PERBANDINGAN ALGORITMA CLASSIFICATION DALAM MENGANALISIS DATASET MULTICLASS

Dikadayanti[1], Sri Damayanti[2], Andi Irmayana[3], Wilem Musu[4]

[1,2]Jurusan Sistem Informasi Universitas Dipa Makassar
Jln. Perintis Kemerdekaan KM. 9 Makassar

e-mail: *[1]dikadayanti15@gmail.com, [2]sryabdullah25@gmail.com, [3]irmayana.andi@undipa.ac.id
[4]wilem.musu@undipa.ac.id

Abstrak

Dataset Penyakit gigi dan mulut adalah sekumpulan data yang memuat informasi penyakit gigi dan mulut dengan ragam atribut yang terkandung didalamnya. Pada penelitian ini, dataset yang dimiliki adalah dataset *multiclass* pasien penyakit gigi dan mulut dengan umur, jenis kelamin, diagnosa, tindakan, dan terapi sebagai atribut yang terkandung di dalamnya dimana terapi dipilih sebagai variable respon. Penetapan suatu penyakit dari hasil Analisa dokter seringkali memerlukan waktu yang cukup lama, begitupun masyarakat awam yang minim pengetahuan akan penyakit gigi dan mulut sehingga menjadi sebab ketidaktahuan terhadap penyakit yang diderita. Pada penelitian ini, akan dilakukan klasifikasi penyakit gigi dan mulut dengan melakukan komparasi *algoritma Knn, C45, dan Naïve Bayes* untuk menentukan algoritma dengan akurasi terbaik dalam mendeteksi terapi yang tepat penyakit gigi dan mulut . Untuk melakukan uji coba, akan digunakan *tools jupyter* sebagai alat untuk melakukan perbandingan akurasi pada algoritma tersebut.Setelah dilakukan penelitian, ditemukan bahwa algoritma c45 memiliki akurasi tertinggi dibanding *algoritma Knn dan Naïve Bayes*,dimana nilai akurasi yang diperoleh algoritma C45 sebesar 0,74%, sehingga algoritma C45 dapat digunakan dan diterapkan untuk mendeteksi terapi yang tepat penyakit gigi dan mulut dengan menggunakan dataset multiclass.

Kata Kunci : *Dataset, Multiclass, Klasifikasi, Algoritma, Python*

Abstract

The Dental and Oral Disease Dataset is a collection of data that contains information on dental and oral diseases with various attributes contained therein. In this study, the dataset owned was a multiclass dataset of dental and oral disease patients with age, sex, diagnosis, treatment, and therapy as the attributes contained therein where therapy was selected as the response variable. Determination of a disease from the results of a doctor's analysis often requires quite a long time, as well as ordinary people who lack knowledge about dental and oral diseases so that it is the cause of ignorance of the disease they suffer. In this study, classification of dental and oral diseases will be carried out by comparing the Knn, C45, and Naïve Bayes algorithms to determine the algorithm with the best accuracy in detecting appropriate therapy for dental and oral diseases. To conduct the trial, jupyter tools will be used as a tool to compare the accuracy of the algorithms. After conducting research, it was found that the c45 algorithm has the highest accuracy compared to the Knn and Naïve Bayes algorithms, where the accuracy value obtained by the C45 algorithm is 0.74%, so that the C45 algorithm can be used and applied to detect appropriate therapy for dental and oral diseases using multiclass datasets.

***Keyword** : Dataset, Multiclass, Classification, Algorithm, Python*

I. PENDAHULUAN

Klasifikasi merupakan teknik dalam *data mining* untuk mengelompokkan data berdasarkan keterikatan data terhadap data sampel. *Data mining* merupakan proses untuk memanipulasi data dengan mengekstraksi informasi yang sebelumnya tidak diketahui dari *dataset* yang berukuran besar . Adapun tujuan klasifikasi pada data mining yaitu dapat menggunakan model tersebut untuk memprediksi kelas dari suatu objek yang mana kelasnya belum diketahui.

Oleh karena itu diperlukan data untuk melakukan proses klasifikasi . Pada penelitian ini dataset yang digunakan merupakan data –data kesehatan rumah sakit berupa penyakit gigi dan mulut yang dikumpulkan secara langsung dari RS. Khusus Daerah Gigi dan Mulut. Dataset yang terkumpul mengandung data mentah sebanyak 7.109 data. Di dalam dataset yang berhasil dikumpulkan memuat beberapa kolom atau variable seperti No, Tanggal, No. RM, Nama Pasien, Alamat, Jenis Kelamin, Kategori, Tindakan, Umur, Terapi, dan Diagnosa. Adapun atribut yang dipilih sebagai variabel terikat yaitu terapi sekaligus menjadi label dari klasifikasi tersebut.

