Jurnal Tekno Insentif | Vol. 18 | No. 1 | April 2024
DOI : https://doi.org/10.36787/jti.v18i1.1263 Halaman 39 - 52

Prediksi Tingkat Atrisi Karyawan menggunakan *Machine Learning*

RESA RIANTI, RONI ANDARSYAH

Universitas Logistik dan Bisnis Internasional, Indonesia
Email : riantiresa224@gmail.com

ABSTRAK

Pengurangan karyawan dapat menjadi masalah serius bagi keunggulan kompetitif suatu organisasi dan mahal dalam hal biaya. Biaya pengurangan karyawan meliputi siklus hidup sumber daya manusia, hilangnya pengetahuan, semangat kerja, dan budaya organisasi. Atrisi karyawan terjadi secara alami dalam suatu organisasi karena berbagai faktor tak terhindarkan. Hal ini dapat menyebabkan kerugian besar bagi organisasi. Untuk mengatasi hal ini, penting bagi perusahaan untuk memahami faktor-faktor yang berpengaruh pada atrisi karyawan. Dalam penelitian ini, digunakan metode pemilihan fitur untuk mengidentifikasi faktor-faktor yang berpengaruh dan menyederhanakan pelatihan data menggunakan *dataset* atrisi *HR-analytics*. Model pembelajaran mesin seperti *Logistic Regression* dan *Support Vector Machine* digunakan untuk melatih dan mengevaluasi data. Tujuan utamanya adalah mendeteksi atrisi dengan akurasi tinggi untuk membantu perusahaan meningkatkan strategi retensi karyawan yang penting dan meningkatkan kepuasan mereka. Hasil penelitian ini dapat membantu manajemen memahami perubahan apa yang harus dilakukan di tempat kerja agar sebagian besar karyawan tetap bertahan. Ini akan membantu perusahaan dalam meramalkan pengurangan karyawan dan mengurangi biaya sumber daya manusia, serta mendorong pertumbuhan ekonomi mereka.

Kata kunci: Pengurangan karyawan, *Dataset HR-analytics*, Metode pemilihan fitur

ABSTRACT

Employee attrition has a serious impact on an organization's competitive advantage and incurs high costs. These costs include the entire human resource life cycle, loss of knowledge, motivation, and organizational culture. Employee attrition occurs naturally within organizations and causes significant losses. To overcome this problem, it is important for companies to understand the factors that influence employee attrition. This research uses feature selection methods and HR analytics attrition datasets to train machine learning models such as Logistic Regression and Support Vector Machine. The goal is to detect attrition with high accuracy to improve critical employee retention and satisfaction strategies. The results assist management in understanding the workplace changes required to retain employees. This helps companies forecast attrition, reduce human resource costs, and drive economic growth.

Keywords: Employee attrition, HR-analytics dataset, Feature selection method

1. PENDAHULUAN

Dalam upaya mencapai kesuksesan, perusahaan perlu memiliki strategi jangka panjang yang mempertimbangkan kekuatan dan kelemahan internal serta ancaman eksternal. Salah satu masalah yang dihadapi adalah pengelolaan tenaga kerja yang kompleks, mulai dari perekrutan hingga pengembangan. Sumber daya manusia sangat penting dalam organisasi dan diperlukan untuk mencapai tujuan perusahaan (Wardhani & Lhaksmana, 2022). Oleh karena itu, manajemen yang baik diperlukan dalam seleksi karyawan agar semua sumber daya manusia memenuhi kualifikasi dan mendukung kemajuan perusahaan.

Dalam suatu perusahaan, atrisi karyawan menjadi isu penting yang perlu ditangani. Proses seleksi ini penting karena investasi waktu dan uang yang dikeluarkan serta potensi kerugian jika sumber daya manusia ini meninggalkan organisasi(Suindari & Juniariani, 2020). Atrisi karyawan merujuk pada jumlah karyawan yang meninggalkan perusahaan dalam periode waktu tertentu. Tingkat atrisi yang tinggi dapat berdampak negatif pada produktivitas, stabilitas, dan biaya perusahaan. Faktor-faktor seperti karakteristik pribadi, faktor sistem perusahaan, dan lingkungan kerja dapat mempengaruhi keputusan karyawan untuk meninggalkan perusahaan (Umami, n.d.). Oleh karena itu, penting bagi divisi SDM untuk memahami faktor-faktor yang berpengaruh dalam atrisi karyawan.

Salah satu pendekatan yang dapat digunakan adalah membangun model pembelajaran mesin untuk memprediksi tingkat atrisi karyawan dengan menggunakan proses *data mining*. Data Mining adalah proses analisis data untuk menemukan hubungan yang jelas dan menghasilkan kesimpulan baru yang berguna bagi pemilik data. Terdapat beberapa bagian dalam Data Mining seperti asosiasi, prediksi, klasifikasi, pengelompokan, dan regresi. Dalam konteks ini, penggabungan pendekatan *data mining* dapat membantu divisi SDM mengambil langkah yang lebih proaktif dalam mengurangi pengurangan karyawan (And & Expert, 2022).

Dengan demikian, perusahaan dapat menghindari biaya dan waktu yang diperlukan untuk merekrut dan melatih karyawan baru. Selain itu, dengan memahami faktor-faktor yang berkontribusi pada atrisi karyawan, perusahaan dapat menerapkan strategi yang lebih efektif untuk meningkatkan kepuasan kerja, memperbaiki sistem perusahaan, dan menciptakan lingkungan kerja yang lebih baik (Arindi et al., 2023). Hal ini akan berdampak positif pada kinerja perusahaan secara keseluruhan dan membantu mencapai tujuan jangka panjang yang telah ditetapkan.

