

## **APPLICATION OF MACHINE LEARNING IN PREDICTING EMPLOYEE DISCIPLINE VIOLATIONS IN FINANCIAL SERVICE COMPANY**

**Muhamad Fadel<sup>1</sup>, Kanasfi<sup>2</sup>, Arief Wibowo<sup>3</sup>**

<sup>1,2,3</sup>Master of Computer Science, Information and Technology Faculty Universitas Budi Luhur, Indonesia  
Email: <sup>1</sup>[2211600081@student.budiluhur.ac.id](mailto:2211600081@student.budiluhur.ac.id), <sup>2</sup>[2211601170@student.budiluhur.ac.id](mailto:2211601170@student.budiluhur.ac.id),  
<sup>3</sup>[arief.wibowo@budiluhur.ac.id](mailto:arief.wibowo@budiluhur.ac.id)

(Article received: July 20, 2023; Revision: August 5, 2023; published: February 13, 2024)

### **Abstract**

*Employee compliance is a commitment to comply with regulations and stay away from matters that are prohibited in the laws and or company regulations which if not obeyed, then employees are given disciplinary sanctions. Employee discipline is an obligation and willingness of employees in obeying all existing rules in a company to achieve its vision and mission, a high-level employee disciplinary violation rate of 38% at PT. HCI who are engaged in financial service sector can have a negative impact on a company's reputation, meanwhile a low level of employee disciplinary violations in a company can have a positive impact on the company's reputation. This paper aims to predict the possibility of employees committing discipline violations and evaluating the performance of accuracy by using Machine Learning Random Forest, Decision Tree, and Naive Bayes techniques. The test results prove that the Machine Learning Random Forest technique is the best model with the highest value in terms of accuracy with a value of 87.30%, while the Machine Learning Decision Tree and Naive Bayes technique has a value of 83.28% and 70.27% respectively, the value from each of the Machine Learning techniques, the comparison was made using majority voting techniques, so as to produce a total accuracy value of 85.31%. With this high accuracy value, the Random Forest model is proven to have better performance individually in analyzing the prediction of disciplinary violations in the application of human resources at company, while the total accuracy value uses a majority voting model of 85.31%, slightly decreased due to the high level of accuracy of the Naive Bayes model compared to other algorithm models.*

**Keywords:** *Decision Tree, Machine Learning, Majority Voting, Naive Bayes, Random Forest.*

## **PENERAPAN MACHINE LEARNING DALAM PREDIKSI PELANGGARAN KEDISIPLINAN KARYAWAN PADA PERUSAHAAN JASA KEUANGAN**

### **Abstrak**

Kepatuhan karyawan merupakan komitmen untuk mematuhi peraturan dan menjauhi hal-hal yang dilarang dalam undang-undang dan peraturan perusahaan yang apabila tidak dipatuhi, maka karyawan diberikan sanksi disiplin. Kedisiplinan karyawan adalah sebuah kewajiban dan kesediaan karyawan dalam mentaati segala aturan yang ada dalam suatu perusahaan dalam upaya pencapaian visi misinya, tingkat pelanggaran kedisiplinan karyawan yang tinggi sebesar 38% pada PT. HCI yang bergerak pada bidang jasa keuangan dapat berdampak negatif pada reputasi perusahaan, sementara itu tingkat pelanggaran kedisiplinan karyawan yang rendah pada sebuah perusahaan dapat berdampak positif pada reputasi perusahaan. Paper ini bertujuan untuk memprediksi kemungkinan karyawan melakukan pelanggaran kedisiplinan, dan mengevaluasi kinerja akurasi dengan menggunakan teknik *Machine learning Random Forest, Decision Tree, dan Naive Bayes*. Hasil pengujian membuktikan bahwa teknik *Machine Learning Random Forest* merupakan model terbaik dengan nilai tertinggi dalam hal akurasi dengan nilai 87,30%, sementara teknik *Machine Learning Decision Tree* dan *Naive Bayes* memiliki nilai masing-masing 83,28% dan 70,27%, nilai dari masing-masing teknik *Machine Learning* tersebut kemudian dilakukan perbandingan menggunakan teknik *Majority Voting*, sehingga menghasilkan nilai akurasi total sebesar 85,31%. Dengan nilai akurasi yang tinggi tersebut, model *Random Forest* terbukti memiliki kinerja yang lebih baik secara individu dalam menganalisis prediksi pelanggaran kedisiplinan pada aplikasi sumber daya manusia di perusahaan, sementara nilai akurasi total menggunakan model *Majority Voting* sebanyak 85,31%, sedikit menurun dikarenakan tingkat akurasi model *Naive Bayes* yang cukup rendah dibandingkan model algoritma lainnya.

**Kata kunci:** *Decision Tree, Machine Learning, Majority Voting, Naive Bayes, Random Forest.*

## 1. PENDAHULUAN

Karyawan berperan penting dalam operasional sebuah perusahaan, karyawan yang memiliki tingkat kedisiplinan yang tinggi dapat menjadi asset yang berharga, dan dapat menjadi contoh serta teladan bagi karyawan yang lain, sementara itu karyawan dengan tingkat kedisiplinan yang rendah dapat menjadi beban, serta menimbulkan resiko bagi perusahaan.

Perusahaan umumnya menggunakan penilaian kedisiplinan sebagai salah satu parameter dalam peningkatan gaji, kenaikan jabatan, tunjangan, atau pun sebagai parameter dalam demosi dan terminasi. Evaluasi kedisiplinan yang tepat dan akuntabel dapat menaikkan kepatuhan, loyalitas, dan semangat karyawan, sehingga objektif yang ditargetkan oleh perusahaan bisa terpenuhi dengan kerja sama semua pihak di antara karyawan dan manajemen.

