

Prediksi *Customer Retention* Perusahaan Asuransi Menggunakan *Machine Learning*

Riyan Pratama^[1], Muhammad Izman Herdiansyah^{[2]*}, Dedy Syamsuar^[3], Ahmad Syazili^[4]

Program Studi Magister Teknik Informatika^{[1], [2], [3], [4]}

Universitas Bina Darma

Palembang, Sumatera Selatan, Indonesia

rpratama.sky@gmail.com^[1], m.herdiansyah@binadarma.ac.id^[2],

dedy_syamsuar@binadarma.ac.id^[3], syazili@binadarma.ac.id^[4]

Abstrak— The application of data mining technique currently is widely used in supporting business activity, especially for insurance company which has to analyze a big number of customer data. The insurance company has to predict its new customer acquisition as well as maintain its existing customers. This paper is focused on how we support insurance companies, especially PT. XYZ, to analyze their customers' characteristics data using the best data mining algorithm technique. The research aims to analyze existing customer data and to predict as well as find optimal patterns of how many of their customers will extend their insurance policies, and how many will not. We also explore the customer retention rate discovering the optimal solution for the company. We applied 4 different algorithms were applied, i.e. support vector machine algorithm, decision tree, k-NN, and random forest algorithm, comparing the results and finding a better solution. From the analysis, we found that the random forest algorithm provides better results in predicting the status of the insurance policy extension of current customers, with an accuracy rate of 91.08% and AUC value of 0.962. This result is quite good for PT XYZ, and could be enhanced in the future by applying a good strategy to increase their customer renewal ratio.

Keywords— *Data Mining, Machine Learning, Customer Retention*

Abstrak— Teknik *data mining* dewasa ini sudah banyak digunakan dalam mendukung aktivitas bisnis, terutama bagi perusahaan asuransi yang perlu menganalisis data nasabah dalam jumlah besar. Perusahaan asuransi harus memprediksi akuisisi pelanggan baru mereka serta untuk mempertahankan pelanggan mereka yang sudah ada. Penelitian ini berfokus pada bagaimana peneliti mendukung perusahaan asuransi, dalam hal ini PT XYZ, untuk menganalisis data karakteristik pelanggan mereka menggunakan teknik algoritma data *mining* yang terbaik. Penelitian ini bertujuan untuk menganalisis data pelanggan yang ada dan untuk memprediksi serta menemukan pola optimal tentang berapa banyak pelanggan mereka yang akan memperpanjang polis asuransi mereka, dan berapa banyak yang tidak akan memperpanjang. Peneliti juga mengkaji tingkat retensi pelanggan untuk menemukan solusi terbaik bagi perusahaan. Peneliti menerapkan 4 algoritma berbeda dalam pemodelan data, yaitu *support vector machine algorithm, decision tree, k-NN, dan random forest algorithm*, membandingkan hasilnya dan menemukan solusi yang lebih

baik. Dari hasil analisis, dapat dinyatakan bahwa algoritma *random forest* memberikan hasil yang lebih baik dalam memprediksi status perpanjangan polis asuransi nasabah saat ini, dengan tingkat akurasi 91,08% dan nilai AUC sebesar 0,962. Hasil ini cukup baik untuk PT XYZ, dan dapat ditingkatkan di masa depan dengan menerapkan strategi yang baik untuk meningkatkan rasio pembaruan pelanggan mereka.

Kata Kunci— *Data Mining, Machine Learning, Customer Retention*

I. PENDAHULUAN

Kemajuan serta perkembangan teknologi akan mengikuti kebutuhan dan aktivitas manusia dalam kesehariannya [1], Kemajuan teknologi telah memberikan dampak besar pada perilaku kegiatan manusia, mempengaruhi mereka untuk semakin menggunakan teknologi dalam kehidupan sehari-hari. Dengan memanfaatkan teknologi, penggunaannya dapat menikmati kemudahan dan kenyamanan pada aktivitas sehari-hari [2, 3]. Menurut [4-6] Teknologi ICT (*Information and Communication Technology*) memainkan peran yang sangat penting dalam dunia modern saat ini dan memiliki manfaat besar dalam kehidupan manusia, dan dapat dirasakan oleh semua orang yang memanfaatkannya. Berkat teknologi ini, menjadikan aktivitas sehari-hari lebih mudah dan efisien

