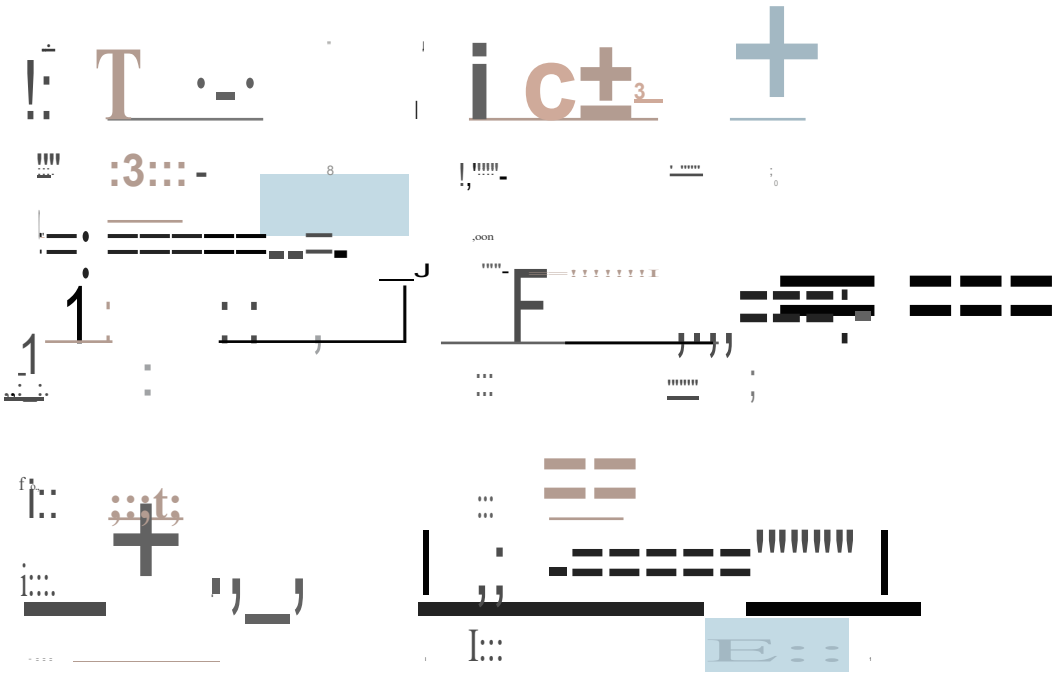
3.4.2 Pie Chart

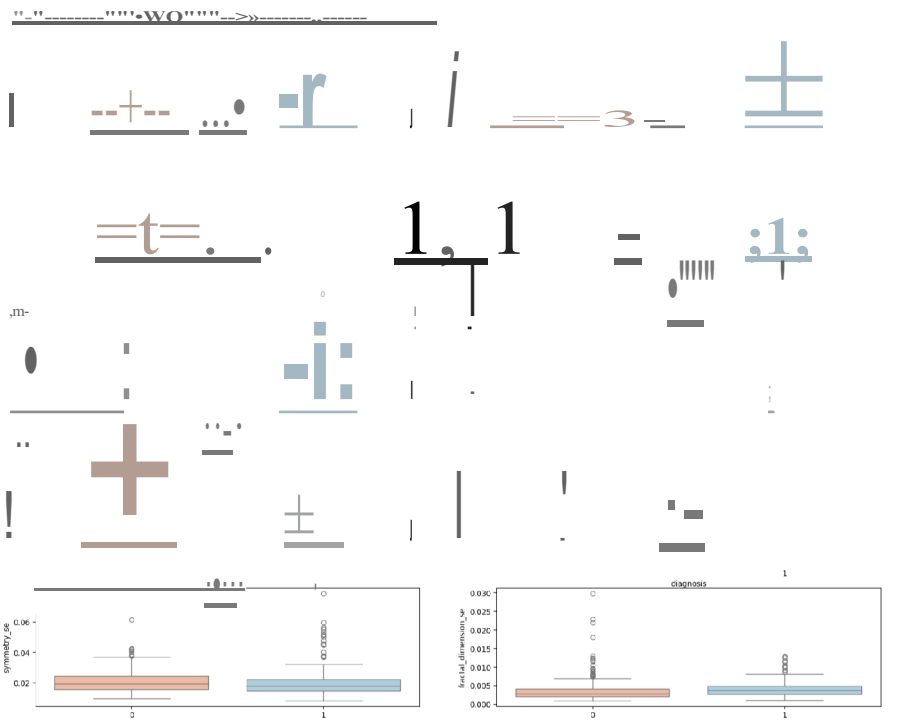


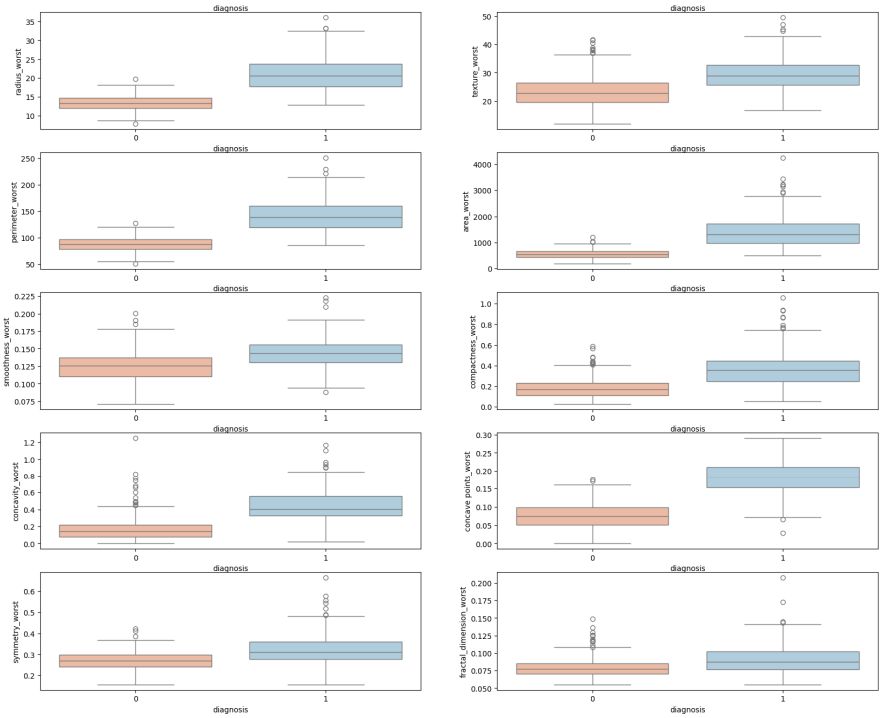
Gambar 3.4 Pie Chart Nilai M = 1 dan B = 0

3.4.3 Box Plots

Visualisasi ini memungkinkan untuk membandingkan bagaimana distribusi nilai fitur berbeda antara dua kelas, memberikan pemahaman yang lebih baik tentang seberapa efektif fitur-fitur numerik dapat membedakan antara kelas-kelas tersebut dalam kasus diagnosis kanker payudara.