Dalam penelitian ini, digunakan data *HR Analytics* dari platform Data World yang mencakup berbagai fitur seperti usia, peran karyawan, tarif harian, kepuasan kerja, tahun di perusahaan, dan lain-lain. Penelitian ini bertujuan untuk menganalisis faktor-faktor penyebab karyawan meninggalkan perusahaan. Penelitian ini, menggunakan pendekatan *machine learning* dengan metode *supervised learning* digunakan untuk mengklasifikasikan data. *Supervised learning* merupakan metode di mana kita dapat memberikan label khusus pada setiap pengamatan. Sebagai contoh, dalam sebuah perusahaan *e-commerce* berusaha meningkatkan layanan pelanggan dengan menerapkan model *machine learning* untuk memprediksi tingkat kepuasan pelanggan. Data transaksi pelanggan, seperti waktu pengiriman, jumlah produk, tingkat diskon, dan umpan balik, digunakan dalam studi ini. Melalui *pre-processing* data, algoritma *Support Vector Machines (SVMs)* dipilih untuk melatih model yang dapat memahami pola hubungan antara fitur-fitur tersebut dan tingkat kepuasan pelanggan. Sejumlah algoritma yang termasuk dalam kelompok *supervised learning* melibatkan teknik-teknik seperti Regresi Linear, Regresi Logistik, Analisis Diskriminan Linear, *k-Nearest Neighbors, Support Vector Machines (SVMs), Random Forest, Decision Tree, dan Naive Bayes* (Santoso et al., 2021).

Teknik Regresi Linear digunakan untuk menemukan hubungan linier antara variabel independen dan dependen, sementara Regresi Logistik lebih cocok untuk memodelkan probabilitas kejadian biner. Analisis Diskriminan Linear berfokus pada pemisahan kelas, k-*Nearest Neighbors* memanfaatkan kedekatan data, SVMs menciptakan batas keputusan linier atau non-linier, *Random Forest* menggabungkan keputusan dari banyak pohon keputusan, *Decision Tree* memetakan keputusan berdasarkan serangkaian aturan, dan Naive Bayes menggunakan teorema Bayes untuk menghitung probabilitas kelas. Gabungan algoritma ini memberikan fleksibilitas dalam menangani berbagai jenis data dan permasalahan prediksi (Raza et al., 2022).

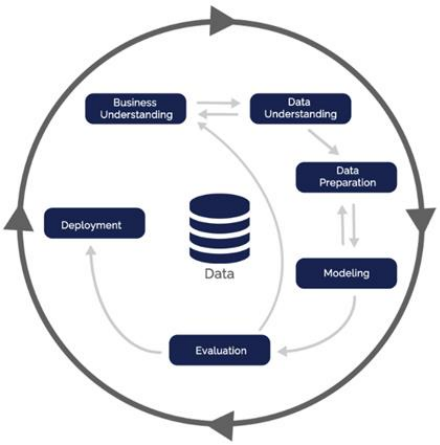
Penelitian memprediksi atrisi karyawan dengan metode *supervised learning* digunakan dengan menggunakan dua algoritma, yaitu *Logistic Regression* dan *Support Vector Machine (SVM),* sebagai perbandingan untuk melakukan prediksi atrisi karyawan dalam data HR Analytics.

Logistic Regression adalah sebuah metode statistik yang digunakan untuk memodelkan hubungan antara variabel dependen (dalam hal ini, tingkat atrisi karyawan) dengan satu atau lebih variabel independen (faktor-faktor yang mempengaruhi atrisi). Tujuan dari *Logistic Regression* adalah untuk memprediksi probabilitas kejadian suatu peristiwa berdasarkan nilai-nilai variabel *independent* (Setiawan et al., 2020). Model *Logistic Regression* menghasilkan output dalam bentuk probabilitas yang dapat diinterpretasikan sebagai kemungkinan terjadinya suatu peristiwa.

Sedangkan *Support Vector Machine (SVM)* adalah sebuah algoritma pembelajaran mesin yang digunakan untuk klasifikasi dan regresi. Dalam konteks ini, SVM digunakan untuk memprediksi tingkat atrisi karyawan berdasarkan data HR Analytics. SVM mencari pemisah optimal antara dua kelas dengan mencari hyperplane terbaik yang memaksimalkan jarak antara kelas-kelas tersebut(Octaviani et al., 2014). SVM juga dapat mengatasi masalah non-linear dengan menggunakan kernel untuk mentransformasi data ke dimensi yang lebih tinggi. Dengan menggunakan kedua algoritma ini, penelitian ini bertujuan untuk membandingkan kinerja *Logistic Regression* dan SVM dalam memprediksi atrisi karyawan dalam dataset HR Analytics (Biswas et al., 2023).

2. METODE PENELITIAN

CRISP-DM (Cross-Industry Standard Process for Data Mining) adalah suatu pendekatan metodologi yang umumnya diterapkan dalam proses data mining. Metodologi ini mengikuti serangkaian langkah-langkah terstruktur yang dimulai dari pemahaman bisnis hingga evaluasi hasil(Purnama et al., n.d.). Pertama, tahap Pemahaman Bisnis melibatkan identifikasi tujuan bisnis, kebutuhan data, dan pemahaman konteks masalah. Selanjutnya, Eksplorasi Data melibatkan pemeriksaan dan pemahaman mendalam terhadap data yang ada. Tahap Persiapan Data melibatkan pembersihan, transformasi, dan penggabungan data untuk memastikan kualitas dan keterpakaiannya. Setelah itu, tahap Pemodelan terlibat dalam pemilihan, pelatihan, dan evaluasi model data mining yang sesuai. Validasi berlanjut melalui Evaluasi, di mana model diuji dan dievaluasi kinerjanya. Terakhir, tahap Penerapan mengimplementasikan solusi data mining dalam lingkungan bisnis. Dengan pendekatan yang terstruktur ini, CRISP-DM memberikan panduan yang jelas untuk mengoptimalkan proses data mining dan meningkatkan pemahaman serta penerapan hasilnya dalam konteks bisnis yang relevan (Fadillah, 2015).



Gambar 1. CRISP-DM

Pada ilustrasi dalam Gambar diatas, tahapan proses CRISP DM dijelaskan sebagai berikut.

