NUSANTARA JOURNAL OF SCIENCE AND TECHNOLOGY (NU-JST)

E-ISSN: 3048-3859
Vol. 1, No. 2, November 2024, hal. 38-47
Tersedia daring pada https://jurnal.unukase.ac.id/nujst
Universitas Nahdlatul Ulama Kalimantan Selatan

Exploratory Data Analysis dan Machine Learning dalam Memprediksi Employee Attrition

Exploratory Data Analysis and Machine Learning in Predicting Employee Attrition

Binti Kholifah*[1], Fendy Bayu Firmansyah, Nafis Sururi[3], dan Danang Satya Nugraha[4]

[1]Universitas Negeri Surabaya, Jl. Ketintang, Ketintang, Kec. Gayungan, Surabaya, Jawa Timur
[2,3,4]Institut Teknologi Mojosari, Dsn. Mojosari, Ds. Ngepeh, Kec. Loceret, Kab. Nganjuk, Jawa Timur
[1]bintikholifah@unesa.ac.id, [2]fendy@itmnganjuk.ac.id, [3]nafissururi@itmnganjuk.ac.id, [4]danangsatyan@itmnganjuk.ac.id

Format Kutipan: Kholifah, B., Firmansyah, F.B., Sururi, N., & Nugraha, D.S. (2024). Exploratory Data Analysis dan Machine Learning dalam Memprediksi Employee Attrition. *Nusantara Journal of Science and Technology*, 1(2), hal. 38-47. https://doi.org/10.69959/nujst.v1i2.118

RIWAYAT ARTIKEL
Dikirim: 18 November 2024
Revisi Akhir: 29 November 2024
Diterbitkan: 30 November 2024
Tersedia Daring Sejak: 30 November 2024

KATA KUNCI
Employee Attrition
Exploratory Data Analysis
Logistic Regression
Support Vector Machine
Naive Bayes

KEYWORDS
Employee Attrition
Exploratory Data Analysis
Logistic Regression
Support Vector Machine
Naive Bayes

ABSTRAK

Employee attrition merupakan salah satu tantangan utama yang dihadapi oleh organisasi dalam mempertahankan sumber daya manusia yang kompeten. Penelitian ini bertujuan untuk mengeksplorasi pola data dan memprediksi employee attrition menggunakan pendekatan *Exploratory Data Analysis* (EDA) dan algoritma *Machine Learning* seperti Logistic Regression, Support Vector Machine (SVM), dan Naive Bayes. Analisis dilakukan pada dataset yang mencakup berbagai faktor seperti demografi, kepuasan kerja, dan performa karyawan. Hasil penelitian menunjukkan bahwa Logistic Regression mencapai akurasi sebesar 87%, namun model ini memiliki kelemahan dalam mendeteksi kelas positif (attrition) yang tercermin dari rendahnya nilai recall. SVM dengan akurasi 85% memberikan precision tinggi untuk kelas positif, tetapi memiliki performa yang kurang baik dalam mendeteksi kasus attrition yang sebenarnya. Sebaliknya, Naive Bayes dengan akurasi 85% menunjukkan kinerja yang lebih seimbang dengan F1-score rata-rata tertimbang yang lebih tinggi dibandingkan kedua model lainnya, meskipun masih terdapat ruang untuk peningkatan, khususnya pada prediksi kelas positif. Dari hasil penelitian, Naive Bayes menjadi model yang lebih andal untuk memprediksi employee attrition dengan performa yang lebih seimbang dibandingkan Logistic Regression dan SVM. Untuk meningkatkan performa prediksi, disarankan untuk mengatasi masalah ketidakseimbangan kelas pada dataset melalui teknik oversampling, undersampling, penggunaan *class weighting*, atau algoritma khusus.

ABSTRACT

Employee attrition is one of the main challenges faced by organizations in retaining competent human resources. This study aims to explore data patterns and predict employee attrition using the Exploratory Data Analysis (EDA) approach and Machine Learning algorithms such as Logistic Regression, Support Vector Machine (SVM), and Naive Bayes. The analysis was conducted on a dataset that includes various factors such as demographics, job satisfaction, and employee performance. The research findings indicate that Logistic Regression achieved an accuracy of 87%, but the model has weaknesses in detecting the positive class (attrition), as reflected by its low recall score. SVM, with an accuracy of 85%, provided high precision for the positive class but performed poorly in detecting actual attrition cases. Conversely, Naive Bayes, with an accuracy of 85%, demonstrated more balanced performance with a higher weighted average F1-score compared to the other models, although there is still room for improvement, particularly in predicting the positive class. Based on the results, Naive Bayes stands out as a more reliable model for predicting employee attrition with more balanced performance compared to Logistic Regression and SVM. To enhance prediction performance, it is recommended to address the class imbalance in the dataset through techniques such as oversampling, undersampling, class weighting, or specialized algorithms.

PENDAHULUAN

Employee attrition atau tingkat pengurangan jumlah karyawan merupakan salah satu tantangan utama yang dihadapi oleh perusahaan di berbagai industri. Tingginya tingkat attrition dapat berdampak pada peningkatan biaya rekrutmen, pelatihan, dan hilangnya pengetahuan organisasi. Pada dasarnya, suatu bisnis/ perusahaan tidak luput dari adanya Sumber Daya Manusia. Sumber Daya Manusia dapat disebut sebagai pemeran utama dalam sebuah organisasi. Sumber Daya Manusia berperan besar dalam menentukan baik tidaknya keberlanjutan

suatu perusahaan mencapai tujuan dan mencapai keberhasilan pelaksanaan tujuan organisasi (Huzain, 2021). Sumber Daya Manusia dalam hal ini karyawan yang dapat memberikan kontribusi maksimal, merupakan impian semua perusahaan. Seorang karyawan yang memberikan kemampuan terbaiknya untuk perusahaan, seyogyanya diberikan timbal balik oleh perusahaan. Hal tersebut membuat perusahaan berupaya makmisal untuk membuat Sumber Daya Manusianya loyal atau tidak berpindah ke perusahaan lain.

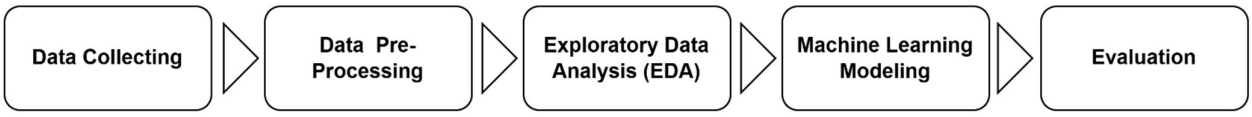
Karyawan akan menunjukkan kemampuan terbaiknya, apabila memiliki kontribusi yang baik sehingga dapat menuju kepada arah tercapainya tujuan suatu perusahaan. Kinerja karyawan sebagai suatu tindakan, perilaku, dan hasil yang dapat diukur sehingga karyawan terikat dengan tujuan organisasi dan berkontribusi pada tujuan perusahaan (Liana & Megantoro, 2023). Faktor Sumber Daya Manusia memegang peranan penting dalam berjalannya suatu perusahaan. Apabila suatu perusahaan sering mengalami keluar masuknya karyawan, maka akan banyak sumber daya lain yang perlu dikeluarkan guna memberikan pelatihan untuk karyawan baru. Dengan demikian, perusahaan harus mampu menciptkan lingkungan kerja yang nyaman guna menumbuhkan kepuasan maupun komitmen karyawan.

