

PREDIKSI KEPUASAN PELANGGAN HOTEL: STUDI PERBANDINGAN ALGORITMA DECISION TREE DAN K-NEAREST NEIGHBOR

HOTEL CUSTOMER SATISFACTION PREDICTION: A COMPARATIVE STUDY OF DECISION TREE AND K-NEAREST NEIGHBOR ALGORITHMS

¹Dwi Ramti Asih*, ²Rianto

^{1,2}Informatika, Universitas Siliwangi

^{1,2}Jl.Siliwangi No.24 Kahuripan, Tasikmalaya, 46115, Indonesia

¹207006029@student.unsil.ac.id, ²rianto@unsil.ac.id

*) Penulis Korespondensi

Abstrak

Kepuasan pelanggan telah menjadi aspek penting bagi setiap bisnis di pasar yang kompetitif saat ini. Memahami kebutuhan, keinginan, dan harapan pelanggan sangat penting bagi suatu bisnis untuk memberikan layanan pelanggan yang luar biasa dan mempertahankan pelanggan. Oleh karena itu pada penelitian kali ini merepresentasikan sebuah studi perbandingan antara dua algoritma machine learning, Decision Tree dan K-Nearest Neighbor, untuk memprediksi kepuasan pelanggan hotel. Studi ini bertujuan untuk mengidentifikasi algoritma mana yang lebih efektif dalam memprediksi kepuasan pelanggan dengan mengevaluasi performa mereka menggunakan berbagai metrik. Metodologi yang digunakan meliputi preprocessing data, seleksi fitur, dan pembuatan model machine learning. Hasil menunjukkan bahwa algoritma Decision Tree lebih unggul dibandingkan dengan K-Nearest Neighbor dalam hal akurasi dan presisi. Temuan dari studi ini memberikan wawasan bagi bisnis di industri perhotelan tentang bagaimana memprediksi kepuasan pelanggan dan meningkatkan layanan mereka.

Kata Kunci: *Decision Tree, Kepuasan Pelanggan, K-Nearest Neighbor, Ulasan Hotel.*

Abstract

Customer satisfaction has become an important aspect for every business in today's competitive market. Understanding customer needs, wants, and expectations is critical for a business to provide outstanding customer service and retain customers. Therefore, this research represents a comparative study between two machine learning algorithms, Decision Tree and K-Nearest Neighbor, to predict hotel customer satisfaction. This study aims to identify which algorithm is more effective in predicting customer satisfaction by evaluating their performance using various metrics. The methodology used includes data preprocessing, feature selection, and machine learning model creation. The results show that the Decision Tree algorithm is superior to the K-Nearest Neighbor in terms of accuracy and precision. The findings from this study provide insights for businesses in the hospitality industry on how to predict customer satisfaction and improve their services.

Keywords: *Customer Satisfaction, Decision Tree, K-Nearest Neighbor, Hotel Reviews.*

PENDAHULUAN

Industri hotel merupakan salah satu industri yang paling kompetitif di seluruh

dunia. Dimana kepuasan pelanggan memainkan peran penting dalam kesuksesan pada bisnis tersebut. Kepuasan pelanggan adalah keterkaitan antara harapan atau ekspektasi pelanggan terhadap layanan yang mereka terima [1]. Kepuasan pelanggan dapat dipengaruhi oleh berbagai faktor seperti kualitas produk atau layanan, harga, pelayanan pelanggan, kemudahan penggunaan, dan lain-lain. Kepuasan pelanggan juga merupakan salah satu faktor utama yang menentukan loyalitas pelanggan dan mempengaruhi keputusan mereka untuk kembali ke hotel. Munculnya ulasan online dan media sosial, pelanggan memiliki lebih banyak kekuatan daripada sebelumnya untuk berbagi pengalaman mereka dengan orang lain. Satu ulasan negatif dapat merusak reputasi hotel dan menyebabkan calon pelanggan memilih pesaing [2].

Pada penelitian ini menggunakan penelitian terdahulu sebagai acuan pengembangan terhadap kasus yang sedang diteliti. Penelitian pertama yaitu Sutan Faisal (2020), penelitian ini membahas mengenai penerapan algoritma K-Nearest Neighbor (K-NN) untuk memprediksi kepuasan pelanggan di industri perhotelan. Penelitian ini dilakukan dengan mengumpulkan data dari survei kepuasan pelanggan dan melakukan analisis menggunakan K-NN untuk memprediksi kepuasan pelanggan. Salah satu kekurangan dari penelitian ini adalah terbatasnya data yang digunakan dalam penelitian [3]. Penelitian kedua yaitu Fadhli Almu'ini Ahda (2019),

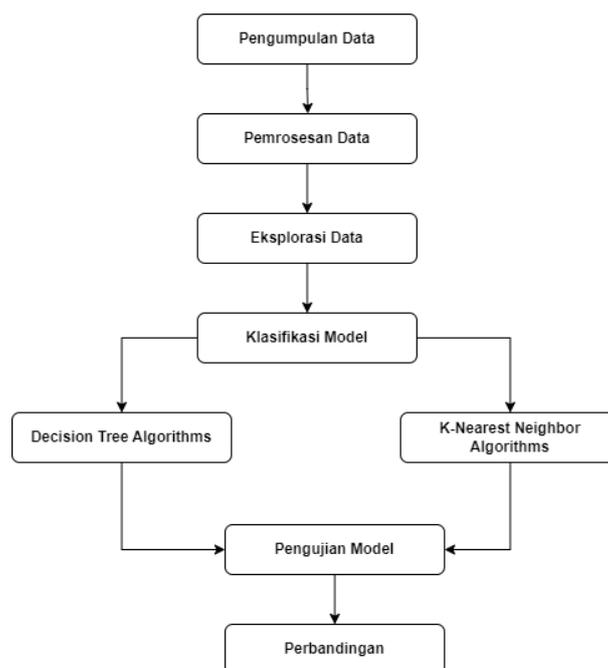
penelitian ini membahas mengenai kepuasan pengunjung perpustakaan terhadap layanan yang diberikan. Penelitian ini dilakukan dengan mengumpulkan data dari survei kepuasan pengunjung dan melakukan analisis menggunakan algoritma Decision Tree (C4.5) untuk memprediksi kepuasan pengunjung. Salah satu kekurangan dari penelitian ini adalah tidak mempertimbangkan faktor-faktor lain yang dapat mempengaruhi kepuasan pengunjung, seperti ketersediaan koleksi, kualitas layanan, dan kenyamanan ruangan [4]. Penelitian ketiga yaitu Safira Amalia (2022), Penelitian tersebut membahas perbandingan tiga algoritma klasifikasi yang digunakan untuk memprediksi kepuasan pelanggan maskapai penerbangan, yaitu Random Forest, SPAARC, dan MLP. Hasil dari penelitian menunjukkan bahwa ketiga algoritma klasifikasi memiliki kinerja yang relatif baik, namun MLP dan SPAARC memberikan hasil yang sedikit lebih baik dibandingkan dengan Random Forest [5].

Berdasarkan literatur diatas penelitian ini dikembangkan dengan cara melakukan pengembangan model prediksi kepuasan pelanggan hotel dengan membandingkan dua algoritma, yaitu Decision Tree dan K-Nearest Neighbor. Penelitian ini memanfaatkan data survei kepuasan pelanggan hotel untuk menguji kinerja kedua algoritma dalam memprediksi kepuasan pelanggan. Dengan demikian, pengembangan yang dilakukan dalam penelitian ini dapat memperoleh model prediksi kepuasan pelanggan hotel yang lebih

akurat. Pada penelitian ini juga mempertimbangkan beberapa faktor yang diduga mempengaruhi kepuasan pelanggan hotel, seperti kualitas layanan, fasilitas hotel, lokasi hotel, dan harga. Faktor-faktor tersebut diperoleh dari data survei kepuasan pelanggan hotel yang dijadikan sebagai sumber data dalam penelitian [6].

