ANALISA PERBANDINGAN KINERJA ALGORITMA KLASIFIKASI UNTUK PREDIKSI PENYAKIT KANKER PAYUDARA

Rudi Hartono, Yusuf Sumaryana, Alwi Nurfaizi

Program Studi Teknik Informatika, Universitas Perjuangan
Jl. Peta No.177, Kahuripan, Kec. Tawang, Tasikmalaya
rudihartono@unper.ac.id, yusufsumaryana@unper.ac.id, alwinurfaizi20@gmail.com

Abstract *- Data uploaded by Globocan in 2020, the number of new cases of breast cancer in Indonesia reached 68,858 cases (16.6%) from a total of 396,914 new cases of cancer, with the number of deaths reaching more than 22 thousand cases. In the health and medical fields, machine learning-based classification is widely used to help doctors and health professionals classify cancer diseases to determine the best course of action, that there are several algorithms, techniques and tools used in predicting breast cancer with different accuracy results. By comparing the classification algorithm and method of Ensemble Learning Bagging and the usual method in this study and by using data derived from SEER NCI uploaded to the Kaggle.com website with the number of attributes that are 16 and the number of records is 4025 records, that the results of the comparison of algorithm models either use the Ensemble Learning Bagging technique or not, The algorithm with the best accuracy performance value is the Random Forest algorithm, all accuracy performance on each algorithm increased by an average increase of 1% using the Ensemble Learning Bagging technique, showing that this technique can be used to improve accuracy performance compared to the default technique In each algorithm or with manual calculations, besides that the number of records can also affect the performance of the algorithm. By adding other algorithm models such as AdaBoost, logistic Regression, XGBClassifier, LGBMClassifier, ExstraTreeClasifier, and Heterogenous Ensemble Learning techniques as well as the process of making Machine Learning with the Streamlit framework, the process for predicting breast cancer will be better.*
Keywords *- accuracy, algorithm, breast cancer, classification.*

Abstrak - Data yang diunggah *Globocan* pada tahun 2020, jumlah kasus baru kanker payudara di Indonesia mencapai 68.858 kasus (16,6%) dari total 396.914 kasus baru kanker, dengan jumlah kematiannya mencapai lebih dari 22 ribu jiwa kasus. dibidang kesehatan dan medis, klasifikasi berbasis pembelajaran mesin banyak digunakan untuk membantu dokter dan profesional kesehatan mengklasifikasikan penyakit kanker untuk menentukan tindakan terbaik, bahwa ada beberapa algoritma, teknik dan alat yang dipakai dalam memprediksi penyakit kanker payudara dengan hasil akurasi yang berbeda-beda. Dengan membandingkan algoritma klasifikasi dan metode *Ensemble Learning Bagging* dan metode biasa pada penelitian ini serta dengan menggunakan data yang berasal dari SEER NCI yang diunggah ke dalam website Kaggle.com dengan jumlah atribut yaitu 16 dan jumlah record nya sebesar 4025 record, bahwa hasil dari perbandingan model algoritma 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. Dengan menambahkan model algoritma lain nya seperti *AdaBoost, logistic Regression, XGBClassifier, LGBMClassifier, ExstraTreeClasifier,* dan teknik *Ensemble Learning Heterogenous* serta proses pembuatan *Machine Learning* dengan *framework Streamlit* maka proses untuk prediksi penyakit kanker payudara akan semakin baik.
Kata kunci - akurasi, algoritma, kanker payudara, klasifikasi.

I. PENDAHULUAN

Kanker adalah proses penyakit yang bermula ketika sel telah mengalami kehilangan pengendalian dan mekanisme normalnya sehingga mengalami pertumbuhan yang tidak normal, cepat, tidak terkendali serta mengabaikan sinyal pengatur pertumbuhan dalam lingkungan sekitar sel tersebut.

Salah satu jenis dari kanker adalah kanker payudara yang merupakan salah satu penyebab kematian tertinggi pada wanita, penyakit ini merupakan penyebab kematian nomor dua setelah kanker paru-paru, menurut organisasi kesehatan dunia, satu juta wanita didiagnosa menderita kanker payudara setiap tahunnya, dan setengah dari mereka akhirnya meninggal.

Sekitar Rp 7,6 triliun dana BPJS digunakan untuk pengobatan anti kanker pada 2019-2020. Menurut data *Globocan* tahun 2020, jumlah kasus kanker payudara naik menjadi 68.858 kasus (16,6%) dari total 396.914 kasus kanker baru di Indonesia dengan jumlah kematian yang kini mencapai lebih dari 22.000 kasus, maka perlu dilakukan skrining dini kanker payudara yang sangat

penting.

Klasifikasi adalah salah satu algoritma *data mining* yang mengklasifikasikan data ke dalam kriteria atau kelas tertentu dengan membaca data yang sudah ada sebelumnya. Beberapa algoritma klasifikasi yang umum digunakan adalah *Support Vector Machine, Decision Tree* dan *Naive Bayes, K-Nearest Neighbors, Random Forest.*

Pada penelitian terdahulu[2] menerapkan penelitian Analisa Perbandingan Algoritma Klasifikasi *support vector machine, decision tree* dan *naive bayes*, menghasilkan nilai akurasi rata-rata untuk algoritma *decision tree* sebesar 87.43 % dan meru- pakan nilai akurasi tertinggi dari 3 perbandingan algoritma untuk algoritma *support vector machine* adalah sebesar 87.16% dan algoritma *naïve bayes* menghasilkan 84.92% perbandingan ini diterapkan pada *chronic kidney* dan *breast cancer dataset.*

Pada penelitian[5] melakukan penelitian komparasi algoritma *data mining* untuk klasifikasi kanker payudara, pada penelitian tersebut menggunakan *dataset public* yaitu *breast cancer wisconsin* yang diambil dari *UCI repository*. *Dataset* ini banyak digunakan oleh peneliti lain, hasil yang diperoleh dalam penelitian ini menunjukan bahwa algoritma *naïve bayes* merupakan algoritma terbaik dengan tingkat akurasi sebesar 95,85% sedangkan untuk algoritma lainnya seperti *k-nearest neighbors* dan *decision tree* C4.5 hanya memperoleh tingkat nilai akurasi masing-masing 94,70% dan 94,71%.