Tujuan dari penelitian ini adalah untuk mengetahui algoritma dengan tingkat performa terbaik dalam memprediksi terapi yang cocok dari dataset penyakit gigi dan mulut. dengan melakukan perbandingan tiga

algoritma klasifikasi yaitu C45, Knn, dan Naïve Bayes .

Tahap terakhir adalah pemberian rekomendasi terhadap aplikasi yang akan dirancang sesuai alur dari analisis yang telah dilakukan. Penulis akan merekomendasikan untuk menggunakan algoritma yang akurasinya terbukti lebih baik berdasarkan hasil uji coba yang telah dilakukan pada tools jupyter notebook dengan menggunakan bahasa pemrograman python.

A. Algoritma

Algoritma berasal dari kata *algoris* dan *ritmi*s yang pertama kali diperkenalkan oleh Abu Ja'far Muhammad Ibn Musa Al Khwarizmi pada 825 M di dalam buku Al-Jabr Wa-al Muqabla. Dalam bidang pemrograman, algoritma didefinisikan sebagai metode yang terdiri dari serangkaian langkah yang terstruktur dan sistematis untuk menyelesaikan masalah dengan bantuan komputer[1].

Algoritma merupakan sekumpulan instruksi atau langkah-langkah yang dituliskan secara sistematis dan digunakan untuk menyelesaikan masalah atau persoalan logika dan matematika dengan bantuan komputer[2].

B. Classification

Classification adalah sebuah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data dengan tujuan untuk memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Classification memiliki banyak aplikasi seperti fraud detection, target marketing,performance prediction, manufacturing, dan medical diagnosis[3].

Classification adalah teknik *data mining* yang paling umum digunakan dengan cara mengklasifikasikan setiap item dari set data ke dalam groupset/kelas-kelas yang telah didefinisikan sebelumnya (*predefined)*[4].

C. Multiclass

Multiclass atau *multinomial classification* adalah masalah dalam contoh klasifikasi ke dalam satu dari tiga atau banyak kelas (klasifikasi *instance* ke dalam salah satu dari dua kelas disebut klasifikasi biner).

D. K- Nearest Neighbor (KNN

K-Nearest Neighbor *Nearest Neighbour* adalah algoritma pengklasifikasian yang didasarkan pada analogi, yaitu membandingkan data uji dengan data pelatihan yang berada dekat dengan dan memiliki kemiripan dengan data uji tersebut. Kemiripan data uji dengan data pelatihan didasarkan pada jaraknya. Adapun rumus KNN sebagai berikut :

$$di = \sqrt{\sum_{i=1}^{p}(x_2 - x_1 i)^2}$$

Dimana :

x_1 = Sampel Data
x_2 = Data Uji/ Testing
i = Variabel Data
d = Jarak
p = Dimensi Data

E. Naïve Bayes

Teorema Bayes adalah perhitungan statistic dengan menghitung probabilitas kemiripan kasus lama yang ada dibasis kasus dengan kasus baru. *Teorema Bayes* memiliki tingkat akurasi yang tinggi dan kecepatan yang baik ketika diterapkan pada *database* yang besar. *Naïve Bayes* termasuk ke dalam pembelajaran *supervised*, sehingga pada tahapan pembelajaran dibutuhkan data awal berupa data pelatihan untuk dapat mengambil keputusan. Pada tahapan pengklasifikasian akan dihitung nilai probabilitas dari masing-masing label kelas yang ada terhadap masukan yang diberikan. *Naïve Bayes* merupakan perhitungan yang paling sederhana, karena mampu mengurangi kompleksitas komputasi menjadi multiplikasi sederhana dari probabilitas. Selain itu, algoritma *naïve bayes* juga mampu menangani set data yang memiliki banyak atribut. Persamaan *naïve bayes* sebagai berikut:

$$P(\mathrm{Ci}|\mathrm{X}) = \frac{P(x\,|Ci)P(Ci)}{P(X)}$$

Dimana :

X = Kriteria sesuai kasus berdasarkan masukan
Ci = Kelas solusi pola ke-i, di mana i adalah label class
P(Ci|X) = Probabilitas kemunculan label kelas Ci dengan kriteria masukan X
P(X|Ci) = Probabilitas kriteria masukan X dengan label kelas Ci
P(Ci) = Probabilitas label Ci

F. C45

Algoritma C4.5 adalah hasil dari pengembangan algoritma ID3 (*Iterative Dichotomiser*)[5]. Algoritma C4.5 atau pohon keputusan mirip sebuah pohon dimana terdapat node internal (bukan daun) yang mendeskripsikan atribut-atribut, setiap cabang menggambarkan hasil dari atribut yang diuji, dan setiap daun menggambarkan kelas. Pohon keputusan dengan mudah dapat dikonversi ke aturan klasifikasi. Secara umum keputusan pengklasifikasi pohon memiliki akurasi yang baik, namun keberhasilan penggunaan tergantung pada data yang akan diolah. Sebelumnya dihitung terlebih dahulu nilai *entropy*, dengan rumus :

$$Entropy(S) = -P+\log_2 P+ -P\text{-}\log_2 P\text{-}$$

Dimana :
S = Ruang data sampel yang digunakan untuk training
P^+ = Jumlah yang bersolusi positif (mendukung) pada data sampel untuk kriteria tertentu.
P- = Jumlah yang bersolusi negative (tidak mendukung) pada data sampel data untuk kriteria tertentu.