Secara keseluruhan, penelitian ini bertujuan untuk membandingkan efektivitas dua algoritma data mining, Decision Tree dan K-Nearest Neighbor, dalam memprediksi kepuasan pelanggan hotel. Penelitian ini akan memberikan kontribusi bagi peneliti lain yang tertarik untuk melakukan penelitian serupa dengan cara secara langsung membandingkan kinerja suatu algoritma dalam memprediksi kepuasan pelanggan hotel [7].

METODE PENELITIAN



Gambar 1. Tahapan Metode Penelitian [7]

Metode penelitian merupakan suatu hal yang penting agar penelitian lebih terarah dan sistematis, metode ini juga diperlukan untuk panduan dalam proses penelitian. Pada penelitian ini menggunakan metode penelitian yang terdiri dari beberapa tahapan yaitu pengumpulan data, pemrosesan data, eksplorasi data, klasifikasi model, pengujian model dan perbandingan [7].

1. Pengumpulan Data

Pengumpulan data dilakukan dengan tujuan untuk mendapatkan informasi yang dibutuhkan dalam suatu penelitian. Pada penelitian ini menggunakan dataset “Hotel Reviews” yang diunduh dari Kaggle. Selanjutnya dataset yang telah diunduh dapat disimpan pada direktori lokal untuk kemudian diolah.

2. Pemrosesan Data

Pemrosesan data merupakan suatu proses pengolahan data mentah (*raw data*) menjadi bentuk yang dapat dipahami dan diinterpretasikan oleh manusia maupun komputer [8]. Pada penelitian ini langkah yang dilakukan untuk melakukan pemrosesan data antara lain dengan melakukan Loading Dataset dengan cara menggunakan *pandas*, membaca file dataset yang telah diunduh dan menyimpannya dalam bentuk *DataFrame*. Kemudian menampilkan ukuran dataset pada *DataFrame* untuk menampilkan jumlah baris dan kolom pada dataset. Selanjutnya menampilkan 5 data pertama dengan cara menggunakan *method head()* pada *DataFrame* untuk menampilkan 5 data pertama pada dataset. Lalu menampilkan informasi dataset dengan cara menggunakan *method info()* pada *DataFrame* untuk menampilkan informasi tentang dataset, termasuk nama kolom, jumlah baris, tipe data kolom, dan jumlah nilai non-null pada setiap kolom. Terakhir melakukan *preprocessing* data dengan cara menghapus kolom yang tidak diperlukan, menghapus baris dengan nilai kosong, dan merubah tipe data kolom yang sesuai dengan tipe data yang diinginkan.

3. Eksplorasi Data

Eksplorasi data merupakan suatu proses analisis data yang dilakukan untuk memahami karakteristik dan pola data secara visual dan deskriptif. Pada penelitian ini eksplorasi data dilakukan dengan cara melakukan visualisasi

data menggunakan histogram, bar plot, dan scatter plot untuk menampilkan distribusi data pada setiap fitur dan hubungan antar fitur. Kemudian dapat pula menampilkan korelasi antar fitur dengan cara menggunakan heatmap pada *seaborn library* untuk menampilkan matriks korelasi antar fitur.

4. Klasifikasi Model

Dalam klasifikasi model untuk penelitian ini, dua algoritma yang dipilih untuk perbandingan adalah Decision Tree dan K-Nearest Neighbor (KNN). Pertama, data latih digunakan untuk melatih kedua model menggunakan metode *fit()*. Setelah melatih model, langkah selanjutnya adalah mengevaluasi performa model menggunakan beberapa metrik evaluasi, seperti akurasi, presisi, recall, dan F1-score. Ini dilakukan dengan menggunakan metode *classification_report()* dari pustaka *scikit-learn*, yang menghasilkan laporan yang menyajikan nilai-nilai metrik evaluasi untuk setiap kelas atau label pada data uji.

Selanjutnya, untuk meningkatkan performa model, digunakan teknik pengoptimalan model dengan menggunakan teknik cross-validation. Dalam langkah ini, digunakan metode *GridSearchCV()* dari pustaka *scikit-learn* untuk menentukan parameter optimal pada model. Proses cross-validation membantu menyesuaikan parameter dan hiperparameter model berdasarkan hasil evaluasi pada subset data latih yang berbeda. Proses pengoptimalan ini bertujuan untuk

meningkatkan kinerja model pada data uji yang tidak terlihat sebelumnya [9]. Berikut adalah pseudocode untuk langkah-langkah yang dijelaskan di atas:

```

Procedure
TrainAndEvaluateModels(training_data,
testing_data):
    // Pelatihan Model
    decision_tree_model =
    DecisionTree.train(training_data)
    knn_model = KNN.train(training_data)
    // Evaluasi Performa Model
    decision_tree_performance =
    EvaluateModel(decision_tree_model,
    testing_data)
    knn_performance =
    EvaluateModel(knn_model, testing_data)
    // Pengoptimalan Model dengan Cross-
    Validation
    optimized_decision_tree_model =
    OptimizeModel(decision_tree_model,
    training_data)
    optimized_knn_model =
    OptimizeModel(knn_model, training_data)
    Return decision_tree_performance,
    knn_performance,
    optimized_decision_tree_model,
    optimized_knn_model
Procedure EvaluateModel(model,
testing_data):
    predicted_labels =
    model.predict(testing_data.features)
    evaluation_report =
    classification_report(testing_data.labels,
    predicted_labels)

```

```

    Return evaluation_report
Procedure OptimizeModel(model,
training_data):
    optimized_model = GridSearchCV(model,
    parameter_grid, cv=5) // 5-fold cross-
    validation
    optimized_model.fit(training_data.features,
    training_data.labels)
    Return optimized_model

```

5. Pengujian Model

Pengujian model merupakan proses untuk mengevaluasi kinerja model yang telah dibangun dengan menguji kemampuan model untuk memprediksi data yang tidak terlihat sebelumnya. Hal ini dapat dilakukan dengan menguji model yang telah dioptimalkan pada data uji menggunakan `predict()` method pada masing-masing model. Mengevaluasi performa model pada data uji menggunakan metrik akurasi, presisi, recall, dan F1-score menggunakan `classification_report()` method pada scikit-learn library.

6. Perbandingan

Kemudian pada tahapan terakhir melakukan perbandingan dari hasil pengujian yang telah dilakukan kepada algoritma Decision Tree dan K-Nearest Neighbor. Hal ini tentunya dimaksudkan agar peneliti dapat mengetahui hasil yang optimal dan efektif dari pengujian yang dilakukan. Dengan begitu penggunaan algoritma dengan akurasi yang lebih tepat diharapkan mampu mempermudah industri perhotelan untuk mengetahui tingkat

kepuasaan pelanggannya secara lebih akurat sehingga mampu menarik pelanggan yang lebih banyak lagi.

HASIL DAN PEMBAHASAN

Pada bagian ini, akan dijelaskan mengenai hasil dan pembahasan dari penelitian yang dilakukan. Hal ini bertujuan untuk mengidentifikasi kepuasan hotel dengan metode yang efektif menggunakan perbandingan dari suatu algoritma.

1. Read Data Frame

Data yang digunakan dalam penelitian ini adalah dataset Hotel Booking Demand, yang berisi informasi tentang pemesanan hotel. Dataset ini memiliki 32 fitur, termasuk nama hotel, tanggal kedatangan dan keberangkatan, jumlah tamu, dan ulasan pelanggan. Sebelum melakukan pemrosesan data dilakukan proses membaca data frame menggunakan Python Pandas. Berikut adalah hasil dari baca data frame, data yang ditampilkan adalah 5 data teratas dari sekitar 10.000 data yang ada.