Perusahaan umumnya menjalankan evaluasi kedisiplinan kepada karyawannya pada setiap periode yang telah ditentukan. Evaluasi kedisiplinan adalah penilaian menyeluruh yang dilaksanakan oleh *HR Compliance* kepada kedisiplinan karyawan guna mendapatkan kondisi karyawan sebenarnya secara utuh. Maka dari itu, evaluasi ini dilaksanakan guna mengukur dan mereview alasan, situasi, dan motivasi karyawan. Kedisiplinan dibuktikan dengan tingkat ketaatan seorang karyawan dalam waktu yang telah ditetapkan saat karyawan melakukan tugasnya sesuai dengan aturan perundang-undangan dan peraturan perusahaan.

Pelanggaran kedisiplinan merupakan pelanggaran yang dilakukan oleh karyawan terhadap aturan perundang-undangan dan peraturan perusahaan, pelanggaran yang dilakukan akan mendapatkan teguran berupa tindakan kedisiplinan.

Tindakan kedisiplinan adalah tindakan yang diambil oleh perusahaan kepada karyawan yang telah melakukan kelalaian atau melanggar aturan yang telah disusun pada peraturan perusahaan, undang-undang ketenagakerjaan Indonesia, etika dan norma sosial, serta setiap kebijakan atau prosedur yang dikeluarkan oleh perusahaan. Hal ini juga dapat dilakukan terhadap karyawan yang telah gagal memenuhi tujuan kerja (KPI, Target, Kinerja) yang ditetapkan oleh atasan. Jenis-jenis tindakan kedisiplinan yang dapat diambil oleh perusahaan adalah surat peringatan pertama, surat peringatan kedua, surat peringatan ketiga, dan terminasi.

Penelitian ini bertujuan untuk melakukan perbandingan dan memprediksi kemungkinan karyawan dalam melakukan pelanggaran dengan

menggunakan teknik *Machine Learning* *Random Forest*, *Decision Tree*, dan *Naive Bayes*, dan hasil akurasi masing-masing teknik algoritma dilakukan perbandingan menggunakan teknik *Majority Voting*.

Penelitian terdahulu mengenai *Majority Voting* yaitu Arief Wibowo yang berjudul penentuan kluster

koridor TransJakarta dengan metode majority voting pada algoritma data mining. Penelitian tersebut diproses melalui Microsoft Excel, Rapid Miner, dan Python menghasilkan algoritma *K-Means* memiliki tiga kluster optimal sebelum masa pandemi, dan dua kluster optimal selama masa pandemi, sementara algoritma *K-Medoids* memiliki tiga kluster optimal sebelum masa pandemi, dan empat kluster optimal selama masa pandemi. Limitasi dari penelitian ini adalah jumlah atribut yang digunakan kurang banyak [1]. Pada kasus penerapan teknik *Vote* menggunakan *C4.5 Naïve Bayes* dan *K-Nearest Neighbor* pada data gangguan autisme yang dibahas pada penelitian yang dilakukan oleh Arifandi Pratama, menghasilkan bahwa perhitungan sesudah menggunakan teknik *Majority Vote* didapatkan nilai akurasi sebesar 88.89%, dimana nilai positif gangguan interaksi sosial memiliki hasil 100%, nilai positif pada gangguan komunikasi sebesar 100%, dan pada nilai positif gangguan perilaku memiliki nilai presisi 66.67%. Oleh karena itu dengan penelitian yang telah dilakukan mendapatkan peningkatan akurasi sebesar 5.56% [2]. Pada kasus pemodelan prediksi status keberlanjutan polis asuransi kendaraan dengan teknik pemilihan mayoritas menggunakan algoritma-algoritma klasifikasi *Data Mining* yang dibahas pada penelitian yang dilakukan oleh Dyah Retno Utari, menghasilkan bahwa model *Heterogeneous Multiple Classifiers* memiliki kinerja yang lebih baik saat dilakukan pemodelan yang terdiri dari tiga jenis pengklasifikasi tunggal yang terdiri dari *C4.5*, *Naive Bayes* ataupun *Support Vector Machine*, sedangkan tingkat akurasi, presisi, dan *recall* dari setiap pengklasifikasi tunggal terbukti tidak lebih akurat dibandingkan performa dari *modeling* yang mengaplikasikan *Heterogeneous Multiple Classifier*, namun demikian secara satuan pengklasifikasi, metode *SVM* terbukti lebih unggul dalam hal kinerja dibanding dua pengklasifikasi lain yaitu *C4.5* maupun *Naive Bayes* [3].

