JSI : Jurnal Sistem Informasi *(E-Journal), VOL.15, NO.1, April 2023*
ISSN Print : 2085-1588
ISSN Online :2355-4614
http://ejournal.unsri.ac.id/index.php/jsi/index
email: jsifasilkom@unsri.ac.id

Uji Akurasi Algoritma Machine Learning Untuk Pemodelan Prediksi Faktor Pendorong Pergantian Karyawan

Ade Surya Budiman[1], Desmulyati[2], Fahrizal[3]
[1,2] Program Studi Teknologi Komputer, Fakultas Teknik & Informatika, Univ. Bina Sarana Informatika
[3]Program Studi Sistem Informasi, Fakultas Teknik & Informatika, Univ. Bina Sarana Informatika
e-mail: [1]ade.aum@bsi.ac.id, [2]desmulyati.dmy@bsi.ac.id, [3]fahrizal.fzl@bsi.ac.id

Abstrak

Tim yang kohesif dan solid mempengaruhi stabilitas proses kerja dalam suatu organisasi. Pergantian anggota tim atau karyawan dalam waktu singkat dapat mempengaruhi bagaimana perusahaan dapat segera mencapai proyek dan target organisasi. Berbagai faktor dapat memicu pergantian karyawan. Dari penelitian ini, ditemukan beberapa faktor pendorong utama pergantian karyawan. Untuk menemukan faktor-faktor pendorong tersebut, dibangun suatu model machine learning. Selanjutnya untuk memastikan akurasi dari model yang dibangun, dilakukan uji akurasi terhadap dua algoritma yang dipergunakan untuk membangun model tersebut, yaitu Logistic Regression dan Random Forest. Pengujian menggunakan dataset publik diperoleh skor akurasi sebesar 0,77 pada Logistic Regression, dan Random Forest memiliki skor akurasi sebesar 0,98. Faktor pendorong turnover karyawan tertinggi adalah tingkat kepuasan sebesar 50,05%, diikuti oleh waktu yang dihabiskan di perusahaan sebesar 27.14%. Faktor pendorong ketiga yang paling signifikan adalah evaluasi terakhir dari pekerja yaitu sebesar 18,27%.

Keywords*: pergantian karyawan, logistic regression, random forest, machine learning, uji akurasi*

Abstract

A cohesive and solid team influences the stability of work processes in an organization. The team member or employee turnover in a short time can affect how the company can promptly achieve a project and organizational targets. Various factors can trigger employee turnover. From this study, we found several main driving factors for employee turnover. To find these driving factors, a machine learning model was built. Furthermore, to ensure the accuracy of the model built, an accuracy test was carried out on the two algorithms used to build the model, namely Logistic Regression and Random Forest. Testing using public datasets obtained an accuracy score of 0.77 on Logistic Regression, and Random Forest has an accuracy score of 0.98. The highest driving factor for employee turnover is the satisfaction level of 50.05%, followed by time spent in the company at 27.14%. The third most significant driving factor is the last evaluation of workers, which is 18.27%.

Keywords*: employee turnover, logistic regression, random forest, machine learning, accuracy test*

1. PENDAHULUAN

Pergantian karyawan dapat berimplikasi negatif pada kinerja sebuah perusahaan atau instansi, dalam mencapai objektif yang telah direncanakan. Keluarnya satu atau lebih karyawan, terutama dengan keahlian yang unik dan spesifik, akan berdampak signifikan terhadap kinerja unit kerja maupun perusahaan atau instansi secara keseluruhan. Gejolak yang terjadi pada karyawan termasuk pengunduran diri karyawan dalam waktu yang singkat dengan jumlah yang besar, akan menjadi masalah berbiaya tinggi (*costly problem*) bagi perusahaan [1].

Tingginya biaya yang harus dikeluarkan untuk mengganti karyawan baru, melalui proses rekrutmen yang terkadang tidak serta merta mendapatkan pengganti yang sepadan dan dapat segera berintegrasi dengan tim untuk melanjutkan proyek yang tengah

berlangsung. Biaya aktual dari proses pergantian karyawan termasuk rekrutmen karyawan baru, akan membutuhkan biaya yang tinggi, bergantung kepada pengalaman dan keahlian yang dimiliki oleh karyawan [2]. Pergantian karyawan yang dilakukan dengan terburu-buru, juga akan menghasilkan permasalahan tersendiri, seperti pekerja yang tidak tepat, akan berdampak pula secara finansial dan produktivitas, termasuk mempengaruhi pula kualitas pekerjaan dari karyawan baru tersebut [3]. Komitmen perusahaan sangat penting untuk memastikan karyawan memiliki ikatan emosional dengan perusahaan untuk mencegah pergantian karyawan yang signifikan [4].

Penelitian terkait dengan pergantian karyawan (*employee turnover*) telah banyak dilakukan dengan menggunakan berbagai metode dan teknik untuk menemukan keterkaitan antara satu faktor dengan faktor lainnya. *Improved Random forest methodology* dipergunakan untuk membantu memprediksi secara akurat *employee turnover* di sebuah cabang perusahaan telekomunikasi di China [5]. Dalam penelitian lainnya, dengan mengadopsi algoritma *Logistic Regression* dan *Gradient-boosted decision trees* (GDBT), dilakukan prediksi pergantian karyawan untuk mendapatkan referensi yang efektif bagi perusahaan dalam mengurangi turnover rate pada karyawan perusahaan [6].