Lalu dalam penelitian[1] penelitian teknik data mining untuk prediksi kanker payudara yang efisien, dimana dalam penelitian ini menggunakan algoritma pengklasifikasian tertentu. Bahwa hasilnya algoritma *support vector machine* mengungguli pengklasifikasi lain dalam hal kinerja akurasi.

Berdasarkan pemahaman dan pengetahuan yang sudah didapatkan oleh penulis melalui studi dari jurnal penelitian terdahulu serta studi Pustaka, bahwa ada beberapa algoritma, teknik dan alat yang dipakai dalam memprediksi penyakit kanker payudara dengan hasil akurasi yang berbeda-beda pada setiap tahun nya.

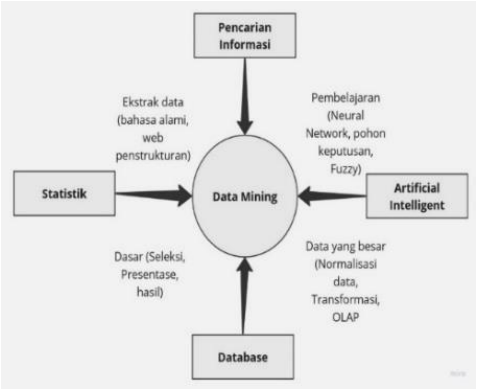
Dengan adanya masalah tersebut serta ada solusi untuk mengatasi penerapan metode penelitian menggunakan data mining terhadap pasien untuk mendiagnosa penyakit serta membandingkan algoritma klasifikasi yaitu *Support Vector Machine, Decision Tree dan Naive Bayes, K-Nearest Neighbors, Random Forest* untuk menghasilkan nilai performa tertinggi dengan bahasa pemrograman *Python* untuk proses perhitungan dengan *dataset* yang digunakan dari basis data kanker payudara *SEER Program of the NCI* nantinya akan dipilih algoritma dengan nilai performa tertinggi untuk digunakan memprediksi penyakit kanker payudara.

Maka tujuan dari penelitian ini adalah untuk menemukan sebuah algoritma Klasifikasi terbaik antara *Support Vector Machine, Decision Tree dan Naive Bayes, K-Nearest Neighbors, Random Forest* untuk memprediksi penyakit kanker payudara menggunakan bahasa pemrograman *Python.*

A. Data Mining

Merujuk pada jurnal yang dibuat oleh [6] Penambangan data atau *Data Mining* merupakan istilah yang digunakan untuk menggambarkan, menemukan informasi dalam basis data. Penambangan data adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan pembelajaran mesin untuk mengekstraksi dan mengidentifikasi informasi yang berguna dan informasi terkait dari basis data besar. Penambangan data bukanlah bidang yang sepenuhnya baru.

Gambar 1 menunjukkan bahwa penambangan data memiliki akar panjang dalam disiplin ilmu seperti kecerdasan buatan, pembelajaran me- sin, statistik, basis data, dan pencarian informasi.



Gambar 1. Akar *datamining*

B. Klasifikasi

Klasifikasi merupakan salah satu peranan utama data mining. Proses klasifikasi adalah proses menghitung data yang sudah ada atau disebut juga data latih dengan data baru atau data uji. Proses ini menciptakan peluang untuk mengambil sampel data. Catatan yang digunakan dalam klasifikasi data harus memiliki pengenal target atau atribut. Beberapa algoritma dapat digunakan untuk menghitung proses klasifikasi. Algoritma ini termasuk *k-neighbors* terdekat, *naïve bayes*, dan pohon keputusan C4.5.

C. Algoritma

Algoritma merupakan urutan langkah-langkah yg wajib diikuti pada matematika atau perhitungan buat memecahkan perkara lain, terutama komputer. Menurut Thomas H Cormen (2009:5) algoritma adalah proses komputasi yang mengambil beberapa nilai atau kumpulan nilai sebagai input dan kemudian memprosesnya sebagai output, jadi algoritma adalah serangkaian langkah komputasi yang mengubah input menjadi output.

Penemu algoritma ini adalah Abu Ja`far Muhammad Ibnu Musa Al-Khawarizmi, seorang ahli geografi, astrologi, astronomi dan matematika sosok yang lahir di Khawarizm, sebuah kota kecil di Khwarezmia (sekarang Uzbekistan), sekitar tahun 780M.

D. Naïve Bayes

Teorema bayes sendiri adalah model matematika yang didasarkan pada statistik dan probabilitas. Algoritma *naive bayes* sering digunakan dalam *filter*

spam, analisis sentimen, analisis prediksi dan sistem rekomendasi. Salah satu alasan utama untuk menggunakan algoritma ini adalah cepat dan mudah diimplementasikan. Tetapi, *naive bayes* membutuhkan fungsi atau prediktor independen.

Korelasi kelas-atribut dianalisis menggunakan teknik klasifikasi ini untuk menentukan probabilitas bersyarat untuk hubungan antara nilai-nilai atribut. Data latih kelas kursus digunakan untuk menghitung setiap jenis kemungkinan. P(C=c) disebut sebagai "probabilitas sebelumnya". Algoritma juga mempertimbangkan apakah x menerima c selain probabilitas sebelumnya, karena karakteristik diasumsikan independen. Kemungkinan setiap fitur dikalikan dengan probabilitas ini untuk sampai pada hasil akhir. Distribusi frekuensi set pelatihan dapat digunakan untuk memperkirakan probabilitas[1].