Penelitian lain membahas penggunaan metode *Decision Tree* yaitu Lemi Iryani yang berjudul penerapan *Machine Learning* dalam klasifikasi kinerja pegawai PT. X. Penelitian tersebut diproses melalui Rapid Miner menghasilkan algoritma *Decision Tree* memiliki nilai akurasi keseluruhan 98.51% untuk nilai rata-rata karyawan untuk mendapatkan penilaian kinerja B atau Baik. Limitasi dari penelitian ini adalah metode yang digunakan hanya satu jenis [4]. Pada kasus *predicting student's satisfaction using a Decision Tree* yang dibahas pada penelitian yang dilakukan oleh Vesna Skrbinjek, menghasilkan algoritma *Decision Tree* memiliki nilai akurasi 64% untuk tingkat kepuasan tinggi dan 61% untuk tingkat kepuasan rendah. Limitasi dari penelitian ini adalah hasil yang bias, yang disebabkan oleh kurangnya data [5]. Pada kasus *decision tree algorithms for prediction of heart disease* yang dibahas pada penelitian yang dilakukan oleh Srabanti

Maji, menghasilkan bahwa algoritma *Hybrid Decision Tree* mendapatkan tingkat akurasi 78%, lebih baik dari tingkat akurasi algoritma C45 sebesar 76.66% dan *Artificial Neural Network* sebesar 77.40% [6]. Pada kasus teknik *Data Mining* menggunakan algoritma *Decision Tree* (C4.5) untuk prediksi seleksi beasiswa jalur KIP pada Universitas Muhammadiyah Kotabumi yang dibahas pada penelitian yang dilakukan oleh Khusnul Khotimah, menghasilkan bahwa klasifikasi data penerimaan beasiswa jalur KIP calon mahasiswa dengan menggunakan teknik *Machine Learning Decision Tree* (C4.5) merupakan teknik *Machine Learning* terbaik. Teknik *Machine Learning Decision Tree* (C4.5) nilai akurasi sebesar 100 %, sementara itu teknik *Machine Learning Naive Bayes* sebesar 90,16%. Limitasi dari penelitian ini adalah kurangnya data [7].

Penelitian lain membahas penggunaan metode *Naive Bayes* yaitu Sonar Priyanka yang berjudul *diabetes prediction using different Machine Learning approaches*. Penelitian tersebut dilakukan menggunakan tools Microsoft Excel dan Rapid Miner memberikan hasil algoritma *Naive Bayes* memiliki nilai akurasi 80% untuk class 0 dan 58% untuk class 1, algoritma *Naive Bayes* mampu menilai *missing values* dengan mengabaikan probabilitas estimasi perhitungan. Limitasi penelitian ini yaitu kurangnya jumlah dataset [8]. Pada kasus prediksi tingkat kelancaran pembayaran kredit bank menggunakan algoritma *Naive Bayes* berbasis *Forward Selection* yang dibahas pada penelitian yang dilakukan oleh Maryam Hasan, menghasilkan teknik *Machine Learning Naive Bayes based on Forward Selection* memiliki tingkat akurasi 71.97%, sementara teknik *Machine Learning Naive Bayes* yang tidak dilandasi *feature selection* memiliki nilai akurasi 68.42% [9]. Pada kasus algoritma *Naive Bayes* untuk prediksi kepuasan pelayanan perekaman e-KTP yang dibahas pada penelitian yang dilakukan oleh Tri Herdiawan Apandi, menghasilkan teknik *Machine Learning Naive Bayes* menghasilkan tingkat akurasi 91.70% dan tingkat pengujian *f-measure* sebanyak 93.92%, sementara tingkat akurasi teknik *Machine Learning Decision Tree* adalah sebanyak 65.90% dan tingkat pengujian *f-measure* sebanyak 79.26% [10]. Pada kasus implementasi *Data Mining* untuk evaluasi kinerja akademik mahasiswa menggunakan algoritma *Naive Bayes Classifier* yang dibahas pada penelitian yang dilakukan oleh M. Syukri Mustafa, menghasilkan teknik *Machine Learning Naive Bayes* menghasilkan tingkat akurasi 92.30%. Limitasi dari penelitian ini adalah metode klasifikasi yang diimplementasikan hanya satu metode [11].

Penelitian lain membahas penggunaan metode *Random Forest* yaitu Joseph Sanjaya yang berjudul prediksi kelalaian pinjaman bank menggunakan *Random Forest* dan *Adaptive Boosting*. Penelitian tersebut diproses melalui Microsoft Excel dan Python menghasilkan algoritma *Random Forest* memiliki

nilai akurasi tertinggi Ketika dilakukan *undersampling*. Limitasi dari penelitian ini adalah data yang digunakan menggunakan data indikasi *risk loan default* tidak real [12]. Pada kasus prediksi harga Bitcoin menggunakan metode *Random Forest* yang dibahas pada penelitian yang dilakukan oleh Siti Saadah, menghasilkan algoritma *Random Forest* memiliki nilai akurasi 98% dengan menggunakan data acak. Sehingga merujuk pada hasil prediksi ini dapat ditarik kesimpulan bahwa untuk memprediksi data yang fluktuatif seperti Bitcoin, *Random Forest* membuktikan sukses memberikan fitting yang sesuai dengan data sesungguhnya [13]. Pada kasus perbandingan kinerja algoritma klasifikasi *Naive Bayes*, *Support Vector Machine (SVM)*, dan *Random Forest* untuk prediksi ketidakhadiran di tempat kerja yang dibahas pada penelitian yang dilakukan oleh Hiyu Natalissifa, menghasilkan teknik *Machine Learning Random Forest* memberikan hasil tingkat akurasi, *recall*, dan presisi yang paling tinggi dibandingkan dengan teknik *Machine Learning Naive Bayes* dan *SVM*, yaitu dengan hasil akurasi sebesar 99.38%, presisi 99.42%, dan *recall* 99.39% [14]. Pada kasus prediksi kemungkinan diabetes pada tahap awal menggunakan algoritma klasifikasi *Random Forest* yang dibahas pada penelitian yang dilakukan oleh Widya Apriliah, menghasilkan teknik *Machine Learning Random Forest* memiliki tingkat akurasi testing yaitu sebesar 97.88%, sedangkan algoritma *Naive Bayes* memiliki nilai akurasi sebesar 91.92%, dan algoritma *SVM* memiliki nilai akurasi sebesar 94.80%. Limitasi dari penelitian ini adalah kurangnya jumlah data [15].