2.1 *Business Understanding*

Langkah awal dalam siklus CRISP-DM adalah Pemahaman Bisnis, yang bertujuan untuk memahami kebutuhan dan tujuan bisnis serta menerjemahkan pemahaman tentang bisnis menjadi definisi masalah dalam konteks data(Wurijanto et al., 2022). Tahap ini mencakup pemahaman mendalam tentang tujuan bisnis, hambatan yang dihadapi, dan kebutuhan informasi yang harus dipenuhi. Dalam penelitian ini, data diperlukan dari berbagai sumber terkait atrisi karyawan, dan dilakukan analisis untuk mengidentifikasi faktor-faktor yang paling berpengaruh terhadap atrisi karyawan. Hal ini bertujuan untuk mengembangkan model prediksi yang dapat memberikan kontribusi signifikan terhadap pemahaman dan penanganan masalah atrisi karyawan(Alsheref et al., 2022).

2.2 *Data Understanding*

Tahap kedua dalam model CRISP-DM adalah Eksplorasi Data, di mana dilakukan pengumpulan data, deskripsi data, dan evaluasi kualitas data. Proses ini melibatkan pencarian, analisis, dan penilaian data yang akan digunakan dalam penelitian. Data yang dikumpulkan berasal dari berbagai sumber, termasuk studi sebelumnya dan wawancara dengan para ahli terkait. Informasi diperoleh dari literatur yang berasal dari beragam media, seperti buku, artikel, atau jurnal, dan karya ilmiah (Qutub et al., 2021).

2.3 *Data Preparation*

Langkah ketiga dalam siklus CRISP-DM adalah Persiapan Data, yang melibatkan proses pembentukan dataset dari data yang telah dikumpulkan. Tahap ini memastikan dataset yang digunakan untuk melatih dan menguji model prediksi memiliki integritas yang baik, variabel yang relevan, dan keseimbangan yang sesuai untuk membangun model yang dapat memberikan hasil prediksi atrisi karyawan yang akurat(Stamolampros et al., 2019).

2.4 *Modelling*

Langkah Pemodelan melibatkan implementasi Machine Learning untuk menentukan teknik data mining, dan algoritma data mining yang sesuai dengan tujuan analisis. Dalam konteks penelitian ini, model regresi logistik dan support vector machine digunakan sebagai metode untuk menganalisis data dan merumuskan prediksi tingkat atrisi karyawan.

2.5 *Evaluation*

Pada tahap ini, model yang telah dirancang akan diujicoba dan dievaluasi guna mengukur tingkat keakuratan.Tahap evaluasi ini penting untuk menilai sejauh mana model yang telah dipilih memenuhi tujuan yang ditetapkan. Selain itu, evaluasi juga digunakan untuk menentukan apakah diperlukan pengembangan model tambahan atau perbaikan untuk mencapai hasil yang lebih baik. Evaluasi keakuratan model membantu dalam mengidentifikasi kelemahan dan kekuatan model, memberikan wawasan yang diperlukan untuk pengambilan keputusan lanjutan, dan memastikan bahwa model dapat diandalkan untuk tujuan prediksi atau analisis yang diinginkan.

2.6 Deployment

Tahapan deployment dalam CRISP-DM adalah langkah terakhir dalam siklus pengembangan model, yang melibatkan implementasi model ke dalam lingkungan produksi. Proses ini mencakup perencanaan deployment, implementasi model, uji coba, dan validasi. Dokumentasi lengkap dibuat, pengguna diberikan pelatihan, dan model diintegrasikan dengan proses bisnis yang relevan. Pemeliharaan dan pembaruan dilakukan secara berkala, sementara monitoring kinerja memastikan model tetap optimal dan relevan. Hasil dari model disampaikan kepada pengguna untuk mendukung pengambilan keputusan, dan evaluasi terus dilakukan untuk memastikan bahwa model terus memenuhi tujuan dan kebutuhan bisnis.

3. HASIL DAN PEMBAHASAN

3.1 Data Understanding

Pada tahap pemahaman data, proses penelitian dan eksplorasi data fokus utama terletak pada pemahaman mendalam terhadap konteks bisnis dengan tujuan mengidentifikasi, mengumpulkan, dan menganalisis dataset untuk mencapai sasaran yang telah dijelaskan dalam Business Understanding. Pada tahap ini, penekanan khusus diberikan pada identifikasi sumber data, yang dilakukan melalui dataset HR Analytics dari Data World. Dengan pemahaman yang lebih dalam terhadap dataset HR Analytics, diharapkan dapat memperoleh informasi yang dapat mendukung pengembangan model analitis yang efektif dan relevan dalam konteks manajemen sumber daya manusia. Data bersifat berlabel dan memiliki jumlah kolom sebanyak 35 kolom dan jumlah total record sebanyak 1,470 record.

3.2 Preprocessing Data

Fungsi IS NULL digunakan untuk mengecek apakah suatu nilai dalam kolom adalah Null atau tidak. Jika nilai tersebut Null, hasilnya adalah TRUE. Jika tidak, hasilnya adalah FALSE.

Age	False
Attrition	False
BusinessTravel	False
DailyRate	False
Department	False
DistanceFromHome	False
Education	False
EducationField	False
EmployeeCount	False
EmployeeNumber	False
EnvironmentSatisfaction	False
Gender	False
HourlyRate	False
JobInvolvement	False
JobLevel	False
JobRole	False
JobSatisfaction	False
MaritalStatus	False
MonthlyIncome	False
MonthlyRate	False
NumCompaniesWorked	False
Over18	False
OverTime	False
PercentSalaryHike	False
PerformanceRating	False
RelationshipSatisfaction	False
StandardHours	False
StockOptionLevel	False
TotalWorkingYears	False
TrainingTimesLastYear	False
WorkLifeBalance	False
YearsAtCompany	False
YearsInCurrentRole	False
YearsSinceLastPromotion	False
YearsWithCurrManager	False

Gambar 2. output nilai null pada dataset

Dalam hasil pada gambar semua nilai adalah False. Ini berarti bahwa tidak ada nilai null dalam setiap kolom DataFrame. Semua kolom dalam data tersebut memiliki nilai yang lengkap (non-null) pada setiap barisnya. Dengan tidak adanya nilai null, data tersebut siap digunakan untuk analisis lebih lanjut.