Perpindahan karyawan di era saat ini semakin banyak ditemui mengingat banyak perusahaan belum bisa menjamin kenyamanan karyawan. Hal tersebut diperkuat dengan adanya UU Cipta Kerja yang memungkinkan suatu perusahaan hanya memberikan kontrak secara berkala kepada karyawan, tanpa ada kejelasan status pasti. Hal serupa juga memungkinkan karyawan dengan masa kerja yang sudah lama mengalami restrukturisasi atau Pemutusan Hubungan Kerja (PHK). Pemutusan hubungan kerja oleh suatu perusahaan memiliki dampak sosial yang jelas mengarah pada konflik antara perusahaan dengan karyawan. Semenjak diberlakukannya Perppu Cipta Kerja terdapat beberapa masalah karena semenjak dikeluarkannya Perppu tersebut, masyarakat mengalami kerugian berupa ketidakpastian hukum (Annisa, 2023). Salah satunya adalah penggunaan ketidakpastian status kepegawaian. Hal tersebut semakin memperkuat adanya perpindahan karyawan dari satu perusahaan ke perusahaan yang lain. Kondisi ekonomi yang ada saat ini juga mempengaruhi kinerja perusahaan, kemudian berdampak pada kesejahteraan karyawannya. Karyawan yang tidak mendapatkan kesejahteraan berpotensi keluar dari tempat kerjanya, untuk mencari peluang di perusahaan lain. Hal tersebut otamatis berdampak pada laju usaha suatu perusahaan. Sehingga penting memberikan jaminan kenyamanan, yang dapat berupa status kepegawaian, peningkatan gaji, insentif, kepastian tunjangan hari tua, fasilitas dan lain lain kepada karyawan guna meningkatkan loyalitasnya pada perusahaan.

Dengan kemajuan teknologi dalam analisis data dan kecerdasan buatan, pendekatan berbasis data seperti Exploratory Data Analysis (EDA) dan Machine Learning dapat memberikan peluang baru untuk mengidentifikasi pola, tren, dan faktor-faktor utama yang memengaruhi employee attrition. Melalui EDA, data karyawan dapat dianalisis untuk memberikan wawasan mendalam mengenai karakteristik demografis, performa kerja, lingkungan kerja, dan variabel lain yang berkontribusi terhadap keputusan karyawan untuk meninggalkan perusahaan (Alsaadi et al., 2022). Penelitian ini bertujuan untuk memanfaatkan pendekatan EDA dalam memahami karakteristik data employee attrition serta membangun model prediktif menggunakan algoritma Machine Learning. Dengan demikian, diharapkan penelitian ini dapat memberikan kontribusi pada pengembangan strategi retensi karyawan yang lebih efektif sekaligus mengurangi dampak negatif dari tingginya tingkat attrition.

METODE PENELITIAN

Jenis penelitian ini adalah penelitian kuantitatif dengan pendekatan eksploratif dan prediktif. Penelitian ini bertujuan untuk mengeksplorasi pola-pola pada data attrition karyawan dan membangun model machine learning untuk memprediksi kemungkinan attrition di masa mendatang. Proses penelitian dilakukan secara daring menggunakan open dataset HR Employee Attrition. Penelitian ini ditujukan untuk membantu perusahaan dalam memahami faktor-faktor yang mempengaruhi keputusan karyawan untuk keluar dari perusahaan. Prosedur dalam penelitian ini dapat dilihat pada Gambar.



Gambar 1. Prosedur Penelitian

Data Collecting

Data collecting adalah langkah awal dalam penelitian ini yang bertujuan untuk mendapatkan informasi yang akan dianalisis dan digunakan dalam membangun model prediktif. Data yang digunakan diunduh dari repositori publik yang menyediakan dataset HR Employee Attrition. Repositori dipilih berdasarkan ketersediaan metadata yang mencakup penjelasan tentang struktur data, variabel, dan konteks pengumpulan data.

Data Pre-Processing

Prapemrosesan data merupakan tahap penting untuk memastikan bahwa data yang digunakan memiliki kualitas yang baik dan representatif. Tahapan ini melibatkan beberapa langkah yang dilakukan secara sistematis untuk membersihkan, mengubah, dan mengoptimalkan data sebelum dianalisis lebih lanjut. Proses ini tidak hanya membantu dalam meningkatkan akurasi model, tetapi juga meminimalkan potensi bias dalam hasil penelitian (Park et al., 2024). Tahapan ini terdiri dari *data cleaning* dan penanganan *outlier*. Pada tahap *data cleaning* dilakukan penghapusan kolom atau atribut yang tidak diperlukan, yang dapat mengurangi kompleksitas dan memfokuskan analisis pada informasi yang relevan. Selain itu, untuk meningkatkan konsistensi, dilakukan perubahan pada nama kolom. Tahap selanjutnya yaitu penanganan *outlier*. *Outlier* merupakan objek data yang menyimpang secara signifikan dari objek data lainnya. Deteksi outlier tersebut penting untuk banyak aplikasi. Pemangkasan *outlier* dapat meningkatkan kinerja dari model (Satapathy et al., 2015).

Exploratory Data Analysis

EDA merupakan pendekatan analisis data yang mengutamakan pola pikir terbuka, kreativitas, serta berbagai sudut pandang. EDA memiliki tujuan untuk menjelajahi data secara menyeluruh tanpa terpaku pada asumsi atau model yang telah ada, hingga terbentuk narasi yang jelas dan terpadu. Pendekatan ini dapat digunakan untuk merumuskan hipotesis baru, mengenali pola dan anomali, serta mengungkap struktur dan keterkaitan mendalam dalam data (Simangunsong et al., 2024). Analisis data dalam manajemen sumber daya manusia memberikan manfaat seperti peningkatan pengambilan keputusan, optimasi rekrutmen, evaluasi kinerja, identifikasi bakat, dan prediksi tren strategis, namun juga menghadirkan tantangan seperti perlindungan data pribadi, risiko diskriminasi, dan keamanan data (Nowicka et al., 2024).

Machine Learning Modeling

Pengembangan model machine learning melibatkan beberapa tahap yang bertujuan untuk menghasilkan model prediktif yang andal. Pertama, dilakukan pemilihan algoritma yang sesuai berdasarkan sifat data dan tujuan penelitian. Dalam penelitian ini, dibandingkan 10 lgoritma klasifikasi dan dipilih 3 metode dengan performa terbaik berdasar akurasi dan ROC AUC. Model-model yang dipilih adalah :

1) Logistic Regression

Logistic Regression adalah metode statistik yang digunakan untuk menganalisis dataset dan menghasilkan hasil biner, di mana hasilnya ditentukan oleh satu atau lebih variabel independen yang bersifat dikotomis, artinya hanya ada dua kemungkinan hasil. Bentuk regresi spesifik digunakan untuk memprediksi output biner dan kategorikal. Metode ini membantu mengatur dampak dari beberapa variabel independen secara bersamaan dan memprediksi salah satu dari dua kategori. *Logistic Regression* menggunakan metode likelihood maksimum untuk merancang fungsi terbaik yang memaksimalkan probabilitas mengklasifikasikan data ke dalam kategori yang benar. Metode ini diterapkan di berbagai bidang, termasuk meramalkan tren pasar, menganalisis tingkat keberhasilan dan kegagalan, perekrutan karyawan, kategorisasi gambar, dan perawatan kesehatan (Celine et al., 2020).