Adapun yang membedakan dari penelitian ini dibandingkan dengan penelitian-penelitian sebelumnya adalah dari sisi dataset yang digunakan yaitu data karyawan sebuah perusahaan jasa keuangan, dan teknik yang digunakan yaitu membandingkan nilai akurasi penerapan algoritma *Decision Tree*, *Naive Bayes*, dan *Random Forest*, lalu hasilnya diterapkan teknik *Majority Voting*, yang kemudian menghasilkan nilai akurasi total.

## 2. METODE PENELITIAN

Dalam tahap penelitian ini, metodologi yang dipakai yaitu SEMMA [4]. Metodologi ini dipakai sebagai system development dan metodologi pengamatan saat data collection.

SEMMA merupakan kepanjangan dari *Sample, Explore, Modify, Model, dan Assess*. Memiliki makna yang mengacu pada langkah-langkah pada objek *data mining*. Metodologi SEMMA ini memiliki lima langkah dalam *system development*. Berikut ini merupakan langkah-langkah pada metodologi SEMMA yang diaplikasikan pada paper ini:

### 1. Data Collection (Sample)

Proses *Data Collection* dilakukan dengan pengunduhan data dari system HR perusahaan ke dalam file Microsoft Excel

### 2. Data Description (Explore)

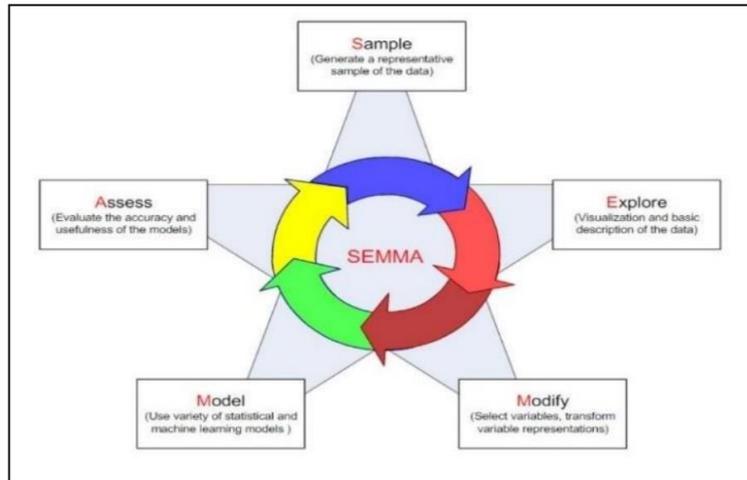
Proses *Data Description* dilakukan dengan cara membagi data karyawan ke dalam beberapa sheet berdasarkan masing-masing atribut yang dibutuhkan, yaitu data gaji terakhir, jenis kelamin, jumlah tanggungan, status pernikahan, Pendidikan terakhir, dan usia.

3. *Data Transformation (Modify)*

Proses *Data Transformation* dilakukan dengan cara menggabungkan beberapa data karyawan

dengan atribut yang berbeda ke dalam sebuah sheet gabungan, nilai yang dijadikan *key* adalah nomor karyawan, selanjutnya dilakukan proses pembersihan untuk data-data yang tidak konsisten atau tidak relevan.

Pada gambar 1 dapat dilihat langkah-langkah pada metodologi SEMMA yang diaplikasikan pada penelitian ini.



Gambar 1. Tahapan Metode SEMMA [4]

4. *Data Modelling (Model)*

Proses *Data Modelling* dilakukan menggunakan tools Rapid Miner, dimana dataset diimport dan dibagi menjadi dua data, yaitu data training dan data testing dengan masing-masing komposisi 80-20.

5. *Data Evaluation (Assess)*

Proses *Data Evaluation* dilakukan menggunakan tools Rapid Miner, dimana kinerja algoritma diukur menggunakan operator *Performance Classification*.

3. HASIL DAN PEMBAHASAN

Langkah-langkah untuk melaksanakan proses data mining mengikuti metodologi SEMMA (*Sample, Explore, Modify, Model, and Assess*), dimulai dari tahap *sampling* dataset sampai dengan *assessment*.

Pada tahap *sampling*, diputuskan menggunakan data karyawan perusahaan per bulan Juni 2023, data diunduh ke dalam format Microsoft Excel.

Tabel 1. Dataset Pribadi Karyawan

Employee No	Nama	Gender	Tanggal Lahir	Status Pernikahan
10000092	I Gusti Ayu Desy	Wanita	11/12/1994	Single
10000093	Aslami	Pria	8/28/1984	Menikah
10000094	Panris	Pria	4/20/1992	Menikah
10000096	Jusniar	Wanita	5/3/1998	Single
10000098	Hadi	Pria	3/19/1976	Menikah
10000100	Abdurrahman	Pria	4/26/1974	Menikah
10000111	Juwita	Wanita	8/16/1989	Single
10000116	Pasini	Wanita	11/25/1989	Single
10000117	Sudiawan	Pria	9/2/1987	Menikah
10000119	Ardila	Wanita	11/27/1994	Menikah

Pada table 1 dapat dilihat contoh dataset pribadi karyawan yang berisi informasi nomor karyawan, nama, jenis kelamin, tanggal lahir, dan status pernikahan.