Untuk menghitung nilai *gain* digunakan rumus sebagai berikut :

$$\text{Gain}(S,A) = Entropy(S) - \sum_{i=1} \frac{|Si|}{|S|} * Entropy(S_i)$$

Dimana :
S = Himpunan kasus
A = Atribut
n = Jumlah partisi atribut A
|*Si*|= Jumlah kasus pada partisi ke-i
|S| = Jumlah kasus dalam S

G. Python

Pada penelitian tersebut, bahasa pemrograman yang kami gunakan untuk mendukung proses analisis perbandingan algoritma *classification* pada dataset yang akan diuji yaitu bahasa pemrograman *Python*, dengan menggunakan *tools jupyter notebook*.

Bahasa python adalah bahasa pemrograman yang memiliki banyak fungsi, interaktif, berorientasi objek dan merupakan bahasa pemrograman tingkat tinggi[6].

Python adalah bahasa pemrograman interpretative, berorientasi objek dan *semantic* yang dinamis[7].

H. Confusion Matrix

Untuk mendapatkan hasil dari penelitian ini, maka diperlukan suatu metode pengujian sebagai langkah-langkah yang digunakan dalam memperoleh kebenaran prediksi terhadap data baru atau terapi sebagai variable respon. Adapun metode yang kami gunakan untuk menghitung *accuracy, recall, precision,* dan *f1-score* yaitu menggunakan pengujian *Confusion Matrix* pada setiap algoritma yang dibandingkan.

Accuracy ialah *Recall* menandakan seberapa baik model yang dihasilkan mencari nilai positif[8].

Precision ialah bobot yang membuktikan seberapa jelas prediskси yang dilaksanakan[8]. *Average precision* mengkalkulasikan bobot rataan *precision* untuk tiap bobot yang terdapat pada jarak 0 sampai 1.

F1-Score adalah salah satu perolehan dari konstanta dua dikali nilai *recall* dan *presisi*, dan dibagi jumlah keduanya[9].

Adapun rumus yang kami gunakan untuk menghitung *accuracy, recall, precision,* dan *f1-score* yaitu menggunakan pengujian *Confusion Matrix* sebagi berikut:

$$Accuracy = \frac{(TP+TN)}{TP+TN+FP+FN}$$

$$Recall = \frac{TP}{TP+FN}$$

$$Precision = \frac{TP}{TP+FP}$$

$$F1\text{-}Score = 2*\frac{Precision*Recall}{Precision+Recall}$$

II. METODOLOGI PENELITIAN

A. Waktu dan Tempat Penelitian

Penelitian ini dilaksanakan selama 2 bulan mulai bulan November 2022 sampai dengan januari 2023 di RS. Khusus daerah Gigi dan Mulut di Jl. Lanto Dg. Pasewang No.286, Maricaya Sel., Kec. Mamajang, Kota Makassar.

B. Jenis Penelitian

Jenis penelitian ini terbagi menjadi dua metode yaitu kualitatif dan kuantitatif. Kualitatif merupakan penelitian dengan metode yang bersifat induktif dan objektif. Sedangkan kuantitatif dikatakan sebagai metode tradisional karena popularitasnya cukup lama. Adapun jenis metode yang kami gunakan dari data yang kami peroleh yaitu Metode kuantitatif.

Berdasarkan data yang kami peroleh Jenis penelitian ini menggunakan metode kuantitatif, karena penelitian ini bertujuan untuk mengetahui tingkat akurasi mana yang paling tinggi dari beberapa *class* atau label.

C. Metode Pengumpulan Data

Dalam penelitian ini teknik pengumpulan data yang peneliti lakukan yaitu dengan mengamati secara langsung pada lokasi penelitian RS. Khusus Daerah Gigi dan Mulut. Salah satu syarat untuk melakukan penelitian di tempat tersebut harus melalui Dinas penanaman modal dan pelayanan terpadu satu pintu (DPM- PTSP) Prov. Sulawesi selatan, untuk memasukkan surat permohonan izin penelitian dari kampus dan satu file pdf judul skripsi yang diupload melalui website *NENI SI- LINCAH* ke DPM-PTSP agar bisa melakukan penelitian di Rs.khusus daerah gigi dan mulut.