2) Support Vector Machine

Support Vector Machine (SVM) merupakan sistem pembelajaran terarah yang menggunakan ruang hipotesis dalam bentuk fungsi linear di ruang fitur berdimensi tinggi lalu dilatih menggunakan algoritma pembelajaran berdasarkan teori optimisasi dengan menerapkan bias pembelajaran. SVM dikembangkan untuk menyelesaikan masalah klasifikasi karena memiliki kemampuan yang lebih baik dalam menggeneralisasi data. Secara dasar, SVM menggunakan prinsip linear, namun telah berkembang untuk dapat bekerja pada masalah non-linear dengan memasukkan konsep kernel dalam ruang berdimensi tinggi (Ovirianti et al., 2022).

3) Naïve Bayes

Algoritma *Naive Bayes* adalah salah satu metode dalam teknik klasifikasi yang menggunakan probabilitas dan statistik. Algoritma ini didasarkan pada konsep yang dikemukakan oleh ilmuwan Inggris, Thomas Bayes, yang memprediksi kemungkinan kejadian di masa depan berdasarkan pengalaman sebelumnya, yang dikenal dengan Teorema Bayes. Teorema tersebut digabungkan dengan asumsi Naive, yang menganggap bahwa atribut-atribut dalam data bersifat independen satu sama lain. Klasifikasi *Naive Bayes* menganggap bahwa keberadaan atau ketidakhadiran suatu fitur dalam suatu kelas tidak berhubungan dengan karakteristik kelas lainnya (Zogara, 2024).

Semua fitur numerik diubah ke skala yang seragam menggunakan teknik *StandardScaler* untuk memastikan bahwa model tidak dipengaruhi oleh skala fitur yang berbeda. Standarisasi dilakukan dengan mengurangi rata-rata dan membagi data dengan standar deviasi masing-masing fitur. Dataset dibagi menjadi dua bagian utama yaitu data pelatihan (70%) dan data uji (30%). Data pelatihan digunakan untuk melatih model, sementara data uji digunakan untuk mengevaluasi performa model. Proses pelatihan mencakup pengoptimalan parameter model dengan teknik seperti grid search atau random search untuk meningkatkan akurasi prediksi.

Evaluation

Setelah model dilatih, evaluasi dilakukan menggunakan berbagai metrik seperti akurasi, precision, recall, dan F1-score. Metrik-metrik ini memberikan gambaran tentang seberapa baik model dalam mengklasifikasikan data dengan benar, termasuk kemampuan dalam menangkap kasus-kasus minoritas (attrition) dan meminimalkan kesalahan prediksi. Proses evaluasi dilakukan dengan mempertimbangkan hasil dari semua metrik tersebut sehingga dapat memastikan bahwa model yang dikembangkan memiliki performa yang stabil dan andal dalam skenario nyata.

HASIL PENELITIAN DAN PEMBAHASAN

Data Collecting

Data yang digunakan merupakan open dataset yaitu HR Employee Attrition. Terdapat 35 variabel dan 1.470 data karyawan. Tujuan dari pengumpulan data ini adalah untuk mengembangkan model dalam memprediksi karyawan yang kemungkinan akan berhenti berdasarkan informasi dari 1.470 data karyawan. Dataset terdiri dari variabel-variabel seperti umur, jenis kelamin, tingkat pendidikan, jabatan, pengalaman kerja, gaji, dan informasi lain yang relevan. Informasi mengenai atribut dataset dapat dilihat pada Tabel 1.

Tabel 1. Informasi Atribut Dataset

Atribut	Deskripsi
Age	Lama bekerja
Attrition	Status
BusinessTravel	Frekuensi perjalanan pekerjaan
DailyRate	Tarif harian

Atribut	Deskripsi
Department	Bagian tempat karyawan
DistanceFromHome	Jarak perjalanan dari rumah ke tempat kerja
Education	Tingkat pendidikan
EducationField	Bidang pendidikan
EmployeeCount	Jumlah karyawan
EmployeeNumber	Nomor karyawan
EnvironmentSatisfaction	Kepuasan lingkungan
Gender	Gender
HourlyRate	Tarif per jam
JobInvolvement	Tingkat keterlibatan
JobLevel	Level pekerjaan
JobRole	Peran pekerjaan
JobSatisfaction	Kepuasan pekerjaan
MaritalStatus	Status pernikahan
MonthlyIncome	Pemasukan bulanan
MonthlyRate	Tarif bulanan
NumCompaniesWorked	Jumlah perusahaan tempat karyawan pernah bekerja sebelumnya
Over18	Apakah usia karyawan lebih dari 18 tahun atau tidak
OverTime	Apakah karyawan bekerja lembur
PercentSalaryHike	Persentase kenaikan gaji
PerformanceRating	Peringkat kinerja
RelationshipSatisfaction	Kepuasan hubungan kerja
StandardHours	Jam kerja standar
StockOptionLevel	Tingkat opsi saham
TotalWorkingYears	Jumlah jam kerja
TrainingTimesLastYear	Jumlah tahun kerja
WorkLifeBalance	Tingkat keseimbangan kehidupan kerja
YearsAtCompany	Tahun masuk di perusahaan
YearsInCurrentRole	Tahun pada peran saat ini
YearsSinceLastPromotion	Tahun saat promosi terakhir
YearsWithCurrManager	Tahun dengan manajer saat ini

Data Pre-Processing

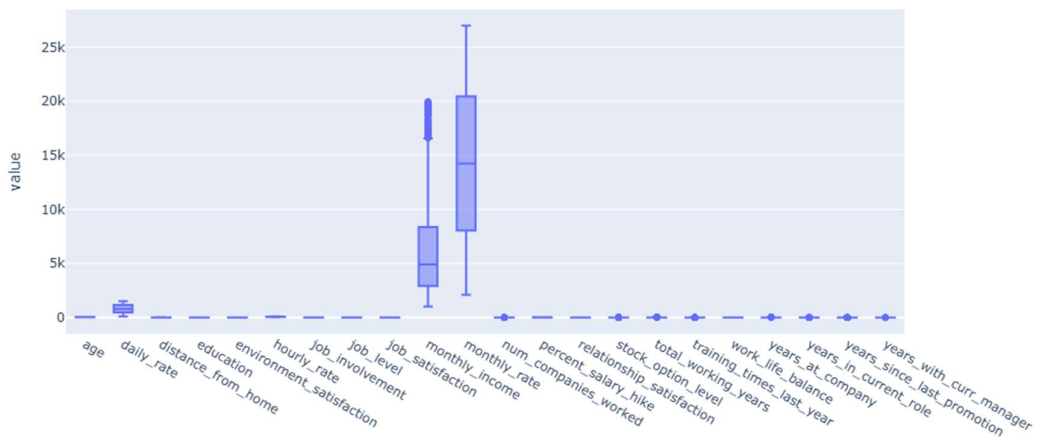
Prapemrosesan data merupakan tahap penting untuk memastikan bahwa data yang digunakan memiliki kualitas yang baik dan representatif. Tahapan ini melibatkan beberapa langkah yang dilakukan secara sistematis untuk membersihkan, mengubah, dan mengoptimalkan data sebelum dianalisis lebih lanjut. Proses ini tidak hanya membantu dalam meningkatkan akurasi model, tetapi juga meminimalkan potensi bias dalam hasil penelitian (Park et al., 2024).