Penggunaan *information technology and communication* memberikan kemudahan bagi *users* dalam menunjang kegiatan kerja. Penggunaan teknologi informasi dan komunikasi memberikan sumbangan besar bagi kemudahan hidup manusia dalam melakukan tugas dan aktivitas sehari-hari, sehingga memudahkan dan memberikan kenyamanan. Seperti dalam bidang asuransi untuk melakukan prediksi (*prediction*) terhadap nasabah/*customer*. Menurut Budiman [7], prediksi merupakan untuk mencari pendekatan yang rasional disertai dengan penumpukan fakta-fakta yang empiris. Sedangkan pada penelitian atau pengolahan data prediksi merupakan teknik menghitung keadaan pada periode selanjutnya [8]. Dari kedua pendapat tersebut dapat dinyatakan bahwa prediksi merupakan sebuah proses perkiraan yang dilakukan terhadap data untuk periode selanjutnya dengan menggunakan pendekatan yang rasional. Hal ini banyak dilakukan di industri untuk menyusun strategi khususnya dalam memprediksi peluang apa saja yang dapat diraih dalam upaya meningkatkan pendapatan atau

keuntungan perusahaan. Pengolahan data *mining* merupakan suatu teknik yang banyak digunakan oleh praktisi dalam menemukan peluang/strategi dimaksud. Perusahaan yang berhasil memanfaatkan teknik dan alat data *mining* dengan baik akan dapat meningkatkan keunggulan strategis dan kompetitif mereka. Dalam industri bisnis, Data Mining memegang peran penting dalam meningkatkan efektivitas proses pengambilan keputusan dan membuka peluang-peluang baru melalui transformasi data menjadi informasi yang berguna dan bermanfaat bagi perusahaan. Penelitian ini berfokus pada bagaimana peneliti mendukung perusahaan asuransi, dalam hal ini PT XYZ, untuk menganalisis data karakteristik pelanggan mereka menggunakan teknik algoritma *data mining* yang terbaik.

Berdasarkan survei yang dilakukan dalam penelitian ini, ditemukan beberapa penelitian yang telah dilakukan oleh peneliti sebelumnya, antara lain penelitian yang dilakukan oleh [9] menggunakan model *Naïve Bayes*, dengan hasil akurasi mencapai 88,00%, presisi 89,19% dan *recall* 100%. Sedangkan penelitian yang dilakukan oleh [10] juga menggunakan model *Naive Bayes*, dengan hasil akurasi sebesar 85,82%, presisi 96,10% dan *recall* 93,55%. Selanjutnya dalam penelitian ini akan menerapkan empat algoritma untuk menemukan solusi yang lebih baik, yaitu *support vector machine*, *decision tree*, k-NN, dan *random forest*.

II. TINJAUAN PUSTAKA

A. Data Mining

Teknik ekstraksi informasi dan penemuan pengetahuan dalam sistem database (*data mining*) juga disebut sebagai *Knowledge Discovery in Database* (KDD) atau *Pattern Recognition*, adalah suatu proses eksplorasi dan analisis data besar dengan tujuan mengidentifikasi informasi baru dan bermanfaat dan menentukan pola dan hubungan kompleks dalam data yang dianalisis. [11-17].

Data *mining* memiliki potensi untuk menjadi alat bantu dalam pengambilan keputusan yang akan datang. Adanya tahapan-tahapan tertentu yang perlu diikuti dalam proses data *mining*, berdasarkan referensi [18]: (1) membersihkan data (untuk menghapus *noise* dan inkonsistensi), (2) integrasi data (menggabungkan beberapa sumber data), (3) seleksi data (memilih hanya data yang berguna untuk analisis), (4) transformasi data (Menyusun data agar lebih terstruktur dapat membantu proses data *mining* menjadi lebih mudah dilakukan), (5) tambang data (langkah utama dalam data *mining*, di mana teknik yang sesuai digunakan untuk menemukan pola data), (6) evaluasi pola, (7) presentasi hasil (memaparkan hasil dalam bentuk visualisasi dan representasi untuk pengguna).