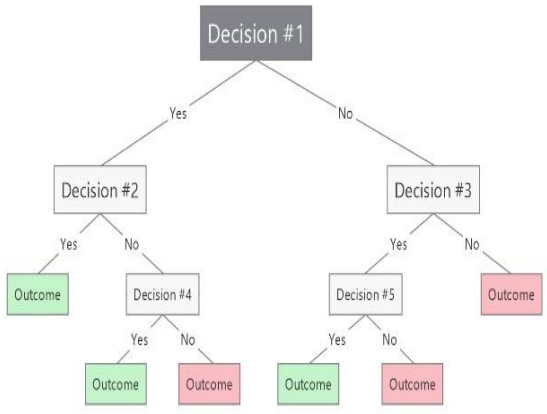
Teorema Bayes dinyatakan secara matematis pada persamaan berikut:

$$P(A|B)\frac{P(B|A)P(A)}{P(B)}$$

Dimana P(B) != 0

1. Pada dasarnya kita mencoba mencari peluang kejadian A jika kejadian B benar. Peristiwa B disebut juga pembuktian.
2. P(A) adalah prior dari A (probabilitas sebelumnya, yaitu probabilitas suatu peristiwa sebelum bukti muncul). Petunjuknya adalah nilai atribut dari entitas yang tidak diketahui (Peristiwa B).
3. P(A|B) adalah probabilitas posterior dari B, yaitu kemungkinan terjadinya setelah bukti disajikan. Ciri utama dari algoritma Naive Bayes Classifier adalah adanya asumsi yang sangat kuat (naif) tentang independensi setiap kondisi/peristiwa.

E. Decision Tree



Gambar 2. Pohon keputusan

Pohon keputusan adalah alat struktur seperti pohon yang memodelkan kemungkinan hasil, biaya sumber daya, manfaat, dan konsekuensi yang mungkin terjadi. Pohon keputusan menyediakan cara untuk merepresentasikan algoritma dengan pernyataan kontrol bersyarat.

Struktur *flowchart* berisi node internal yang mewakili tes atau atribut di setiap langkah.

Pada setiap cabang mewakili hasil atribut, sedangkan jalur dari daun ke akar mewakili aturan klasifikasi tersebut. Pohon keputusan adalah salah satu bentuk terbaik dari algoritma pembelajaran berdasarkan metode pembelajaran yang berbeda. Algoritma ini secara umum:

1. Pilih atribut sebagai *root*
2. Buat cabang untuk setiap nilai
3. Bagi kasus dalam cabang
4. Ulangi setiap cabang sampai semua kasus dalam cabang memiliki kelas yang sama pilih atribut berdasarkan mengulangi proses untuk nilai "*gain*" tertinggi dari atribut yang ada.

Perhitungan gain

$$\text{Gain } (S,\text{A}) = Entropy(S) - \sum_{i=1}^{n} * Entropy(Si)$$

Keterangan
S : himpunan
A : atribut
n : jumlah partisi atribut A
| Si | : jumlah kasus pada partisi ke-i
| S |: jumlah kasus dalam S

Menghitung nilai Entropy

$$\text{Entropy(S) - } \sum_{i=1}^{n} - pi * log_2 pi$$

S=Himpunan Kasus
A : fitur
n : jumlah partisi S
pi : proporsi dari Si terhadap S

Perincian algoritma

1. Menghitung jumlah kasus seluruhnya, jumlah berkeputusan “Yes” maupun “No”.
2. Menghitung *entropy* dari semua kasus yg terbagi berdasarkan atribut.
3. Lakukan penghitungan *gain* utk setiap atributnya.

F. Random Forest

Merujuk pada jurnal yang ditulis oleh [8] *Random Forest* merupakan salah satu metode algoritma yang digunakan untuk klasifikasi dan regresi. Metode ini merupakan sebuah *ensemble* (kumpulan) metode pembelajaran menggunakan pohon keputusan sebagai *base classifier* yang dibangun dan dikombinasikan, ada tiga aspek penting dalam metode *random forest*, yaitu:

1. Melakukan *bootstrap sampling* untuk membangun pohon prediksi.
2. Masing-masing pohon keputusan memprediksi dengan prediktor acak.
3. Random Forest kemudian membuat prediksi dengan menggabungkan hasil dari setiap pohon keputusan menggunakan mayoritas klasifikasi atau mean regresi.

G. Support Vector Machine

Support Vector Machine atau disebut SVM merupakan metode *Machine Learning* yang memungkinkan analisis dan penyortiran data ke dalam

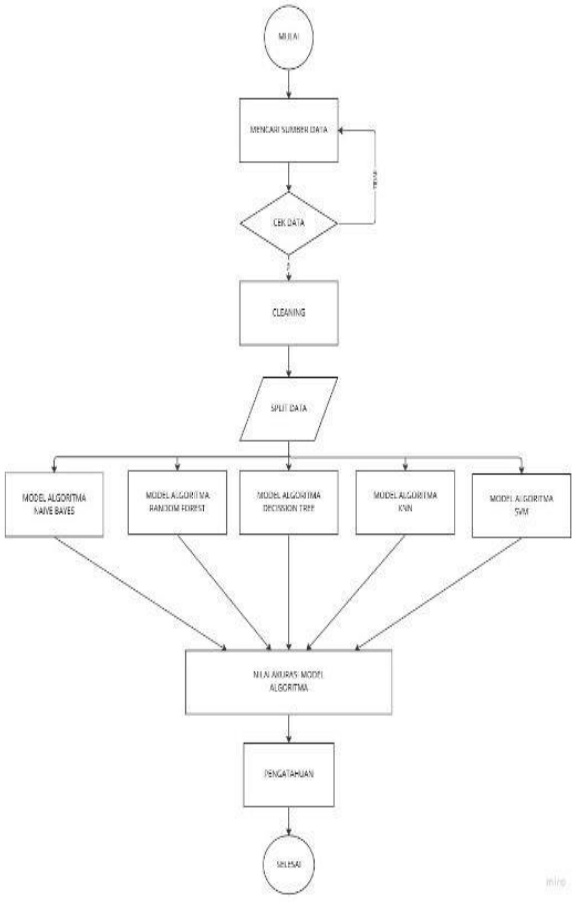
dua kategori.

SVM yang bertujuan untuk menemukan *hyperplane* atau fungsi yang diskriminan (batas keputusan) terbaik untuk memisahkan antara dua atau lebih kelas di ruang input. Sebuah *hyperplane* dapat berupa garis atau garis dalam dua dimensi dan bidang datar dalam beberapa bidang yaitu kanker payudara. Kanker payudara merupakan salah satu penyebab kematian tertinggi pada wanita, penyakit ini merupakan penyebab kematian kedua setelah kanker paru-paru.