Pada tahap *exploring*, diputuskan untuk mengunduh data karyawan dari masing-masing table

ke dalam beberapa sheet berdasarkan masing-masing atribut, yaitu data gaji terakhir, jenis kelamin, jumlah tanggungan, status pernikahan, Pendidikan terakhir, dan usia.



Gambar 2. Sheet Pembagian Data

Pada gambar 2 dapat dilihat bahwa dataset karyawan sudah dibagi sesuai dengan atributnya

masing-masing, yaitu data keluarga, gaji terakhir, data pribadi, data pendidikan, data tanggal

bergabung, data surat peringatan kesatu, surat peringatan kedua, dan surat peringatan ketiga.

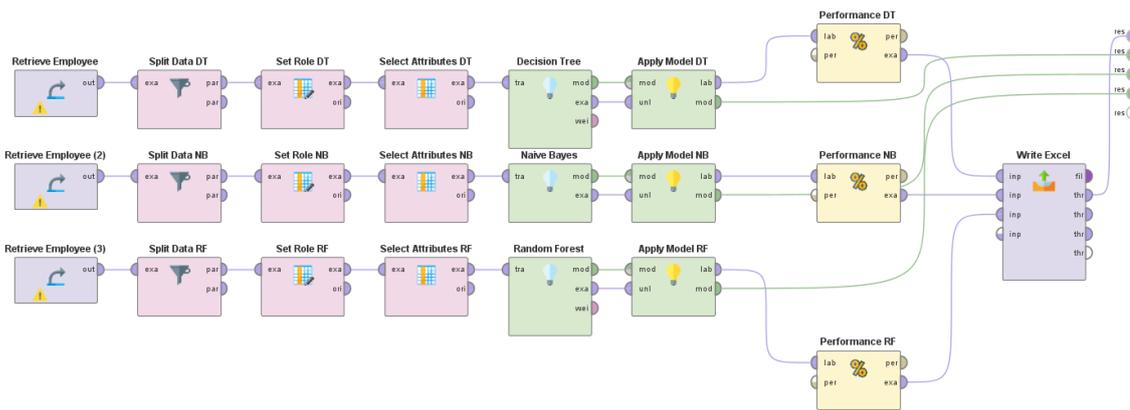
Pada tahap *modifying*, dilakukan transformasi data menjadi sebuah format yang bisa diterima oleh Rapid Miner untuk kebutuhan analisa.

Tabel 2. Dataset Pegawai Setelah dilakukan Transformasi

Employee No	Nama	Posisi	Join Date	Pendidikan	Gender	Status	Tanggungan	Gaji	Usia	SP 1	SP 2	SP 3
10000092	I Gusti	Sales Agent	6/1/2018	SMA	Wanita	Single	0	3,617,236	29	X		X
10000093	Aslami	Collection	6/19/2019	S1	Pria	Menikah	4	10,150,000	39	X		
10000094	Panris	Sales Agent	4/1/2018	SMA	Pria	Menikah	4	4,383,612	31	X	X	
10000096	Jusniar	Sales Agent	6/1/2019	SMA	Wanita	Single	0	4,095,197	25	X		X
10000098	Hadi	Territory	6/8/2015	S1	Pria	Menikah	3	20,296,841	47	X		
10000100	Rahman	Regional	8/14/2017	S1	Pria	Menikah	4	7,148,869	49			
10000111	Juwita	Sales Agent	6/1/2018	D3	Wanita	Single	0	3,967,555	34	X		X
10000116	Pasini	Sales Agent	2/1/2019	SMA	Wanita	Single	0	6,487,929	34	X		
10000117	Sudiawan	Investigator	1/25/2018	S1	Pria	Menikah	4	7,852,464	36			
10000119	Ardila	Sales Agent	3/1/2018	SMA	Wanita	Menikah	2	4,409,568	29			

Pada tabel 2 dapat dilihat bahwa dataset karyawan sudah digabung menjadi sebuah dataset gabungan yang berisi semua informasi yang dibutuhkan untuk dianalisa menggunakan Rapid Miner, proses penggabungan menggunakan fungsi *vlookup* pada Microsoft Excel, dimana nomor karyawan dijadikan sebagai *unique identifier* antar sheet.

Pada tahap *modelling*, dilakukan perancangan dan pengunggahan dataset karyawan perusahaan ke dalam tools Rapid Miner, rancangan dimulai dari pembagian data latih dan data uji dengan masing-masing komposisi 80-20, selanjutnya diaplikasikan algoritma *Decision Tree*, *Naïve Bayes*, dan *Random Forest*.