B. Machine Learning

Menurut IBM, *machine learning* termasuk dalam cabang Kecerdasan Buatan (*Artificial Intelligence/AI*). Ini adalah salah satu cabang ilmu komputer yang memfokuskan pada penggunaan data dan algoritma untuk memecahkan masalah dan membuat keputusan untuk menirukan cara manusia belajar dan meningkatkan kemampuan belajar komputer secara bertahap, sehingga meningkatkan akurasi [19], [20-

23]. *Machine learning* memfokuskan diri pada pengembangan sistem yang memiliki kemampuan untuk belajar dan membuat keputusan secara mandiri, tanpa perlu diprogram ulang oleh manusia. Tidak hanya dapat menentukan aturan untuk tindakan yang optimal dalam pengambilan keputusan, mesin juga mampu beradaptasi terhadap perubahan yang terjadi. [24].

C. Customer Retention

Customer retention, upaya atau bagaimana untuk dapat meningkatkan penjualan dengan memuaskan dan memberikan pelayanan kepada *customer*, agar mereka terus kembali [25, 26]. Menurut studi yang dilakukan oleh Anderson & Mittal dan dikutip dalam [27], retensi pelanggan dihubungkan dengan loyalitas dan profitabilitas. Sementara itu, studi lain oleh Buttle yang dikutip dalam Bramulya, et al. [28], menyatakan bahwa retensi pelanggan merupakan bentuk dari loyalitas berdasarkan perilaku (*behavioral loyalty*), yang dapat diamati dari frekuensi membeli produk yang tinggi oleh pelanggan. Sebagai tambahan, loyalitas merupakan suatu perasaan dan sikap dari konsumen terhadap suatu produk, diukur melalui komponen-komponen seperti sikap, keyakinan, perasaan, dan keinginan untuk membeli. Menurut Ennew dan Binks yang dikutip dalam Bramulya, et al. [28] loyalitas dan retensi *customer* memiliki perbedaan. Loyalitas berkaitan dengan sikap konsumen terhadap perusahaan, sementara retensi terkait dengan perilaku konsumen terhadap perusahaan.

Menurut Sembiring, et al. [29] berpendapat bahwa sebagian besar dari pihak perusahaan lebih memfokuskan strategi untuk mendapatkan *customer* baru (*customer acquisition*) dibandingkan dengan strategi untuk mempertahankan *customer* (*customer retention*). Menurut penelitian Ghavami dan Olyaei yang dikutip oleh Gengeswari, et al. [30], Biaya untuk mencari dan mendapatkan *new customer* ternyata lebih jauh tinggi bila dibandingkan dengan mempertahankan pelanggan *existing*. Oleh karena itu, perusahaan dapat menghemat biaya secara signifikan dengan memfokuskan upaya pada retensi pelanggan. Menurut Gengeswari, et al. [30], Reichheld dan Scheffer memaparkan bahwa perusahaan bisa mengalami kenaikan profit sebanyak 25 hingga 95 persen melalui peningkatan 5 persen pada tingkat retensi *customer*. Konsekuensinya, dapat dikatakan bahwa kenaikan kecil pada tingkat retensi *customer* akan mempercepat pertumbuhan laba perusahaan.

Pihak perusahaan akan mendapatkan keuntungan dari *customer retention* yakni memiliki *customer* yang loyal terhadap *company* [31-38]. Berdasarkan hasil *research marketing metrics*, tingkat keberhasilan penjualan yang dilakukan terhadap *customer retention* sebesar 60-70% sedangkan tingkat keberhasilan penjualan terhadap *new customer* itu di angka 5-20% Charlton (2012) dikutip dalam [39, 40].

Data yang diolah dalam penelitian ini merupakan *customer* yang dimiliki PT. XYZ. sehingga diperlukan suatu analisis yang tepat untuk menentukan tingkat *customer retention* terhadap penggunaan produk yang dijual PT. XYZ. Dari analisis tersebut diperlukan suatu pendekatan model,

sehingga didapatkan hasil yang terbaik dengan melakukan algoritma terhadap data yang ada seperti *support vector machine*, *decision tree*, *k-NN*, *random forest*. Tujuan penelitian ini adalah menganalisis dan memprediksi data *existing customer* untuk menemukan pola dari data *existing customer* tersebut yang melakukan perpanjangan asuransi dan tidak perpanjangan asuransi serta untuk mengetahui *customer retention rate*.