Menurut organisasi kesehatan dunia, satu juta wanita didiagnosis menderita kanker payudara setiap tahun dan separuh dari mereka akhirnya meninggal, pada umumnya hal ini disebabkan penanganan dini serta pengobatan yang lambat mengakibatkan kanker baru terdeteksi setelah memasuki stadium akhir.

II. METODE PENELITIAN

Database kanker payudara ini diperoleh dari basis data kanker payudara *SEER Program of the NCI*. Data yang digunakan berjumlah 4025 *record* yang terdiri dari 16 atribut, atribut class digunakan sebagai label yang memiliki dua kemungkinan kanker tersebut *Dead* atau *Alive.* Langkah-langkah yang digunakan dalam penelitian ini adalah sebagai berikut:



Gambar 3. Metode penelitian

Penelitian menggunakan dataset yang diperoleh dari basis data kanker payudara *SEER Program of the NCI*, yang diupload ke *website www.kaggle.com*, dengan melalui tahapan pengecekan apakah *dataset* dapat digunakan serta memiliki sumber yang jelas, serta atribut dan *record* yang telah memenuhi untuk melakukan penelitian dan menghilangkan dataset yang kosong dan duplikat, maka dari total 4024 data menjadi 4004 setelah me- lalui proses *cleaning*.

Setelah melalui tahap *cleaning* data akan dibagi menjadi dua bagian yaitu untuk *data training* akan diberi 80% dari total *record* dan *data testing* akan diberi 20% dari total *record* untuk perhitungan menggunakan *library scikit learn.*

Untuk perhitungan secara manual menggunakan sampel data dengan jumlah 10 record data akan dibagi menjadi dua bagian yaitu data *training* akan diberi 9 dari total *record* dan *data testing* akan diberi 1 dari total *record*. setelah data sudah melalui proses *split data* maka akan memasuki proses pemodelan pada setiap algoritma yaitu *naïve bayes, random forest, decision tree, support vector machine, k-nearest neighbors,* setelah itu hasil akurasi pemodelan dari setiap akurasi akan dibandingkan dan dipilih mana yang lebih layak untuk digunakan.

Setelah melalui beberapa tahapan yang telah dilalui maka hasil dari pemodelan dari setiap algoritma yang dibandingkan akan menghasilkan sebuah kesimpulan.

III. HASIL DAN PEMBAHASAN

Berikut ini merupakan hasil implementasi dari algortima klasifikasi dengan hasil sebagai berikut.

Tabel 1. Hasil akurasi

No	Algoritma	Nilai
1	*Random Forest*	0.927591 %
2	*Decision Tree*	0.921348 %
3	*K-Nearest*	0.897628 %
4	*Support Vector Machine*	0.876404 %
5	*Naïve Bayes*	0.823970 %

Dari tabel di atas bahwa Random Forest memiliki hasil tertinggi yaitu 0.927591 %, dalam penelitian ini penulis akan membandingkan algoritma Decision Tree dengan K-Nearest Neighbor dengan perhitungan manual dan perhitungan menggunakan Python, dikarenakan Random Forest merupakan sebuah teknik ensemble learning dari decision tree yaitu dengan cara menggabungkan beberapa metode hingga mencapai hasil yang maksimal.

A. Perhitungan manual algoritma Decision Tree

Dalam perhitungan manual ini penulis menggunakan 10 data sample dengan instance yang sesuai dengan data asli, untuk menentukan Alive atau Dead kriteria yang diperlukan meliputi:

Berdasarkan tabel diatas dibuat tabel keputusan untuk menentukan Alive atau Dead dengan melihat atribut yang diperlukan dengan urutan secara umum sebagai berikut:

1. Pilih atribut sebagai akar.
2. Membuat cabang untuk tiap-tiap nilai.
3. Bagi kasus dalam cabang.

4. Ulangi proses untuk setiap cabang.
5. Memilih atribut berdasarkan nilai gain tertinggi dari atribut yang ada.

Untuk menentukan Node pertama dalam sebuah pohon keputusan maka harus dilakukan perhitungan untuk mencari nilai Entropy dengan rumus berikut.

$$\text{Entropy(S)} - \sum_{i=1}^{n} - pi * log_2 pi$$

Setelah nilai Entropy di dapatkan makan dilanjutkan untuk mencari nilai Gain, nilai Gain dapat diketahui dengan menggunakan rumus berikut.

$$\text{Gain } (S,\text{A}) = Entropy(S) - \sum_{i=1}^{n} * Entropy(Si)$$

Berikut merupakan tabel hasil perhitungan dari seluruh variabel yang digunakan dalam perhitungan manual dengan menggunakan dataset yang berjumlah 10 record dengan instance yang sesuai.

Berikut hasil dari perhitungan semua variabel untuk menemukan nilai Gain dari semua variabel. Variabel 6th Stage memiliki nilai Gain terbesar dari semua variabel sehingga variabel 6th Stage dapat ditetapkan sebagai node 1.

Tabel 2. Nilai *gain* semua variabel

No	Variabel	Gain
1	6TH STAGE	1,058594937
2	MARITAL STATUS	0,448244104
3	N STAGE	0,417091187
4	DIFFERNTIATED	0,373935113
5	T STAGE	0,348244104
6	A STAGE	0,19130425
7	GRADE	0,171230407
8	RACE	0,112404521
9	ESTROGEN STATUS	0,112404521
10	PROSTRONE STATUS	0,004237907
11	SURVIVAL MONTS	-4,963799332
12	REGIONAL NODE POSITIVE	-6,492234443
13	TUMOR SIZE	-6,752385775
14	AGE	-7,154336287
15	REGIONAL NODE EXIMINED	-7,386339327

Variabel 6th *Stage* memiliki nilai Gain terbesar dari semua variabel sehingga variabel 6th Stage dapat ditetapkan sebagai node 1 pada pembuatan pohon keputusan. Berikut node ke 1 pada proses perhitungan Decision tree secara manual dengan variabel *6th Stage*.

Berikut hasil dari perhitungan variabel *6th Stage* untuk menemukan nilai *Gain* dari semua variabel. Variabel *Marital Status* memiliki nilai *Gain* terbesar dari semua variabel sehingga variabel *Marital Status* dapat ditetapkan sebagai *node* ke 2.