Data yang peneliti dapatkan hasil pengumpulan data sebanyak 7.110 yang akan peneliti gunakan sebagai class atau label untuk mengetahui berapa persen tingkat akurasi yang lebih tinggi.

D. Metode Pengujian

Untuk mendapatkan hasil dari penelitian ini, maka diperlukan suatu metode pengujian sebagai langkah-langkah yang digunakan dalam memperoleh kebenaran prediksi terhadap data baru atau terapi sebagai variable respon.

1) Metode Pengujian Confusion Matrix

Untuk mendapatkan hasil dari penelitian ini, maka diperlukan suatu metode pengujian sebagai langkah-langkah yang digunakan dalam memperoleh kebenaran prediksi terhadap data baru atau terapi sebagai variable respon. Adapun metode yang kami gunakan untuk menghitung *accuracy, recall, precision,* dan *f1-score* yaitu menggunakan pengujian *Confusion Matrix* pada setiap algoritma yang dibandingkan.

Tabel 1. Confusion Matrix Multiclass

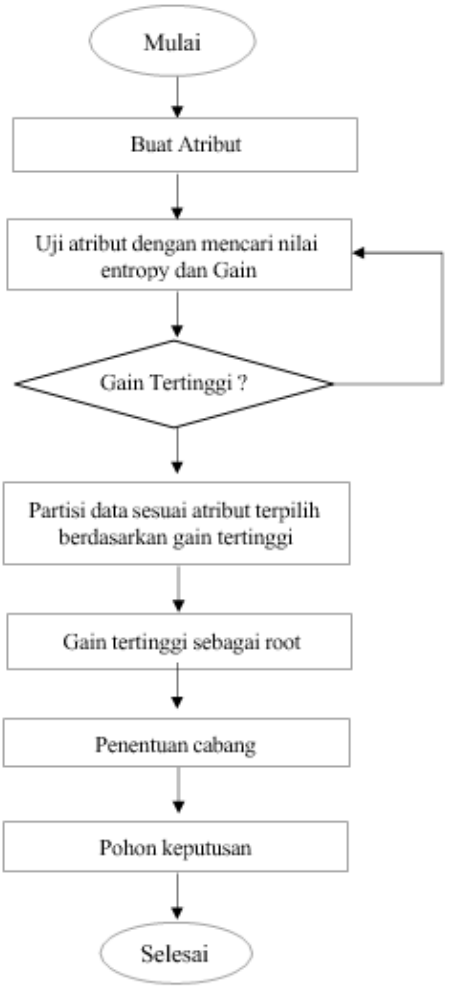
	Prediksi		
Aktual	Jeruk	Pepaya	Anggur
Jeruk	7	8	9
Pepaya	1	2	3
Anggur	3	2	1

2) Algoritma *Knn*

Gambar 1. Alur Pengujian *Knn*

Gambar 1 di atas merupakan alur pengujian knn dimana menentukan parameter K sebagai banyaknya jumlah tetangga terdekat dengan objek baru. Menghitung jarak antar objek/data baru terhadap semua objek/data yang telah di training. Mengurutkan hasil perhitungan tersebut. Tentukan tetangga terdekat berdasarkan jarak minimum ke K. Tentukan kategori dari tetangga terdekat dengan objek/data. Gunakan kategori mayoritas sebagai klasifikasi objek/data baru.

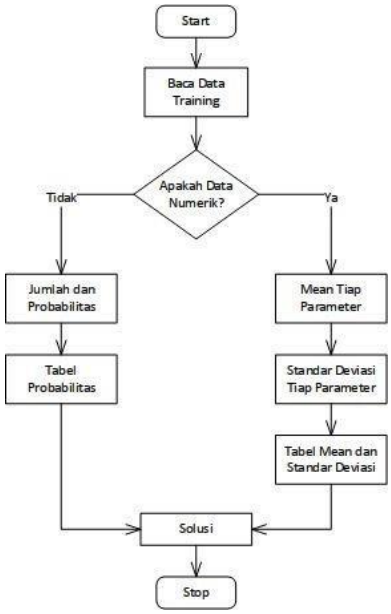
2) Alur Pengujian *C45*



Gambar 2. Alur Pengujian *C45*

Gambar 2 di atas diketahui alur algoritma *C4.5* yang digunakan. Pada persiapan awal ditentukan atribut yang digunakan kemudian melakukan uji atribut dengan mencari nilai *gain* tertinggi berdasarkan perhitungan *entropy* dari masing-masing atribut. Apabila ditemukan *gain* tertinggi maka *gain* tersebut akan menjadi *root* awal. Selanjutnya dilakukan penentuan cabang dengan cara yang sama dengan melihat *gain* tertinggi dari tiap hasil partisi.