1) *Data cleaning*

Pada tahap *data cleaning* dilakukan penghapusan kolom atau atribut yang tidak diperlukan, yang dapat mengurangi kompleksitas dan memfokuskan analisis pada informasi yang relevan. Pada tahap ini, kolom "*employee_count*", "*over18*" & "*standard_hours*" dihapus, karena kolom hanya terdiri dari 1 nilai untuk semua entri, sehingga tidak memiliki pengaruh besar terhadap proses analisis dan prediksi. Selain itu, untuk meningkatkan konsistensi, nama kolom dapat diubah menjadi huruf kecil semua dan mengganti pemisah antar suku kata dengan tanda *underscore* (misalnya, "*HourlyRate*" menjadi "*hourly_rate*").

2) Penanganan *outliers*

Outlier merupakan objek data yang menyimpang secara signifikan dari objek data lainnya. Deteksi *outlier* tersebut penting untuk banyak aplikasi. Pemangkasan *outlier* dapat meningkatkan kinerja dari model (Satapathy et al., 2015). Terdapat *outlier* dalam jumlah kecil (<5%) di beberapa kolom atau nilai outlier berada dalam kisaran yang realistis dan wajar, kecuali outlier pada kolom "*monthly_income*". hasil deteksi outlier dapat dilihat pada Gambar 2.

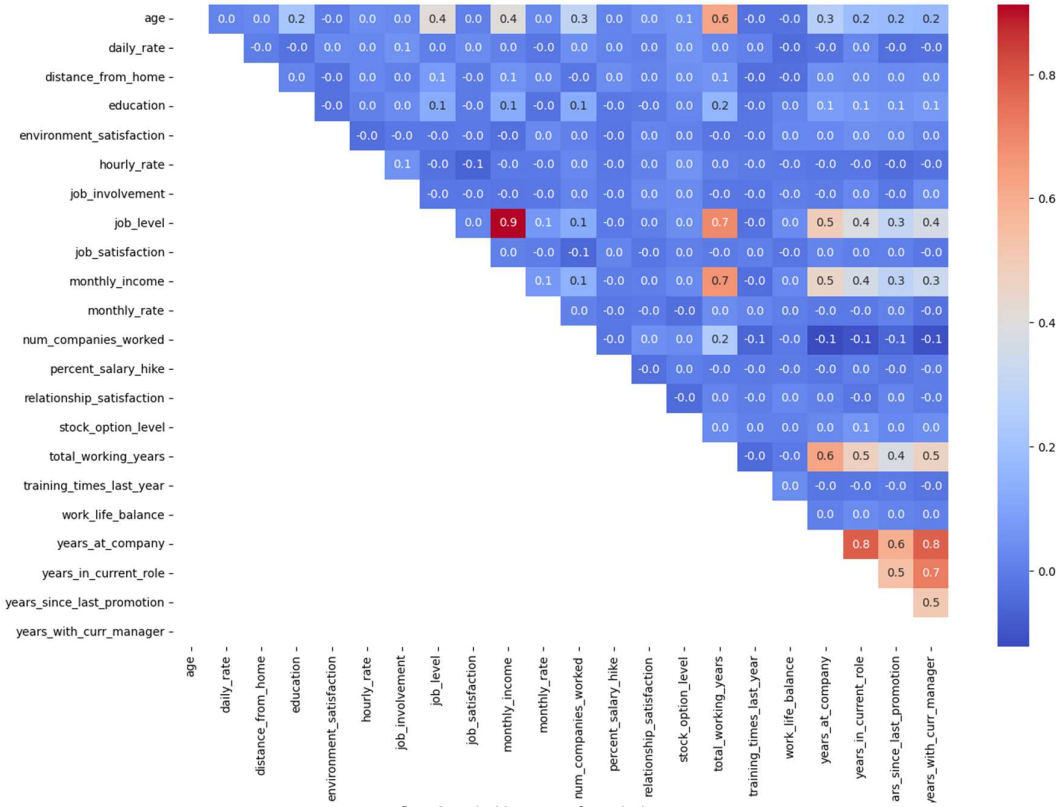


Gambar 2. Deteksi *Outlier*

Selanjutnya dilakukan penghapusan beberapa data outlier sesuai pada kolom "*monthly_income*". Jumlah data awal yaitu 1.470 data dan jumlah data setelah melalui tahap penanganan outlier adalah 1.356 data.

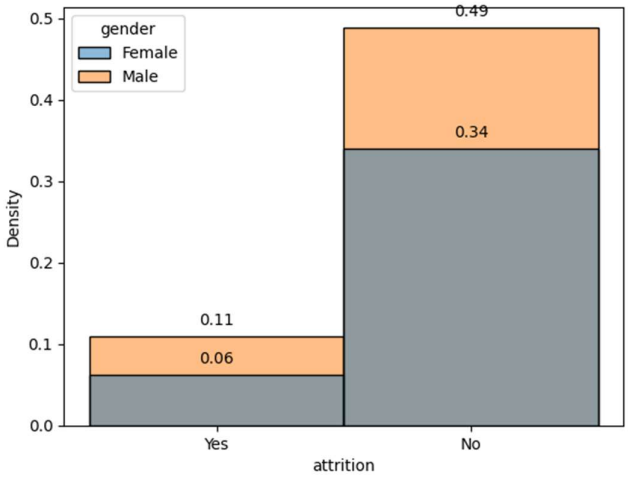
Exploratory Data Analysis

Pada bagian ini, pertama dilakukan *heatmap analysis* untuk menentukan hubungan atau korelasi antara setiap kolom pada dataset. Adapun *heatmap* dapat dilihat pada Gambar 3.