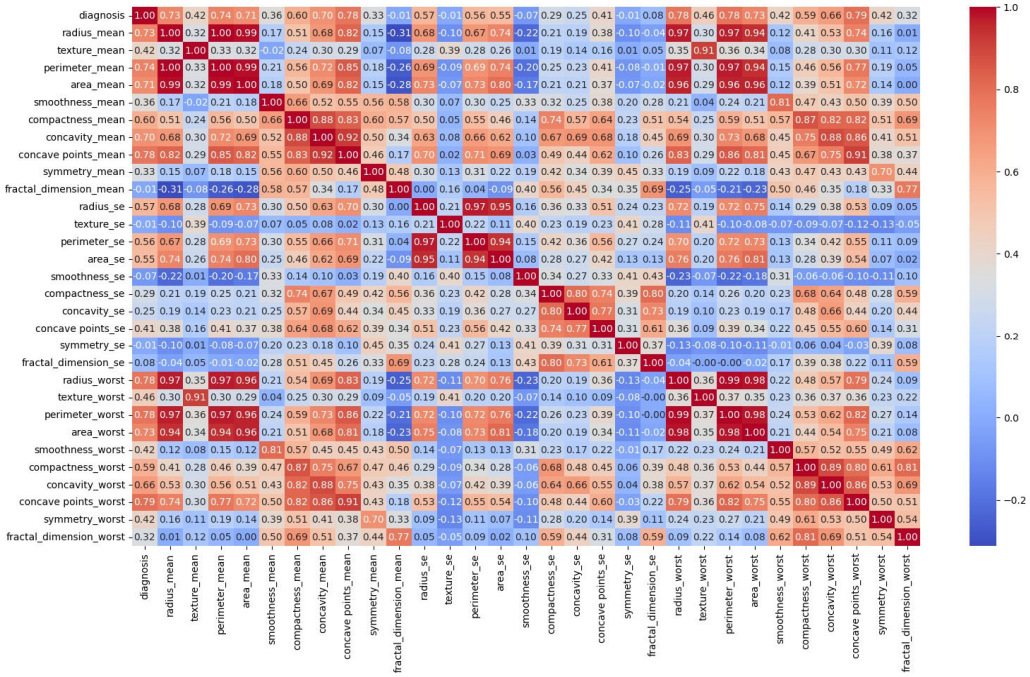



symmetry_se
0.06
0.04
0.02
0
1
diagnosis
fractal_dimension_se
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
1



Gambar 3.5 Box Plot Distribusi Setiap Variabel

3.5 Korelasi Antar Variabel



Gambar 3.6 Heatmap Korelasi Antar Variabel

Nilai korelasi yang ada pada *heatmap* tersebut memiliki rentang -1 hingga 1 yang menunjukan hubungan antara variabel tersebut akan berbanding lurus atau berlawanan. Nilai korelasi yang mendekati nilai 1 memiliki arti bahwa hubungan antara kedua variabel semakin kuat, baik positif maupun negatif. Dibawah ini nilai dari korelasi antara variabel independen dan variabel y yang merupakan variabel dependen yakni diagnosis.

Variabel	**Nilai Korelasi**
diagnosis	1
concave points_worst	0.793566
perimeter_worst	0.782914
concave points_mean	0.776614
radius_worst	0.776454
perimeter_mean	0.742636
area_worst	0.733825
radius_mean	0.730029
area_mean	0.708984
concavity_mean	0.69636
concavity_worst	0.65961
compactness_mean	0.596534
compactness_worst	0.590998
radius_se	0.567134
perimeter_se	0.556141
area_se	0.548236
texture_worst	0.456903
smoothness_worst	0.421465

symmetry_worst	0.416294
texture_mean	0.415185
concave points_se	0.408042
smoothness_mean	0.35856
symmetry_mean	0.330499
fractal_dimension_worst	0.323872
compactness_se	0.292999
concavity_se	0.25373
fractal_dimension_se	0.077972
symmetry_se	-0.00652
texture_se	-0.0083
fractal_dimension_mean	-0.01284
smoothness_se	-0.06702

Berdasarkan nilai tersebut, variabel seperti radius_mean, perimeter_mean, area_mean, concave points_mean, radius_worst, perimeter_worst, area_worst, concave points_worst. Fitur-fitur ini memiliki nilai korelasi tinggi (0.70 ke atas), menunjukkan bahwa mereka berkaitan erat dengan diagnosa.

3.6 Pembuatan Model Artificial Intelligence

Prediksi kanker payudara dilakukan dengan menggunakan algoritma K-Nearest Neighbour (KNN). Setelah mendapatkan nilai akurasi dengan algoritma KNN, kami melakukan analisis dengan melakukan perbandingan dengan algoritma yang lain yakni Artificial Neural Network (ANN) dan Bagging Random Forest.

