BAB II

TINJAUAN PUSTAKA

2.1 Penilitian Terkait

Penelitian yang merujuk pada [5] oleh Louis Madaerdo Sotarjua & Dian Budhi Santoso pada tahun 2022 dengan Judul "Perbandingan Algoritma *KNN, Decision Tree, dan Random Forest* pada Data Kelas Tidak Seimbang untuk Klasifikasi Promosi Karyawan" mengkaji penerapan teknik over-sampling SMOTE bersama dengan algoritma klasifikasi yang telah disebutkan. Penelitian ini bertujuan untuk mengklasifikasikan karyawan ke dalam dua kategori: karyawan yang dipromosikan dan karyawan yang tidak dipromosikan.

Penelitian berjudul "Klasifikasi Citra Eurosat Menggunakan Algoritma *KNN, Decision Tree, dan Random Forest*" yang dilakukan oleh Yusril Iza Fajarendra, Yulis Rizal Fauzan, dan Shofwatul'Uyun pada tahun 2024 menunjukkan bahwa penggunaan algoritma *Random Forest* dapat meningkatkan akurasi dan performa dalam klasifikasi citra satelit. Dalam studi ini, ketiga algoritma—*KNN, Decision Tree, dan Random Forest*—dibandingkan untuk mengevaluasi kinerjanya dalam mengklasifikasikan data citra EuroSat. Data yang telah diekstraksi kemudian digunakan sebagai input untuk ketiga algoritma tersebut. Pengujian performa model dilakukan dengan metode K-Fold Cross Validation. Hasil penelitian menunjukkan bahwa Random Forest memberikan kinerja terbaik, dengan akurasi sebesar 81,32%, recall 81,33%, precision 81,18%, dan F1 score 80,82%.. [6]

Penelitian yang dilakukan oleh Muhammad Akil Hi Umar dan Bangkit Sitohang pada tahun 2024 berjudul "Analisis Faktor-Faktor Yang Memengaruhi Keputusan Pembelian Paket Wisata Menggunakan Model Klasifikasi *Decision Tree*,

Random Forest Dan K-Nearest Neighbours" menerapkan metode *Decision Tree*. Metode ini mampu memberikan prediksi dengan akurasi yang cukup tinggi. Temuan ini memberikan wawasan penting bagi sektor perjalanan untuk meningkatkan strategi pemasaran dan penjualan mereka. Perusahaan travel dapat memanfaatkan hasil penelitian ini untuk merancang pemasaran yang lebih efisien dan menyesuaikan produk yang mereka tawarkan sesuai dengan preferensi dan karakter pelanggan. Dengan demikian, penelitian ini tidak hanya memberikan pemahaman yang lebih baik tentang perilaku konsumen dalam industri perjalanan, tetapi juga memberikan panduan praktis bagi perusahaan untuk meningkatkan kepuasan pelanggan dan hasil bisnis secara optimal.[7]

Penelitian yang merujuk pada [8] oleh Nichola Charles, Aldi Bernard Wijaya, Marvin, dan Denny Jollyta pada tahun 2024, berjudul "Analisis Data Penyewaan Sepeda di Seoul Menggunakan Berbagai Metode Klasifikasi." Penelitian ini menghadapi tantangan terkait kemacetan lalu lintas dan mendukung alternatif transportasi yang ramah lingkungan. Fokus dari studi ini adalah pada pentingnya ketepatan dan konsistensi dalam klasifikasi sesuai dengan musim, dengan menerapkan teknik analisis data seperti Principal Component Analysis (PCA), *K-Nearest Neighbor (KNN), Naïve Bayes, Support Vector Machine (SVM), Decision Tree, Random Forest, Regresi Logistik, dan Gradient Boosting*. Temuan dari analisis klasifikasi yang dilakukan berdasarkan bulan dan tanggal menunjukan bahwa metode Random Forest mencatatkan tingkat akurasi tertinggi, yakni 0.709, dengan pembagian data sebanyak 80% untuk pelatihan dan 20% untuk pengujian.