Tabel 1 . Tabel Read Data Frame

id	Gender	Age	purpose_of_travel	Type of Travel	Type Of Booking	Hotel wifi service	Departure	Ease of online booking	Hotel Location	...	
0	70172	Male	13	aviation	Personal Travel	Not defined	3	4	3	1	...
1	5047	Male	25	tourism	Group Travel	Group bookings	3	2	3	3	...
2	110028	Female	26	tourism	Group Travel	Group bookings	2	2	2	2	...
3	24026	Female	25	tourism	Group Travel	Group bookings	2	5	5	5	...
4	119299	Male	61	aviation	Group Travel	Group bookings	3	3	3	3	...

2. Exploratory Data Analysis (EDA)

Selanjutnya dilakukan eksplorasi data (EDA) untuk memahami karakteristik dataset dan memeriksa apakah terdapat nilai yang hilang atau outlier.

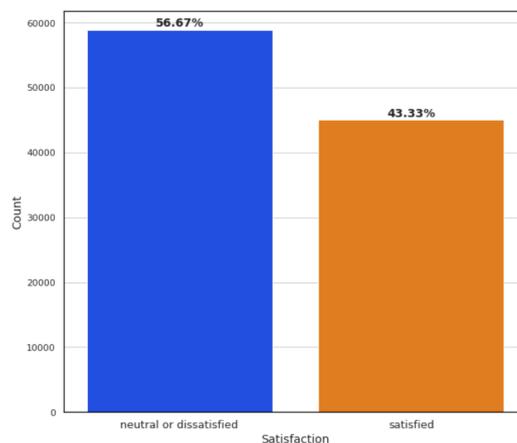
EDA menjadi sangat penting sebelum melakukan feature engineering dan modeling karena dalam tahap ini harus dilakukan pemahaman data terlebih dahulu [10]. Pada tahap ini digunakan teknik visualisasi data seperti histogram, boxplot, dan scatterplot untuk memperoleh gambaran yang lebih jelas tentang distribusi dan hubungan antara variabel dalam dataset.

2.1. Jumlah Kepuasan Pelanggan Hotel

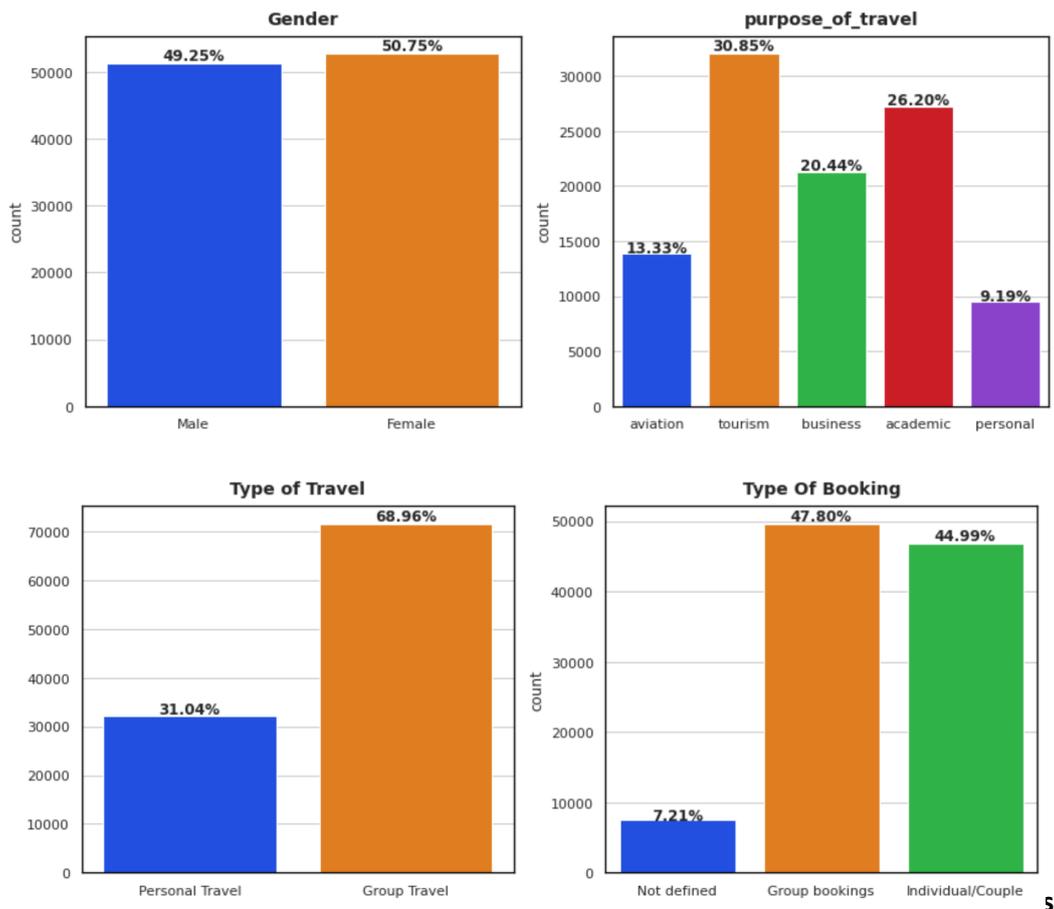
Grafik jumlah kepuasan pengguna hotel menunjukkan data “less imbalance” dengan "neutral or dissatisfied" sebesar 56.67% dan "satisfied" sebesar 43.33%, oleh karena itu akan dilakukan 3 percobaan untuk mencari model yang terbaik diantaranya data tanpa resampling, oversampling, dan undersampling.

2.2. Distribusi Data pada Kolom Kategorikal

Grafik distribusi data kategorikal pada kolom 'Gender' menunjukkan data hampir seimbang, lalu kolom 'Type of Booking' pada grafik "Group bookings" dan "Individual" menunjukkan data seimbang sedangkan grafik "Not defined" hanya terdapat sedikit data, untuk kolom lainnya adalah data tidak seimbang. Distribusi data juga dilakukan pada data kategorikal dalam bentuk skala ordinal, mayoritas orang memberikan nilai 4 pada kolom 'Departure/Arrival convenience'; 'Stay comfort'; 'Common Room Entertainment'; dan 'Other service', untuk nilai hampir seimbang 2 dan 3 pada kolom 'Hotel wifi service' dan 'Ease of Online booking', untuk nilai hampir seimbang 3 dan 4 pada kolom 'Checkin/Checkout service' dan 'Cleanliness', kemudian untuk nilai yang hampir seimbang 2 hingga 5 pada kolom 'Food and drink'.



Gambar 2. Grafik Jumlah Kepuasan Pengguna



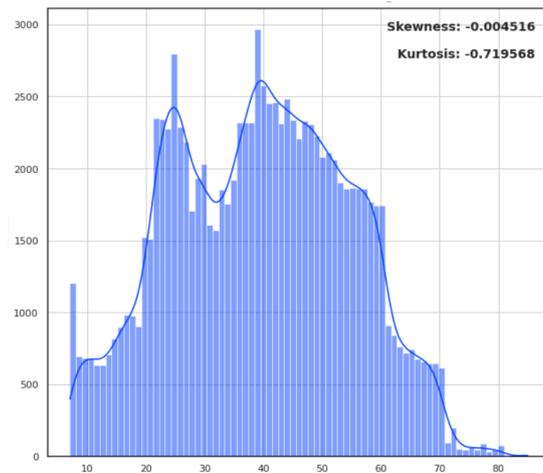
Gambar 3. Grafik Distribusi Data Pada Kolom Kategorikal