Dengan berbagai pendekatan algoritma dan teknik *machine learning* dalam memprediksi pergantian karyawan, tentunya perlu pula untuk menguji seberapa akurat algoritma-algoritma tersebut dalam memprediksi tingkat pergantian karyawan (*employee turnover rate*) ini. Untuk itu dalam penelitian ini, dilakukan uji akurasi antara algoritma *Linear Regression* dan algoritma *Random Forest*. Sekaligus pula untuk memetakan faktor apa saja yang mendorong terjadinya pergantian karyawan.

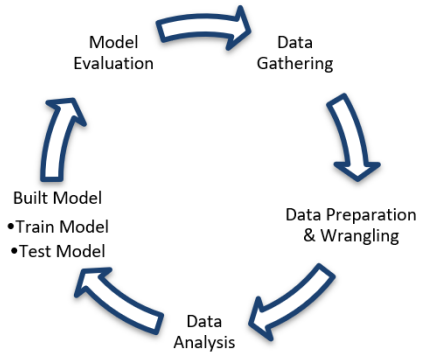
Model *machine learning* yang memiliki akurasi yang tinggi akan sangat bermanfaat membantu bagian terkait di sebuah organisasi atau perusahaan dalam menentukan faktor-faktor apa saja yang dapat mendorong terjadinya pergantian karyawan tersebut. Dengan mengetahui faktor yang mendorong terjadinya pergantian karyawan, suatu organisasi terutama pada bidang sumber daya manusia, dapat mempersiapkan langkah antisipatif dan secara efektif mengelola sumber daya manusia yang dimilikinya. Sehingga berdasarkan hasil pengujian dari model *machine learning* ini dapat menjadi dasar rekomendasi bagi organisasi atau perusahaan dalam mengambil suatu keputusan terkait dengan pengelolaan sumber daya manusia, seperti rencana pelatihan, tata pemberian insentif, perencanaan lingkungan kerja hingga kepada program perekrutan karyawan.

2. METODE PENELITIAN

2.1. Pemodelan dan Siklus Data (*Data Lifecycle*)

Dalam penelitian ini, dilakukan beberapa perlakuan terhadap data sebelum dilakukan pembangunan model *machine learning* untuk memprediksi pergantian karyawan. Proses *end to end* dalam pemodelan prediksi pergantian karyawan ini, dideskripsikan sebagai *data lifecycle*, sebagaimana diperlihatkan pada gambar 1.

JSI : Jurnal Sistem Informasi *(E-Journal), VOL.15, NO.1, April 2023*
ISSN Print : 2085-1588
ISSN Online :2355-4614
http://ejournal.unsri.ac.id/index.php/jsi/index
email: jsifasilkom@unsri.ac.id



Gambar 1. *Data Lifecycle* Dalam Pemodelan Prediksi Pergantian Karyawan

a. Data Gathering.

Dataset yang digunakan untuk membangun model prediksi pergantian karyawan diambil dari dataset publik di www.kaggle.com, "HR_comma_sep". *Dataset* ini tersedia untuk diunduh secara bebas lisensi dalam format *comma-separated value* (CSV) yang terdiri dari 14999 record dalam 10 (sepuluh) kolom. *Dataset* terdiri atas sejumlah variabel, yaitu *satisfaction_level* (dengan nilai antara 0-1), *last_evaluation* (dengan nilai antara 0-1), *number_of_project, average_monthly_hours, time_spend_company* (dalam tahun), *Work_accident, decision, promotion_last_5years, department,* dan *salary*. Atribut masing-masing *field* atau kolom dalam *dataset* yang digunakan dapat dilihat di Tabel 1.

Tabel 1**.** *Dataset Overview*

Nama Kolom / *Field*	***Field Rename***	**Atribut**		
		Data Type	**Nilai**	**Deskripsi**
satisfaction_level	-	float	Antara 0.09 – 1.00	Tingkat kepuasan kerja karyawan
last_evaluation	-	float	Antara 0.36 – 1.00	Hasil evaluasi terbaru terhadap karyawan
number_project	number_of_project	integer	Antara 2 – 7	Jumlah proyek yang dikerjakan oleh karyawan
average_monthly_hours	-	integer	Antara 96 – 310	Rata-rata jam kerja bulanan karyawan
time_spend_company	-	integer	Antara 2 – 10	Masa kerja karyawan di perusahaan
work_accident	-	integer	0 (tidak pernah) and 1 (pernah)	Kecelakaan kerja yang pernah

JSI : Jurnal Sistem Informasi *(E-Journal), VOL.15, NO.1, April 2023*
ISSN Print : 2085-1588
ISSN Online :2355-4614
http://ejournal.unsri.ac.id/index.php/jsi/index
email: jsifasilkom@unsri.ac.id