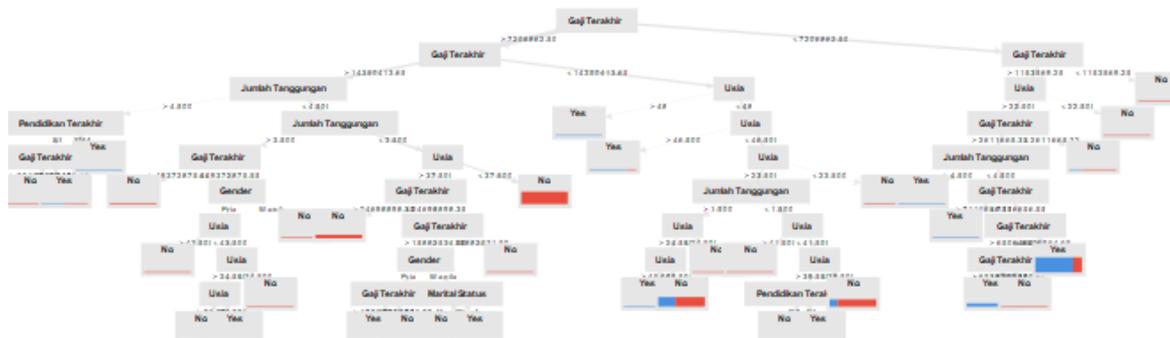


Gambar 3. Desain Rapid Miner

Pada gambar 3 dapat dilihat bahwa rancangan Rapid Miner dimulai dari proses *retrieve employee*, lalu data dibagi menjadi data latih dan data uji dengan komposisi 80-20, selanjutnya ditentukan *set role*, yaitu target yang ingin diprediksi “memiliki surat peringatan”, kemudian ditentukan attribute yang mempengaruhi yaitu gaji terakhir, jenis kelamin, jumlah tanggungan, status pernikahan, Pendidikan

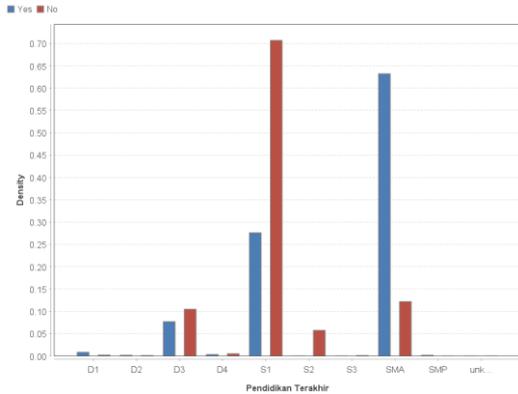
terakhir, dan usia, lalu data akan dihubungkan ke setiap algoritma *Machine Learning* yang digunakan, yaitu algoritma *Random Forest*, *Decision Tree*, dan *Naive Bayes*.

Pada tahap *assessing*, dilakukan pembahasan tentang hasil dan temuan percobaan, kemudian dilakukan pengukuran kinerja menggunakan operator *performance classification*.



Gambar 4. Hasil Algoritma Decision tree

Pada gambar 4 dapat dilihat ternyata karyawan yang diprediksi mendapatkan surat peringatan adalah karyawan dengan gaji  $\leq 6,506,464$ , usia  $> 22$  tahun, dan jumlah tanggungan  $\leq 4$ .



Gambar 5. Hasil Algoritma Naïve Bayes Pendidikan Terakhir

Pada gambar 5 dapat dilihat ternyata latar belakang pendidikan karyawan yang diprediksi mendapatkan surat peringatan mayoritas adalah lulusan SMA, sementara itu latar belakang pendidikan karyawan yang diprediksi paling tinggi tidak mendapatkan surat peringatan adalah lulusan S1.

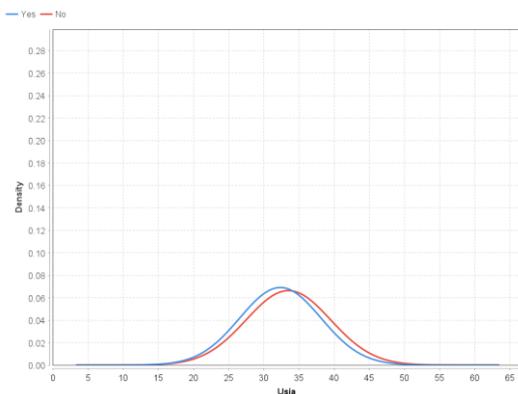


Table View  Plot View

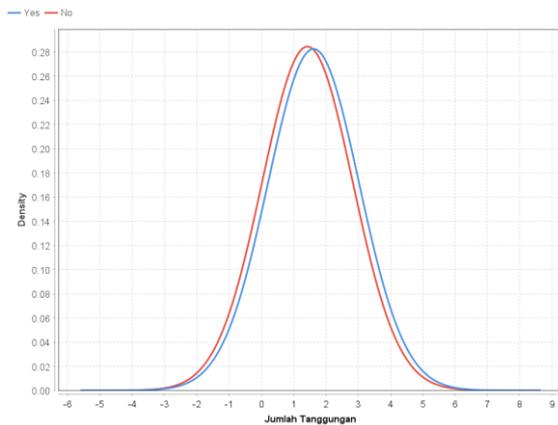
accuracy: 87.30%

	true Yes	true No	class precision
pred. Yes	481	85	84.98%
pred. No	117	908	88.59%
class recall	80.43%	91.44%	

Gambar 8. Hasil Akurasi Algoritma Random Forest

Gambar 6. Hasil Algoritma Naïve Bayes Usia

Pada gambar 6 dapat dilihat ternyata latar belakang usia karyawan yang diprediksi mendapatkan surat peringatan mayoritas berusia antara 30-35 tahun, sementara itu latar belakang usia karyawan yang diprediksi paling tinggi tidak mendapatkan surat peringatan adalah berusia 32-37 tahun



Gambar 7. Hasil Algoritma Naïve Bayes Jumlah Tanggungan

Pada gambar 7 dapat dilihat ternyata latar belakang jumlah tanggungan karyawan yang diprediksi mendapatkan surat peringatan mayoritas sebanyak antara 1-2 orang, cukup seimbang dengan latar belakang jumlah tanggungan karyawan yang diprediksi paling tinggi tidak mendapatkan surat peringatan adalah sebanyak 1-2 orang.