Tabel 3. Nilai *gain* variabel *6th Stage*

No	Variabel	*Gain*
1	*MARITAL STATUS*	-0,435839583
2	*N STAGE*	-0,4669925
3	*DIFFERENTIATED*	-0,510148574
4	*T STAGE*	-0,535839583
5	*A STAGE*	-0,692779438
6	*GRADE*	-0,71285328
7	*RACE*	-0,771679166
8	*ESTROGEN STATUS*	-0,771679166
9	*PROSTRONE STATUS*	-0,87984578
10	*SURVIVAL MONTS*	-5,847883019
11	*REGIONAL NODE POSITIVE*	-7,37631813
12	*TUMOR SIZE*	-7,636469462
13	*AGE*	-8,038419974
14	*REGIONAL NODE EXIMINED*	-8,270423014

Variabel Marital Status memiliki nilai Gain terbesar dari semua variabel sehingga variabel Marital Status dapat ditetapkan sebagai node ke 2 pada pembuatan pohon keputusan. Berikut node ke 1 pada proses perhitungan Decision tree secara manual dengan variabel Marital Status.

Berikut hasil dari perhitungan variabel Marital Status untuk menentukan node ke 3. Variabel N Stage memiliki nilai Gain terbesar dari semua variabel sehingga variabel N Stage dapat ditetapkan sebagai node ke 3.

Tabel 4. Nilai gain variabel Marital Status

No	Variabel	*Gain*
1	*N STAGE*	0,0330075
2	*DIFFERENTIATED*	-0,010148574
3	*T STAGE*	-0,035839583
4	*A STAGE*	-0,192779438
5	*GRADE*	-0,21285328
6	*RACE*	-0,271679166
7	*ESTROGEN STATUS*	-0,271679166
8	*PROSTRONE STATUS*	-0,37984578
9	*SURVIVAL MONTS*	-5,347883019
10	*REGIONAL NODE POSITIVE*	-6,87631813
11	*TUMOR SIZE*	-7,136469462
12	*AGE*	-7,538419974
13	*REGIONAL NODE EXIMINED*	-7,770423014

Variabel N Stage memiliki nilai Gain terbesar dari semua variabel sehingga variabel N Stage dapat ditetapkan sebagai node ke 3 pada pembuatan pohon keputusan. Berikut node ke 3 pada proses perhitungan Decision tree secara manual dengan variabel N Stage.

Berikut hasil dari perhitungan variabel N Stage

untuk menentukan node ke 4. Variabel Differentiate memiliki nilai Gain terbesar dari semua variabel sehingga variabel Differentiate dapat ditetapkan sebagai node ke 4.

Tabel 5. Nilai *gain* variabel *N Stage*

No	Variabel	*Gain*
2	*DIFFERENTIATE*	-0,010148574
3	*T STAGE*	-0,035839583
4	*A STAGE*	-0,192779438
5	*GRADE*	-0,21285328
6	*RACE*	-0,271679166
7	*ESTROGEN STATUS*	-0,271679166
8	*PROSTRONE STATUS*	-0,37984578
9	*SURVIVAL MONTS*	-5,347883019
10	*REGIONAL NODE POSITIVE*	-6,87631813
11	*TUMOR SIZE*	-7,136469462
12	*AGE*	-7,538419974
13	*REGIONAL NODE EXIMINED*	-7,770423014

Variabel *Differentiate* memiliki nilai *Gain* terbesar dari semua variabel sehingga variabel *Differentiate* dapat ditetapkan sebagai *node* ke 4 pada pembuatan pohon keputusan. Berikut *node* ke 4 pada proses perhitungan *Decision tree* secara manual dengan variabel *Differentiate.*

Berikut hasil dari perhitungan variabel *Differentiate* untuk menentukan node ke 5. Variabel *T Stage* memiliki nilai *Gain* terbesar dari semua variabel sehingga variabel *T Stage* dapat ditetapkan sebagai *node* ke 5.

Tabel 6. Nilai *gain* variabel *Differentiate*

No	Variabel	*Gain*
2	*DIFFERENTIATE*	-0,010148574
3	*T STAGE*	-0,035839583
4	*A STAGE*	-0,192779438
5	*GRADE*	-0,21285328
6	*RACE*	-0,271679166
7	*ESTROGEN STATUS*	-0,271679166
8	*PROSTRONE STATUS*	-0,37984578
9	*SURVIVAL MONTS*	-5,347883019
10	*REGIONAL NODE POSITIVE*	-6,87631813
11	*TUMOR SIZE*	-7,136469462
12	*AGE*	-7,538419974
13	*REGIONAL NODE EXIMINED*	-7,770423014

Variabel *T Stage* memiliki nilai *Gain* terbesar dari semua variabel sehingga variabel *T Stage* dapat ditetapkan sebagai *node* ke 5 pada pembuatan pohon keputusan. Berikut *node* ke 5 pada proses perhitungan *Decision tree* secara manual dengan variabel *T Stage.*

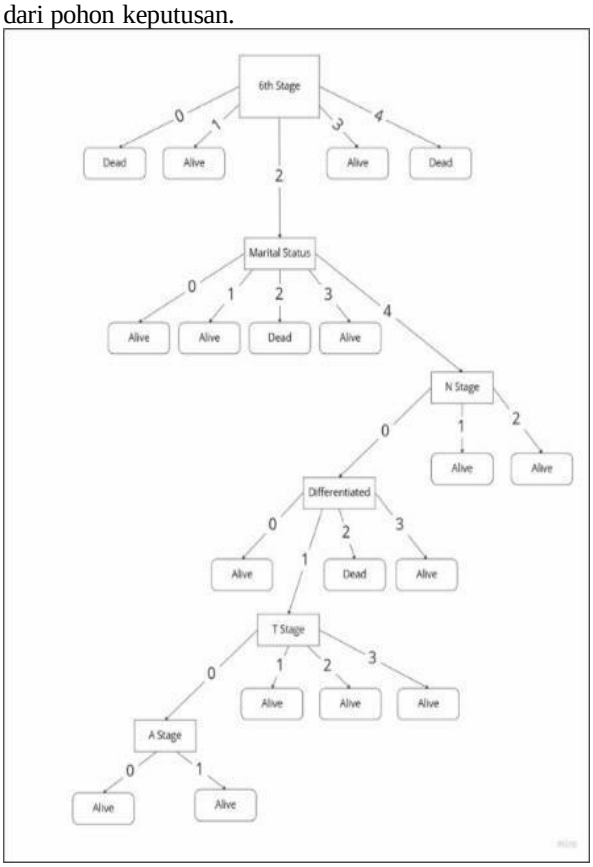
Berikut hasil dari perhitungan variabel *Differentiate* untuk menentukan node ke 5. Variabel *T Stage* memiliki nilai *Gain* terbesar dari semua variabel sehingga variabel *T Stage* dapat ditetapkan sebagai *node* ke 5.