2.3. Distribusi Data Kolom Numerik

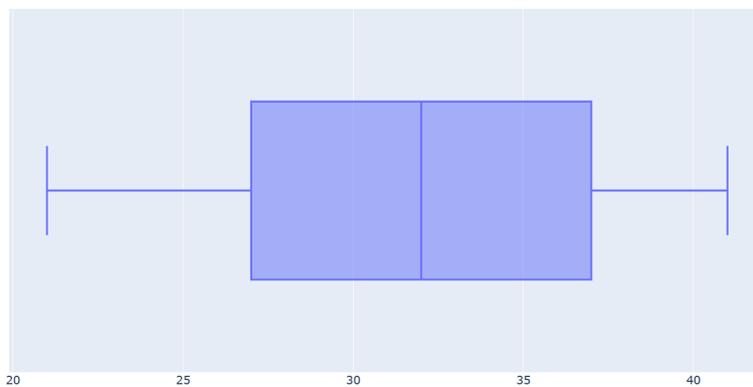
Nilai kurtosis menunjukkan -0,7 puncak tumpul, yang artinya bahwa data terdistribusi dengan merata dan tidak ada data ekstrim (outlier). Untuk membuktikan bahwa tidak ada data ekstrim dapat dilihat pada grafik boxplot berikut. Grafik menunjukkan bahwa dikolom numerik 'Age' tidak ada yang 1,5 lebih besar dari kuartil 3, maka data pada kolom 'Age' bukan data ekstrim (outlier).

Selanjutnya untuk menguji normalitas distribusi data, dapat menggunakan

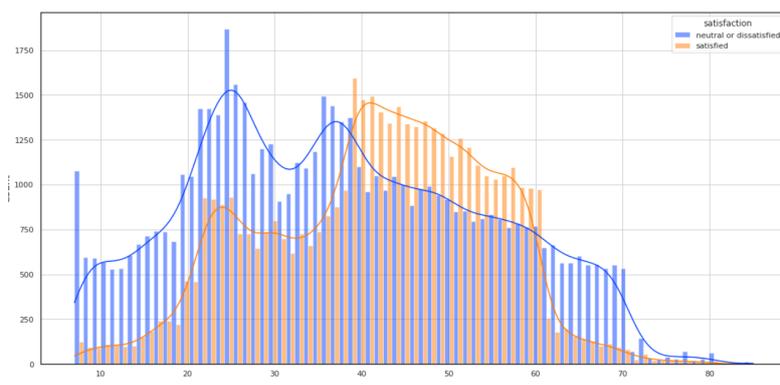
perbandingan z *critical value* atau nilai kritis z dihitung menggunakan $Z(\alpha/2)$. Pada kasus ini data diuji menggunakan α atau probabilitas 5%, jadi $Z(2,5\%)$ atau $Z(0,025)$. Untuk mencari nilai $Z(0,025)$ menggunakan probabilitas 0,975 yang didapatkan dari $1 - 0,025$. Berdasarkan tabel z nilai kritis z dengan probabilitas 0,975 adalah 1,96. Jadi jika dibandingkan nilai skewness dengan nilai kritis z didapatkan bahwa nilai skewness (-0,0045) lebih kecil dari nilai kritis z (1,96), maka distribusi data tersebut adalah normal.



Gambar 4. Grafik Distribusi Data Kolom Age



Gambar 5. Grafik Boxplot Age



Gambar 6. Grafik Perbandingan Kepuasan Hotel Berdasarkan Umur

2.4. Perbandingan Kepuasan Pengguna Hotel Berdasarkan Umur

Grafik menunjukkan rata-rata pengguna hotel yang puas pada rentan umur 40 hingga 60.

2.5. Korelasi Masing-masing Fitur Terhadap Kepuasan Peleanggan

Grafik heatmap menunjukkan fitur yang berhubungan kuat dengan 'satisfaction' adalah

'Common Room entertainment' dengan nilai korelasi sebesar 0,4.

Terdapat fitur yang berbanding terbalik dengan 'satisfaction' yaitu 'Departure/Arrival convenience' dengan nilai korelasi -0,052. Fitur yang tidak ada hubungannya dengan 'satisfaction' diantaranya 'id' dan 'Hotel location' dengan nilai korelasi yang mendekati 0, maka dari itu kedua fitur tersebut tidak dipakai dalam proses pembuatan model nantinya.

3. Data Pre-Processing

Setelah selesai melakukan EDA selanjutnya ke tahapan data pre-processing, pada tahap ini terdapat beberapa langkah yaitu

3.1. Menghapus Kolom yang Tidak Diperlukan

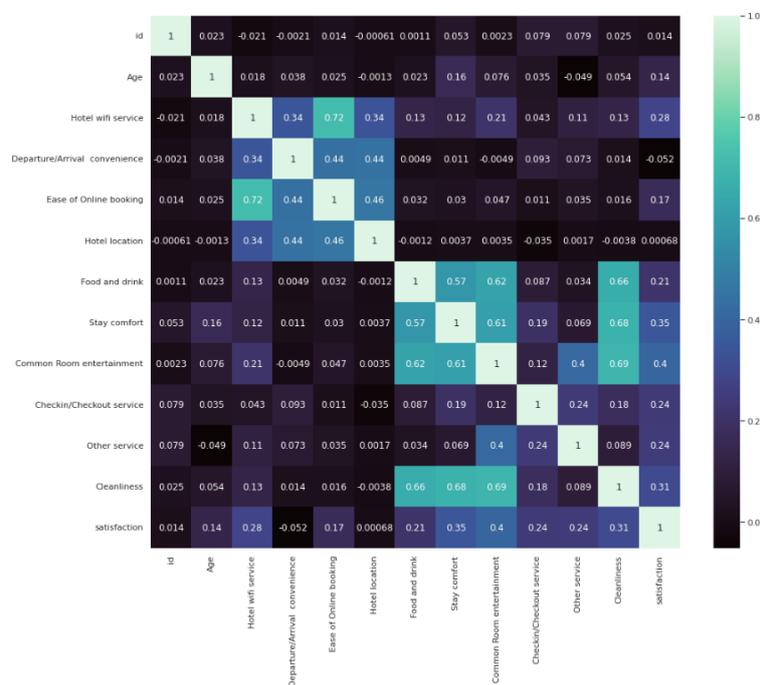
Pada tahap ini, kolom yang tidak diperlukan seperti 'Reservation status', 'Reservation status date', 'Country', 'Market Segment', dan 'Distribution Channel' dihapus dari dataset.

3.2. Menghapus duplikat

Kemudian dilakukan penghapusan baris yang sama di dalam dataset untuk menghindari pengaruh data duplikat pada hasil akhir.

3.3. Feature Engineering

Feature engineering dilakukan pada kolom 'Reservation status' dengan mempertimbangkan nilai yang berarti bagi analisis yaitu 'Canceled' atau 'Not Canceled'.