Penelitian yang dilakukan oleh Miftah Raka Sujono, Agus Bahtiar, dan Bambang Irawan pada tahun 2023 berjudul "Analisis Model Machine Learning Untuk Jenis Aspal Di Jawa Barat Menggunakan Algoritma *Decision Tree* Dan *Random Forest*" menunjukkan hasil yang menggembirakan. Model berbasis algoritma *Random Forest* mencapai akurasi sebesar 90,79% pada dataset yang mencakup jenis aspal di Jawa Barat. Dalam evaluasi model, *Random Forest* menunjukkan kinerja sempurna

dengan akurasi 100% pada data pelatihan dan 90,79% pada data pengujian. Selain itu, metrik evaluasi seperti presisi, recall, dan F1-score juga menunjukkan hasil yang sangat baik untuk setiap kelas.. [9]

Tabel 2.1 Penelitian Terkait

No	Judul, Penulis, Tahun	Dataset	Metode	Hasil
1	Perbandingan *Algoritma KNN, Decision Tree, Dan Random Forest* pada data imbalanced class untuk klasifikasi promosi karyawan. Louis Madaerdo, Dian Budhi Santoso 2022	Data pada penelitian ini diperoleh melalui kaggle (https://www.kaggle.com/rsnayak/wns-analytics-wizard-2018-ml-hackathon).	*K-Nearest Neighbor(KNN), Decision Tree dan Random Forest.*	Hasil perfomal model klasifikasi dari model algoritma yang digunakan pada penelitian ini, maka model *KNN* memiliki hasil perfoma yang terbaik nilai metriks evaluasinya.
2	Klasifikasi Citra Eurosat menggunakan algoritma *KNN, Decision Tree, dan Random Forest.*	Data yang digunakan pada penelitian ini yaitu citra satelit aero sat yang	*K-Nearest Neighbor(KNN), Decision Tree dan Random Forest.*	dapat disimpulkan bahwa algoritma *Random Forest* mendapatkan performa terbaik dibandingkan

	Yusril Iza Fajarendra, Yulis Rizal Fauzan, Shofwatul ‘uyun 2024	diperoleh dari link https://www.kaggle.com/datset/apollo2506/eurosat-dataset/data.		*algoritma KNN dan Decision Tree.*
3	Analisis faktor-faktor yang memengaruhi keputusan pembelian paket wisata menggunakan model klasifikasi *Decision Tree, Random Forest, Dan K-Nearest Neighbours*. Muhammad Akil Hi Umar, Bangkit Sitohang 2024	Data set yang dipergunakan pada proses pengolahan data kali ini diambil dari Kaggle. Data set yang digunakan berisi informasi tentang pelanggan yang berencana membeli paket	*Decision Tree, Random Forest, Dan K-Nearest Neighbours.*	Hasil eksperimen menunjukkan model *KNN* memiliki performa yangbaik dalam mengklasifikasikan pelanggan berdasarkan peluang pembelian paket wisata baru.

		wisata. Dataset ini terdiri dari 4888 baris dan 20 kolom.		
4	Analisis data sewa sepeda di seoul dengan variatif metode klasifikasi Nichola Charel, Aldi Bernard Wijaya, Marvin, Deny Jollyta 2024	Data penelitian ini penulis menggunakan dataset Rental Sepeda di Seoul yang dapat diunduh dari situs UCI Machine Learning Repository dalam bentuk format CSV(Comma Separated Values), data	*Principal Component Analysis (PCA), K-Nearest Neighbor (KNN), NaïveBayes, Support Vector Machine (SVM), Decision Tree, Random Forest, Regresi Logistik, danGradient Boosting.*	Hasil penelitian dalam analisis klasifikasi berdasarkan musimam per jam dan tanggal, mendapatkan keakurasian data yang paling tinggi sebesar 0.709 oleh *metode Random Forest* dengan menggunakan data training 80% dan data testing 20%

		tersebut memiliki 8760 baris data		
5	Analisis model machine learning untuk aspal di jawa barat menggunakan *algoritma Decision Tree Dan Random Forest.* Mitah Raka Sujono, Agus Bahtiar, Bambang Irawan 2023	Data set yang digunakan terdiri dari 1.191 data yang dikumpulkan dari opendata.jabarprov.go.id dari tahun 2019 hingga 2022. Dengan menggunakan aplikasi Google Colab, analisis dilakukan berdasarkan tahapan Knowledge	*K-Nearest Neighbor(KNN), Decision Tree dan Random Forest.*	Hasil penelitian menunjukkan bahwa model yang menggunakan algoritma *Random Forest* memiliki akurasi sebesar 90,79% pada dataset yang mengandung aspal di Jawa Barat.