III. METODOLOGI

Research ini menggunakan *experimental research* dengan melakukan uji coba terhadap *dataset* yang telah dikumpulkan. Penelitian eksperimental merupakan jenis penelitian kuantitatif di mana peneliti melakukan manipulasi pada Satu atau lebih variabel yang tidak terikat, mengontrol variabel lain yang terikat, dan mengamati dampak manipulasi tersebut pada variabel terikat [41]. Menurut studi yang dikutip dari [42] dan [43, 44] dapat diterangkan bahwa metode penelitian kuantitatif menggunakan data numerik dalam proses pengukuran untuk memperoleh analisis yang akurat dari fenomena yang terjadi, sehingga menghasilkan kesimpulan secara umum.

A. Metode Pengumpulan Data

Menurut pendapat Sugiyono (2016:156) seperti yang dikutip dari [45], terdapat beberapa alternatif metode dan sumber yang dapat digunakan dalam proses pengumpulan data. Sumber data dimiliki PT.XYZ berjumlah 15945 dari kurun waktu 2012-2021. Menurut Sugiyono (2011), pengambilan sampel data dilakukan dengan menggunakan teknik tertentu berdasarkan pertimbangan yang sesuai. Referensi ini dikutip oleh [46]. Sampel data PT. XYZ ini berjumlah 6197 baris, 25 fitur dengan 1 *class/label*, periode tahun 2012-2021 dari data *customer* yang melakukan perpanjangan asuransi dan tidak perpanjangan asuransi yang berakhir di tahun 2021. Atribut dari *dataset* dapat diidentifikasi dalam tabel 1 yang terlampir berikut:

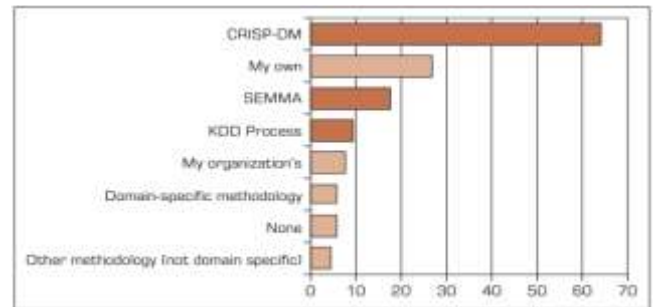
TABEL 1. DETAIL DAN KETERANGAN DATASET

No	Atribut	Tipe Data	Keterangan
1	No ID	Integer	Nomor ID
2	Perpanjangan	Binominal	YA TIDAK
3	Pekerjaan	Polynomial	Pekerjaan Tertanggung
4	Penghasilan	Integer	Penghasilan
5	Sumber Dana	Polynomial	Sumber Dana Penghasilan
6	Urutan Penutupan	Integer	1 - 10
7	Kelas Bangunan	Integer	1 2 3
8	Luas Bangunan	Integer	Luas Bangunan dalam Meter Persegi
9	Tingkat Daerah	Binominal	Kab Kota
10	Okupasi	Polynomial	Jenis Okupasi
11	Periode Awal	Date	Periode Awal Pertanggungan Asuransi
12	Periode Akhir	Date	Periode Akhir Pertanggungan Asuransi

No	Atribut	Tipe Data	Keterangan
13	Waktu Pertanggungan	Integer	Waktu Pertanggungan Dalam Tahun
14	HP Total	Integer	Harga Total Pertanggungan (HP) Objek Bangunan
15	Harga Pertanggungan	Integer	Harga Pertanggungan (HP) Objek Bangunan
16	HP Perabot	Integer	Harga Perabot Dalam Rumah
17	HP Lainnya	Integer	Harga Pertanggungan Lainnya
18	Rate	Real	Rate Asuransi
19	Premi Bayar	Integer	Jumlah Premi yang Harus Dibayar
20	Tipe Bayar	Binominal	Refund KPR Tunai
21	Premi Non Tunai	Integer	Jumlah Premi Refund KPR
22	Premi Tunai	Integer	Jumlah Premi Tunai
23	Perluasan Gempa	Binominal	Apakah Ada Perluasan Gempa