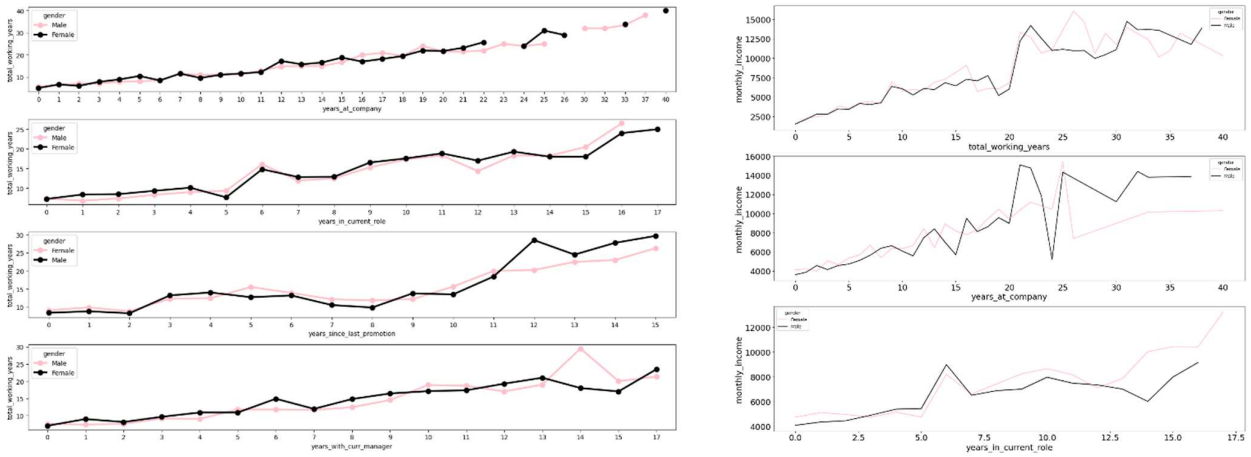
Gambar 3. *Heatmap Correlation*

Berdasarkan analisis menggunakan heatmap, hasil menunjukkan adanya hubungan yang signifikan dengan nilai koefisien korelasi yang kuat antara variabel-variabel berikut: *education, age, job_level, monthly_income, num_companies_worked, years_at_company, years_in_current_role, years_since_last_promotion and years_with_curr_manager columns* (tingkat pendidikan, usia, tingkat pekerjaan, pendapatan bulanan, jumlah perusahaan tempat bekerja sebelumnya, durasi bekerja di perusahaan saat ini, masa kerja dalam jabatan saat ini, waktu sejak promosi terakhir, dan durasi kerja bersama manajer saat ini).



Gambar 4. Distribusi Gender

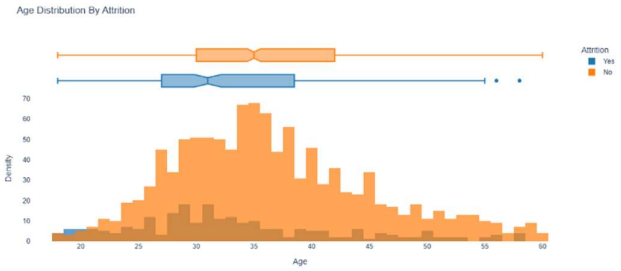
Dari seluruh data, terdapat 59,8% laki-laki dan 40,2% perempuan. Berdasar Gambar 4, kedua kelompok gender memiliki distribusi usia yang sama, yaitu berkisar antara 18 - 60 tahun. Tingkat attrition antara karyawan laki-laki dan perempuan juga tetap sama yaitu 15%. Hal ini membuktikan bahwa data tersebut tidak bias.



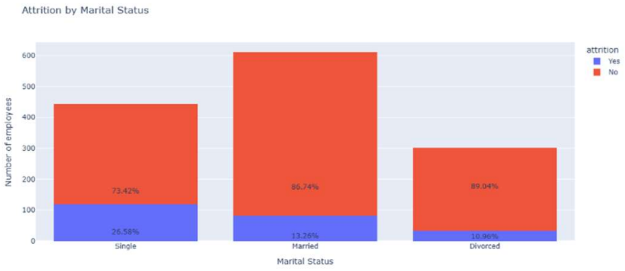
(a) (b)

Gambar 5. Korelasi Lama Waktu Bekerja dengan Beberapa Variabel

Grafik pada Gambar 5(a) menunjukkan bahwa lama bekerja karyawan (sumbu y) memiliki korelasi erat dengan lama bekerja karyawan di perusahaan dan posisi saat ini (sumbu x) baik pria maupun wanita. Korelasi erat antara lama bekerja karyawan dengan lama bekerja di perusahaan dan posisi saat ini menunjukkan bahwa peningkatan dalam satu variabel diikuti dengan peningkatan pada variabel lainnya. Grafik pada Gambar 6(b) menunjukkan bahwa terdapat korelasi linier antara pendapatan bulanan dengan total lama kerja dan lama bekerja karyawan di perusahaan, tetapi hanya dari 0 - 20 tahun.Pendapatan melonjak pada tanda 20 tahun kerja atau lebih dan berfluktuasi secara fluktuatif sejak saat itu. Namun, data membuktikan bahwa semakin banyak seseorang bekerja dalam 1 peran, semakin banyak pula penghasilannya setiap bulan.

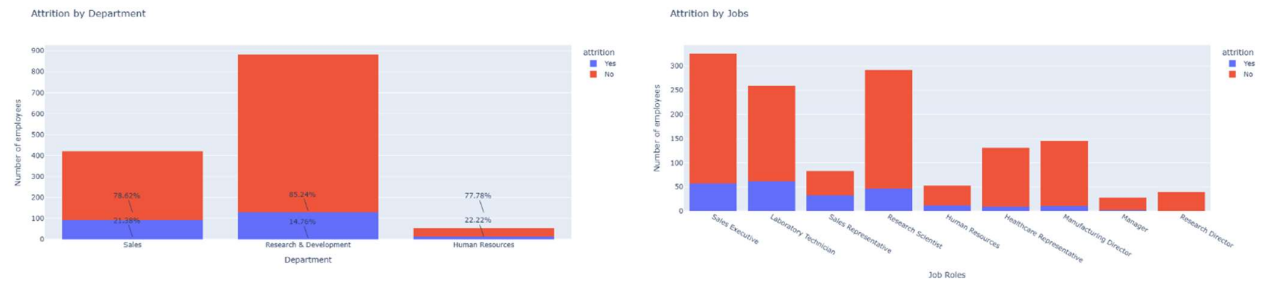


Gambar 6. Korelasi *Attrition* dengan Usia



Gambar 7. Korelasi *Attrition* dengan Status Pernikahan

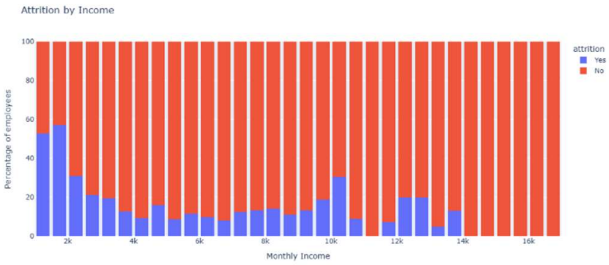
Gambar 6 menunjukkan bahwa karyawan dengan rentang usia 25 hingga 35 tahun memiliki tingkat pergantian karyawan yang paling menonjol serta menunjukkan kecenderungan yang lebih tinggi untuk keluar secara sukarela di antara semua kelompok usia. Sebaliknya, terjadi tren penurunan pada karyawan dari usia sekitar 40 tahun ke atas, yang menunjukkan kemungkinan lebih kecil untuk keluar. Hal ini juga terjadi pada karyawan dari usia 25 tahun ke bawah, yang kemungkinan besar dapat dikaitkan dengan masa percobaan atau rentang waktu perkenalan bagi perusahaan. Sementara Gambar 7 menunjukkan korelasi pergantian karyawan dengan status pernikahan. Status pernikahan karyawan terbagi antara menikah, lajang, dan bercerai. Grafik menunjukkan tidak ada pengaruh yang erat antara status perkawinan dengan pergantian karyawan.