Nama Kolom / *Field*	***Field Rename***	**Atribut**		
		Data Type	**Nilai**	**Deskripsi**
				dialami/dilakukan oleh karyawan
left	decision	integer	0 (Tidak Mengundurkan diri) and 1 (Mengundurkan diri)	Status/keputusan karyawan
promotion_last_5year	-	integer	0 (tidak pernah) and 1 (pernah)	Promosi yang didapatkan karyawan dalam 5 tahun terakhir
sales	department	string	accounting; engineering & it support; human resources; management; product mng; research and development; sales and marketing;	Divisi atau departemen tempat karyawan ditempatkan. Beberapa departemen telah digabung/bergant i nama untuk menyederhanaka n analisis data
salary	-	string	low; medium; high	Tingkat gaji karyawan

b. Data Preparation & Wrangling

Sebelum dilakukan proses lebih lanjut dari dataset yang didapatkan, *dataset* tersebut perlu divalidasi terlebih dahulu. Validasi ini diantaranya adalah pemeriksaan nilai kosong (*null*) dan verifikasi tipe data. Dari hasil pemeriksaan, dataset tidak memiliki nilai kosong dan sudah tidak diperlukan konversi tipe data sebelum pengolahan lebih lanjut. Disamping itu, beberapa atribut dari *dataset* awal, memerlukan beberapa penyesuaian terkait dengan penamaan atribut. Penyesuaian ini diperlukan untuk menghindari mispersepsi terhadap penamaan atribut. Rincian perubahan tersebut dapat dilihat pada Tabel 1.

c. Data Analysis

Dalam tahap ini dipilih algoritma analisa data yang akan digunakan. Untuk penelitian ini, penulis mempergunakan dan membandingkan algoritma *Logistic Regression* dan *Random Forest Classification* sebagai teknik *machine learning* dalam memprediksi faktor pendorong terjadinya pergantian karyawan.

JSI : Jurnal Sistem Informasi *(E-Journal), VOL.15, NO.1, April 2023*
ISSN Print : 2085-1588
ISSN Online :2355-4614
http://ejournal.unsri.ac.id/index.php/jsi/index
email: jsifasilkom@unsri.ac.id

d. Built Model

Dalam tahap ini, dataset akan dibagi kedalam dua kelompok utama yaitu *training set* dan *testing set*. *Training set* diperlukan untuk memeriksa dan memahami *pattern*, *features* dan *rule* dari *dataset* yang dipergunakan. Sedangkan *testing set* diperlukan untuk menentukan seberapa besar keakuratan (dalam persen) model yang dibuat sesuai dengan kebutuhan.

e. Deployment.

Dalam tahap ini penulis mengetengahkan simpulan dari model yang dibangun berupa apa saja faktor utama yang menjadi pendorong terjadinya pergantian karyawan, sekaligus rekomendasi bagi pemegang kepentingan dan pemegang keputusan dalam sebuah organisasi atau instansi.

2.2. Persiapan *Dataset*

Data yang diperoleh dari dataset terlebih dahulu disiapkan untuk proses pemodelan. Perbaikan nama variabel atau kolom pada dataset dilakukan untuk memastikan dataset memiliki nama variabel yang akurat dan ringkas. Variabel dalam kumpulan data sumber diberi nama "*Sales*". Dimana variabel ini berisi *record* dari divisi atau departemen tempat karyawan tersebut bekerja. Dalam penelitian ini, variabel tersebut berganti nama menjadi "department". Selanjutnya pada variabel '*left*' yang menunjukkan status dan keputusan karyawan, 0 untuk "Mengundurkan diri", dan 1 untuk "Tidak Mengundurkan diri", nama variabel tersebut diubah menjadi '*decision*' atau status pergantian (*turnover*) karyawan.

Seperti terlihat pada Tabel 1, untuk mempermudah analisis data, tiga divisi yang membidangi aspek teknis, yaitu IT, Support, dan departemen teknis, digabungkan menjadi satu divisi, Engineering dan Technical Departments. Gambar 2 menunjukkan gambaran umum *dataset* yang telah disiapkan dan dibersihkan sebelum masuk ke analisis data.

Sample data:

	satisfaction_level	last_evaluation	number_of_project	average_monthly_hours	time_spend_company	Work_accident	decision	promotion_last_5years	department	salary
0	0.38	0.53	2	157	3	0	1	0	sales and marketing	low
1	0.80	0.86	5	262	6	0	1	0	sales and marketing	medium
2	0.11	0.88	7	272	4	0	1	0	sales and marketing	medium
3	0.72	0.87	5	223	5	0	1	0	sales and marketing	low
4	0.37	0.52	2	159	3	0	1	0	sales and marketing	low
...	...	...	...	...	...	...	...	...	...	...
14994	0.40	0.57	2	151	3	0	1	0	Engineering and IT Support	low
14995	0.37	0.48	2	160	3	0	1	0	Engineering and IT Support	low
14996	0.37	0.53	2	143	3	0	1	0	Engineering and IT Support	low
14997	0.50	0.96	6	280	4	0	1	0	Engineering and IT Support	low
14998	0.37	0.52	2	158	3	0	1	0	Engineering and IT Support	low