B. Teknik Analisis Data

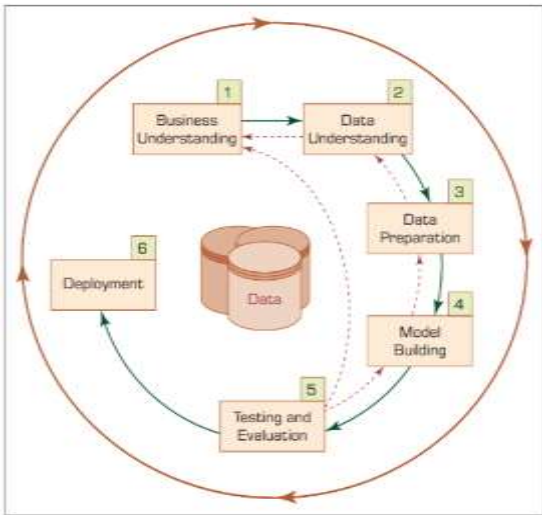
Dalam proses pengolahan data *mining* yang dilakukan tersebut terdapat beberapa model yang dapat digunakan. Berdasarkan penelitian oleh Sharda, et al. [47], terdapat beberapa tipe model pemrosesan data *mining* yang dapat diterapkan, seperti (1) *CRISP-DM*, (2) *My Own*, (3) *SEMMA*, (4) *KDD Process*, (5) *My Organization's*. Menurut [47], model *CRISP-DM* sering digunakan oleh para peneliti dalam proses pengolahan data *mining*. Hal ini dapat ditemukan pada ilustrasi pada gambar 1.



Gambar 1. Peringkat model data mining yang digunakan

Menurut model (*CRISP-DM*), terdapat enam tahapan yang dapat dilakukan (1) *business understanding*, merupakan tahapan awal yang dilakukan dengan tujuan memahami dari proses data *mining* yang akan dilakukan, (2) *data understanding*, tahap data *understanding* membutuhkan identifikasi data yang relevan dari berbagai sumber *database*. Dalam melakukan identifikasi dan seleksi data, analisis harus memiliki pemahaman jelas dan ringkas mengenai tugas data *mining* yang akan dilakukan. Hal ini memastikan bahwa data yang paling relevan dapat diidentifikasi dan sumber data yang cermat dan variabel yang paling relevan dapat dipilih, sehingga mempermudah algoritma data *mining* dalam menemukan pola pengetahuan yang berguna dengan cepat, (3) *data preparation*, *data preparation* melibatkan tahap pengambilan dan pembersihan data yang telah diidentifikasi pada tahap sebelumnya untuk mempersiapkannya untuk

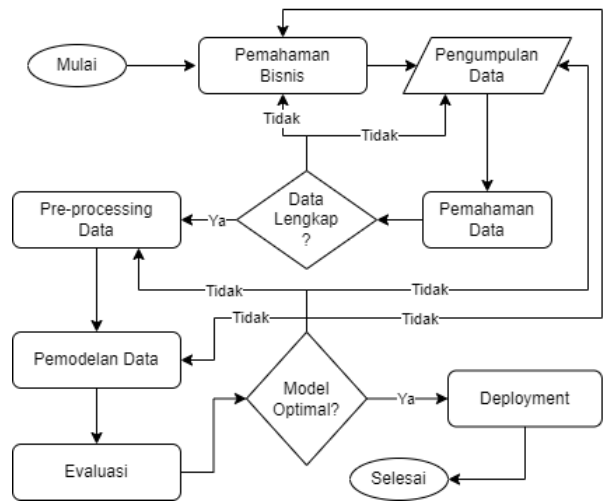
analisis menggunakan teknik data mining, (4) *model building*, tahap pembangunan model melibatkan pemilihan teknik pemodelan yang sesuai dan penerapannya pada data yang sudah dipersiapkan, untuk memenuhi tujuan bisnis tertentu, (5) *testing and evaluation*, tahap pemodelan yang dikembangkan melalui proses penilaian dan evaluasi terhadap tingkat akurasi dan generalitasnya, (6) Tahap *deployment*, Tahap *deployment* dalam CRISP-DM adalah tahap di mana hasil model data mining dipresentasikan dan diterapkan dalam lingkungan bisnis untuk mengatasi masalah dan memenuhi tujuan bisnis. Gambar 2 menunjukkan tahapan-tahapan dari model CRISP-DM.