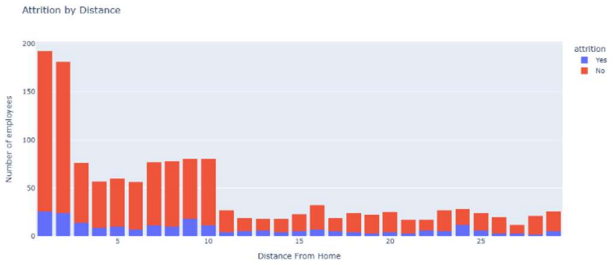
(a) (b)

Gambar 8. Korelasi *Attrition* dengan Penempatan Kerja (1)

Gambar 8(a) menunjukkan hubungan *attrition* antar departemen. Tingkat pergantian karyawan konsisten antar departemen berkisar pada level 80 - 85%. Gambar 8(b) menunjukkan hubungan *attrition* dengan peran di tempat kerja. Tingkat pergantian karyawan secara khusus lebih rendah pada posisi Direktur Riset dan Manufaktur, Manajer, SDM, dan perwakilan Layanan Kesehatan.



Gambar 9. Korelasi *Attrition* dengan Pendapatan



Gambar 10. Korelasi *Attrition* dengan Jarak Tempat Kerja

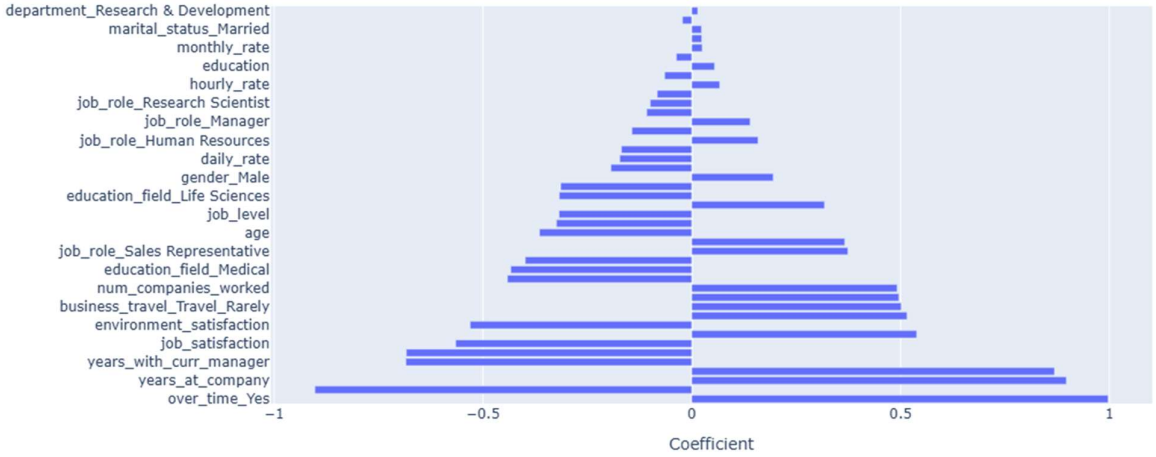
Gambar 9 menunjukkan hubungan *attrition* dengan pendapatan karyawan. Grafik menunjukkan bahwa tingkat pergantian karyawan lebih tinggi pada karyawan yang memiliki pendapatan lebih rendah dari $2000 hingga $4000. Gambar 10 menunjukkan hubungan *attrition* dengan jarak rumah ke tempat kerja karyawan. Mayoritas karyawan memiliki tempat tinggal dalam jarak 10 km dari tempat kerja. Namun, tidak ada bukti yang mendukung bahwa tingkat pergantian karyawan lebih tinggi pada orang yang tinggal jauh dari tempat kerja. Hal ini dapat dilihat dari Gambar 10.

Machine Learning Evaluation

Penelitian ini membandingkan 3 metode klasifikasi yaitu Logistic Regression, Support Vector Machine (SVM), dan Naïve Bayes. Penilaian dilakukan menggunakan confussion matriks untuk mengukur keandalan dan kemampuan model membedakan kelas.

1) *Logistic Regression*

Logistic Regression tidak hanya menghasilkan prediksi, tetapi juga memberikan interpretasi terhadap pengaruh masing-masing atribut melalui koefisien model. Koefisien ini menunjukkan hubungan antara setiap fitur dengan probabilitas terjadinya attrition. Nilai positif menunjukkan bahwa peningkatan pada atribut meningkatkan kemungkinan karyawan mengalami attrition, sedangkan nilai negatif menunjukkan bahwa atribut menurunkan kemungkinan attrition. Gambar 12 menunjukkan pengaruh masing-masing atribut terhadap model *Logistic Regression*.



Gambar 11. *Feature Coefficients* pada *Logistic Regression*

Gambar 13 menunjukkan *Confussion Matrix* berdasar hasil prediksi dari model *Logistic Regression*. *Precision False* (Tidak *Attrition*) 0.90 menunjukkan bahwa model dengan benar mengidentifikasi 90% karyawan yang tidak mengalami *attrition* dari seluruh prediksi "*False*" yang dibuat. *PrecisionTrue* (*Attrition*) 0.62 menunjukkan bahwa 62% dari prediksi karyawan yang diprediksi mengalami *attrition* benar-benar mengalami *attrition*. Angka ini lebih rendah, yang berarti terdapat banyak *false positives* (karyawan yang diprediksi *attrition* padahal tidak).

Recall False (Tidak *Attrition*) 0.95 menunjukkan bahwa model berhasil menangkap 95% karyawan yang sebenarnya tidak mengalami *attrition*. Model sangat baik dalam mengidentifikasi kelas ini. *Recall True* (*Attrition*) 0.46 menunjukkan bahwa hanya 46% dari karyawan yang benar-benar mengalami *attrition* berhasil diprediksi oleh model. Artinya, banyak *false negatives* (karyawan yang sebenarnya *attrition* tetapi tidak terdeteksi). F1-Score False (Tidak *Attrition*) 0.92 mencerminkan keseimbangan antara *precision* dan *recall* untuk kelas *False*. Model sangat baik dalam mengidentifikasi karyawan yang tidak mengalami *attrition*. *F1-Score True* (*Attrition*) 0.53 menunjukkan bahwa model tidak seimbang dalam memprediksi karyawan yang mengalami *attrition*. Ini menunjukkan bahwa model lebih cenderung untuk mengklasifikasikan karyawan sebagai *False* meskipun mereka mungkin akan meninggalkan perusahaan.