3) Algoritma *Naïve Bayes*



Gambar 3. Alur Pengujian *Naïve Bayes*

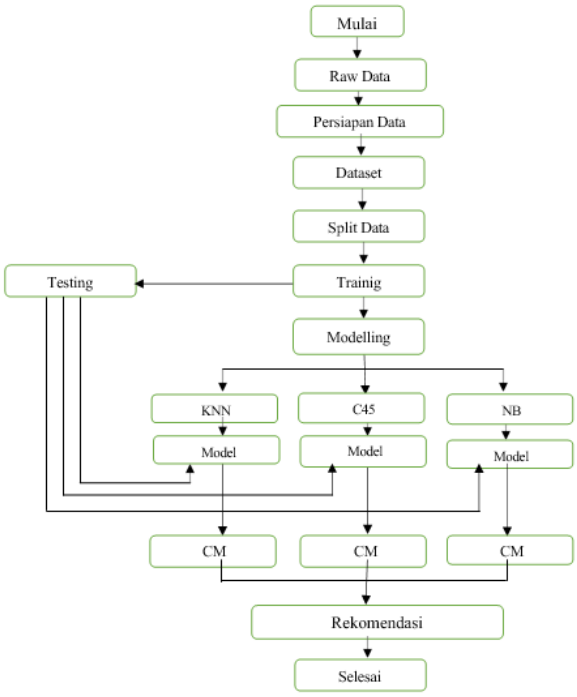
Gambar 3 diatas adalah alur pengujian *naïve bayes* , dimana tahap dimulai dengan membaca jumlah data training dan probabilitas, namun jika data numerik maka menghitung nilai *mean* dan standar deviasi dari masing- masing parameter yang merupakan numerik. Menghitung nilai probabilistik dengan cara menghitung jumlah data yang sesuai dari kategori yang sama dibagi dengan jumlah data pada kategori tersebut. Mendapatkan nilai dalam tabel mean, standar deviasi dan probabilitas menghasilkan solusi dan selesai.

E. Penelitian Terkait

1. Penelitian terkait algoritma klasifikasi yang dilakukan oleh (Tri Retnasari, Eva Rahmawati 2017) menggunakan Algoritma *Naive bayes* dan *C4.5* dalam mendiagnosa prediksi penyakit jantung. Hasil penelitian bahwa nilai accuracy dan AUC Naive Bayes lebih tinggi dibandingkan *C4.5*. Penerapan Naive Bayes untuk prediksi jantung menghasilkan selisih nilai akurasi sebesar 2.97%.
2. pada penelitian (Mei Lestari, 2014) menggunakan Algoritma *K- Nearest Neighbors* untuk mendeteksi penyakit jantung. Hasil penelitian bahwa untuk melakukan kinerja algoritma tersebut dilakukan dengan menggunakan *confusion matrix* dan kurva ROC, diperoleh nilai akurasi 70% dan termasuk klasifikasi baik karena memiliki nilai AUC 0.875.
3. Perbandingan penelitian dari (Dian prajarini, 2016) menggunakan algoritma *C45*, *Naive Bayes*, *KNN*, dan *SVM* dalam membandingkan algoritma tersebut untuk memprediksi penyakit kulit. Hasil penggunaan algoritma klasifikasi data mining (*C4.5, Naive Bayes, KNN, SVM*) pada prediksi penyakit kulit dengan data set diambil dari UCI serta pengujiannya menunjukkan bahwa Algoritma *C4.5*, *Naive Bayes*, *KNN*, SVM bisa diterapkan untuk prediksi penyakit kulit dengan nilai akurasi, presisi dan recall di atas 94%. Hasil pengujian menunjukkan algoritma *naive bayes* dan SVM memiliki nilai akurasi dan recall yang sama dan yang paling tinggi yaitu 98,1%, sedangkan nilai presisinya berbeda. *Naive bayes* memberikan nilai presisi sebesar 98,3% dan *SVM* memiliki nilai presisi sebesar 98,2%, hanya selisih 0,1%. Jika dibandingkan berdasarkan nilai, maka *Naive Bayes* dianggap sebagai algoritma klasifikasi yang lebih baik. Tetapi jika dibandingkan secara global, *Naive Bayes* dan *SVM* merupakan algoritma klasifikasi yang lebih baik dibandingkan *C4.5* dan *KNN* untuk prediksi penyakit kulit.
4. Penelitian yang dilakukan oleh (Wildan Budiman Zulfikar, Nur Lukman, 2016) dalam membandingkan *naïve bayes classifier* dengan *nearest neighbor* untuk identifikasi penyakit mata. Hasil dari penelitian ini yaitu algoritma *Naïve Bayes* dan *nearest neighbor* memiliki keakuratan yang sebanding. Dalam hal kecepatan, *nearest neighbor* memiliki rata-rata catatan lebih cepat 0.027 detik dari pada *naïve bayes classifier.*
5. Pada penelitian yang dilakukan oleh (Mila Listiana, Sudjalwo, Dedi Gunawan, 2015) dalam perbandingan algoritma *decision tree (c4.5)* dan *naïve bayes* pada data mining untuk identifikasi tumbuh kembang anak balita (studi kasus puskesmas kartasura). Adapun hasil dari penelitian ini bahwa berdasarkan dari nilai *accuracy* maupun *recallnya naive bayes* lebih tinggi dibandingkan dengan *decision tree* yaitu dengan nilai *accuracy* 75,66% untuk *decision tree* dan 76,97% untuk *naive bayes.* Untuk nilai *recall*-nya *naive bayes* lebih unggul yaitu 96,89% dibandingkan *decision tree* 89,78%. Meskipun dalam penelitian ini tingkat *Precision*-nya lebih tinggi *decision tree* yaitu 85,23% dibandingkan *naive bayes* 84,17%.
6. Pada penelitian yang dilakukan oleh (Mursyid Ardansya , Andi Sunyanto ,Emha Taufiq Luthfi 2021) dalam Analisis Perbandingan Akurasi Algoritma *naïve bayes* dan *C4.5* untuk Klasifikasi Diabetes, hasil penelitian perbandingan performa algoritma *naïve bayes* dan *C4.5* untuk klasifikasi penyakit diabetes yang telah dilakukan dapat diambil kesimpulan bahwa algoritma *C4.5* memiliki nilai performa yang baik dibandingkan algoritma *naïve bayes* dimana algoritma *C4.5* unggul di setiap skenario dengan skenario 4 sebagai skenario yang tinggi dengan hasil *accuracy* 99,03%, *precision* 100%, dan *recall* 98,18% dimana 3,88% peningkatan *accuracy* dari *accuracy* skenario 1.