		Discovery in Database (KDD).		

Dari hasil beberapa review jurnal yang saya baca, saya menyimpulkan bahwa agar dataset dapat terakurasi dengan baik maka dibutuhkan suatu pengembangan model algoritma untuk mengklasifikasikasi faktor-faktor karyawan meninggalkan perusahaan atau tidak agar mendapatkan akurasi tertinggi. Dan metode yang digunakan untuk bebrapa jurnal yang saya review tingkat akurasi cenderung lebih tinggi dengan menggunakan algoritma *K-Neighbors Classifer (KNN)* dan *Random Forest.* Hasil dapat dilihat dari tabel berikut :

Tabel 2.2 Hasil Akurasi Review Jurnal

No. Jurnal	**Metode Terbaik Dan Akurasi**
1	*K-Nearest Neighbor (KNN)* : akurasi 86.57%
2	*Random Forest* : akurasi 81,32%
3	*K-Nearest Neighbor (KNN)* : akurasi 96%
4	*Random Forest* : akurasi 90,79%
5	*Random Forest* : akurasi 70,95%

2.2 Landasan Teori

a. Perbandingan

Perbandingan adalah proses membandingkan dua nilai atau lebih dari suatu besaran sejenis, yang disampaikan dengan cara yang jelas dan sederhana. [10].

b. Klasifikasi

Klasifikasi adalah suatu proses yang bertujuan untuk menemukan model atau fungsi yang dapat menjelaskan dan membedakan antara berbagai konsep atau kelas data. Tujuannya adalah untuk memperkirakan kelas suatu objek yang labelnya belum diketahui. [11]. Klasifikasi adalah aspek krusial dalam proses penambangan data komunitas. Sebagai teknik penambangan data yang bersifat prediktif, klasifikasi memanfaatkan informasi yang telah diketahui dari berbagai kumpulan data untuk membuat prediksi mengenai nilai-nilai yang belum terungkap.[12].

c. Turnover Karyawan

Turnover karyawan merujuk pada proses pengunduran diri permanen yang dilakukan oleh karyawan, baik secara sukarela maupun tidak. [13]. Turnover karyawan adalah suatu kondisi yang sebaiknya dihindari oleh perusahaan atau organisasi. Biasanya, turnover menjadi pilihan terakhir bagi seorang karyawan ketika mereka merasa bahwa kondisi kerja yang ada tidak lagi sejalan dengan harapan dan keinginan mereka.[14]. Turnover merujuk pada fenomena keluarnya karyawan dari suatu organisasi atau tempat mereka bekerja. [15]. Turnover merujuk pada rencana karyawan untuk meninggalkan organisasi, yang sering kali dipicu oleh penurunan produktivitas perusahaan atau perubahan dalam tugas kerja. Niat karyawan untuk melakukan turnover biasanya ditandai dengan keinginan untuk mencari pekerjaan di tempat lain yang dianggap lebih mudah diakses.[16].