	precision	recall	f1-score	support
False	0.90	0.95	0.92	342
True	0.62	0.46	0.53	65
accuracy			0.87	407
macro avg	0.76	0.70	0.73	407
weighted avg	0.86	0.87	0.86	407

Gambar 12. *Confussion Matrix* dari Model *Logistic Regression*

Akurasi model sebesar 0.87 menunjukkan bahwa 87% dari prediksi model (baik untuk *True* maupun *False*) adalah benar. Meskipun angka ini terlihat tinggi, imbalance class (perbedaan jumlah antara kelas True dan False) membuat evaluasi lebih mendalam menggunakan metrik lain seperti *recall* dan *f1-score* lebih penting. *Macro average* menunjukkan nilai rata-rata dari *precision*, *recall*, dan *f1-score* untuk kedua kelas, tanpa mempertimbangkan ukuran kelas. Nilai 0.76 untuk *precision*, 0.70 untuk *recall*, dan 0.73 untuk *f1-score* memberikan gambaran umum tentang performa model yang seimbang di kedua kelas. *Weighted Average* menghitung rata-rata berdasarkan proporsi kelas dalam dataset. Dengan nilai 0.86 untuk *precision*, 0.87 untuk *recall*, dan 0.86 untuk *f1-score*, ini menunjukkan bahwa model memiliki performa yang sangat baik secara keseluruhan, meskipun ada ketidakseimbangan kelas. Model ini menunjukkan kesulitan dalam memprediksi kelas *True* (*Attrition*), yang lebih sedikit jumlahnya dalam dataset. Hal ini menyebabkan *recall* yang lebih rendah untuk kelas True.

2) Support Vector Machine

Gambar 14 menunjukkan *Confussion Matrix* berdasar hasil prediksi dari model *Support Vector Machine*. *Precision False* (Tidak *Attrition*) 0.85 menunjukkan bahwa model dengan benar mengidentifikasi 85% karyawan yang tidak mengalami attrition dari seluruh prediksi "*False*" yang dibuat. *PrecisionTrue* (*Attrition*) 0.92 menunjukkan bahwa 62% dari prediksi karyawan yang diprediksi mengalami *attrition* benar-benar mengalami *attrition*. Prediksi *true* maupun *false* menunjukkan hasil precision yang lebih seimbang dibanding *logistic regression*

Recall False (Tidak *Attrition*) 1.00 menunjukkan bahwa model berhasil mengidentidikasi semua karyawan yang sebenarnya tidak mengalami attrition. Model sangat baik dalam mengidentifikasi kelas ini. *Recall True* (*Attrition*) 0.17 menunjukkan bahwa hanya 17% dari karyawan yang benar-benar mengalami *attrition* berhasil diprediksi oleh model. Artinya, sangat banyak *false negatives* (karyawan yang sebenarnya *attrition* tetapi tidak terdeteksi). *F1-Score False* (Tidak *Attrition*) 0.92 mencerminkan keseimbangan antara *precision* dan *recall* untuk kelas *False*. Model sangat baik dalam mengidentifikasi karyawan yang tidak mengalami *attrition*. *F1-Score True* (*Attrition*) 0.29 menunjukkan bahwa model tidak seimbang dalam memprediksi karyawan yang mengalami *attrition*. Ini menunjukkan bahwa model lebih cenderung untuk mengklasifikasikan karyawan sebagai *False* meskipun mereka mungkin akan meninggalkan perusahaan.

	precision	recall	f1-score	support
0	0.85	1.00	0.92	336
1	0.92	0.17	0.29	71
accuracy			0.85	407
macro avg	0.89	0.58	0.60	407
weighted avg	0.86	0.85	0.81	407

Gambar 13. *Confussion Matrix* dari Model *Support Vector Machine*

Akurasi model sebesar 0.85 menunjukkan bahwa 85% dari prediksi model (baik untuk *True* maupun *False*) adalah benar. Meskipun angka ini terlihat tinggi, imbalance class (perbedaan jumlah antara kelas *True* dan *False*) membuat evaluasi lebih mendalam menggunakan metrik lain seperti *recall* dan *f1-score* lebih penting. *Macro average* menunjukkan nilai rata-rata dari *precision*, *recall*, dan *f1-score* untuk kedua kelas, tanpa mempertimbangkan ukuran kelas. Nilai 0.89 untuk *precision*, 0.58 untuk *recall*, dan 0.60 untuk *f1-score* memberikan gambaran umum tentang performa model yang seimbang di kedua kelas. *Weighted Average* menghitung rata-rata berdasarkan proporsi kelas dalam dataset. Dengan nilai 0.86 untuk *precision*, 0.85 untuk *recall*, dan 0.81 untuk *f1-score*, ini menunjukkan bahwa model memiliki performa yang sangat baik secara keseluruhan, meskipun ada ketidakseimbangan kelas. Model ini menunjukkan kesulitan dalam memprediksi kelas *True* (*Attrition*), yang lebih sedikit jumlahnya dalam dataset. Hal ini menyebabkan *recall* yang lebih rendah untuk kelas True.

3) Naïve Bayes

Gambar 15 menunjukkan *Confussion Matrix* berdasar hasil prediksi dari model *Naïve Bayes*. *Precision False* (Tidak *Attrition*) 0.91 menunjukkan bahwa model dengan benar mengidentifikasi 91% karyawan yang tidak mengalami *attrition* dari seluruh prediksi "*False*" yang dibuat. *PrecisionTrue* (*Attrition*) 0.56 menunjukkan bahwa 56% dari prediksi karyawan yang diprediksi mengalami *attrition* benar-benar mengalami *attrition*. Angka ini lebih rendah, yang berarti terdapat banyak *false positives* (karyawan yang diprediksi *attrition* padahal tidak).

Recall False (Tidak *Attrition*) 0,90 menunjukkan bahwa model berhasil mengidentidikasi 90% karyawan yang sebenarnya tidak mengalami *attrition*. Model sangat baik dalam mengidentifikasi kelas ini. *Recall True* (*Attrition*) 0.59 menunjukkan bahwa hanya 59% dari karyawan yang benar-benar mengalami *attrition* berhasil diprediksi oleh model. Artinya, cukup banyak *false negatives* (karyawan yang sebenarnya *attrition* tetapi tidak terdeteksi). *F1-Score False* (Tidak *Attrition*) 0.91 mencerminkan keseimbangan antara *precision* dan *recall* untuk kelas *False*. Model sangat baik dalam mengidentifikasi karyawan yang tidak mengalami *attrition*. *F1-Score True* (*Attrition*) 0.58

menunjukkan bahwa model tidak seimbang dalam memprediksi karyawan yang mengalami *attrition*. Ini menunjukkan bahwa model lebih cenderung untuk mengklasifikasikan karyawan sebagai *False* meskipun mereka mungkin akan meninggalkan perusahaan..

	precision	recall	f1-score	support
0	0.91	0.90	0.91	336
1	0.56	0.59	0.58	71
accuracy			0.85	407
macro avg	0.74	0.75	0.74	407
weighted avg	0.85	0.85	0.85	407