Education	EmployeeCount	EmployeeNumber	EnvironmentSatisfaction	HourlyRate	JobInvolvement	JobLevel	JobSatisfaction
3	1	1875	2	78	3	1	2
3	1	174	3	77	3	3	4
2	1	733	4	32	3	3	1
3	1	1499	3	46	3	1	3
4	1	1880	2	63	2	2	3

```
]: employee_data["Education"] = employee_data["Education"].replace({1:"Below College",2:"College",3:"Bache
]: employee_data["EnvironmentSatisfaction"] = employee_data["EnvironmentSatisfaction"].replace({1:"Low",2:
]: employee_data["JobInvolvement"] = employee_data["JobInvolvement"].replace({1:"Low",2:"Medium",3:"High",
]: employee_data["JobLevel"] = employee_data["JobLevel"].replace({1:"Entry Level",2:"Junior Level",3:"Mid
                                                    5:"Executive Level"})
]: employee_data["JobSatisfaction"] = employee_data["JobSatisfaction"].replace({1:"Low",2:"Medium",3:"High
]: employee_data["PerformanceRating"] = employee_data["PerformanceRating"].replace({1:"Low",2:"Good",3:"Ex
]: employee_data["RelationshipSatisfaction"] = employee_data["RelationshipSatisfaction"].replace({1:"Low",
]: employee_data["WorkLifeBalance"] = employee_data["WorkLifeBalance"].replace({1:"Bad",2:"Good",3:"Better
```

Gambar 3. mengubah kategori label dalam fitur angka pada dataset

Code tersebut menggunakan metode `replace` dari pandas untuk mengganti nilai pada kolom "Education" dalam DataFrame `employee_data`. Penggantian dilakukan dengan menggunakan dictionary yang menentukan pemetaan antara nilai awal dan nilai baru. Dalam hal ini, kolom "Education" awalnya mungkin berisi angka-angka 1, 2, 3, 4, dan 5, yang masing-masing mewakili tingkat pendidikan tertentu.

Dictionary yang diberikan dalam metode `replace` menyatakan bahwa:

a. 1 akan diganti dengan "Below College"

b. 2 akan diganti dengan "College"

c. 3 akan diganti dengan "Bachelor"

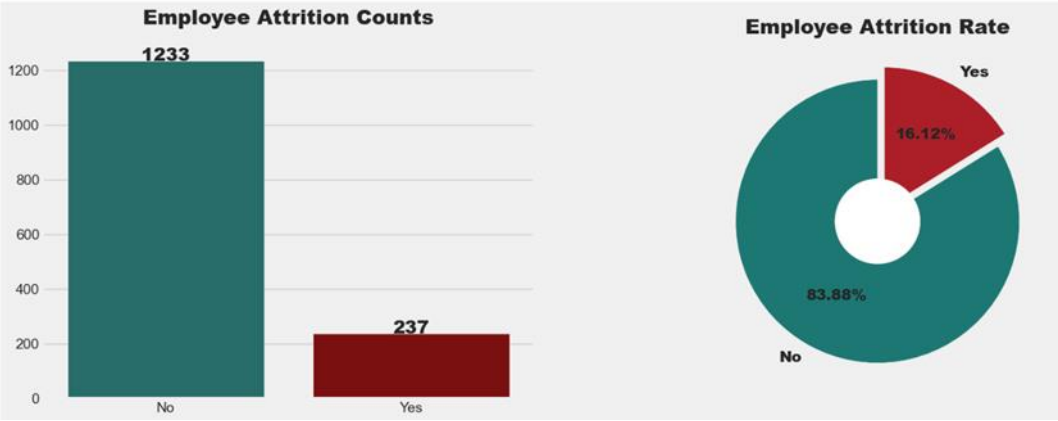
d. 4 akan diganti dengan "Master"

e. 5 akan diganti dengan "Doctor"

Sehingga, setelah eksekusi code tersebut, nilai-nilai pada kolom "Education" akan diganti sesuai dengan mapping yang diberikan. Tujuan umumnya adalah untuk membuat data lebih deskriptif dan memudahkan interpretasi, terutama ketika bekerja dengan kategori atau faktor. Hasil setelah dikategorikan maka sebagai berikut :

Education	EducationField	EnvironmentSatisfaction	Gender	JobInvolvement	JobLevel	JobRole	JobSatisfaction	MaritalStat
Master	Medical	Very High	Male	High	Junior Level	Manufacturing Director	High	Marr
Bachelor	Life Sciences	Low	Male	High	Mid Level	Sales Executive	High	Divorc
Bachelor	Life Sciences	High	Female	Low	Mid Level	Human Resources	High	Marr
Bachelor	Medical	Low	Male	High	Entry Level	Research Scientist	Very High	Marr
College	Medical	High	Male	Medium	Mid Level	Healthcare Representative	Low	Marr