2.3 Metode klasifikasi turnover karyawan karyawan

1. Decision Tree

Decision Tree adalah metode yang terkenal dan efektif untuk melakukan prediksi dan klasifikasi. Metode ini mengubah data menjadi struktur pohon keputusan, yang kemudian dapat disederhanakan menjadi serangkaian aturan. Dengan demikian, *Decision Tree* memungkinkan kita untuk memahami informasi dengan lebih jelas dan membuat keputusan yang lebih baik.[17]. *Decision tree* adalah sebuah teknik model prediktif yang dapat digunakan untuk tugas klasifikasi dan prediksi. [18]. Algoritma

pohon keputusan C4.5 merupakan salah satu metode klasifikasi data yang memiliki struktur sederhana dan mudah diinterpretasikan..[19]. *Decision Tree* adalah sebuah diagram alir yang mirip dengan struktur pohon. Di dalamnya, setiap node internal mewakili pengujian terhadap suatu atribut, sementara setiap cabang menggambarkan hasil dari pengujian tersebut. Di akhir, leaf node menunjukkan kelas-kelas atau distribusi kelas yang terkait.[20]. *Decision Tree* adalah algoritma yang sering digunakan dalam proses pengambilan keputusan. Algoritma ini mencari solusi dari suatu permasalahan dengan menggunakan kriteria sebagai node yang saling terhubung, membentuk struktur menyerupai pohon. Sebagai model prediksi untuk menentukan suatu keputusan, *Decision Tree* mengandalkan struktur hierarki atau pohon. Setiap pohon memiliki cabang-cabang yang mewakili atribut-atribut yang harus dipenuhi untuk melangkah ke cabang berikutnya, hingga akhirnya mencapai bagian daun.[21]. C4.5 adalah kumpulan algoritma untuk teknik klalsifikaasi dalam pembelajaran mesin dan penambangan data. Tujuannya adalah pembelajaran terawasi, dimana setiap tupel dalam kumpulan data dapat dijelaskan oleh sekumpulaln nilai atribut, dan setiap tupel milik salah satu dalri banyak kelas yalng berbeda dan tidak kompatibe [22]. Algoritma C4.5 adalah sebuah algoritma yang berfungsi untuk membangun decision tree (pohon keputusan). Algoritma C4.5 dan pohon keputusan merupakan dua model yang tidak terpisahkan. Algoritma C4.5 adalah salah satu dari algoritma klasifikasi yang kuat dan banyak digunakan atau di implementasikan untuk pengklasifikasian dalam berbagai hal [23].

2. Random Forest

Random forest adalah metode pembelajaran mesin yang mengimplementasikan konsep pembelajaran terawasi untuk membangun kelas klasifikasi. [24]. Algoritma *Random Forest Classifier* adalah perpanjangan dari model *Decision Tree*, di mana setiap pohon keputusan dilatih dengan menggunakan contoh yang berbeda. Dalam model ini, sejumlah pohon dibuat dengan metode yang mirip, dan seiring bertambahnya data, pohon-pohon itu juga akan tumbuh. *Random Forest Classifier* adalah algoritma yang

menciptakan pohon untuk klasifikasi dan regresi, di mana setiap node dibagi dengan algoritma yang disesuaikan untuk mengurangi kehilangan squared-error. [25]. *Random Forest* adalah salah satu metode ensemble learning yang menggabungkan beberapa *decision tree*. Metode ini beroperasi dengan membangun sejumlah *decision tree* berdasarkan subset data latih yang dipilih secara acak. Hasil prediksi dari setiap *decision tree* kemudian digabungkan melalui proses voting untuk menghasilkan prediksi akhir. [26]. *Random Forest (RF)* adalah sebuah algoritma yang memanfaatkan metode pemisahan biner rekursif untuk mencapai node akhir dalam struktur pohon klasifikasi dan regresi. Algoritma ini memiliki sejumlah keunggulan, seperti menghasilkan tingkat kesalahan yang relatif rendah, performa yang baik dalam klasifikasi, serta kemampuan untuk menangani data pelatihan dalam jumlah besar secara efisien. Selain itu, *Random Forest* juga merupakan metode yang efektif untuk mengestimasi data yang hilang. Algoritma ini membentuk banyak pohon independen, yang diperoleh dari subset data yang dipilih secara acak melalui bootstrap dari sampel pelatihan, serta dari variabel input yang dipilih di setiap node.[21]. *Random Forest* merupakan metode yang menggabungkan berbagai teknik pohon keputusan ke dalam suatu model yang harmonis. Terdapat tiga poin utama yang menjadi inti dari metode *Random Forest*: pertama, proses pengambilan sampel menggunakan bootstrapping untuk membangun pohon-pohon prediksi; kedua, setiap pohon keputusan membuat prediksi dengan menggunakan sejumlah prediktor yang dipilih secara acak; dan ketiga, *Random Forest* menghasilkan prediksi akhir dengan mengombinasikan hasil dari setiap pohon keputusan melalui mekanisme voting mayoritas untuk klasifikasi, atau dengan menghitung rata-rata untuk regresi.[27].