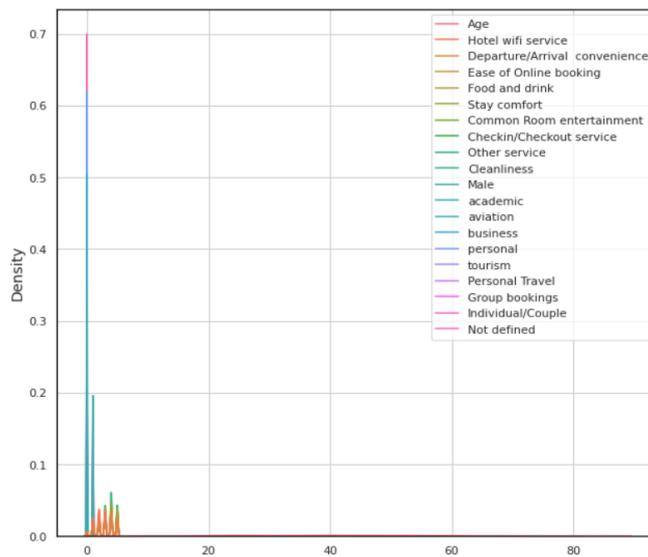
Gambar 7. Korelasi Fitur

3.5. Standarisasi Data

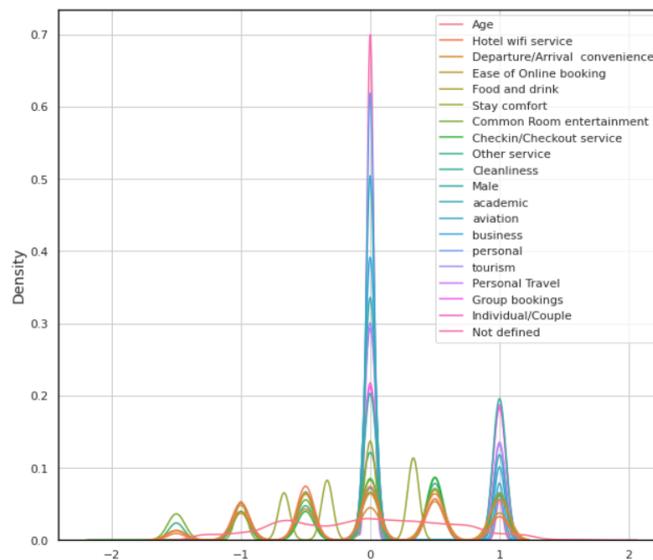
Terakhir, data numerik dalam dataset diubah menjadi standar skala agar tidak mempengaruhi hasil akhir dari model. Standarisasi dilakukan dengan menggunakan fungsi StandardScaler dari library sklearn. Berdasarkan seluruh grafik kernel density estimation (KDE) data lebih dominan skew ke kiri, maka dari itu standarisasi diperlukan. Metode standarisasi yang digunakan adalah

Robust Scaler, dimana metode ini cocok untuk menangani data yang skew ke kiri dan juga data kategorikal yang sudah diencoding [11]. Robust Scaler ini tidak mempertimbangkan median data dan menskalakan data berdasarkan rentang antara kuartil ke-1 (25%) dan kuartil ke-3 (75%).

Setelah proses preprocessing data selesai dilakukan, data siap digunakan dalam proses modelling.



Gambar 8. Grafik Sebelum Standarisasi



Gambar 9. Sesudah Standarisasi Robust Scaler

4. Modeling

Modelling merupakan tahapan yang dilakukan untuk membangun model prediksi atau klasifikasi dengan menggunakan data yang telah diproses pada tahap sebelumnya. Pada bagian ini dilakukan prediksi kepuasan tamu hotel dengan membandingkan dua algoritma yaitu Decision Tree dan KNN [12]. Dengan menentukan metode terbaik yang akan digunakan diantaranya tanpa resampling, undersampling dan oversampling.

Kemudian akan dilakukan evaluasi performa model dengan menghitung metrik evaluasi seperti akurasi, presisi, recall, dan F1-score [13]. Dengan cara perhitungan yaitu pertama akurasi, akurasi dari model yang dihitung dengan menggunakan jumlah prediksi yang benar dibagi dengan total jumlah prediksi.

Kedua presisi, kemampuan model untuk memprediksi hasil yang benar dari semua hasil yang diprediksi positif. Ketiga recall, Kemampuan model untuk menemukan kembali semua hasil positif yang benar dari semua hasil yang benar positif. Keempat F1 Score, rata-rata harmonis antara precision dan recall.

Pada saat melakukan modeling akan dilakukan tuning hyperparameter pada setiap metodenya hal ini untuk meningkatkan kinerja model. Sehingga diharapkan mampu menemukan kombinasi nilai hyperparameter yang optimal pada algoritma yang digunakan untuk memprediksi kepuasan tamu hotel.

4.1. Metode Tanpa Resampling

Metode Tanpa Resampling adalah salah satu metode dalam mengatasi ketidakseimbangan kelas pada data yang tidak melakukan manipulasi pada jumlah data pada setiap kelasnya. Dalam metode ini, data pada kelas mayoritas tidak dihapus sedangkan pada kelas minoritas juga tidak ditambah. Sehingga, jumlah data pada setiap kelas tetap sama seperti data asli. Metode Tanpa Resampling pada penelitian ini dilakukan pada saat membagi data menjadi data latih dan data uji dengan melakukan stratified sampling. Metode Tanpa Resampling mempertahankan distribusi kelas yang ada pada data asli tanpa melakukan manipulasi tambahan pada jumlah data di setiap kelasnya. Dalam konteks penelitian ini, teknik ini digunakan untuk mengatasi ketidakseimbangan kelas pada dataset hotel booking tanpa mengubah jumlah data pada setiap kelasnya.

Penggunaan teknik ini memiliki beberapa keuntungan. Pertama, karena tidak ada manipulasi pada jumlah data, proses ini mempertahankan informasi yang ada pada data asli tanpa memperkenalkan bias yang tidak diinginkan. Kedua, dengan mempertahankan distribusi kelas yang ada, teknik ini dapat membantu model untuk belajar dari sebagian besar kasus yang mungkin terjadi di dunia nyata. Namun, metode ini juga memiliki beberapa kelemahan. Salah satunya adalah jika kelas minoritas memiliki representasi yang sangat rendah dibandingkan dengan kelas mayoritas, model yang dihasilkan mungkin

cenderung tidak mampu mempelajari pola yang ada dalam kelas minoritas dengan baik. Hal ini dapat mengakibatkan performa model kurang optimal dalam mengklasifikasikan data dari kelas minoritas [14].

Dalam penelitian ini, kinerja penggunaan metode tanpa resampling dievaluasi dengan mengukur akurasi, presisi, recall, dan F1-score dari model Decision Tree dan K-Nearest Neighbor yang dilatih menggunakan data yang tidak mengalami resampling. Evaluasi ini membantu dalam memahami seberapa baik model dapat memprediksi kepuasan tamu hotel tanpa

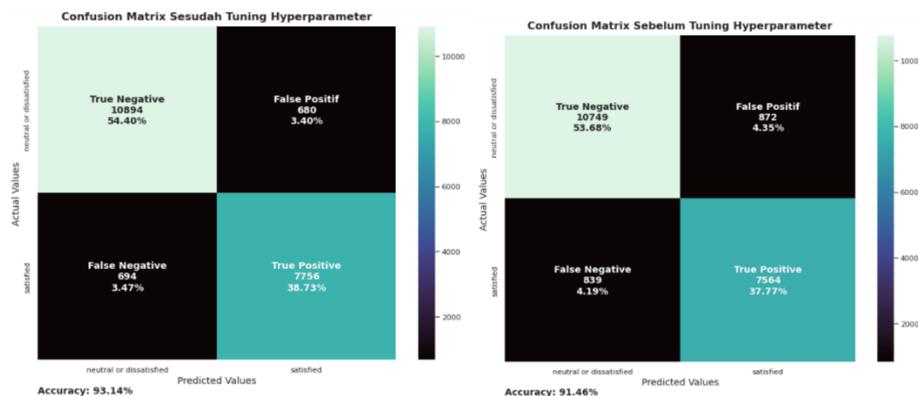
memanipulasi jumlah data pada setiap kelasnya.