Gambar 2. tahapan-tahapan dari model CRISP-DM

C. Tahapan Penelitian

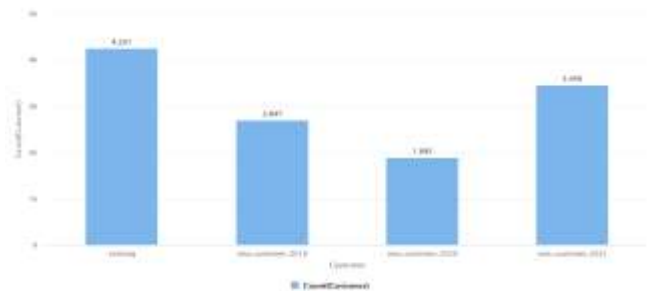
Penelitian ini menggunakan alur penelitian yang berdasarkan pada konsep CRISP-DM. Tahapan pertama adalah pemahaman bisnis, di mana permasalahan yang ingin dipecahkan adalah memahami dan menemukan pola dari data *existing customer* yang melakukan perpanjangan polis asuransi dan tidak melakukan perpanjangan. Pola tersebut digunakan untuk menentukan strategi yang tepat dalam meningkatkan rasio perpanjangan asuransi. Tahapan kedua adalah pemahaman dan identifikasi data, yang dilakukan berdasarkan hasil pengumpulan data. Tahap *pre-processing* data dilakukan pada tahap selanjutnya, dengan melakukan *filtering* data, menghapus data *noise*, mengubah nilai *inconsistent*, *replace missing value* dan *attribute selection*. Tahap modeling data dilakukan selanjutnya, dengan menggunakan empat algoritma klasifikasi. Model-model yang dihasilkan dari keempat algoritma dievaluasi pada tahap berikutnya, termasuk nilai akurasi dan AUC, uji *t-test*, dan komparasi kurva ROC. Tahap terakhir adalah penerapan model yang telah dikembangkan pada tahap sebelumnya dalam lingkungan bisnis.



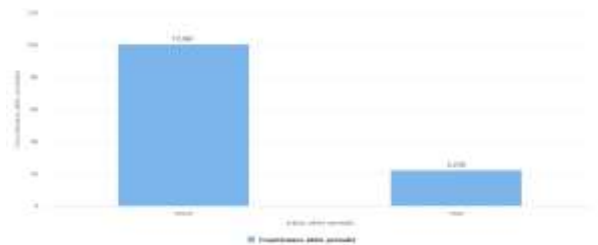
Gambar 3. Tahapan Penelitian

IV. HASIL DAN PEMBAHASAN

Periode yang di ambil untuk menghitung tingkat pemeliharaan pelanggan adalah tahun 2019 hingga 2021. *Customer existing* berjumlah 4251 yang masih aktif periode asuransi sampai dengan akhir tahun 2018, total *new customer* periode tahun 2019-2021 berjumlah 8048. Pada akhir tahun 2021 jumlah *customer* yang hilang sebanyak 2238 *customer* dengan periode asuransi yang telah berakhir dan tidak melakukan perpanjangan asuransi. Selanjutnya sebanyak 10061 *customer* yang masih melakukan perpanjangan asuransi atau masa asuransi yang masih aktif. Gambar 4 dan 5 menunjukkan hasil visual dari analisis.



Gambar 4. Grafik Customer Retention Tahun 2019-2021



Gambar 5. Grafik Customer Retention akhir periode 2021

CRR dapat dihitung dalam berbagai periode fleksibel, seperti mingguan, triwulanan, bulanan, atau tahunan, tergantung pada strategi, misi, dan visi perusahaan [48, 49].

$$CRR = \frac{E-N}{S} \cdot 100 \quad (1)$$

Teridentifikasi bahwa E adalah jumlah kumulatif pelanggan pada periode tertentu, N adalah jumlah pelanggan baru yang dicapai perusahaan selama periode tertentu, dan S adalah jumlah pelanggan pada awal periode tertentu.