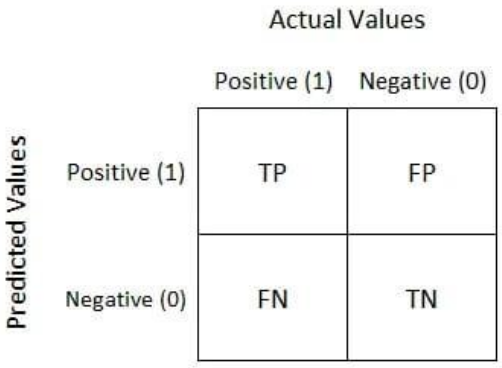
3. KNN (K-Nearest Neighbor)

Algoritma *K-Nearest Neighbor* adalah metode klasifikasi yang mengelompokkan objek berdasarkan data pelatihan yang memiliki jarak terdekat dengan objek tersebut. [28]. Algoritma *K-Nearest Neighbor (KNN)* merupakan sebuah teknik untuk mengklasifikasikan sekumpulan data dengan dasar pembelajaran dari data yang telah

diklasifikasi sebelumnya. [29]. *K-Nearest Neighbor (KNN)* adalah sebuah algoritma pengklasifikasi data yang cukup sederhana. Algoritma ini menggunakan perhitungan jarak terpendek sebagai ukuran untuk mengklasifikasikan kasus baru, berdasarkan tingkat kemiripannya dengan data yang sudah ada. [30]. Proses perancangan model algoritma *K-Nearest Neighbor* dimulai dengan membangun model tersebut menggunakan beberapa nilai k yang umum digunakan. Selanjutnya, nilai k yang memberikan performa terbaik akan dipilih. Dalam pengukuran jarak, digunakan dua metode, yaitu Euclidean Distance dan Manhattan Distance. Hasil dari kedua metode ini kemudian dibandingkan untuk menentukan model yang paling optimal dalam menangani data terkait penyakit jantung.[31]. Algoritma *KNN* adalah metode klasifikasi atau regresi nonparametrik yang digunakan dalam pengenalan pola. [32]. Algoritma *K-Nearest Neighbor (KNN)* adalah metode dalam supervised learning yang mengklasifikasikan sebuah instance baru berdasarkan data dari kategori yang paling banyak hadir di sekitarnya. Kelas yang paling sering muncul di antara tetangga terdekat tersebut akan menjadi hasil klasifikasi. Data yang digunakan dalam proses ini adalah hasil perhitungan rata-rata dari kumpulan data awal. [30]. *K-NN* adalah teknik klasifikasi yang digunakan untuk mengklasifikasikan dataset yang telah ditandai sebelumnya. Keakuratan algoritma *K-NN* sangat dipengaruhi oleh variasi fitur, terutama ketika nilainya tidak sesuai dengan nilai yang diperkirakan. Beberapa penelitian yang memanfaatkan algoritma *K-NN* fokus secara khusus pada pemilihan fitur dan pembobotan, dengan tujuan meningkatkan efisiensi algoritma dalam proses klasifikasi.[33]. *K-Nearest Neighbor (KNN)* adalah salah satu algoritma klasifikasi yang sederhana dan populer dalam Machine Learning. Algoritma *KNN* digunakan untuk mengklasifikasikan data berdasarkan kategori atau label yang dimiliki oleh tetangga terdekatnya [34].