III. HASIL DAN PEMBAHASAN

A. *Perancangan Solusi*



Gambar 3. Alur Perancangan Solusi

Pembuatan perancangan solusi diharapkan dapat menyelesaikan masalah- masalah yang ditemukan pada penelitian ini. Gambar perancangan solusi diatas memiliki beberapa tahap, mulai dari tahap awal proses *raw data* hingga pemberian rekomendasi yang menandakan sebagai tahap akhir dari proses penelitian tersebut. Tahap-tahap yang tersaji pada bagan di atas akan dijelaskan secara ringkas sebagai berikut.

Tahap pertama ini adalah proses pengumpulan atau pengambilan data mentah (*raw data*) yang dilakukan oleh penulis. Adapun teknik yang dilakukan dalam pengambilan data tersebut yaitu *field research* , yaitu dengan mengamati secara langsung pada tempat penelitian yakni Rumah Sakit Khusus Daerah Gigi dan Mulut (RSKDGM) yang berlokasi di Jl. Lanto Dg. Pasewang No.286, Maricanaya Sel., Kec. Mamajang, Kota Makassar.

Data informasi pasien penyakit gigi dan mulut tersebut memiliki beberapa variabel yang terkandung di dalamnya seperti No, Tgl, No.RM, Nama Pasien, Alamat, Jenis Kelamin, Kategori, Tindakan, Regio/umur, Terapi, dan Diagnosa.

Raw data atau data mentah yang berhasil dikumpulkan dari tempat penelitian, selanjutnya akan diproses lagi pada tahap persiapan data sebelum menjadi dataset yang siap pakai untuk tahap-tahap selanjutnya.

Data cleansing adalah kegiatan menganalisa kualitas data dengan cara memodifikasi, mengubah, atau menghapus data-data yang dianggap tidak perlu, tidak lengkap, data tidak akurat, atau memiliki format data atau file yang salah dalam basis data guna menghasilkan data berkualitas tinggi. Data cleansing juga biasa disebut data cleaning atau data scrubbin.

Seperti definisi diatas, proses *cleansing* dilakukan untuk mengeliminasi atau membuang data yang bersifat *nose* dan variabel atau kolom yang tidak dibutuhkan seperti Tanggal, No.RM, Nama Pasien, Alamat, dan Kategori sehingga hanya menyisakan lima kolom saja yaitu Jenis Kelamin, Tindakan, Usia, dan terapi. Selain kolom yang dihilangkan, data-data yang bersifat kosong atau *nose* juga dihapus yang menyebabkan terjadinya pengurangan data. Sehingga diperoleh 6.202 data dari total data yang mencapai 7.109 yang berarti data-data yang terbuang sebanyak 907 buah data.

Setelah melalui tahap persiapan data dengan melakukan eliminasi atau *cleansing* data, maka diperolehlah dataset yang siap untuk digunakan dalam proses *modelling*. Dataset yang semula berbentuk file *excel*, diubah menjadi file *csv* di mana data tersusun secara terstruktur dan data informasi yang tersimpan dengan tanda pemisah koma bukan dengan kolom.