Gambar 14. *Confussion Matrix* dari Model *Naïve Bayes*

Akurasi model sebesar 0.85 menunjukkan bahwa 85% dari prediksi model (baik untuk *True* maupun *False*) adalah benar. Meskipun angka ini terlihat tinggi, *imbalance class* (perbedaan jumlah antara kelas *True* dan *False*) membuat evaluasi lebih mendalam menggunakan metrik lain seperti *recall* dan *f1-score* lebih penting. *Macro average* menunjukkan nilai rata-rata dari *precision*, *recall*, dan *f1-score* untuk kedua kelas, tanpa mempertimbangkan ukuran kelas. Nilai 0.74 untuk *precision*, 0.75 untuk *recall*, dan 0.74 untuk *f1-score* memberikan gambaran umum tentang performa model yang seimbang di kedua kelas. *Weighted Average* menghitung rata-rata berdasarkan proporsi kelas dalam dataset. Dengan nilai 0.85 untuk *precision*, 0.85 untuk *recall*, dan 0.85 untuk *f1-score*, ini menunjukkan bahwa model memiliki performa yang sangat baik secara keseluruhan, meskipun ada ketidakseimbangan kelas. Model ini menunjukkan kesulitan dalam memprediksi kelas *True* (*Attrition*), yang lebih sedikit jumlahnya dalam dataset. Hal ini menyebabkan *recall* yang lebih rendah untuk kelas True.

Logistic Regression memiliki akurasi yang baik, tetapi lebih cenderung memprediksi kelas negatif, dengan *recall* dan *F1-score* yang lebih rendah untuk kelas positif. *Support Vector Machine* (SVM) memiliki *precision* yang sangat baik untuk kelas positif, tetapi *recall*-nya sangat rendah untuk kelas positif, sehingga kurang dapat diandalkan dalam memprediksi kelas positif. *Naive Bayes* menunjukkan kinerja yang lebih seimbang antara kedua kelas, dengan *precision* dan *recall* yang lebih merata, serta F1-score rata-rata tertimbang yang lebih tinggi dibandingkan kedua model lainnya.

SIMPULAN DAN SARAN

Simpulan

Berdasarkan hasil analisis dan penerapan metode *Machine Learning* pada prediksi *Employee Attrition*, dapat disimpulkan bahwa setiap model yang diuji memiliki kekuatan dan kelemahan masing-masing. *Logistic Regression* menunjukkan akurasi yang baik dengan 87%, namun model ini cenderung lebih banyak memprediksi kelas negatif (tidak terjadi *attrition*) dengan *recall* yang lebih rendah pada kelas positif. *Support Vector Machine* (SVM) memiliki precision yang sangat baik untuk kelas positif, namun mengalami kesulitan dalam mendeteksi kasus attrition yang sebenarnya, yang tercermin pada rendahnya *recall* untuk kelas positif. Sementara itu, *Naive Bayes* menawarkan kinerja yang lebih seimbang antara kelas positif dan negatif, dengan *F1-score* rata-rata tertimbang yang lebih tinggi, meskipun *precision* dan *recall* untuk kelas positif masih dapat ditingkatkan. Secara keseluruhan, *Naive Bayes* memberikan hasil yang lebih seimbang dan dapat diandalkan dalam memprediksi attrition dibandingkan dengan kedua model lainnya. Meskipun demikian, perlu dilakukan peningkatan lebih lanjut pada setiap model untuk mendapatkan hasil yang lebih optimal, khususnya pada kelas positif, guna meminimalkan kehilangan karyawan yang berpotensi.

Saran

Logistic Regression, *Support Vector Machine*, *Naive Bayes* menunjukkan kesulitan dalam memprediksi kelas *True* (*Attrition*), yang lebih sedikit jumlahnya dalam dataset. Hal ini menyebabkan *recall* yang lebih rendah untuk kelas *True*. Strategi untuk mengatasi ketidakseimbangan kelas dapat mencakup:

1) Penggunaan teknik *oversampling* atau *undersampling* pada data pelatihan.
2) Penggunaan *class weighting* untuk memberikan bobot lebih pada kelas yang kurang terwakili.
3) Penerapan algoritma khusus yang lebih tahan terhadap ketidakseimbangan kelas seperti SMOTE (*Synthetic Minority Over-sampling Technique*).

DAFTAR PUSTAKA

Alsaadi, E. M. T. A., Khlebus, S. F., & Alabaichi, A. (2022). Identification of human resource analytics using machine learning algorithms. *Telkomnika (Telecommunication Computing Electronics and Control)*, *20*(5), 1004–1015. https://doi.org/10.12928/TELKOMNIKA.v20i5.21818

Annisa, O. C. N. (2023). Analisis Dampak Peraturan Pemerintah Pengganti Undang-Undang Cipta Kerja Terhadap Hak Pesangon Pemutusan Hubungan Kerja. *Journal Equitable*, *8*(1), 129–143. https://doi.org/10.37859/jeq.v8i1.4494

Celine, S., Dominic, M. M., & Devi, M. S. (2020). Logistic Regression for Employability Prediction. *International Journal of Innovative Technology and Exploring Engineering*, *9*(3), 2471–2478. https://doi.org/10.35940/ijitee.c8170.019320

Huzain, H. (2021). *Pengelolaan Sumber Daya Manuasia*. Universitas Islam Negeri Alauddin Makassar.

Liana, Y., & Megantoro, W. (2023). Loyalitas Karyawan dan Kompensasi Terhadap Kinerja Karyawan Pada CV. Lensa Informatika Surabaya. *INSPIRASI: JURNAL ILMU-ILMU SOSIAL*, *20*(1), 821–829.

Nowicka, J., Pauliuchuk, Y., Ciekanowski, Z., Falda, B., & Sikora, K. (2024). The Use of Data Analytics in Human Resource Management. *European Research Studies Journal*, *XXVII*(Issue 2), 203–215. https://doi.org/10.35808/ersj/3380

Ovirianti, N. H., Zarlis, M., & Mawengkang, H. (2022). Support Vector Machine Using A Classification Algorithm. *SinkrOn*, *7*(3), 2103–2107. https://doi.org/10.33395/sinkron.v7i3.11597

Park, H. J., Koo, Y. S., Yang, H. Y., Han, Y. S., & Nam, C. S. (2024). Study on Data Preprocessing for Machine Learning Based on Semiconductor Manufacturing Processes. *Sensors*, *24*(17), 1–14. https://doi.org/10.3390/s24175461

Satapathy, S. C., Govardhan, A., Raju, K. S., & Mandal, J. K. (2015). Improving Classification by Outlier Detection and Removal. *Advances in Intelligent Systems and Computing*, *338*, I–IV. https://doi.org/10.1007/978-3-319-13731-5

Simangunsong, J., Simanjuntak, M. S., & Simanjuntak, N. D. (2024). *Mental disorder classification with exploratory data analysis (EDA)*. *7*(3), 210–217.

Zogara, L. U. (2024). Detect Classification of Employees Tending to Move Work With the Naive Bayes Algorithm. *Jurnal Ilmiah Sistem Informasi (JISI)*.