14999 rows × 10 columns

Gambar 2. *Dataset Overview* Menampilkan Lima Data Teratas dan Data Terbawah Setelah Tahap Persiapan *Dataset*

JSI : Jurnal Sistem Informasi *(E-Journal), VOL.15, NO.1, April 2023*
ISSN Print : 2085-1588
ISSN Online :2355-4614
http://ejournal.unsri.ac.id/index.php/jsi/index
email: jsifasilkom@unsri.ac.id

3. HASIL DAN DISKUSI

3.1 Pembuatan Model

Recursive Feature Elimination (RFE) sebagai teknik pemilihan *subset*, diperlukan untuk memastikan hanya fitur (kolom) yang paling relevan yang dipergunakan untuk membangun model. Dengan menggunakan fitur yang paling relevan, algoritma pada model *machine learning* yang dibangun akan berjalan lebih efektif dan efisien. Terhadap kondisi *overfitting* pada data, RFE memiliki kemampuan yang lebih baik dibandingkan metode lainnya [7]. Pada penelitian ini untuk penggunaan estimator *Logistic Regression* ditentukan jumlah fitur terpilih sebanyak 10 fitur.

Logistic Regression adalah metode klasifikasi tradisional yang melibatkan diskriminan linier dengan keluaran utama berupa nilai probabilitas. Kemudian model yang dibangun akan membentuk batas linier yang memisahkan ruang input menjadi dua bagian [8]. Dalam membangun model menggunakan algoritma *Logistic Regression* ini, kami menggunakan *test size* 30% atau 0.3 dengan menerapkan *random test* = 0 sehingga *train set* dan *test set* akan selalu sama setiap kali dieksekusi. Dari pengujian tersebut, akurasi model yang dibangun dengan *Logistic Regression* adalah sebesar 0,770.

Dengan menggunakan algoritma *Random Forest* akan membentuk pohon keputusan secara acak untuk setiap iterasi yang terdapat pada metode pembelajaran ini. Hasilnya, *Random Forest* akan meningkatkan performa dan akurasi algoritma *machine learning* ini, termasuk menghindari *overfitting* pada data. *Random Forest* memiliki kemampuan yang sangat baik untuk menangani *dataset* yang memiliki variabel prediktor dalam jumlah besar [9]. Berikut langkah-langkah yang dilakukan saat menggunakan algoritma *Random Forest*:

a. *Selection*. Pemilihan dilakukan pada sampel acak dari *dataset* yang dimiliki.
b. *Decision Tree Construction*. Dengan menggunakan algoritma *Random Forest*, untuk setiap sampel yang dipilih akan dibuat pohon keputusan (*Decision Tree*).
c. *Vote*. Dari setiap hasil prediksi, kemudian dilakukan *voting*.
d. *Final Results*. Hasil prediksi dengan *voting* terbanyak akan dijadikan sebagai hasil akhir.

Pada penelitian ini komposisi pemisahan data *training* dan *testing* adalah 70% : 30%, atau *test size* = 0,30. Sehingga, diperoleh akurasi sebesar 0,980 dari model yang dibangun dengan menggunakan algoritma *Random Forest*.

3.2 Confussion Matrix

Confusion matrix digunakan untuk mengevaluasi kinerja model pengklasifikasi (*the classification model*) atau algoritma yang digunakan dalam penelitian ini. Kinerja *train classifier* dapat dievaluasi berdasarkan tingkat akurasi yang dapat dihitung dengan menggunakan *confusion matrix* [10]. Dengan menggunakan *confusion matrix*, kami membandingkan antara nilai aktual dengan nilai prediksi dari model *machine learning* yang dibuat dalam penelitian ini. Untuk menilai kualitas model yang dibangun juga dapat mengacu pada 3 metriks yaitu:

a. *Precision.* Sebagai representasi dari nilai prediksi positif yang benar relatif terhadap total prediksi positif;

JSI : Jurnal Sistem Informasi *(E-Journal), VOL.15, NO.1, April 2023*
ISSN Print : 2085-1588
ISSN Online :2355-4614
http://ejournal.unsri.ac.id/index.php/jsi/index
email: jsifasilkom@unsri.ac.id

b. *Recall.* Sebagai persentase dari prediksi nilai positif yang benar, relatif terhadap total nilai positif aktual;
c. *F1-Score.* F1-Score menunjukkan nilai rata-rata *Precision* dan *Recall*, dimana nilai yang terbaik mendekati 1 (satu).