Tabel 7. Nilai *gain* variabel *T Stage*

No	Variabel	*Gain*
4	*A STAGE*	-0,192779438
5	*GRADE*	-0,21285328
6	*RACE*	-0,271679166
7	*ESTROGEN STATUS*	-0,271679166
8	*PROSTRONE STATUS*	-0,37984578
9	*SURVIVAL MONTS*	-5,347883019
10	*REGIONAL NODE POSITIVE*	-6,87631813
11	*TUMOR SIZE*	-7,136469462
12	*AGE*	-7,538419974
13	*REGIONAL NODE EXIMINED*	-7,770423014

Variabel *A Stage* memiliki nilai *Gain* terbesar dari semua variabel sehingga variabel *A Stage* dapat ditetapkan sebagai *node* ke 6 pada pembuatan pohon keputusan. Karena hasil dari perhitungan untuk mencari node telah selesai maka pohon keputusan yang dibuat secara manual memiliki jumlah node sebanyak 6.

Setelah tahap perhitungan *entropy* dan *gain* maka dilakukan pembuatan pohon keputusan yang digambar secara manual, berikut merupakan gambaran manual dari pohon keputusan.



Gambar 4. Pohon keputusan manual

Setelah pohon keputusan diketahui selanjutnya masukan data testing untuk mengukur apakah hasil sesuai.

Tabel 8. Data uji

Age	*Race*	*Marital Status*	*T Stage*	*N Stage*	*6th Stage*	*differentiate*	*Grade*	*A Stage*	*Tumor Size*	*Estrogen Status*	*Progesterone*	*Regional Node*	*Reginol Node*	*Survival*	*Status*
60	2	0	1	0	2	0	2	1	29	0	0	20	1	77	0

Pada data uji ini memiliki atribut Status yang bernilai 0 (Alive), selanjut masukan tiap record kedalam pohon keputusan, diketahui bahwa 6Th stage bernilai 2 dan Marital Status bernilai 0 maka hasilnya alive, maka hasil dari perhitungan tersebut, bahwa dari data uji yang digunakan yaitu 0 (Alive).

A. Perhitungan manual K-Nearest Neighbors

Dalam perhitungan manual ini penulis menggunakan 10 data sample dengan instance yang sesuai dengan data asli, salah satu atribut merupakan data solusi per item data yaitu status dengan instance "Alive" atau "Dead".

Urutan perhitungan algoritma K-Nearest Neighbors adalah sebagai berikut:

1. Menentukan parameter K sebagai banyaknya jumlah tetangga terdekat dengan objek dahulu baru.
2. Menghitung jarak antar data baru dengan data yang telah di latuh
3. Urutkan hasil perhitungan data tersebut
4. Tentukan tetangga terdekat berdasarkan jarak minimun ke K.
5. Tentukan kategori dari tetangga terdekat dengan data.
6. Gunakan kategori mayoritas sebagai klasifikasi data baru

Tentukan data training untuk menghasilkan apakah tergolong Alive atau Dead, tentukan parameter K dengan jumlah tetangga terdekat yaitu K = 3, lalu hitung jarak antara data baru dengan semua data training menggunakan Euclidean.

$$(40-66)^2+(0-0)^2+(1-4)^2+(3-0)^2+(1-0)^2+(3-0)^2+(3-0)^2$$
$$+(1-2)^2+(0-1)^2+(60-19)^2+(1-1)^2+(1-1)^2$$

$$+(7-2)^2+(5-1)^2+(88-77)^2=210{,}498953$$

$$(53-66)^2+(1-0)^2+(3-4)^2+(1-0)^2+(1-0)^2+(1-0)^2+(1-0)^2$$
$$+(3-2)^2+(1-1)^2+(23-19)^2+(1-1)^2+(0-1)^2$$

$$+(15-2)^2+(6-1)^2+(84-77)^2=257.8564065$$

$$(68-66)^2+(2-0)^2+(4-4)^2+(0-0)^2+(0-0)^2+(0-0)^2+(0-0)^2$$
$$+(2-2)^2+(1-1)^2+(13-19)^2+(1-1)^2+(1-1)^2$$

$$+(9-2)^2+(1-1)^2+(64-77)^2=224{,}6332496$$

$$(58-66)^2+(2-0)^2+(0-4)^2+(2-0)^2+(2-0)^2+(4-0)^2+(0-0)^2$$
$$+(2-2)^2+(1-1)^2+(63-19)^2+(1-1)^2+(1-1)^2$$

$$+(14-2)^2+(7-1)^2+(75-77)^2=229{,}2106182$$

$$(57-66)^2+(2-0)^2+(3-4)^2+(1-0)^2+(1-0)^2+(1-0)^2+(1-0)^2$$
$$+(36-2)^2+(35-1)^2+(48-77)^2=3178{,}452079$$

$$(60-66)^2+(2-0)^2+(0-4)^2+(1-0)^2+(0-0)^2+(2-0)^2+(0-0)^2$$
$$+(2-2)^2+(1-1)^2+(29-19)^2+(0-1)^2+(0-1)^2$$

$$+(20-2)^2+(1-1)^2+(77-77)^2=257.8564065$$

$$(63-66)^2+(2-0)^2+(1-4)^2+(0-0)^2+(2-0)^2+(4-0)^2+(0-0)^2$$
$$+(2-2)^2+(0-1)^2+(18-19)^2+(1-1)^2+(1-1)^2$$

$$+(19-2)^2+(19-1)^2+(51-77)^2=1295{,}63325$$

$$(51-66)^2+(0-0)^2+(0-4)^2+(1-0)^2+(1-0)^2+(1-0)^2+(0-0)^2$$
$$+(2-2)^2+(1-1)^2+(23-19)^2+(1-1)^2+(1-1)^2$$

$$+(15-2)^2+(5-1)^2+(105-77)^2=985{,}1245155$$

Gambar 5. Rumus algoritma KNN

Distance urutkan jarak dari data training dengan data baru berdasarkan jarak minimum K.