Pada gambar 8 dihasilkan *scoring class recall* pada nilai “Yes” sebanyak 80,43%, *class recall* untuk nilai “No” sebesar 91,44%. *Class precision* untuk nilai “Yes” sebesar 84,98%, *class precision* untuk nilai “No” sebesar 88,59%. Dan diperoleh nilai akurasi secara keseluruhan sebesar 87,30%.

Tabel 3. Hasil Perbandingan Akurasi Algoritma dan Voting

Pendidikan	Tanggungan	Gaji	Usia	Peringatan	Decision Tree	Naïve Bayes	Random Forest	Majority Voting
SMA	4	4,383,612	31	Yes	Benar	Benar	Benar	3
SMA	0	4,095,197	25	Yes	Benar	Benar	Benar	3
S1	3	20,296,841	47	Yes	Benar	Salah	Benar	2
D3	0	3,967,555	34	Yes	Benar	Benar	Benar	3
SMA	2	4,409,568	29	No	Salah	Salah	Salah	0

S1	2	4,529,971	32	Yes	Benar	Benar	Benar	3
S1	0	10,150,000	34	Yes	Salah	Benar	Salah	1
S1	0	4,383,612	29	Yes	Benar	Benar	Benar	3
S1	0	4,383,612	34	Yes	Benar	Benar	Benar	3
S1	1	6,382,607	28	No	Salah	Salah	Salah	0

Pada table 3 dapat dilihat perbandingan prediksi masing-masing algoritma dan nilai akurasi dari teknik *Majority Voting* terhadap dataset pegawai yang memiliki surat peringatan, jika prediksi ketiga algoritma menyatakan hasil prediksi bernilai benar, maka akan menghasilkan nilai *Majority Voting* sebanyak 3. Bila hasil prediksi memberikan keputusan dua nilai benar dan satu hasil salah, maka akan menghasilkan nilai *Majority Voting* sebanyak 2. Bila hanya satu prediksi algoritma yang menghasilkan nilai benar, maka akan menghasilkan nilai *Majority Voting* sebanyak 1, sedangkan jika tidak ada satu pun prediksi algoritma memberikan keputusan nilai benar, maka akan menghasilkan nilai *Majority Voting* sebanyak 0.

Selanjutnya nilai *Majority Voting* keseluruhan dihitung dengan kondisi  $\geq 2$ , yaitu prediksi *Majority Voting* benar, lalu jumlah record dengan prediksi *Majority Voting* benar dibagi dengan jumlah record yang ada pada sample dataset, sehingga didapatkan nilai persentase total akurasi *Majority Voting* sebesar 85,31%.

#### 4. DISKUSI

Kedisiplinan karyawan merupakan hal penting dalam sebuah perusahaan, tingkat kepatuhan yang tinggi akan membantu perusahaan dalam menggapai objektif dan target yang telah ditetapkan oleh manajemen. Namun, tidak bisa dipungkiri bahwa ada karyawan yang melakukan pelanggaran kedisiplinan, oleh karena itu menjadi hal yang sangat esensial bagi perusahaan untuk melakukan manajemen kedisiplinan karyawan yang efektif, sehingga tingkat pelanggaran kedisiplinan karyawan dapat ditekan, dan operasional perusahaan dapat berjalan dengan baik.

Penelitian sebelumnya yang membahas teknik *Majority Voting* seperti yang sudah dijelaskan di bagian pendahuluan, telah mengeksplorasi beberapa metode seperti *Clustering K-Means*, *K-Medoids*, *C.45*, *Naïve Bayes*, *K-nearest Neighbor (KNN)*. Namun, terdapat limitasi penelitian pada area dataset yang digunakan, secara spesifik terkait dengan pengelompokan data koridor TransJakarta sebelum masa pandemi Covid-19 dan selama masa pandemi Covid-19, dengan hanya metode klasterisasi *K-Means* dan *K-Medoids*.

Penelitian ini mengadopsi metodologi SEMMA (*Sample, Explore, Modify, Model, and Assess*), yang memiliki tahap penelitian dimulai dari pengumpulan data, eksplorasi data, transformasi data, data modeling, sampai dengan pengukuran kinerja, serta memanfaatkan data yang diperoleh dari sistem HR perusahaan.

Teknik algoritma yang digunakan melibatkan prediksi *Decision Tree*, *Naive Bayes*, dan *Random Forest*, untuk memperkirakan kemungkinan karyawan melakukan pelanggaran kedisiplinan, dan teknik *Majority Voting* digunakan membandingkan nilai akurasi masing-masing algoritma, dan mendapatkan nilai akurasi total. Dataset awal 29.241 baris data dan 130 kolom kemudian dibersihkan dan dianalisis menggunakan Rapid Miner.

Hasil pemodelan yang berasal dari analisis dievaluasi secara menyeluruh, yang menghasilkan perumusan hasil yang konklusif. Dari hasil analisis dengan menggunakan teknik *Decision Tree*, *Naive Bayes*, *Random Forest*, dan *Majority Voting*, diperoleh prediksi untuk pelanggaran kedisiplinan karyawan berdasarkan pendidikan terakhir, gender, status pernikahan, jumlah tanggungan, gaji terakhir, dan usia.

Hasil prediksi teknik *Random Forest* memiliki nilai akurasi tertinggi 87.30%. Sedangkan setelah diaplikasikan teknik majority voting, nilai akurasi total menjadi 85.31% sedikit lebih rendah dikarenakan nilai akurasi *Naive Bayes* yang rendah sebanyak 70.27%..