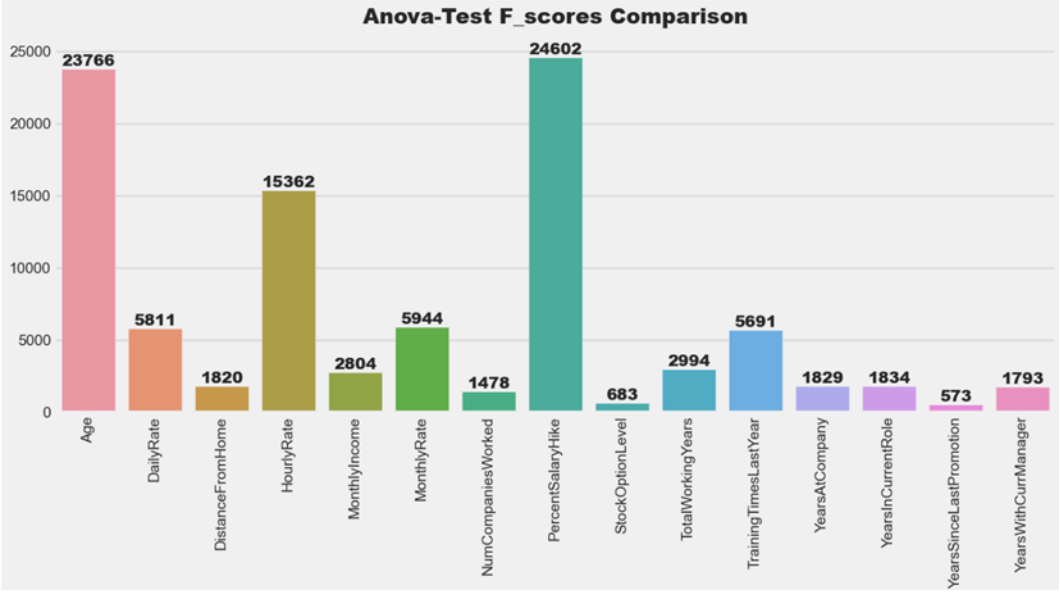
Gambar 4. Daftar fitur kategorik



Gambar 5. Visualisasi Data Atrisi Karyawan

Berdasarkan gambar diatas maka Tingkat atrisi karyawan di organisasi ini mencapai 16,12%, sebuah angka yang jauh melampaui tingkat normal yang disarankan oleh para ahli di bidang

Sumber Daya Manusia, yakni antara 4% hingga 6%. Dengan demikian, organisasi perlu mengambil langkah-langkah proaktif untuk mengurangi tingkat atrisi yang berbahaya ini. Tindakan pencegahan dan perbaikan kondisi kerja, peningkatan kepuasan karyawan, serta strategi retensi bakal menjadi kunci dalam memitigasi dampak negatif yang dapat timbul akibat tingkat atrisi yang tinggi ini.



Gambar 6. Visualisasi F_score dari uji anova

Analisis Variansi (ANOVA) adalah suatu metode statistik yang digunakan untuk membandingkan rata-rata antara tiga atau lebih kelompok yang independen. Tujuan utama dari ANOVA adalah untuk menilai apakah ada perbedaan signifikan di antara rata-rata kelompok-kelompok tersebut. Analisis hasil dengan memeriksa nilai p-value. Jika nilai p-value cukup rendah (biasanya di bawah tingkat signifikansi yang ditentukan, misalnya 0.05), maka kita dapat menolak hipotesis nol dan menyimpulkan bahwa terdapat perbedaan signifikan antara rata-rata kelompok-kelompok tersebut.

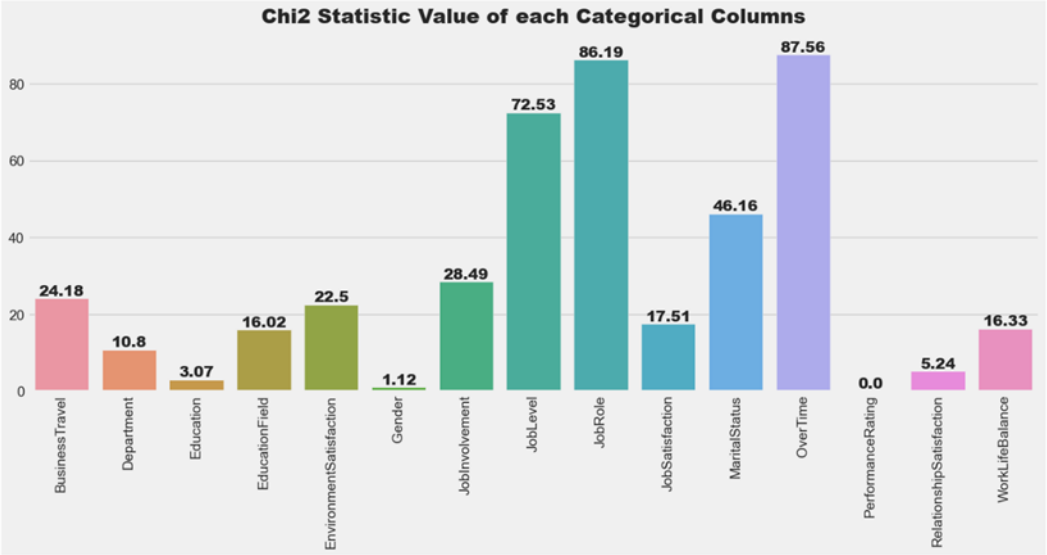
Pembahasan :

1. Nilai F yang Lebih Besar:
    - Nilai F adalah statistik uji dalam uji ANOVA. Jika nilai F lebih besar, itu menunjukkan adanya perbedaan yang lebih besar antara rata-rata kelompok-kelompok yang diuji. Dengan kata lain, semakin tinggi nilai F, semakin kuat hubungan antara variabel independen (kelompok) dan variabel dependen (data pengamatan).
2. Nilai p di Bawah Tingkat Signifikansi:
    - Nilai p adalah ukuran signifikansi statistik. Jika nilai p kurang dari tingkat signifikansi yang telah ditentukan (biasanya 0,05), kita dapat menolak hipotesis nol. Hipotesis nol dalam konteks uji ANOVA menyatakan bahwa tidak ada perbedaan signifikan antara rata-rata kelompok-kelompok yang diuji.
    - Jadi, jika nilai p < 0,05, kita dapat menyimpulkan bahwa terdapat perbedaan signifikan antara setidaknya dua kelompok yang diuji.