Tabel 9. Jarak minimum K

Status	*Euclidean Distance*
1	3178,452079
0	1295,63325
0	985,1245155
0	338,6885775
0	257,8564065
0	229,2106182
1	224,6332496
0	210,948953
1	165,3725219

Tabel 10. Hasil jarak minimum K

No	***Euclidean Distance***	**Apakah termasuk 3-NN**
5	3178,452079	*NO* (K>3)
7	1295,63325	*NO* (K>3)
8	985,1245155	*NO* (K>3)
6	338,6885775	*NO* (K>3)
2	257,8564065	*YES* (K<3)
4	229,2106182	*YES* (K<3)
3	224,6332496	*YES* (K<3)
1	210,948953	*YES* (K<3)
9	165,3725219	*YES* (K<3)

Tentukan kategori dari tetangga terdekat terdapat pada baris ke 5,6,7,8,9 kategori YES diambil jika nilai K<= 3, jadi bari ke 5,6,7,8,9 termasuk kategori YES dan sisanya NO.

Kategori YES untuk KNN mencakup baris 5,6,7,8,9 lalu diberikan status berdasarkan tabel di awal, baris ke 5,6,8 memiliki Status 0 dan bari ke 7.9 memiliki status NO, gunakan status mayoritas dari tetangga terdekat sebagai nilai prediksi data yang baru.

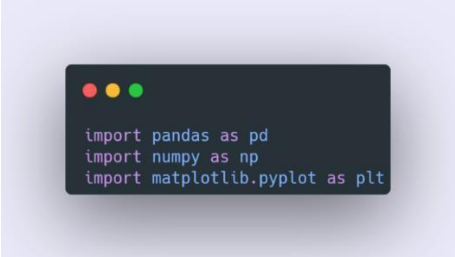
Hasil dari perhitungan manual diatas maka data testing yang digunakan memiliki hasil 0 jadi prediksi dari perhitungan tersebut akurat.

Tabel 11. Kesimpulan nilai K

No	Euclidean Distance	Apakah termasuk 3-NN	Status
5	3178,452079	NO (K>3)	1
7	1295,63325	NO (K>3)	0
8	985,1245155	NO (K>3)	0
6	338,6885775	NO (K>3)	0
2	257,8564065	YES (K<3)	0
4	229,2106182	YES (K<3)	0
3	224,6332496	YES (K<3)	1
1	210,948953	YES (K<3)	0
9	165,3725219	YES (K<3)	1

B. *Model algoritma decision tree dengan python*

Pemanggilan library dasar Pandas, Numpy, dan Matplotlib sebagai langkah awal dalam pembuatan model algoritma decision tree. Cek apakah dalam dataset masih terdapat missing value



Gambar 6. Import library

Gambar 7. Syntax missing values

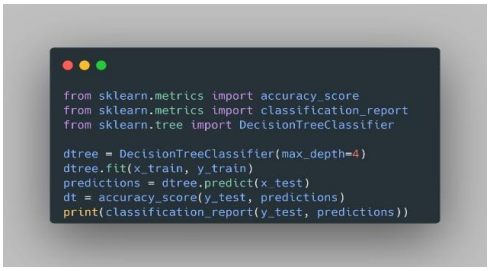
Dari output di atas bahwa dataset tidak memiliki missing values, jadi dapat dilanjutkan ke tahap selanjutnya. Pada tahap ini dataset di cek kembali apakah terdapat data yang kosong atau memiliki data duplikat.

Gambar 8. Syntax data duplikat

Pada tahap dataset akan dibagi menjadi untuk memisahkan antara data latih dan data uji menggunakan metode Split data lalu melakukan standarskaler agar data yang digunakan tidak memiliki penyimpangan besar.

Gambar 9. Membagi data

Untuk data latih atribut status dilepas lalu dimasukan kedalam variabel x, karena atribut "Status" merupakan label dari dataset sehingga tidak digunakan kedalam data uji, lalu untuk variabel y semua atribut digunakan untuk data latih, selanjut- nya digunakan fungsi test-size = 0.2 dimana dataset dibagi menjadi data latih 80% dan data uji 20%. Pada tahap ini model algoritma decision tree diterapkan menggunakan library sklearn.



Gambar 10. Model algoritma decision tree

Model algoritma decision tree dipanggil menggunakan perintah "from sklearn.tree import DecisionTreeClassifier" dengan max_depth = 4, setelah itu model tersebut dimasukan kedalam variabel "dtree",

Untuk melihat nilai akurasi dari model algoritma decision tree dengan menggunakan accuracy_score dan classifiacation_report, berikut untuk hasil nilai akurasi.

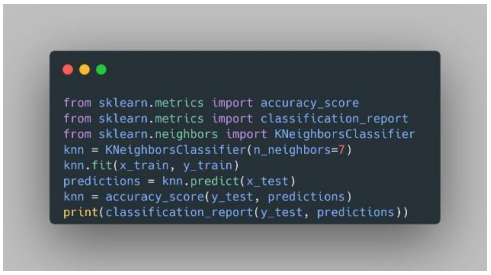
	precision	recall	f1-score	support
0	0.93	0.98	0.95	683
1	0.82	0.59	0.69	118
accuracy			0.92	801
macro avg	0.88	0.79	0.82	801
weighted avg	0.92	0.92	0.92	801

Gambar 11. Hasil model decision tree

C. *Model algoritma KNN dengan python*

Untuk data latih atribut "Status" dilepas lalu dimasukan kedalam variabel x, karena atribut status merupakan label dari dataset sehingga tidak digunakan kedalam data uji, lalu untuk variabel y semua atribut digunakan untuk data latih, selanjutnya digunakan fungsi test-size = 0.2 dimana dataset dibagi menjadi data latih 80% dan data uji 20%.