Tahap selanjutnya adalah pembagian data atau *split data.* Pada penelitian ini split data digunakan sebagai teknik membagi dataset menjadi dua subhimpunan yaitu data latih (*train*) dan data tes (*test*), untuk percobaan yang akan dilakukan, diberikan sebanyak 80 persen untuk *data training* dan *data testing* sebanyak 20 persen dari dataset yang digunakan.

Tahap selanjutnya yaitu *modelling*, tahapa ini digunakan untuk menemukan model pada algoritma yang dibandingkan yaitu *Knn*, *C45*, dan *Naive bayes*. Untuk proses modelling dan semua proses yang disebutkan akan dilakukan *di tools jupyter notebook* dengan menggunakan bahasa pemrograman *pyton.*

Model yang telah didapatkan pada proses modelling, selanjutnya akan diuji kinerjanya melalui *confusion matrix* untuk megetahui keakuratan model dari algoritma *Knn, C45*, dan *Naive bayes.* Proses ini akan menggunakan performa matriks yang umum digunakan yaitu *accuracy*, *precission*, dan *recall.*

Tahap terakhir adalah pemberian rekomendasi terhadap aplikasi yang akan dirancang sesuai dengan alur dari analisis pada penelitian ini. Penulis akan merekomendasikan untuk menggunakan algoritma yang akurasinya terbukti lebih baik berdasarkan hasil dari serangkaian uji coba yang telah dilakukan pada *tools jupyter notebook* dengan menggunakan bahasa pemrograman *python*.

B. *Hasil dan Pembahasan*

Berikut hasil terbaik yang didapatkan dari hasil klasifikasi algorima K-Nearest Neighbor, algorima C45 (*decision tree)*.Dan algorima *Naive Bayes* yatu:.

1) Algoritma KNN (K-Nearest Neighbor)

accuracy			0.69	1241
macro avg	0.26	0.21	0.22	1241
weighted avg	0.64	0.69	0.64	1241

Gambar 4. Hasil Pengujian *Algoritma KNN*

Berdasarkan hasil gambar 4 di atas, setelah dilakukan beberapa kali percobaan pada Algoritma *Knn* dan menghasilkan nilai rata-rata berikut nilai tertinggi dari confusion matrix yang meliputi precision, recall, F1 Score dan support dari K-Nearest Neighbor.Maka hasil dari percobaan tersebut menhasilkan nilai *Accuracy* sebesar 0,69% dari jumlah support 1241.

2) Algoritma C45 (decision tree)

accuracy			0.74	1241
macro avg	0.29	0.29	0.28	1241
weighted avg	0.71	0.74	0.71	1241

Gambar 5. Hasil Pengujian *Algoritma C45*

Berdasarkan hasil gambar 5 di atas, Setelah dilakukan beberapa kali percobaan pada Algoritma *C45* (*decision tree).* dan menghasilkan nilai rata-rata berikut nilai tertinggi dari confusion matrix yang meliputi precision, recall, F1 Score dan support dari K-Nearest Neighbor.Maka hasil dari `percobaan tersebut menhasilkan nilai *Accuracy* sebesar 0,74% dari jumlah support 1241.

3) Algoritma Naïve bayes

accuracy			0.49	1241
macro avg	0.12	0.11	0.10	1241
weighted avg	0.38	0.49	0.40	1241

Gambar 6. Hasil Pengujian *Algoritma Naïve bayes*

Berdasarkan tabel report 4.4 di atas,setelah dilakukan beberapa kali percobaan pada Algoritma *Naive Bayes* dan menghasilkan nilai rata-rata berikut nilai tertinggi dari confusion matrix yang meliputi precision, recall, F1 Score dan support dari K-Nearest Neighbor.Maka hasil dari percobaan tersebut menhasilkan nilai *Accuracy* sebesar 0,49% dari jumlah support 1241.

Seperti yang telah disebutkan di atas bahwa untuk melakukan proses perbandingan algoritma *c45,knn,* dan *naive bayes*, digunakan sebuah tools *jupyter notebook* pada setiap tahap-tahapnya. Adapun tahap yang dilakukan seperti pemanggilan library setiap algoritma untuk memudahkan dalam penulisan fungsi dalam menganalisis dataset.

Penetapan data *testing* dan data *training* juga dilakukan pada proses pengujian data , di mana data testing yang digunakan 1.240 dari dataset yang ada dan data training sebanyak 4.962 dari dataset penyakit gigi dan mulut.