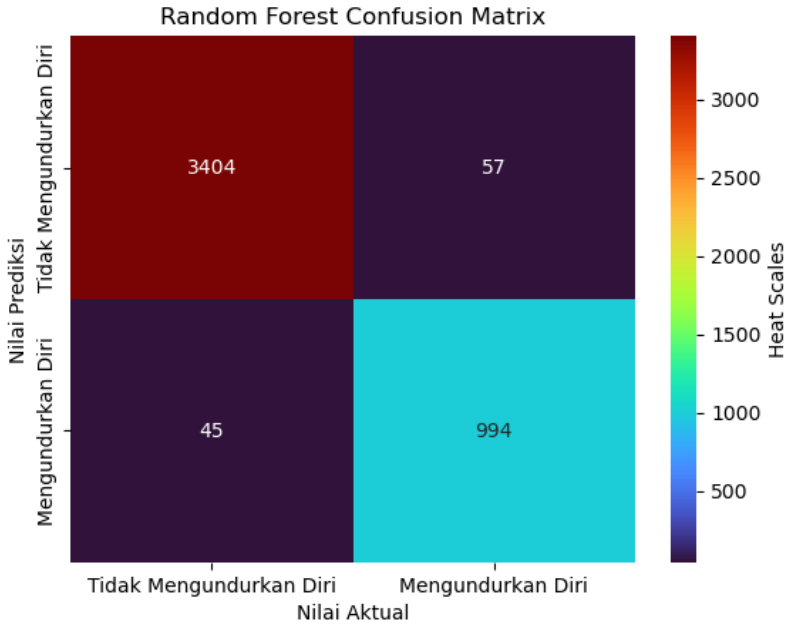
Precision dan *Recall* dapat menunjukkan anomali dalam klasifikasi data [11]. Keakuratan model yang dibangun dapat dihitung dengan menggunakan rumus berikut:

$$\text{Accuracy} = \frac{TN+TP}{TN+FP+FN+TP} \quad (1)$$

True Negative (TN) menunjukkan berapa banyak karyawan yang diprediksi "Tidak Mengundurkan Diri", dan menurut prediksi, para karyawan tersebut memang "Tidak Mengundurkan Diri". *True Positive* (TP) menunjukkan berapa banyak pegawai yang "Mengundurkan Diri" dan ternyata mengikuti prediksi akan "Mengundurkan Diri". *False Negative* (FN) menunjukkan berapa banyak karyawan yang "Tidak Mengundurkan Diri" tetapi diprediksi "Mengundurkan Diri". Terakhir, *False Positive* (FP) menunjukkan berapa banyak karyawan yang "Mengundurkan Diri", tetapi berdasarkan prediksi, mereka ternyata "Tidak Mengundurkan Diri".

Menggunakan fungsi Confusion_matrix yang terdapat pada *Scikit-Learn library* (Sklearn) Python terhadap algoritma *Random Forest*, dengan 4500 sampel digunakan sebagai *dataset* klasifikasi. Hasilnya, didapatkan 3404 data dengan nilai *True Positive* (TP), atau data yang terklasifikasi dengan benar menjadi data karyawan "Tidak Mengundurkan diri" pada model yang dibangun. Selanjutnya didapatkan 995 data dengan nilai *True Negative* (TN) atau data yang diklasifikasikan dengan benar menjadi data karyawan "Mengundurkan diri" pada model yang dibangun. Sebaliknya, terdapat 44 data yang terklasifikasi sebagai *False Positive* (FP) dan 57 data terklasifikasi sebagai *False Negative* (FN) merupakan bagian dari data yang salah klasifikasi. Deskripsi diatas dapat dilihat pada Gambar 3.

JSI : Jurnal Sistem Informasi *(E-Journal), VOL.15, NO.1, April 2023*
ISSN Print : 2085-1588
ISSN Online :2355-4614
http://ejournal.unsri.ac.id/index.php/jsi/index
email: jsifasilkom@unsri.ac.id



Gambar 3. *Confusion Matrix* Pada Algoritma *Random Forest*

Dengan menggunakan fungsi *classification_report* pada *sklearn library*, untuk keputusan 0 (nol) atau "Mengundurkan Diri" dengan *Precision*-nya 0,81, *Recall* 0,92 dan *F1-Score* 0,86 dengan *data support* sebanyak 3461. Sedangkan untuk keputusan 1 (satu) atau "Tidak Mengundurkan Diri", dengan nilai *Precision* 0,95, *Recall* 0,96, dan *F1-Score* 0,95, dengan *data support* sebanyak 1039 data. Secara lengkap *classification report* yang diperoleh dengan menggunakan sklearn library pada algoritma Random Forest, dapat dilihat pada Tabel 2.

Tabel 2. *Classification Report* Algoritma *Random Forest*

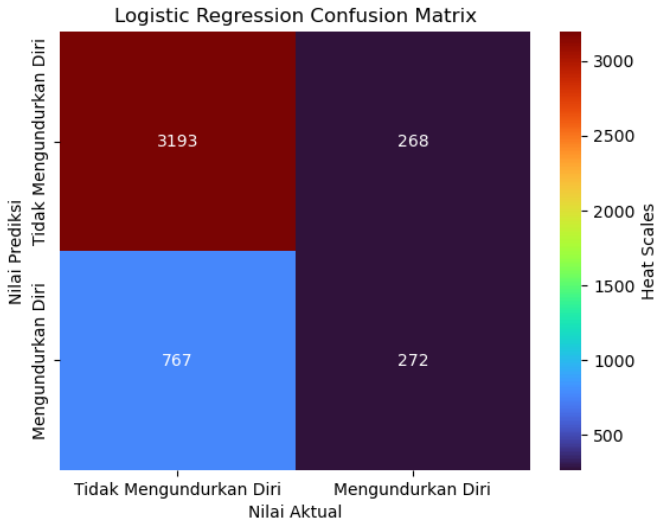
	precision	*recall*	*f1-score*	*support*
0	0.99	0.98	0.99	3461
1	0.95	0.96	0.95	1039
Accuracy			**0.98**	4500
Macro avg	0.97	0.97	0.97	4500
Weighted avg	0.98	0.98	0.98	4500