	Features	F_Score	P_value
0	Age	23766.934042	0.00000000000000000000
1	DailyRate	5811.796569	0.00000000000000000000
2	DistanceFromHome	1820.614585	0.00000000000000000000
3	HourlyRate	15362.122371	0.00000000000000000000
4	MonthlyIncome	2804.459632	0.00000000000000000000
5	MonthlyRate	5944.089071	0.00000000000000000000
6	NumCompaniesWorked	1478.188633	0.00000000000000000000
7	PercentSalaryHike	24602.507947	0.00000000000000000000
8	StockOptionLevel	683.069576	0.00000000000000000000
9	TotalWorkingYears	2994.906310	0.00000000000000000000
10	TrainingTimesLastYear	5691.401732	0.00000000000000000000
11	YearsAtCompany	1829.442766	0.00000000000000000000
12	YearsInCurrentRole	1834.262264	0.00000000000000000000
13	YearsSinceLastPromotion	573.896430	0.00000000000000000000
14	YearsWithCurrManager	1793.291314	0.00000000000000000000

Gambar 7. membandingkan f_score dan p_value dari uji anova.

Hasil analisis menunjukkan bahwa beberapa fitur memiliki korelasi yang kuat dengan tingkat atrisi karyawan, sementara yang lainnya tidak menunjukkan hubungan yang signifikan. Fitur-fitur seperti usia karyawan, serta aspek kompensasi seperti DailyRate, HourlyRate, MonthlyIncome, dan MonthlyRate, menunjukkan korelasi yang tinggi dengan tingkat atrisi. Hal ini mengindikasikan bahwa faktor-faktor ini mungkin memiliki peran penting dalam keputusan karyawan untuk tetap atau meninggalkan organisasi. Selain itu, fitur-fitur terkait pengalaman kerja, kenaikan gaji, dan durasi waktu di perusahaan atau dengan manajer saat ini juga menunjukkan korelasi yang kuat. Di sisi lain, beberapa fitur seperti DistanceFromHome, StockOptionLevel, YearsInCurrentRole, dan YearsSinceLastPromotion, meskipun memiliki variasi, tidak menunjukkan hubungan yang signifikan dengan tingkat atrisi. Mungkin terdapat faktor-faktor lain yang tidak tercakup dalam data yang turut berperan dalam keputusan atrisi karyawan.



Gambar 8. Uji chi-squre

	Features	Chi_2 Statistic	P_value
0	BusinessTravel	24.182414	0.00000560861447644993
1	Department	10.796007	0.00452560657447963286
2	Education	3.073961	0.54552533765659494414
3	EducationField	16.024674	0.00677398013902521211
4	EnvironmentSatisfaction	22.503881	0.00005123468906289433
5	Gender	1.116967	0.29057244902890855265
6	JobInvolvement	28.492021	0.00000286318063671342
7	JobLevel	72.529013	0.00000000000000663468
8	JobRole	86.190254	0.00000000000000275248
9	JobSatisfaction	17.505077	0.00055630045103875563
10	MaritalStatus	46.163677	0.00000000009455511060
11	OverTime	87.564294	0.00000000000000000001
12	PerformanceRating	0.000155	0.99007454659345761616
13	RelationshipSatisfaction	5.241068	0.15497244371052629197
14	WorkLifeBalance	16.325097	0.00097256988453488236

Gambar 9. membandingkan chi-square dan P_value

Analisis deskriptif atribut kategoris mengungkapkan hubungan yang signifikan antara beberapa fitur dan tingkat atrisi karyawan. Departemen, EducationField, EnvironmentSatisfaction, JobInvolvement, JobLevel, JobRole, JobSatisfaction, MaritalStatus, OverTime, dan WorkLifeBalance adalah atribut yang memiliki korelasi yang signifikan dengan atrisi, menyoroti aspek-aspek seperti kepuasan kerja, keterlibatan, level pekerjaan, dan faktor-faktor lain yang mempengaruhi keputusan karyawan. Sementara itu, atribut seperti Gender, Education, PerformanceRating, dan RelationshipSatisfaction tidak menunjukkan korelasi yang signifikan, menandakan bahwa variabilitas dalam atribut-atribut ini mungkin tidak berperan sentral dalam keputusan atrisi karyawan.

3.3 Implementasi Algoritma

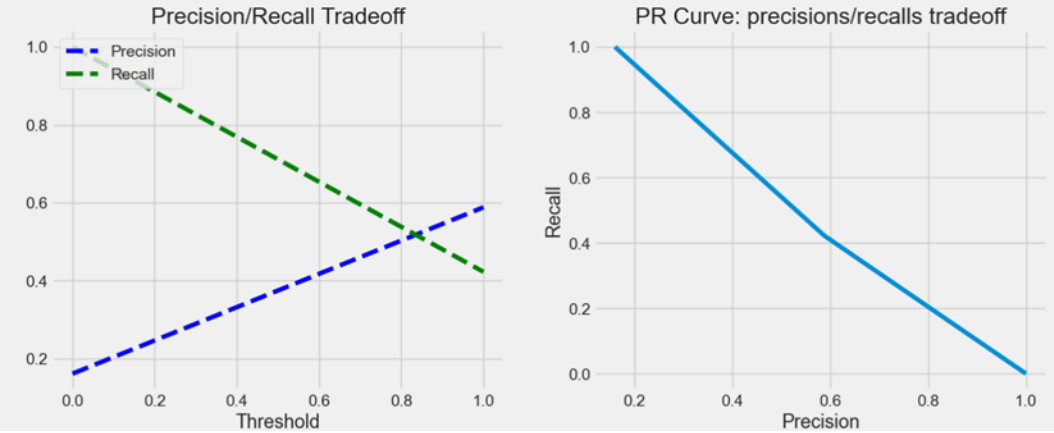
Setelah menyelesaikan tahap preprocessing, langkah berikutnya adalah memasuki tahap pemodelan karena data telah siap untuk digunakan. Pada tahap pemodelan ini, dilakukan serangkaian pengujian model machine learning pada dataset dengan menerapkan algoritma Regresi Logistik dan Support Vector Machine. Untuk mengevaluasi akurasi model, dataset dibagi menjadi dua bagian, dengan 80% digunakan untuk melatih model (train) dan 20% sisanya untuk menguji model (test).