3.6.1 K-Nearest Neighbour (KNN)

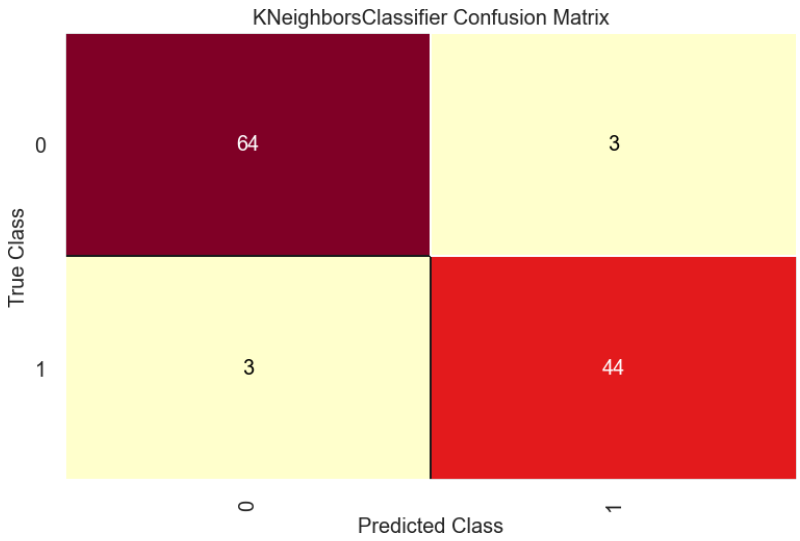
```
Accuracy : 94.73684210526315%

Confusion Matrix :

[[64  3]
 [ 3 44]]

Classification Report :

              precision    recall  f1-score   support

           0       0.96      0.96      0.96        67
           1       0.94      0.94      0.94        47

    accuracy                           0.95       114
   macro avg       0.95      0.95      0.95       114
weighted avg       0.95      0.95      0.95       114
```

Gambar 3.7 Evaluasi Model KNN

Hasil Evaluasi model K-Nearest Neighbour (KNN) menunjukan bahwa:

- Precision: Model memiliki tingkat keakuratan sekitar 94% dalam mengklasifikasikan data positif.
- Recall: Model berhasil mengidentifikasi sekitar 94% dari keseluruhan data positif yang tersedia.
- F1-score: Merupakan rata-rata harmonis antara precision dan recall, yang dalam kasus ini sekitar 94%.

Hal tersebut juga dapat terlihat dari Confusion Matrix dibawah ini yang dapat mengevaluasi model KNN yang dibuat.



Gambar 3.8 Confusion Matrix Model KNN

Dari Confusion Matrix tersebut dapat diinterpretasikan sebagai berikut:

- True Positives (TP): Model KNN berhasil mengidentifikasi 44 sampel sebagai positif yang benar (kelas 1).
- True Negatives (TN): Model KNN berhasil mengidentifikasi 64 sampel sebagai negatif yang benar (kelas 0).
- False Positives (FP): Model KNN salah mengklasifikasikan 3 sampel negatif sebagai positif (kelas 0 diprediksi sebagai kelas 1).
- False Negatives (FN): Model KNN salah mengklasifikasikan 3 sampel positif sebagai negatif (kelas 1 diprediksi sebagai kelas 0).

Sehingga dapat diambil kesimpulan yang sama dengan hasil dari model KNN yang dibuat.

Dengan akurasi keseluruhan sekitar 95%, model ini menunjukkan kemampuan yang sangat baik dalam melakukan prediksi pada data yang diberikan.