Dengan nilai *Precision* adalah 0,95, hal ini menunjukkan bahwa dari prediksi "Tidak Mengundurkan Diri" oleh model yang dibangun, hanya 5% yang tidak sesuai dengan prediksi ("Mengundurkan Diri"). Sedangkan untuk *recall* sebesar 0,96% menunjukkan bahwa model yang dibangun hanya gagal memprediksi secara tepat sebesar 4% dari data yang digunakan untuk evaluasi. Sementara itu nilai *F1-Score* diperoleh dari rumus:

JSI : Jurnal Sistem Informasi *(E-Journal), VOL.15, NO.1, April 2023*
ISSN Print : 2085-1588
ISSN Online :2355-4614
http://ejournal.unsri.ac.id/index.php/jsi/index
email: jsifasilkom@unsri.ac.id

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

Pada *F1-Score* yang menunjukkan angka mendekati 1 (satu), yaitu 0,95. Hal ini berarti bahwa model tersebut telah mampu memprediksi dengan baik karyawan yang "Mengundurkan Diri" atau "Tidak Mengundurkan Diri". Dengan *data support* sebesar 1039, hal ini menunjukkan bahwa jumlah set data uji dari 4500 set data uji yang termasuk dalam kelompok "Tidak Mengundurkan Diri". Sedangkan apabila menggunakan algoritma *Logistic Regression*, seperti yang terlihat pada gambar 4, Nilai Prediksi pada kondisi "Tidak Mengundurkan Diri" adalah 3193 sebagai *True Positive* dan *True Negative* adalah 272 pada kondisi "Mengundurkan Diri".



Gambar 4. *Confusion Matrix* Pada Algoritma *Logistic Regression*

Oleh karena itu, algoritma *Logistic Regression* menghasilkan akurasi yang lebih rendah dibandingkan dengan algoritma *Random Forest*. Tabel 3 memperlihatkan secara rinci *classification report* yang dihasilkan pada algoritma *Logistic Regression.*

Tabel 3. *Classification Report* Algoritma *Logistic Regression*

	precision	*recall*	*f1-score*	*support*
0	0.81	0.92	0.86	3461
1	0.50	0.26	0.34	1039
Accuracy			**0.77**	4500
Macro avg	0.66	0.59	0.60	4500
Weighted avg	0.74	0.77	0.74	4500

Berdasarkan *classification report*, nilai *Precision* sebesar 0,81 pada keputusan 0 (nol) dan 0,50 pada keputusan 1 (satu). Sedangkan *Recall* dan *F1-Score* pada keputusan 0

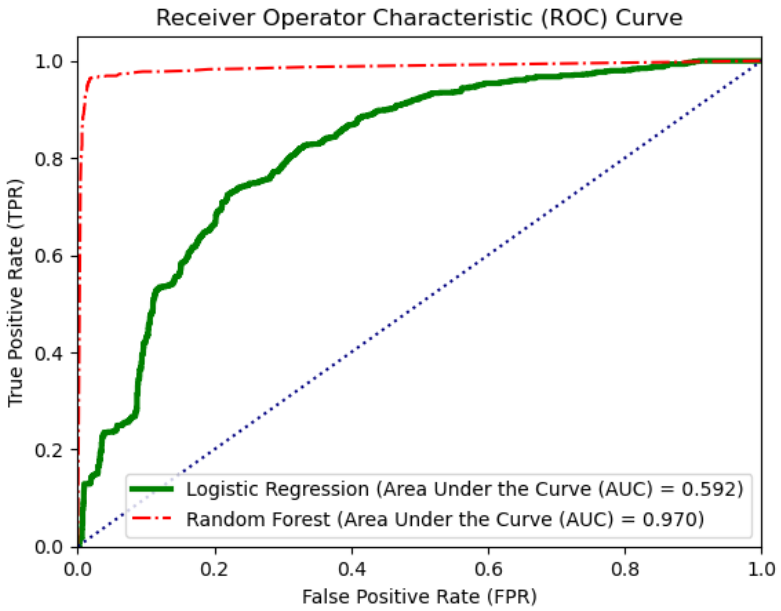
JSI : Jurnal Sistem Informasi *(E-Journal), VOL.15, NO.1, April 2023*
ISSN Print : 2085-1588
ISSN Online :2355-4614
http://ejournal.unsri.ac.id/index.php/jsi/index
email: jsifasilkom@unsri.ac.id

(nol) masing-masing sebesar 0,92 dan 0,86. Sedangkan untuk keputusan 1 (satu), nilai *Recall* dan *F1-Score* masing-masing adalah 0,26 dan 0,34.