$$CRR = \frac{10061-8048}{4251} \cdot 100$$

$$CRR = \frac{2013}{4251} \cdot 100$$

$$CRR = 0,4735 \times 100$$

$$CRR = 47,35$$

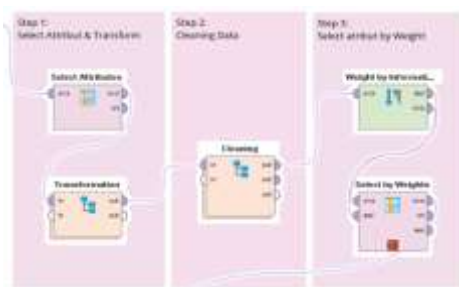
Dalam hal ini perusahaan mengawali periode dengan 4251 customer, mendapatkan 8048 new customer kemudian kehilangan 2,238 lalu menutup dengan 10061 customer. Tingkat retensi menunjukkan 47,35% yang berarti cukup rendah. Namun, setiap perusahaan atau industri memiliki tingkat retensi unik. Menghitung tingkat retensi merupakan hal yang sama pentingnya dengan mengestimasi pertumbuhan pendapatan perusahaan. Dalam tahapan *data preparation*, tool yang digunakan adalah RapidMiner terlihat pada Gambar 6 berikut:



Gambar 6. Attribute Selection pada RapidMiner

Transformation dilakukan terhadap data dan atribut. Hal ini dilakukan untuk memodifikasi sumber data ke format yang berbeda sehingga dapat diproses *data mining* ke tahap selanjutnya dengan melakukan okupasi, luas bangunan, transformasi waktu pertanggungangan, transformasi penghasilan.

Untuk tahapan *cleaning* merupakan penghapusan data yang tidak diperlukan. Tahapan prosesnya adalah: (1) *filtering data*, (2) menghapus data *noise*, (3) mengubah nilai *inconsistent*, (4), *replace missing value*. Kemudian melakukan *select attribute by weight*. Gambar 7 merupakan proses *cleaning* data.



Gambar 7. Data Cleaning Process

A. Support Vector Machine

Saat proses modeling menggunakan SVM, semua atribut *dataset* yang sudah dilakukan *pre-processing* ditransformasikan ke tipe data *numerical*. Kemudian dilakukan *cross validation* dan optimasi parameter. Dalam penelitian ini, penggunaan algoritma SVM dengan kombinasi parameter *number of fold cross validation* sebesar 10, *kernel radial* dengan γ 10 dan C bernilai 100 menghasilkan nilai akurasi sebesar 90.07% dan nilai AUC sebesar 0,937. Nilai akurasi dan AUC yang tinggi ini menunjukkan bahwa model SVM yang dihasilkan mampu dengan baik dalam memprediksi apakah seorang pelanggan akan memperpanjang polis asuransi atau tidak.

Dalam penggunaan SVM, parameter *number of fold cross validation* yang tinggi dapat menghasilkan model yang lebih akurat, karena pembagian data menjadi data training dan data testing dilakukan sebanyak 10 kali. Penggunaan *kernel radial* dengan γ yang cukup tinggi dapat membantu model dalam menangani kasus dengan dimensi yang tinggi dan kompleksitas yang tinggi. Sedangkan penggunaan nilai C yang tinggi dapat membantu model untuk meminimalkan nilai *error* pada data training, sehingga dapat meningkatkan performa pada data testing.

```

Parameter set:
Performance:
#PerformanceMetric 1
----accuracy: 90.07% +/- 0.03% (micro average) 90.07%
ConfusionMatrix:
True:  TIDAM  YA
TIDAM: 3570  446
YA:    260  3028
----precision: 92.18% +/- 2.13% (micro average) 92.05% (positive class: YA)
ConfusionMatrix:
True:  TIDAM  YA
TIDAM: 3570  446
YA:    260  3028
----recall: 87.15% +/- 2.78% (micro average) 87.15% (positive class: YA)
ConfusionMatrix:
True:  TIDAM  YA
TIDAM: 3570  446
YA:    260  3028
----F1: 89.66% +/- 0.88% (micro average) 89.93% (positive class: YA)
----AUC: 0.937 +/- 0.038 (micro average) 0.937 (positive class: YA)
----GINI (information): 0.977 +/- 0.008 (micro average) 0.977 (positive class: YA)
}
Cross Validation: number_of_folds = 10
SVM: kernel_type = radial
SVM: kernel_gamma = 10
SVM: C = 100
    
```