3.6.2 Artificial Neural Network (ANN)

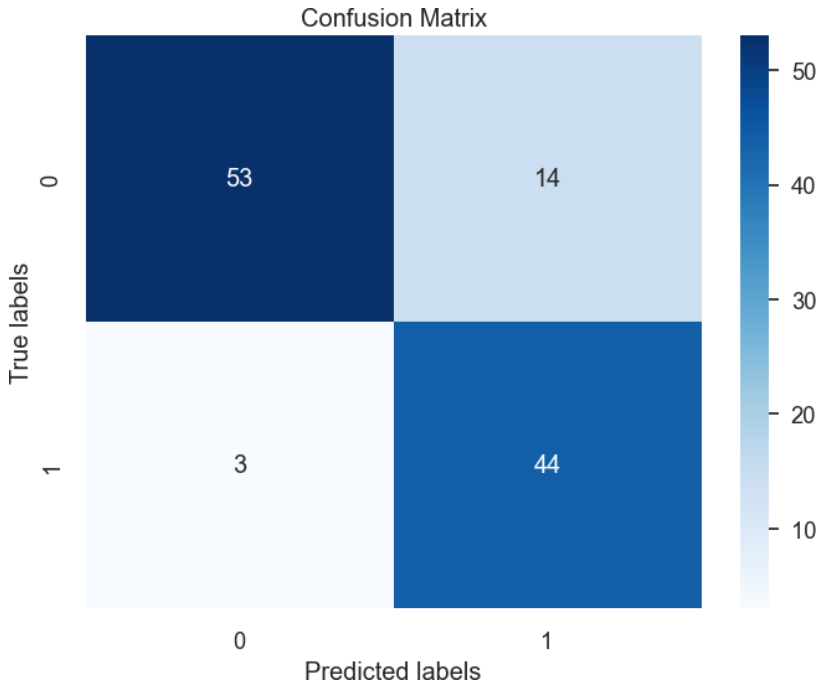
```
Precision, Recall, F1-score for ANN:
============================================================
- Precision: 0.8689990061360298
- Recall: 0.8508771929824561
- F1-score: 0.8520202946390367
============================================================


Classification Report for ANN:
============================================================
              precision    recall  f1-score   support

           0       0.95      0.79      0.86        67
           1       0.76      0.94      0.84        47

    accuracy                           0.85       114
   macro avg       0.85      0.86      0.85       114
weighted avg       0.87      0.85      0.85       114

============================================================
Accuracy of ANN: 85.08771929824562
============================================================
```

Gambar 3.9 Evaluasi Model ANN

Hasil Evaluasi model Artificial Neural Network (ANN) menunjukan bahwa:

- Precision: Model memiliki tingkat keakuratan sekitar 76% dalam mengklasifikasikan data positif.
- Recall: Model berhasil mengidentifikasi sekitar 94% dari keseluruhan data positif yang tersedia.
- F1-score: Merupakan rata-rata harmonis antara precision dan recall, yang dalam kasus ini sekitar 85%.

Hal tersebut juga dapat terlihat dari Confusion Matrix dibawah ini yang dapat mengevaluasi model ANN yang dibuat.



Gambar 3.10 Confusion Matrix Model ANN

Dari Confusion Matrix tersebut dapat diinterpretasikan sebagai berikut:

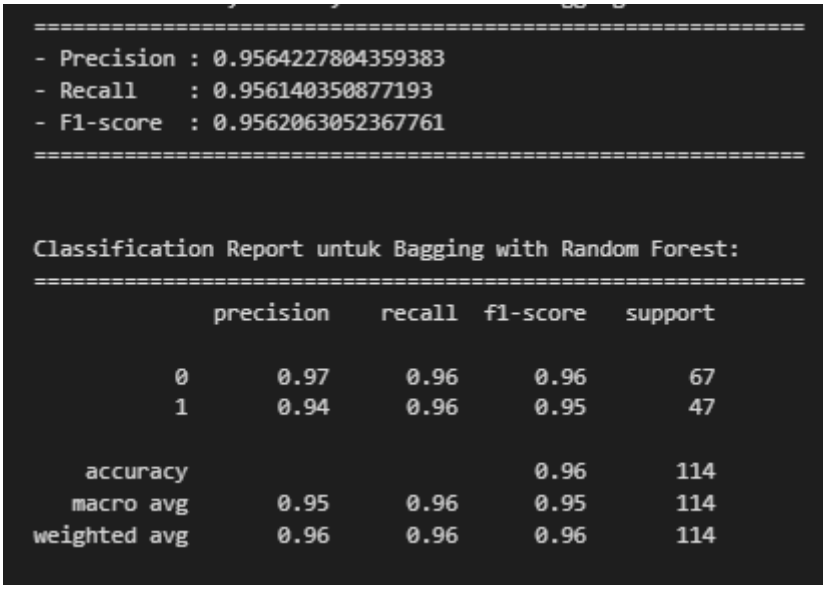
- True Positives (TP): Model ANN berhasil mengidentifikasi 44 sampel sebagai positif yang benar (kelas 1).
- True Negatives (TN): Model ANN berhasil mengidentifikasi 53 sampel sebagai negatif yang benar (kelas 0).
- False Positives (FP): Model ANN salah mengklasifikasikan 14 sampel negatif sebagai positif (kelas 0 diprediksi sebagai kelas 1).
- False Negatives (FN): Model ANN salah mengklasifikasikan 3 sampel positif sebagai negatif (kelas 1 diprediksi sebagai kelas 0).