4.1.1. Decision Tree Classifier

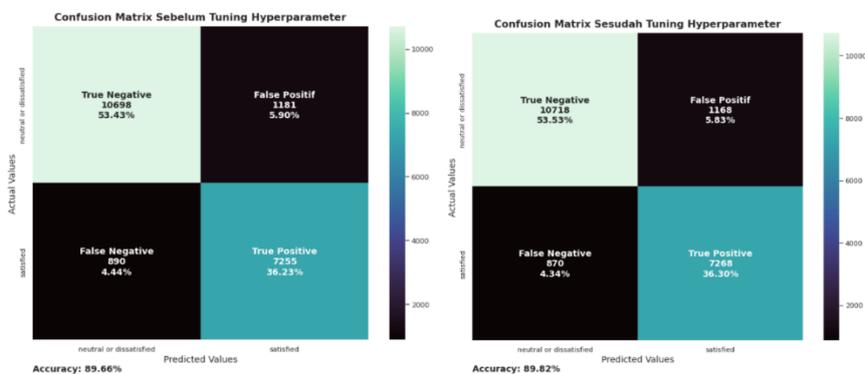
Dapat dilihat pada matrik diatas bahwa setelah dilakukan tuning hyperparameter terdapat peningkatan akurasi sebesar 1.68% yaitu dari nilai akurasi 91,46% menjadi 93,14%.

4.1.2. K-Nearest Neighbor Classifier

Dapat dilihat pada matrik diatas bahwa setelah dilakukan hyperparameter tuning terdapat peningkatan akurasi sebesar 0.16% yaitu dari nilai akurasi 89,66% menjadi 89,82%.



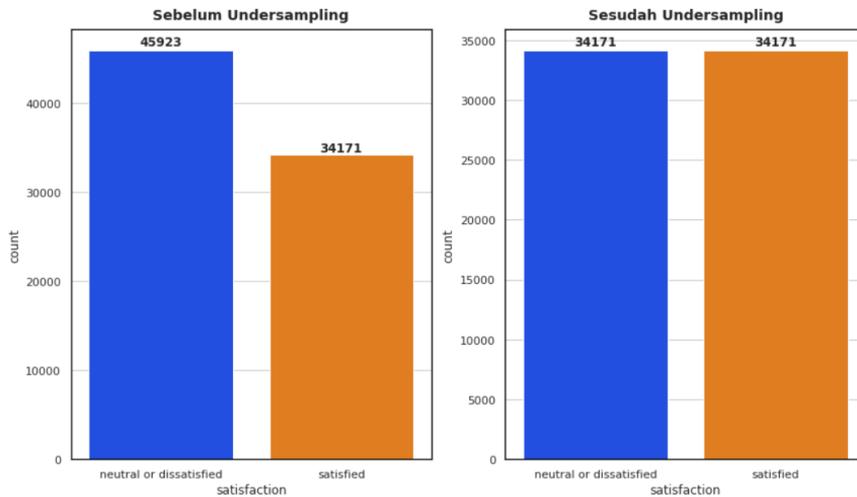
Gambar 10. Matrik Sebelum dan Sesudah *Tuning Hyperparameter*



Gambar 11. Matrik Sebelum dan Sesudah *Tuning Hyperparameter*

Tabel 2. Metode Tanpa Resampling

Model	Accuracy	Precision	Recall	F1 Score
Decision Tree	93,14%	0,91	0,91	0,91
KNN	89,82%	0,89	0,89	0,89



Gambar 12. Melakukan Undersampling

Hasil dari penggunaan metode tanpa resampling pada project ini adalah model Decision Tree dan K-Nearest Neighbor masing-masing mendapatkan akurasi sebesar 93,14% dan 89,82%.

4.2. Metode Undersampling

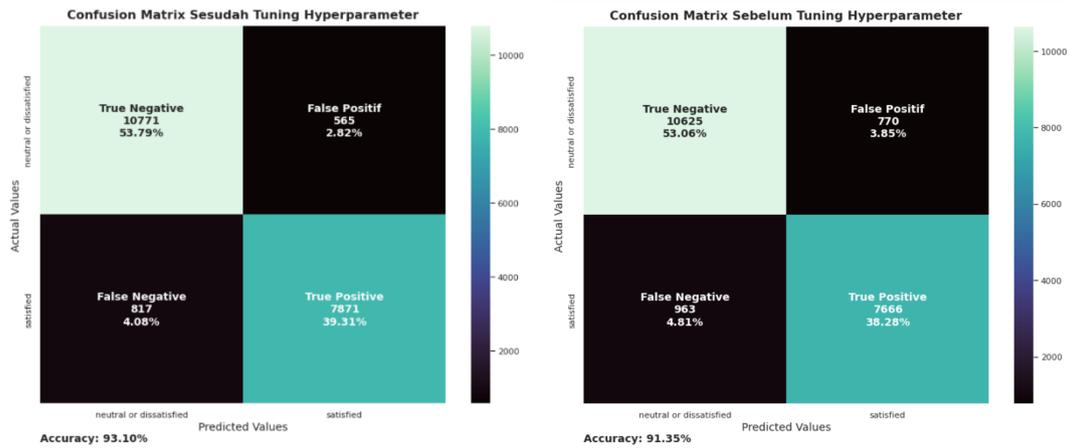
Pada penelitian ini metode undersampling diterapkan pada data training dengan menggunakan teknik NearMiss. NearMiss adalah teknik undersampling yang mengurangi jumlah sampel pada kelas mayoritas dengan cara memilih sampel dari kelas mayoritas yang paling dekat dengan sampel kelas minoritas [15]. Dengan melihat perbandingan dua grafik tersebut terdapat

perubahan pada jumlah sampel setiap kelas setelah dilakukan undersampling.

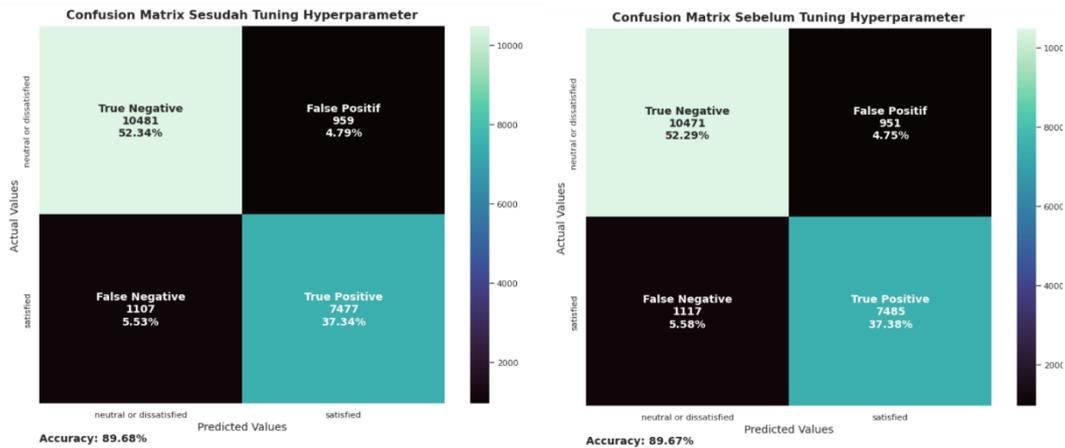
Tujuan dari undersampling adalah untuk menyeimbangkan jumlah sampel pada setiap kelas sehingga model dapat mempelajari fitur-fitur yang ada pada setiap kelas dengan lebih baik. Selain itu, undersampling juga dapat membantu menghindari overfitting pada model.

4.2.1. Decision Tree Classifier

Dapat dilihat pada matrik diatas bahwa setelah dilakukan tuning hyperparameter terdapat peningkatan akurasi sebesar 1.75% yaitu dari nilai akurasi 91,35% menjadi 93,10%.



Gambar 13. Matrik Sebelum dan Sesudah *Tuning Hyperparameter*



Gambar 14. Matrik Sebelum dan Sesudah *Tuning Hyperparameter*

Tabel 3. Metode *Undersampling*

Model	Accuracy	Precision	Recall	F1 Score
Decision Tree	93,10%	0,92	0,90	0,91
KNN	89,68%	0,89	0,89	0,89

4.2.2. *K-Nearest Neighbor Classifier*

Dapat dilihat pada matrik diatas bahwa setelah dilakukan hyperparameter tuning terdapat peningkatan akurasi sebesar 0.01% yaitu dari nilai akurasi 89,67% menjadi 89,68%.