#### 5. KESIMPULAN

Penerapan beberapa teknik algoritma pada data karyawan pada perusahaan yang bergerak pada bidang jasa keuangan dapat menghasilkan tingkat akurasi yang berbeda-beda, yaitu teknik *Machine Learning Naïve Bayes* sebanyak 70,27%, *Decision Tree* sebanyak 83,28%, dan *Random Forest* sebanyak 87,30%. Hasil akurasi antar teknik algoritma tersebut kemudian dilakukan perbandingan, selanjutnya dilakukan evaluasi penilaian menggunakan teknik *Majority Voting* untuk mendapatkan nilai akurasi total terhadap seluruh teknik algoritma. Hasilnya didapatkan nilai akurasi total sebanyak 85,31%.

Pada masa yang akan datang, penelitian ini dapat dilanjutkan dengan penambahan attribute lain misalnya pangkat, golongan, dan masa kerja karyawan, sehingga data yang dihasilkan bisa lebih optimal.

Dengan penambahan attribute tersebut, diharapkan perusahaan dapat mendapatkan analisis yang lebih tajam, terkait dengan evaluasi tindakan pencegahan terhadap karyawan, pembinaan, dan menjaga tingkat kedisiplinan.

#### DAFTAR PUSTAKA

- [1] R. Sistem, A. Wibowo, M. Makruf, I. Virdyna, and F. C. Venna, "Penentuan Klaster Koridor TransJakarta dengan Metode Majority Voting," vol. 1, no. 10, pp. 565–

- 575, 2021.
- [2] A. Pratama *et al.*, “Jurnal Smart Teknologi Penerapan Teknik Vote Menggunakan C4 . 5 Naïve Bayes Dan K-nearest Neighbor Pada Data Gangguan Autisme Application of Vote Technique Using C4 . 5 Naïve Bayes and K-nearest Neighbor on Autism Disorder Data Jurnal Smart Teknologi,” vol. 3, no. 6, pp. 638–645, 2022.
- [3] D. R. Utari and A. Wibowo, “Pemodelan Prediksi Status Keberlanjutan Polis Asuransi Kendaraan dengan Teknik Pemilihan Mayoritas Menggunakan Algoritma-Algoritma Klasifikasi Data Mining,” vol. 5, no. 2502, pp. 19–24, 2020, doi: 10.22236/teknoka.v5i.391.
- [4] L. Iryani, “Penerapan Machine Learning Dalam Klasifikasi Kinerja Pegawai Pt X,” *J. Informanika*, vol. 09, no. 01, pp. 1–6, 2023, doi: <https://doi.org/10.52233/informanika.v9i01.391>.
- [5] V. Skrbinjek and V. Dermol, “Predicting students’ satisfaction using a decision tree,” no. 1994, 2019.
- [6] S. Maji and S. Arora, *Decision Tree Algorithms for Prediction of Heart Disease*. Springer Singapore, 2019. doi: 10.1007/978-981-13-0586-3.
- [7] K. Khotimah, “Teknik Data Mining menggunakan Algoritma Decision Tree (C4.5) untuk Prediksi Seleksi Beasiswa Jalur KIP pada Universitas Muhammadiyah Kotabumi,” *J. SIMADA (Sistem Inf. dan Manaj. Basis Data)*, vol. 4, no. 2, pp. 145–152, 2022, doi: 10.30873/simada.v4i2.3064.
- [8] P. Sonar and K. Jayamalini, “Diabetes Prediction Using Different Machine Learning Approaches,” *2019 3rd Int. Conf. Comput. Methodol. Commun.*, no. Iccmc, pp. 367–371, 2019.
- [9] M. Hasan, “Prediksi Tingkat Kelancaran Pembayaran Kredit Bank Menggunakan Algoritma Naive Bayes Berbasis Forward Selection,” vol. 9, pp. 317–324, 2017.
- [10] C. A. S. Tri Herdiawan Apandi, “Algoritma Naive Bayes untuk Prediksi Kepuasan Pelayanan Perekaman e-KTP,” *JUITA J. Inform. e-ISSN 2579-9801, Vol. 7, Nomor 2*, vol. 7, 2019.
- [11] M. Syukri Mustafa, M. Rizky Ramadhan, and A. P. Thenata, “Implementasi Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier,” *Citec J.*, vol. 4, no. 2, pp. 151–162, 2017.
- [12] J. Sanjaya, E. Renata, V. E. Budiman, F. Anderson, and M. Ayub, “Prediksi Kelalaian Pinjaman Bank Menggunakan Random Forest dan Adaptive Boosting,” vol. 6, no. April, pp. 50–60, 2020.
- [13] T. Online, S. Kasus, D. Acak, P. Awal, and M. Pandemic, “Jurnal Politeknik Caltex Riau Prediksi Harga Bitcoin Menggunakan Metode Random Forest,” vol. 7, no. 1, pp. 24–32, 2021.
- [14] H. Nalatissifa, W. Gata, S. Diantika, and K. Nisa, “Perbandingan Kinerja Algoritma Klasifikasi Naive Bayes , Support Vector Machine ( SVM ), dan Random Forest untuk Prediksi Ketidakhadiran di Tempat Kerja,” vol. 5, no. 4, pp. 578–584, 2021.
- [15] W. Apriliah, I. Kurniawan, M. Baydhowi, and T. Haryati, “SISTEMASI: Jurnal Sistem Informasi Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest,” *J. Sist. Inf.*, vol. 10, pp. 163–171, 2021, [Online]. Available: <http://sistemasi.ftik.unisi.ac.id>.