Sehingga dapat diambil kesimpulan yang sama dengan hasil dari model KNN yang dibuat.

Dengan akurasi keseluruhan sekitar 85%, model ini menunjukkan

kemampuan yang sangat baik dalam melakukan prediksi pada data yang diberikan.

3.6.3 Bagging with Random Forest (Bagging RF)



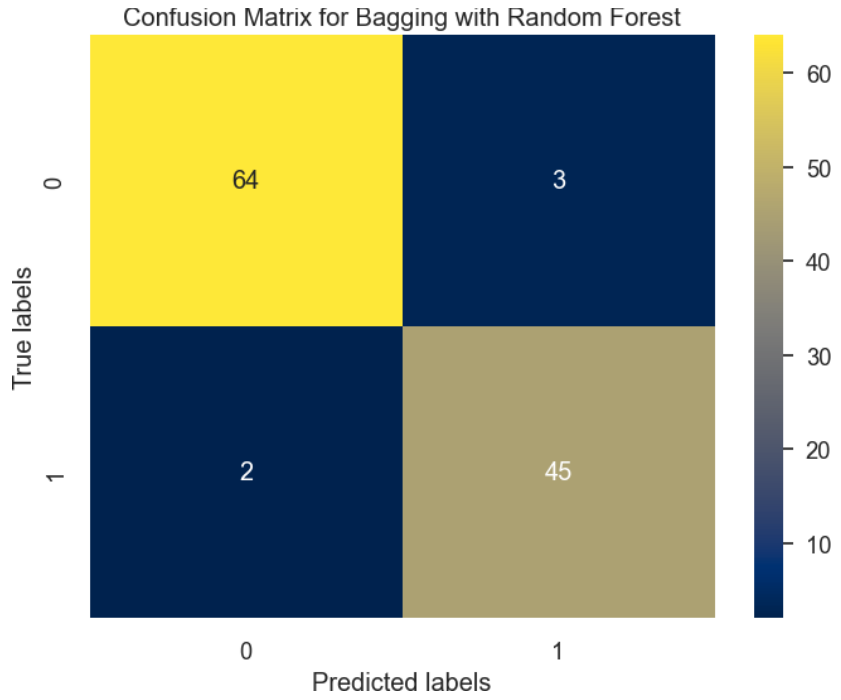
```
============================================================
- Precision : 0.9564227804359383
- Recall    : 0.956140350877193
- F1-score  : 0.9562063052367761
============================================================


Classification Report untuk Bagging with Random Forest:
============================================================
              precision    recall  f1-score   support

           0       0.97      0.96      0.96        67
           1       0.94      0.96      0.95        47

    accuracy                           0.96       114
   macro avg       0.95      0.96      0.95       114
weighted avg       0.96      0.96      0.96       114
```

Gambar 3.11 Evaluasi Model Bagging RF

Hasil evaluasi model Bagging dengan Random Forest menunjukkan bahwa:

- Precision: Model memiliki tingkat keakuratan sekitar 94% dalam mengklasifikasikan data positif.
- Recall: Model berhasil mengidentifikasi sekitar 96% dari keseluruhan data positif yang tersedia.
- F1-score: Merupakan rata-rata harmonis antara precision dan recall, yang dalam kasus ini sekitar 95%.

Hal tersebut juga dapat terlihat dari Confusion Matrix dibawah ini yang dapat mengevaluasi model Bagging with Random Forest yang dibuat.