a. Algoritma Regresi Logistik

```
ACCURACY SCORE:
0.9291
CLASSIFICATION REPORT:
                     0           1  accuracy    macro avg  weighted avg
precision     0.935982    0.878049  0.929057     0.907016      0.926636
recall        0.982619    0.650602  0.929057     0.816611      0.929057
f1-score      0.958734    0.747405  0.929057     0.853069      0.924642
support     863.000000  166.000000  0.929057  1029.000000   1029.000000
```

Gambar 10. output algoritma regresi logistik

Akurasi adalah rasio dari jumlah prediksi yang benar (True Positives + True Negatives) terhadap total jumlah sampel. Nilai akurasi di sini adalah 0.9291, atau 92.91%, yang menunjukkan bahwa sekitar 92.91% dari semua prediksi yang dilakukan oleh model adalah benar.



Gambar 11. Precision algoritma regresi logistik

Visualisasi evaluasi model klasifikasi, khususnya model regresi logistik (`lr_clf`). Pertama, terdapat grafik tradeoff presisi-recall yang menunjukkan bagaimana presisi dan recall berubah dengan variasi nilai threshold. Garis biru putus-putus mewakili presisi, sementara garis hijau putus-putus mewakili recall, dengan sumbu x menunjukkan nilai threshold. Grafik ini membantu memahami keseimbangan antara presisi dan recall saat threshold berubah, memungkinkan pemilihan nilai threshold yang sesuai dengan kebutuhan aplikasi tertentu.
Selanjutnya, terdapat grafik kurva presisi-recall yang menampilkan hubungan antara presisi dan recall tanpa mempertimbangkan nilai threshold tertentu. Grafik ini memberikan gambaran menyeluruh tentang kinerja model tanpa fokus pada

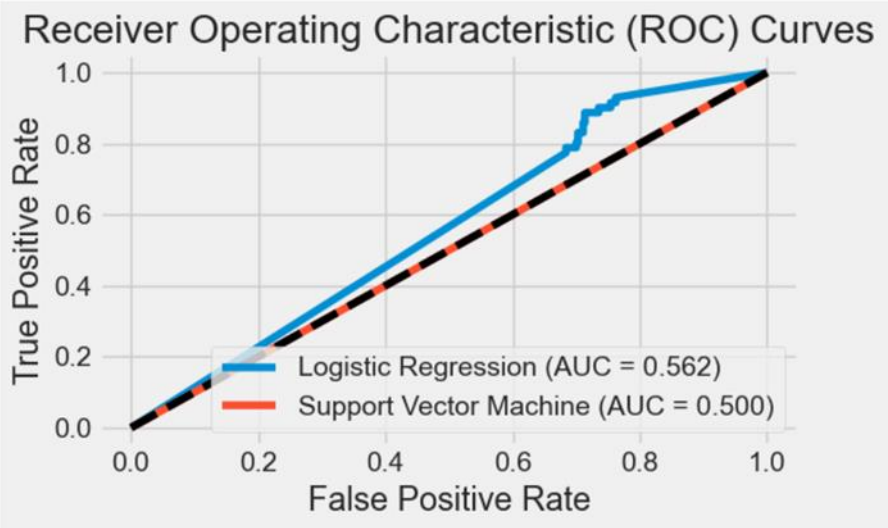
pengaturan threshold khusus. Hal ini bermanfaat untuk mengevaluasi seberapa baik model dapat mengklasifikasikan kelas positif dan negative.

b. Algoritma Support vector machine

```
TRAINIG RESULTS:
===============================
CONFUSION MATRIX:
[[860   3]
 [ 63 103]]
ACCURACY SCORE:
0.9359
CLASSIFICATION REPORT:
                    0           1  accuracy    macro avg  weighted avg
precision    0.931744    0.971698   0.93586     0.951721      0.938190
recall       0.996524    0.620482   0.93586     0.808503      0.935860
f1-score     0.963046    0.757353   0.93586     0.860199      0.929863
support    863.000000  166.000000   0.93586  1029.000000   1029.000000
```

Gambar 12. output algoritma SVM

Hasil output dari model yang menggunakan algoritma Support Vector Machine (SVM) telah menghasilkan nilai akurasi sebesar 94.17%. Akurasi dihitung sebagai rasio antara jumlah prediksi yang benar (True Positives + True Negatives) terhadap total jumlah sampel. Dengan akurasi sebesar 94.17%, dapat disimpulkan bahwa sekitar 94.17% dari semua prediksi yang dibuat oleh model SVM ini adalah benar. Tingkat akurasi yang tinggi ini menandakan bahwa model SVM mampu melakukan prediksi dengan akurasi yang baik pada dataset yang digunakan.



Gambar 13. Perbandingan kedua algoritma

Berdasarkan evaluasi dua model klasifikasi, regresi logistik dan Support Vector Machine (SVM), terdapat perbedaan signifikan dalam performa keduanya. Model SVM menunjukkan akurasi lebih tinggi (94.17%) dibandingkan dengan regresi logistik (92.91%). Kedua model cenderung memiliki presisi yang lebih tinggi untuk kelas negatif, namun SVM memiliki tingkat presisi keseluruhan yang lebih tinggi untuk kedua kelas. Dalam hal recall, SVM unggul dalam mengidentifikasi instance kelas negatif,

sementara regresi logistik memiliki recall yang lebih tinggi untuk kelas positif. Secara keseluruhan, SVM menunjukkan performa lebih baik dengan akurasi yang tinggi dan nilai F1-score yang seimbang untuk kedua kelas. Pemilihan model yang paling sesuai harus disesuaikan dengan konteks aplikasi dan kebutuhan bisnis spesifik, dengan kemungkinan perlu dilakukan evaluasi lebih lanjut untuk validasi hasil.