Pada tahap ini model algoritma K-NN diterapkan menggunakan library sklearn,



Gambar 12. Model algoritma KNN

Model algoritma KNN dipanggil menggunakan perintah "*from sklearn.tree import KNeighbor Classifier*" dengan n_neighbor= 7, setelah itu model tersebut dimasukan kedalam variabel "knn".

Untuk melihat nilai akurasi dari model algoritma KNN dengan menggunakan *accuracy_score* dan *classifiacation_report*, berikut untuk hasil nilai akurasi.

	precision	recall	f1-score	support
0	0.90	0.99	0.94	683
1	0.85	0.37	0.52	118
accuracy			0.90	801
macro avg	0.87	0.68	0.73	801
weighted avg	0.89	0.90	0.88	801

Gambar 13. Hasil algoritma KNN

D. Hasil

Hasil dari perhitungan manual dan menggunakan bahasa pemrograman python model algoritma decision tree dan KNN, diketahui pada perhitungan manual pada kedua model menghasilkan nilai akurasi benar dan sesuai dengan data testing sedangkan dari perhitungan model menggunakan bahasa pemrograman python dan library Scikit Learn sebagai berikut:

Tabel 12. Perbandingan dua algoritma

No	Algoritma	Nilai Akurasi
1	*Decision Tree*	0.921348%
2	*K-Nearest Neighbors*	0.897628%

Bisa dilihat bahwa perbandingan algoritma *Decision tree* dan KNN, *Decision tree* memiliki nilai akurasi diatas KNN yaitu sebesar 0.921348 dan KNN sebesar 0.897628.

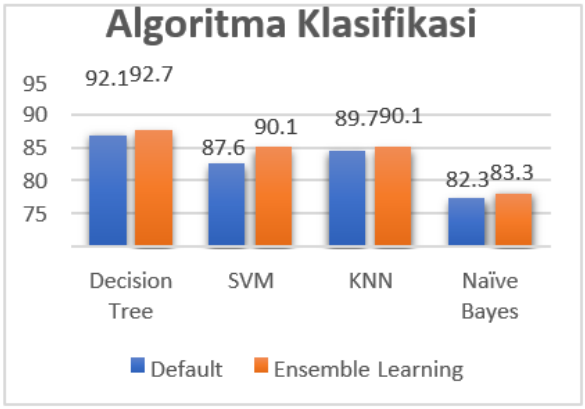
Dan untuk keseluruhan model algoritma klasifikasi menghasilkan nilai akurasi sebagai berikut:

Tabel 13. Hasil akhir

No	Algoritma	Nilai Akurasi
1	*Random Forest*	0.927591 %
2	*Decision Tree*	0.921348 %
3	*KNN*	0.897628 %
4	*SVM*	0.876404 %
5	*Naïve Bayes*	0.823970 %

Bisa dilihat bahwa Random Forest yang memiliki nilai akurasi tertinggi dari ke 4 algoritma yang lainnya, karena Random Forest menggunakan teknik Ensemble Learning dari algoritma Decision Tree yang merupakan sekumpulan dari pohon keputusan.

Agar perbandingan antar algoritma seimbang model algoritma KNN akan menggunakan teknik *ensemble learning* dalam pemodelannya, berikut dari hasilnya.



IV. KESIMPULAN

Dengan menggunakan data yang berasal dari SEER NCI yang diunggah ke dalam website Kaggle.com dengan jumlah atribut yaitu 16 dan jumlah record nya sebesar 4025 record, bahwa hasil dari perbandingan model algoritma baik menggunakan teknik *Ensemble Learning Bagging* atau tidak, algoritma dengan nilai kinerja akurasi terbaik yaitu algoritma Random Forest.

Semua kinerja akurasi pada setiap algo- ritma 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.

DAFTAR PUSTAKA

[1] I. Solikin, R. P. Bhumi, and J. Power, "Teknik Data Mining untuk Prediksi Kanker Payudara yang Efisien".

[2] A. H. Yunial, "Prosiding Seminar Nasional Informatika dan Sistem Informasi Analisa Perbandingan Algoritma Klasifikasi Support Vector Machine, Decession Tree Dan Naive Bayes".

[3] H. Oktavianto and R. P. Handri, "Analisis Klasifikasi Kanker Payudara Menggunakan Algoritma Naive Bayes," 2019. [Online]. Available: https://archive.ics.uci.edu/ml/.

[4] L. Indah Prahartiwi, W. Dari, and S. Nusa Mandiri, "Komparasi Algoritma Naive Bayes, Decision Tree dan Support Vector Machine untuk Prediksi Penyakit Kanker Payudara," Jurnal Teknik Komputer AMIK BSI, vol. 7, no. 1, 2021, doi: 10.31294/jtk.v4i2.

[5] M. Faizal Kurniawan, "Komparasi Algoritma Data Mining Untuk Klasifikasi Penyakit Kanker Payudara," 2017.

[6] A. H. Yunial, "Prosiding Seminar Nasional Informatika dan Sistem Informasi Analisa Perbandingan Algoritma Klasifikasi Support Vector Machine, Decision Tree Dan Naive Bayes".

[7] I. Solikin, R. P. Bhumi, and J. Power, "Teknik Data Mining untuk Prediksi Kanker Payudara yang Efisien".

[8] A. Primajaya and B. N. Sari, "Random For- est Algorithm for Prediction of Precipitation," 2018.

[9] R. Thaniket and E. Taufik Luthf, "Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Algoritma Support Vector Machine," 2020.

[10] J. Perintis Kemerdekaan Km, M. Syukri Mustafa, and I. Wayan Simpen, "Prosiding Seminar Ilmiah Sistem Informasi Dan Teknologi Informasi Pusat Penelitian dan Pengabdian Pada Masyarakat (P4M) STMIK Dipanegara Makassar Implementasi Algoritma K-Nearest Neighbor (KNN) Untuk Memprediksi Pasien Terkena Penyakit Diabetes Pada Puskesmas Manyampa Kabupaten Bulukumba."