Gambar 3.12 Confusion Matrix Model Bagging RF

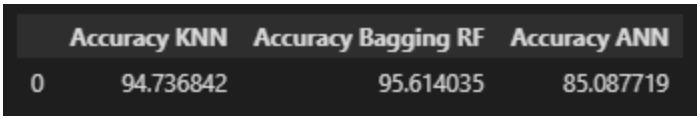
Dari Confusion Matrix tersebut dapat diinterpretasikan sebagai berikut:

- True Positives (TP): Model Bagging dengan Random Forest berhasil mengidentifikasi 45 sampel sebagai positif yang benar (kelas 1).
- True Negatives (TN): Model Bagging dengan Random Forest berhasil mengidentifikasi 64 sampel sebagai negatif yang benar (kelas 0).
- False Positives (FP): Model Bagging dengan Random Forest salah mengklasifikasikan 3 sampel negatif sebagai positif (kelas 0 diprediksi sebagai kelas 1).
- False Negatives (FN): Model Bagging dengan Random Forest salah mengklasifikasikan 2 sampel positif sebagai negatif (kelas 1 diprediksi sebagai kelas 0).

Sehingga dapat diambil kesimpulan yang sama dengan hasil dari model Bagging with Random Forest yang dibuat.

Dengan akurasi keseluruhan sekitar 96%, model ini menunjukkan kemampuan yang sangat baik dalam melakukan prediksi pada data yang diberikan.

3.7 Perbandingan Evaluasi Model



	Accuracy KNN	Accuracy Bagging RF	Accuracy ANN
0	94.736842	95.614035	85.087719

Gambar 3.13 Perbandingan Evaluasi Model

Setelah prediksi menggunakan beberapa model artificial intelligence atau machine learning, didapatkan hasil sebagai berikut:

- K-Nearest Neighbors (KNN) : 94.74%
- Bagging Random Forest (Bagging RF): 95.61%
- Artificial Neural Network (ANN): 85.09%

BAB IV

KESIMPULAN DAN SARAN

4.1 Kesimpulan

Dalam analisis dataset kanker payudara, visualisasi data sangat penting karena membantu kita memahami bagaimana kelas kanker (benign atau malignant) didistribusikan di dalam dataset, serta fitur-fitur apa yang paling relevan untuk prediksi. Kami menggunakan metode seperti Random Forest dan K-Nearest Neighbor (KNN) untuk membandingkan seberapa baik mereka dalam memprediksi jenis kanker.

Khususnya, KNN menonjol karena tingkat keakuratannya yang tinggi dalam membedakan kanker sebagai benign atau malignant. Dengan menerapkan visualisasi seperti histogram dan analisis `value_counts()`, kami mendapatkan pemahaman yang lebih dalam tentang bagaimana kelas kanker didistribusikan dan karakteristik numerik fitur-fitur dalam dataset. Informasi ini krusial dalam memilih model yang paling tepat untuk membangun prediksi yang akurat. Pendekatan ini tidak hanya membantu meningkatkan efisiensi penggunaan model dalam praktek klinis, tetapi juga mendukung upaya kami dalam menyumbangkan wawasan berharga terkait manajemen dan pemahaman lebih baik tentang kanker payudara di bidang medis.

4.2 Saran

Dari analisis yang telah kami lakukan, kami memiliki beberapa saran untuk mengembangkan pendekatan analisis data kanker payudara. Pertama, gunakan teknik ensemble seperti Bagging dan Boosting untuk meningkatkan akurasi prediksi. Selanjutnya, penting untuk mempertimbangkan penyeimbangan data agar model dapat mengatasi perbedaan jumlah sampel antara kelas malignan dan benign dengan lebih baik. Kolaborasi antara tim data science dan profesional medis juga sangat diperlukan untuk memastikan implementasi model yang efektif dalam praktik klinis sehari-hari. Dengan mengikuti saran ini, diharapkan pengelolaan kanker payudara melalui pendekatan analisis data dapat lebih terarah dan berhasil.