4. KESIMPULAN

Penelitian ini memprediksi tingkat atrisi karyawan menggunakan machine learning, hasil evaluasi dua model klasifikasi, yaitu regresi logistik dan Support Vector Machine (SVM), menunjukkan perbedaan signifikan dalam performa keduanya. Model SVM mampu mencapai tingkat akurasi yang lebih tinggi (94.17%) dibandingkan dengan regresi logistik (92.91%), dengan presisi yang lebih tinggi untuk kedua kelas. Meskipun cenderung memiliki presisi lebih tinggi untuk kelas negatif, SVM juga menunjukkan kemampuan unggul dalam mengidentifikasi instance dari kelas negatif berdasarkan nilai recall. Kesimpulannya, model SVM menawarkan performa yang lebih baik secara keseluruhan, diukur dari akurasi dan nilai F1-score yang seimbang untuk kedua kelas. Dengan mengimplementasikan rekomendasi ini, diharapkan perusahaan dapat meningkatkan retensi karyawan, menciptakan lingkungan kerja yang lebih baik, dan mencapai tujuan jangka panjang yang telah ditetapkan.

DAFTAR RUJUKAN

Alsheref, F. K., Fattoh, I. E., & Mead, W. (2022). Automated Prediction of Employee Attrition Using Ensemble Model Based on Machine Learning Algorithms. *Computational Intelligence and Neuroscience*, *2022*. https://doi.org/10.1155/2022/7728668

And, I., & Expert, D. (2022). *Sistem Prediksi Awal Terhadap Atrisi Karyawan Menggunakan Algoritma C4.5 Informasi Artikel* (Vol. 4, Issue 1). https://e-journal.unper.ac.id/index.php/informatics

Biswas, A. K., Seethalakshmi, R., Mariappan, P., & Bhattacharjee, D. (2023). An ensemble learning model for predicting the intention to quit among employees using classification algorithms. *Decision Analytics Journal*, *9*. https://doi.org/10.1016/j.dajour.2023.100335

Fadillah, A. P. (2015). Penerapan Metode CRISP-DM untuk Prediksi Kelulusan Studi Mahasiswa Menempuh Mata Kuliah (Studi Kasus Universitas XYZ). In *Jurnal Teknik Informatika dan Sistem Informasi* (Vol. 1).

Octaviani, P. A., Wilandari, Y., & Ispriyanti, D. (2014). *Penerapan Metode Klasifikasi Support Vector Machine (Svm) Pada Data Akreditasi Sekolah Dasar (Sd) Di Kabupaten Magelang*. *3*(4), 811–820. Http://Ejournal-S1.Undip.Ac.Id/Index.Php/Gaussian

Purnama, I., Saputra, R., & Wibowo, A. (N.D.). *Implementasi Data Mining Menggunakan Crisp-Dm Pada Sistem Informasi Eksekutif Dinas Kelautan Dan Perikanan Provinsi Jawa Tengah*.

Qutub, A., Al-Mehmadi, A., Al-Hssan, M., Aljohani, R., & Alghamdi, H. S. (2021). Prediction of Employee Attrition Using Machine Learning and Ensemble Methods. *International Journal of Machine Learning and Computing*, *11*(2), 110–114. https://doi.org/10.18178/ijmlc.2021.11.2.1022

Raza, A., Munir, K., Almutairi, M., Younas, F., & Fareed, M. M. S. (2022). Predicting Employee Attrition Using Machine Learning Approaches. *Applied Sciences (Switzerland)*, *12*(13). https://doi.org/10.3390/app12136424

Santoso, P., Abijono, H., & Anggreini, N. L. (2021). Algoritma Supervised Learning Dan Unsupervised Learning Dalam Pengolahan Data. *Unira Malang |*, *4*(2).

Setia Arindi, A., Mirza, A., Darma Palembang Jl Jenderal Ahmad Yani No, B., Seberang Ulu, K. I., Palembang, K., & Selatan, S. (2023). *Model Klasifikasi Kinerja Pegawai Dengan Penerapan Machine Learning Menggunakan Tools Python*. *8*(1).

Setiawan, I., Suprihanto, S., Nugraha, A. C., & Hutahaean, J. (2020). HR analytics: Employee attrition analysis using logistic regression. *IOP Conference Series: Materials Science and Engineering*, *830*(3). https://doi.org/10.1088/1757-899X/830/3/032001

Stamolampros, P., Korfiatis, N., Chalvatzis, K., & Buhalis, D. (2019). Job satisfaction and employee turnover determinants in high contact services: Insights from Employees'Online reviews. *Tourism Management*, *75*, 130–147. https://doi.org/10.1016/j.tourman.2019.04.030

Suindari, N. M., & Juniariani, N. M. R. (2020). Pengelolaan Keuangan, Kompetensi Sumber Daya Manusia Dan Strategi Pemasaran Dalam Mengukur Kinerja Usaha Mikro Kecil Menengah (UMKM). *KRISNA: Kumpulan Riset Akuntansi*, *11*(2), 148–154. https://doi.org/10.22225/kr.11.2.1423.148-154

Umami, A. (n.d.). *Klasifikasi Faktor-faktor yang Mempengaruhi Pengurangan Karyawan Pada Perusahaan "XYZ."*

Wardhani, F. H., & Lhaksmana, K. M. (2022). Predicting Employee Attrition Using Logistic Regression With Feature Selection. *Sinkron*, *7*(4), 2214–2222. https://doi.org/10.33395/sinkron.v7i4.11783

Wurijanto, T., Setiawan, H. B., & Tjandrarini, A. B. (2022). Penerapan Model CRISP-DM pada Prediksi Nasabah Kredit yang Berisiko Menggunakan Algoritma Support Vector Machine. *Jurnal Ilmiah Scroll: Jendela Teknologi Informasi*, *10*(1). https://univ45sby.ac.id/ejournal/index.php/informatika