Hasil dari penggunaan metode undersampling pada project ini adalah model Decision Tree dan K-Nearest Neighbor masing-masing mendapatkan akurasi sebesar 93,10% dan 89,68% dengan parameter-

parameter yang telah dioptimalisasi. Akurasi ini sedikit lebih rendah dibandingkan dengan metode tanpa resampling namun dapat dianggap cukup baik mengingat kelas-kelas pada data cukup tidak seimbang. Selain itu, teknik undersampling juga dapat mengurangi risiko overfitting pada model.

4.3. Metode *Oversampling*

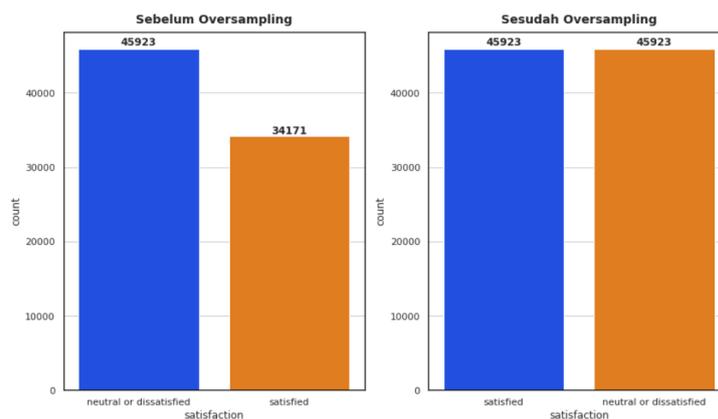
Pada penelitian ini, teknik oversampling digunakan dengan menggunakan Syntetic

Minority Oversampling (SMOTE) pada data training. Teknik SMOTE ini mampu membuat sampel sintetis pada kelas minoritas dengan cara mengambil tetangga terdekat dan membuat titik sintetis di antara titik-titik tersebut [16]. Tujuan dari penggunaan teknik oversampling ini adalah untuk meningkatkan kinerja model dalam mengenali kelas minoritas (yaitu, data yang menunjukkan ketidakpuasan pelanggan). Dengan melihat perbandingan dua grafik tersebut terdapat perubahan pada jumlah sampel setiap kelas setelah dilakukan oversampling. Tujuan dari

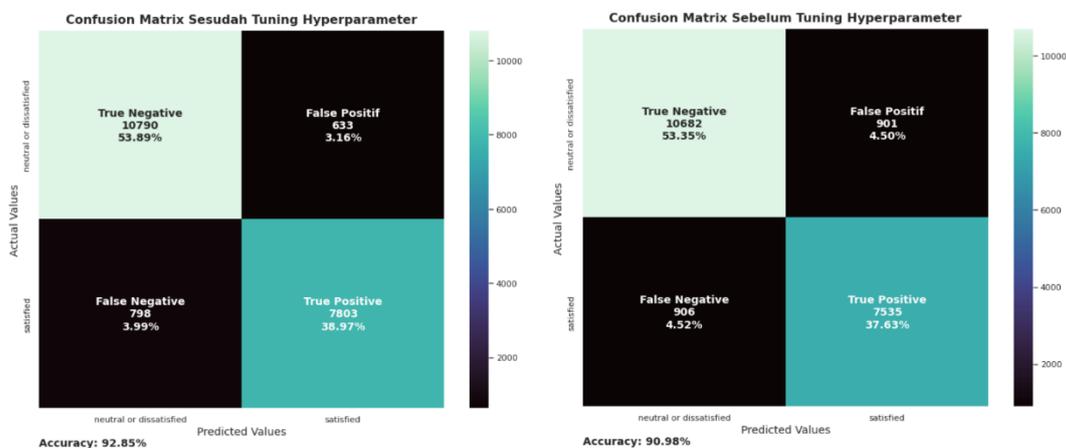
oversampling adalah untuk menyeimbangkan jumlah sampel pada setiap kelas sehingga model dapat mempelajari fitur-fitur yang ada pada setiap kelas dengan lebih baik. Selain itu, undersampling juga dapat membantu menghindari overfitting pada model.

4.3.1. Decision Tree Classifier

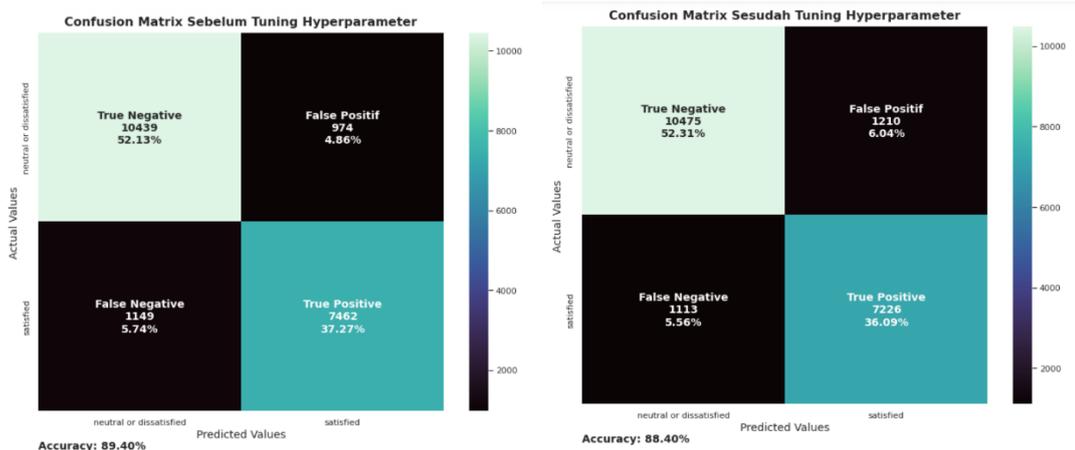
Dapat dilihat pada matrik diatas bahwa setelah dilakukan tuning hyperparameter terdapat peningkatan akurasi sebesar 1.85% yaitu dari nilai akurasi 90,98% menjadi 92,85%.



Gambar 15. Melakukan *Oversampling*



Gambar 16. Matrik Sebelum dan Sesudah *Tuning Hyperparameter*



Gambar 17. Matrik Sebelum dan Sesudah *Tuning Hyperparameter*

Tabel 4. Metode *Oversampling*

Model	Accuracy	Precision	Recall	F1 Score
Decision Tree	92,85%	0,92	0,90	0,91
KNN	88,40%	0,88	0,89	0,89

Tabel 5. Fitur yang Perlu Diperhatikan untuk Meningkatkan Kepuasan Pelanggan

0	label
0 0.342813	Hotel wifi service
1 0.286417	Group bookings
2 0.134389	Common Room entertainment
3 0.049493	Personal Travel
4 0.039887	Stay comfort
5 0.030905	Age
6 0.022963	Other service
7 0.021736	Departure/Arrival convenience
8 0.020492	Cleanliness
9 0.019444	Checkin/Checkout service

4.3.2. K-Nearest Neighbor Classifier

Dapat dilihat pada matrik diatas bahwa setelah dilakukan hyperparameter tuning terdapat penurunan akurasi sebesar 0.01% yaitu dari nilai akurasi 89,40% menjadi 88,40%. Teknik oversampling ini diaplikasikan pada model Decision Tree dan K-Nearest Neighbors, dengan menghasilkan akurasi sebesar 92,85% dan 88,40% secara berturut-turut seperti pada tabel 4.