Pada pengujian yang dilakukan, setiap algoritma memiliki karakteristik pengujiannya masing-masing seperti algortima *c45* dilakukan perhitungan *entrophy* dan nilai *gain* pada setiap atribut, algoritma *knn* yang menghitung jarak atau nilai *euclidean metric* tetangga terdekat, di mana pada penelitian ini tetangga terdekat dipilih sebanyak 11. Algoritma *naive bayes* yang menghitung probabilitas setiap atribut di mana pada algoritma ini juga dilakukan normalisasi data untuk menskalakan data suatu atribut sehingga berada dalam rentang yang lebih kecil, seperti -1 hingga 1 atau 0 hingga 1.

Berdasarkan hasil perhitungan *confusion matrix* pada Algoritma *Knn, C45, dan Naive bayes* diatas pada dataset multiclas penyakit gigi dan mulut menggunakan *tools jupyter notebook*, dapat dilihat perbandingan Accuracy, Precision, dan Recall *Knn, C45, dan Naive bayes.* Didapatkan hasil algoritma yang paling terbaik kesatu c45 dengan akurasi 0,74%. hasil algoritma yang paling terbaik kedua *Knn* dengan akurasi 0,69% dan hasil algoritma yang paling terbaik ketiga *Naive Bayes* dengan akurasi 0.49%.

Pada penelitian yang dilakukan oleh (Mursyid Ardansya , Andi Sunyanto ,Emha Taufiq Luthfi 2021) dalam Analisis Perbandingan Akurasi Algoritma *naïve bayes* dan *C4.5* untuk Klasifikasi Diabetes, hasil penelitian perbandingan performa algoritma *naïve bayes* dan *C4.5* untuk klasifikasi penyakit diabetes yang telah dilakukan dapat diambil kesimpulan bahwa algoritma *C4.5* memiliki nilai performa yang baik dibandingkan algoritma *naïve bayes* dimana algoritma *C4.5* unggul di setiap skenario dengan skenario 4 sebagai skenario yang tinggi dengan hasil *accuracy* 99,03%, *precision* 100%, dan *recall* 98,18% dimana 3,88% peningkatan *accuracy* dari *accuracy* skenario 1.

IV. KESIMPULAN

Setelah melakukan analisis untuk mengetahui performance terbaik dari tiga algoritma klasifikasi terhadap dataset multiclass di dapatkan kesimpulan bahwa berdasarkan tabel hasil perbandingan Algoritma *Knn, C45, dan Naive bayes* diatas, untuk dataset Multiclas penyakit gigi dan mulut menggunakan *tools jupyter notebook* , dapat dilihat perbandingan Accuracy, Precision, dan Recall *Knn, C45, dan Naive bayes.* Hasil algoritma yang terbaik kesatu *c45* dengan akurasi 0,74%, hasil algoritma terbaik kedua yaitu *Knn* dengan akurasi 0,69% dan hasil algoritma terbaik ketiga yaitu *Naive Bayes* dengan akurasi 0,49%.

V. SARAN

Adapun saran yang ingin disampaikan penulis Penelitian ini juga masih jauh dari kata sempurna oleh karena itu penulis menyarankan agar penelitian selanjutnya dapat mengembangkan dengan sistem web ataupun android untuk dapat memprediksi dengan menggunakan aplikasi *mechine learning*.

REFERENSI

[1] Jando, E., & Nani, P. A. (2018). Algoritma dan Pemrograman dengan Bahasa Java. Penerbit Andi.

[2] Sismoro, H. (2005). Pengantar Logika Informatika. Algoritma dan Pemrograman Komputer. Penerbit Andi.

[3] Han, J., Kamber, M., & Pei, J. (2012). Classification. In Data Mining (pp. 393-442). Elsevier.

[4] Aamer, M., Abbott, D., Abdullaev, A., Abe, Y., Abreu-Afonso, J., Absil, P., ... & Alic, N. (2012). 2012 Index IEEE Photonics Technology Letters Vol. 24. IEEE Photonics Technology Letters, 24(24), 2309.

[5] Han, J., & Kamber, M. (2006). Classification and prediction. Data mining: Concepts and techniques, 2006, 347-50.

[6] Wallentin, L., Becker, R. C., Budaj, A., Cannon, C. P., Emanuelsson, H., Held, C., ... & Harrington, R. A. (2009). Ticagrelor versus clopidogrel in patients with acute coronary syndromes. New England Journal of Medicine, 361(11), 1045-1057.

[7] Python Software Foundation. (2016). Python Software Foundation. Python Language Reference, version, 3, 6.

[8] Zulkiflie, M. A. (2021). IMPLEMENTASI ALGORITMA OBJECT DETECTION YOLOV4 DAN EUCLIDEAN DISTANCE DALAM MENDETEKSI PELANGGARAN SOCIAL DISTANCING (Doctoral dissertation, Universitas Hasanuddin).

[9] Agastya, I. M. A., & Setyanto, A. (2018, November). Classification of Indonesian batik using deep learning techniques and data augmentation. In 2018 3rd international conference on information technology, information system and electrical engineering (ICITISEE) (pp. 27-31). IEEE.