3 Metrik Evaluasi

Pada tahap ini, dilakukan evaluasi terhadap performa model algoritma yang diterapkan dalam metode pembelajaran klasifikasi, yaitu *KNN*, *Decision Tree,* dan *Random Forest*. Evaluasi ini didasarkan pada analisis objek yang diklasifikasikan dengan benar dan yang salah. Untuk mengukur performa klasifikasi dalam penelitian ini, digunakan confusion matrix, yang menyajikan perhitungan hasil klasifikasi aktual yang dapat diprediksi.[5].



Gambar 2.3 Metrik Evaluasi

Keterangan :

1. True Positive (TP)**:** Kasus di mana model memprediksi **positif** dan hasil sebenarnya juga **positif**. Artinya, model berhasil mengidentifikasi dengan benar kelas positif.
2. False Positive (FP)**:** Kasus di mana model memprediksi **positif**, tetapi hasil sebenarnya adalah **negatif**. Ini disebut juga dengan *type I error*, yaitu ketika model salah mendeteksi positif.
3. False Negative (FN)**:** Kasus di mana model memprediksi **negatif**, tetapi hasil sebenarnya adalah **positif**. Ini disebut juga dengan *type II error*, yaitu ketika model gagal mendeteksi positif.

4. True Negative (TN)**:** Kasus di mana model memprediksi **negatif** dan hasil sebenarnya juga **negatif**. Artinya, model berhasil mengidentifikasi dengan benar kelas negatif.

Pada Tabel 1 TP adalah True Positive, TN adalah True Negative, FP adalah *False Positive* dan FN adalah *False Negative*. Pada penelitian ini performa klasifikasi yang akan dihitung adalah *accuracy*, *precision*, *recall* dan F1 *score*. rumus perhitungan *accuracy*, *precision*, *recall* dan F1 *score*[5].

a) Accuracy merupakan perhitungan untuk mengukur persentase prediksi yang benar dari keseluruhan data. [5]

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

b) Precision merupakan perhitungan untuk mengukur akurasi dan prediksi positif.[5]

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

c) Recall adalah mengukur sensitivitas dari model, yaitu seberapa baik model dalam mendeteksi kelas positif.[5]

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

d) F1-Score adalah perhitungan rata-rata harmonis dari precision dan recall, yang berguna jika terdapat ketidakseimbangan kelas.[5]

$$\textbf{F1 Score} = \textbf{2} \times \frac{\textbf{Precision} \times \textbf{Recall}}{\textbf{Precision} + \textbf{Recall}}$$

2.4 Tools Yang Digunakan Untuk Mengolah Data

a. Google Colaboratory

Colaboratory, atau yang lebih dikenal dengan sebutan "Colab," adalah sebuah produk dari Google Research. Colab memungkinkan siapa saja untuk menulis dan menjalankan kode Python secara langsung melalui browser. Alat ini sangat ideal untuk keperluan machine learning, analisis data, dan juga sebagai sarana pendidikan.[35]. Google Colab adalah layanan gratis yang disediakan oleh Google, dirancang untuk memudahkan pengguna dalam memanfaatkan Python dan Pandas. Salah satu keunggulannya adalah pengguna tidak perlu menginstal Python atau Pandas di komputer mereka, serta tidak memerlukan editor yang diinstal. Cukup dengan memiliki akun Google yang aktif, Anda sudah dapat memanfaatkan semua fitur yang ditawarkan.[36].

2.3 Bahasa Yang Digunakan Untuk Mengolah Data

a. Python

Python adalah salah satu bahasa pemrograman tingkat tinggi yang terkenal dengan sifatnya yang interpretatif, interaktif, dan berorientasi objek. Keunggulannya terletak pada kemampuannya untuk beroperasi di berbagai platform, termasuk Linux, Windows, Mac, dan lainnya. Salah satu alasan Python sangat populer adalah kemudahan dalam mempelajarinya, berkat sintaks yang jelas dan elegan. Selain itu, Python dilengkapi dengan berbagai modul yang memiliki struktur data tingkat tinggi, efisien, dan siap digunakan. Kode sumber aplikasi yang ditulis dalam Python biasanya dikompilasi menjadi format perantara yang disebut byte code, yang kemudian dieksekusi untuk menjalankan program [37].