Dari hasil tersebut, dapat dilihat bahwa model yang menggunakan metode tanpa

resampling memberikan performa yang cukup baik dibandingkan dengan metode undersampling dan oversampling. Decision Tree menjadi model terbaik dengan akurasi 93,18% pada metode tanpa resampling, 93,10% pada metode undersampling, dan 92,85% pada metode oversampling. Hal ini menunjukkan bahwa model Decision Tree dapat mengenali pattern pada data dengan baik dan cocok untuk digunakan dalam melakukan klasifikasi pada dataset hotel satisfaction. Setelah tahap pemodelan, khususnya setelah

menggunakan model Decision Tree tanpa resampling. Tabel ini memberikan informasi tentang urutan fitur yang perlu diperhatikan untuk meningkatkan kepuasan pelanggan berdasarkan hasil dari model Decision Tree tersebut. Fitur-fitur ini diperoleh dari analisis data yang dilakukan menggunakan teknik pembelajaran mesin. Berikut berdasarkan hasil Decision Tree tanpa resampling urutan fitur yang perlu diperhatikan jika ingin meningkatkan kepuasan pelanggan dapat dilihat pada tabel 5.

Tabel 5 di atas menunjukkan urutan fitur yang perlu diperhatikan untuk meningkatkan kepuasan pelanggan berdasarkan hasil dari model Decision Tree tanpa resampling. Fitur-fitur ini diperoleh dari analisis data yang dilakukan menggunakan teknik pembelajaran mesin yang telah dilakukan. Urutan fitur tersebut diperoleh dari hasil analisis pentingnya masing-masing fitur dalam melakukan klasifikasi kepuasan pelanggan. Nilai yang tercantum pada tabel merupakan pentingnya setiap fitur dalam mempengaruhi keputusan klasifikasi yang dihasilkan oleh model Decision Tree.

KESIMPULAN DAN SARAN

Berdasarkan hasil modelling menggunakan tiga metode yaitu tanpa resampling, undersampling, dan oversampling dengan tuning hyperparameter dapat disimpulkan bahwa, diperoleh hasil terbaik menggunakan Algoritma Decision Tree tanpa

resampling dengan akurasi 93,14%. Kompleksitas waktu tercepat diantara kedua algoritma yang digunakan adalah Decision Tree. Berdasarkan Decision Tree tanpa resampling, faktor yang paling berpengaruh dalam meningkatkan kepuasan pelanggan adalah hotel wifi services, group bookings, dan common room entertainment. Hal ini dapat menjadi panduan bagi manajemen hotel dalam meningkatkan kualitas pelayanan dan kepuasan pelanggan.

DAFTAR PUSTAKA

- [1] S. Rizal, A. R. Rahim, and E. Wardiana, "PENGARUH KUALITAS PELAYANAN TERHADAP KEPUASAN NASABAH PADA PT. BANK RAKYAT INDONESIA (PERSERO) Tbk.UNIT BENGKABANG WATAMPONE," *J. Ilmu Manaj. Profitab.*, vol. 4, no. 1, pp. 98–113, 2020, doi: 10.26618/profitability.v4i1.3051.
- [2] W. B. Yang and M. R. A. Campos, "Factors Affecting Customer Satisfaction in International Hotels in Chenzhou City, China," *PEOPLE Int. J. Soc. Sci.*, vol. 5, no. 3, pp. 925–939, 2020, doi: 10.20319/pjss.2020.53.925939.
- [3] S. F. Sutan Faisal and N. Sutan Faisal, "Implementation of K-Nearest Neighbor Algorithm for Customer Satisfaction," *Buana Inf. Technol.*

- Comput. Sci. (BIT CS)*, vol. 1, no. 2, pp. 27–32, 2020, doi: 10.36805/bit-cs.v1i2.886.
- [4] Fadhli Almu’ini Ahda and Mohammad Zainuddin, “Prediksi Kepuasan Pelayanan Perpustakaan Menggunakan Algoritma Decision Tree (C4.5),” *J. Teknol. Inf.*, vol. 10, no. 2, pp. 143–150, 2019.
- [5] S. Amalia, I. Deborah, and I. N. Yulita, “Comparative analysis of classification algorithm: Random Forest, SPAARC, and MLP for airlines customer satisfaction,” *Sinergi*, vol. 26, no. 2, p. 213, 2022, doi: 10.22441/sinergi.2022.2.010.
- [6] B. Noori, “Classification of Customer Reviews Using Machine Learning Algorithms,” *Appl. Artif. Intell.*, vol. 35, no. 8, pp. 567–588, 2021, doi: 10.1080/08839514.2021.1922843.
- [7] A. K. Febrian, Y. H. Chrisnanto, D. Pupita, N. Sabrina, and J. Achmad Yani, “SNESTIK Seminar Nasional Teknik Elektro, Sistem Informasi, dan Teknik Informatika Studi Komparasi Metode Klasifikasi K-Nearest Neighbor dan Naïve Bayes dalam Mengidentifikasi Kepuasan Pelanggan Terhadap Produk,” *Semin. Nas. Tek. Elektro, Sist. Informasi, dan Tek. Inform.*, p. 333, 2022, [Online]. Available: <https://ejurnal.itats.ac.id/snestikdanhttps://snestik.itats.ac.id>
- [8] B. A. C. P., “DOI: 10.29408/jit.v1i1.892,” *Baiq Andriskha Candra*, vol. 1, no. 1, pp. 32–39, 2018.
- [9] P. Pangestu and R. Setyadi, “Penerapan Metode K-Nearest Neighbor Untuk Pemilihan Rekomendasi Game FPS Pada Aplikasi Google Play Store,” vol. 4, no. 2, pp. 742–747, 2023, doi: 10.47065/josh.v4i2.3006.
- [10] A. Suwarno *et al.*, “Jurnal Teknologi Pelita Bangsa,” *J. Teknol. Pelita Bangsa*, vol. 12, no. 4, pp. 33–40, 2021.
- [11] C. L. Lin, T. P. Chen, K. C. Fan, H. Y. Cheng, and C. H. Chuang, “Radar high-resolution range profile ship recognition using two-channel convolutional neural networks concatenated with bidirectional long short-term memory,” *Remote Sens.*, vol. 13, no. 7, 2021, doi: 10.3390/rs13071259.
- [12] A. N. Yuliarina and H. Hendry, “Comparison of Prediction Analysis of Gofood Service Users Using the Knn & Naive Bayes Algorithm With Rapidminer Software,” *J. Tek. Inform.*, vol. 3, no. 4, pp. 847–856, 2022, doi: 10.20884/1.jutif.2022.3.4.294.
- [13] D. McCashin and C. M. Murphy, “Using TikTok for public and youth mental health – A systematic review and content analysis,” *Clin. Child Psychol. Psychiatry*, vol. 28, no. 1, pp. 279–306, 2023, doi: 10.1177/13591045221106608.
- [14] Y. Hartiwi, E. Rasywir, Y. Pratama, and P. A. Jusia, “Eksperimen Pengenalan

- Wajah dengan fitur Indoor Positioning System menggunakan Algoritma CNN,” *Paradig. - J. Komput. dan Inform.*, vol. 22, no. 2, pp. 109–116, 2020, doi: 10.31294/p.v22i2.8906.
- [15] M. R. Raharjo and A. P. Windarto, “Penerapan Machine Learning dengan Konsep Data Mining Rough Set (Prediksi Tingkat Pemahaman Mahasiswa terhadap Matakuliah),” *J. Media Inform. Budidarma*, vol. 5, no. 1, p. 317, 2021, doi: 10.30865/mib.v5i1.2745.
- [16] A. Sanjaya, E. Setyati, and H. Budianto, “Modeling of Convolutional Neural Network Architecture for Recognizing The Pandava Mask,” *Inf. J. Ilm. Bid. Teknol. Inf. dan Komun.*, vol. 5, no. 2, pp. 99–103, 2020, [Online]. Available: <https://doaj.org/article/1078dd6621ea4b7c9e06bb47b0455d11>