3.3 Evaluasi Model

Untuk mengevaluasi model machine learning yang telah dibuat dapat digunakan *Area Under the Curve* (AUC) yang dikalkulasi dari *Receiver Operator Characteristic* (ROC). Grafik ROC digunakan untuk memvisualisasikan, mengatur, dan memilih algoritma klasifikasi berdasarkan kinerja masing-masing algoritma. AUC-ROC digunakan sebagai visualisasi performa model *machine learning*. Dengan membandingkan *True Positive Rate* (TPR) dan *False Positive Rate* (FPR) dari kedua algoritma yang digunakan pada suatu model *machine learning* [12]. TPR diperoleh dari rumus TP/TP+FN dari *Confusion Matrix* pada gambar 3 dan gambar 4.

TPR menunjukkan probabilitas bahwa nilai aktual yang positif juga akan menghasilkan nilai positif selama pengujian. Sebaliknya, FPR digunakan untuk mengukur akurasi pengujian berdasarkan model *machine learning*. Dimana FPR menunjukkan probabilitas atau proporsi dimana nilai negatif yang salah diidentifikasi sebagai nilai positif dari data. FPR diperoleh dari rumus FP/(FP+TN). Dengan menggunakan *matplotlib library* pada Python, diperoleh perbandingan AUC-ROC untuk algoritma *Logistic Regression* dengan nilai area dibawah kurva (*Area Under the Curve*/AUC) sebesar 0,592, sedangkan AUC untuk algortima *Random Forest* sebesar 0,971.



Gambar 5. Kurva AUC-ROC Pada Algoritma *Logistic Regression* dan Algoritma *Random Forest*

Berdasarkan kurva pada Gambar 5 terlihat bahwa AUC pada kurva ROC algoritma *Random Forest* lebih tinggi daripada kurva ROC pada algoritma *Logistic Regression*. AUC mewakili kemampuan model *machine learning* untuk memisahkan keputusan "Mengundurkan Diri" dan "Tidak Mengundurkan Diri". Nilai AUC maksimum adalah 1 (satu), yang berarti model yang diuji idealnya dapat memisahkan dua kelas yang berbeda; sebaliknya, AUC = 0 (nol) berarti model gagal dalam proses klasifikasi atau salah dalam mengklasifikasikan [13].

Dengan demikian, dapat dikatakan bahwa *Random Forest* memiliki kemampuan klasifikasi yang lebih baik daripada *Logistic Regression* terkait dengan *positive class* pada *dataset* yang diberikan. Dengan demikian, algoritma yang lebih akurat digunakan untuk memprediksi faktor pendorong pergantian karyawan, dengan melihat pada keputusan “Mengundurkan Diri” atau “Tidak Mengundurkan Diri” adalah algoritma *Random Forest*.

Selanjutnya, adalah penggunaan *feature selection* sebagai teknik untuk mereduksi dimensi data. Penggunaan *feature selection* yang relevan akan membantu peneliti menyingkirkan atribut yang tidak diperlukan sesuai dengan tujuan penelitian sehingga dapat memberikan hasil yang lebih cepat dan lebih baik [14]. Dengan menggunakan *feature importance* di *sklearn library*, diperoleh lima faktor pendorong teratas (*top driving factors*) berikut:

a. Tingkat kepuasan kerja (*satisfaction_level*), merupakan *driving factor* tertinggi dengan nilai 50.05%
b. Durasi masa kerja di perusahaan *(time_spend_company*), menjadi *driving factor* dengan nilai 27.14%.
c. Hasil evaluasi terakhir (*last_evaluation*), menjadi faktor yang memicu atau mendorong terjadinya pergantian karyawan dengan nilai 18.27%.
d. Kecelakaan kerja (*work_accident*). Menjadi *driving factor* lainnya dengan nilai 1.46%.
e. Gaji (*salary*). Gaji yang rendah (*low salary*) menjadi faktor pendorong dengan nilai 1.24%, dimana nilai ini lebih tinggi dari pada menjadi faktor pemicu pada karyawan dengan gaji yang tinggi (*high salary*) yang hanya bernilai 0.67% sebagai faktor pendorong terjadinya pergantian karyawan.

Dengan demikian model *machine learning* yang dibangun dengan algoritma *Random Forest* memiliki tingkat akurasi yang lebih tinggi dibandingkan dengan algoritma *Logistic Regression*, sehingga hasil pengujian dengan algoritma *Random Forest* dapat menjadi dasar dalam pengambilan keputusan terkait dengan faktor pendorong terjadinya pergantian karyawan. Dimana faktor utama penyebab terjadinya pergantian karyawan adalah tingkat kepuasan karyawan terhadap tempat kerja atau lingkungan kerja.

4. KESIMPULAN

Dengan pemodelan *machine learning* untuk klasifikasi, beberapa faktor yang mendorong pergantian karyawan terjadi dari *sample dataset* yang digunakan. Dengan membandingkan kedua model yang dibangun dengan algoritma *Logistic Regression* dan algoritma *Random Forest*, didapatkan akurasi yang lebih baik pada model yang dibangun dengan algoritma *Random Forest*. Berdasarkan model yang dibangun dari *Random*

Forest ini, diketahui bahwa faktor pemicu pergantian karyawan yang paling tinggi adalah tingkat kepuasan kerja karyawan. Kesimpulan ini sejalan dengan hasil beberapa penelitian terkait yang menunjukkan bahwa tingkat kepuasan kerja merupakan salah satu faktor utama yang mendorong pergantian karyawan secara masif [15]–[18]. Untuk penelitian selanjutnya, akurasi algoritma *Random Forest* juga dapat dibandingkan dengan algoritma *machine learning* lainnya seperti *Support Vector Machine* (SVM) untuk kondisi dengan jumlah data yang besar dan variabel *dataset* yang lebih banyak.

DAFTAR PUSTAKA

[1] W. A. Al-Suraihi, S. A. Samikon, A.-H. A. Al-Suraihi, and I. Ibrahim, "Employee Turnover: Causes, Importance and Retention Strategies," *European Journal of Business and Management Research*, vol. 6, no. 3, pp. 1–10, Jun. 2021, doi: 10.24018/ejbmr.2021.6.3.893.

[2] A. C. Kae, C. Xinying, J. O. Victor, and K. K. Wah, "Employee Turnover Prediction by Machine Learning Techniques," *Journal of Telecommunication, Electronic and Computer Engineering*, vol. 13, no. 4, pp. 49–56, 2021.

[3] M. Holliday, "What Is Employee Turnover & Why It Matters for Your Business," *Oracle / NetSuite*, Jan. 14, 2021. Accessed: Feb. 07, 2023. [Online]. Available: https://www.netsuite.com/portal/resource/articles/human-resources/employee-turnover.shtml

[4] A. Živković, J. Franjković, and D. Dujak, "The Role Of Organizational Commitment In Employee Turnover In Logistics Activities Of Food Supply Chain," *Logforum*, vol. 17, no. 1, pp. 25–36, 2021, doi: 10.17270/J.LOG.2021.536.

[5] X. Gao, J. Wen, and C. Zhang, "An Improved Random Forest Algorithm for Predicting Employee Turnover," *Math Probl Eng*, vol. 2019, 2019, doi: 10.1155/2019/4140707.

[6] H. Zhang, L. Xu, X. Cheng, K. Chao, and X. Zhao, "Analysis and Prediction of Employee Turnover Characteristics based on Machine Learning," in *ISCIT 2018 - 18th International Symposium on Communication and Information Technology*, Dec. 2018, pp. 433–437. doi: 10.1109/ISCIT.2018.8587962.

[7] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines," *Mach Learn*, vol. 46, pp. 389–422, 2002.

[8] Y. Zhao, M. K. Hryniewicki, F. Cheng, B. Fu, and X. Zhu, "Employee Turnover Prediction With Machine Learning: A Reliable Approach," in *Advances in Intelligent Systems and Computing*, 2018, vol. 869, pp. 737–758. doi: 10.1007/978-3-030-01057-7_56.

[9] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A Comparison Of Random Forest Variable Selection Methods For Classification Prediction Modeling," *Expert Syst Appl*, vol. 134, pp. 93–101, Nov. 2019, doi: 10.1016/j.eswa.2019.05.028.

[10] S. N. Khera and Divya, "Predictive Modelling of Employee Turnover in Indian IT Industry Using Machine Learning Techniques," *Vision*, vol. 23, no. 1, pp. 12–21, Mar. 2019, doi: 10.1177/0972262918821221.

[11] N. Tatbul, T. J. Lee, S. Zdonik, M. Alam, and J. Gottschlich, "Precision and Recall for Time Series," in *32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada*, 2018.

[12] S. A. Khan and Z. Ali Rana, "Evaluating Performance of Software Defect Prediction Models Using Area under Precision-Recall Curve (AUC-PR)," in *2019 2nd International Conference on Advancements in Computational Sciences, ICACS 2019*, Apr. 2019. doi: 10.23919/ICACS.2019.8689135.

[13] K. Hajian-Tilaki, "Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation," *Caspian J Intern Med*, vol. 4, no. 2, pp. 627–635, 2013.

[14] D. Jain and V. Singh, "Feature Selection And Classification Systems For Chronic Disease Prediction: A Review," *Egyptian Informatics Journal*, vol. 19, no. 3, pp. 179–189, Nov. 2018, doi: 10.1016/j.eij.2018.03.002.

[15] A. H. Khan and M. Aleem, "Impact Of Job Satisfaction On Employee Turnover: An Empirical Study Of Autonomous Medical Institutions Of Pakistan," *Journal of International Studies*, vol. 7, no. 1, pp. 122–132, 2014.

[16] A. Musawer, D. K. Amarkhil, and M. Laiq, "Factors Influencing Employees' Intention To Leave Job," *International Journal of Innovation in Engineering Research and Technology*, vol. 8, no. 2, 2021.

[17] V. K. Asimah, "Factors That Influence Labour Turnover Intentions In The Hospitality Industry In Ghana," 2018. [Online]. Available: http//:www.ajhtl.com

[18] C. A. al Mamun and M. N. Hasan, "Factors Affecting Employee Turnover and Sound Retention Strategies In Business Organization: A Conceptual View," *Problems and Perspectives in Management*, vol. 15, no. 1, pp. 63–71, 2017, doi: 10.21511/ppm.15(1).2017.06.