Gambar 8. Best Parameter & Performance Vector – Training (SVM)

B. Decision Tree

Saat proses modeling menggunakan *decision tree*, semua atribut *dataset* yang sudah dilakukan *pre-processing* kemudian dilakukan *cross validation* dan optimasi parameter. Hasil yang didapatkan menggunakan *decision tree* adalah nilai akurasi sebesar 91,45% dan AUC sebesar 0,912 dengan parameter yang dioptimalkan menggunakan teknik *cross validation*. *Criterion* yang digunakan adalah *Gini Index* (CART) dengan *pruning* dan *pre-pruning* yang bernilai *false* serta *maximal depth* sebesar 15. Hasil model ini menunjukkan bahwa *decision tree* dengan parameter yang dioptimalkan tersebut dapat digunakan untuk memprediksi target variabel dengan tingkat akurasi yang cukup tinggi. Meskipun nilai AUC sedikit rendah dibandingkan dengan nilai akurasi, namun hal ini masih dalam kisaran yang dapat diterima untuk digunakan sebagai model prediksi.

```

Parameter set:

Performance:
PerformanceVector [
----accuracy: 91.45% +/- 0.44% (micro average: 91.45%)
ConfusionMatrix:
True:  TIDAK  YA
TIDAK: 3341  311
YA:    297  3239
----precision: 91.42% +/- 1.07% (micro average: 91.42%) (positive class: YA)
ConfusionMatrix:
True:  TIDAK  YA
TIDAK: 3341  311
YA:    297  3239
----recall: 91.04% +/- 0.34% (micro average: 91.04%) (positive class: YA)
ConfusionMatrix:
True:  TIDAK  YA
TIDAK: 3341  311
YA:    297  3239
----AUC (optimistic): 0.974 +/- 0.003 (micro average: 0.974) (positive class: YA)
----AUC: 0.962 +/- 0.003 (micro average: 0.962) (positive class: YA)
----AUC (pessimistic): 0.955 +/- 0.015 (micro average: 0.955) (positive class: YA)
]
Cross Validation.number_of_folds = 10
Decision Tree.criterion = gini_index
Decision Tree.apply_pruning = false
Decision Tree.apply_prepruning = false
Decision Tree.maximal_depth = 15
    
```

Gambar 9. Best Parameter & Performance Vector – Training (DT)

C. k-Nearest Neighbor

Saat proses modeling menggunakan k-NN, semua atribut dataset yang sudah dilakukan pre-processing ditransformasikan ke dalam tipe data numerical lalu dilakukan training dan dilakukan optimasi parameter untuk mendapatkan hasil yang maksimal. Dalam penelitian ini, penggunaan algoritma k-NN dengan kombinasi parameter cross validation dengan jumlah fold sebanyak 10. Parameter k pada algoritma k-NN dipilih dengan nilai 3, dan weighted vote bernilai true, yang berarti bobot untuk kelas tetangga berdasarkan jaraknya, dan measure type bernilai numerical measures, yaitu menggunakan jarak Euclidean sebagai metrik pengukuran. Berdasarkan hasil kombinasi parameter tersebut, model ini menghasilkan nilai akurasi sebesar 87,28% dan AUC sebesar 0,906. Dari nilai akurasi dan AUC yang dihasilkan, dapat dikatakan bahwa model ini cukup baik dalam melakukan klasifikasi data. Hal ini menunjukkan bahwa model dapat mengklasifikasikan data dengan baik dan memiliki kemampuan yang cukup baik dalam membedakan antara kelas yang berbeda.

```

Parameter set:

Performance:
PerformanceVector [
----accuracy: 87.28% +/- 1.70% (micro average: 87.28%)
ConfusionMatrix:
True:  TIDAK  YA
TIDAK: 3244  313
YA:    392  2908
----precision: 88.32% +/- 2.06% (micro average: 88.30%) (positive class: YA)
ConfusionMatrix:
True:  TIDAK  YA
TIDAK: 3244  313
YA:    392  2908
----recall: 89.25% +/- 1.04% (micro average: 89.24%) (positive class: YA)
ConfusionMatrix:
True:  TIDAK  YA
TIDAK: 3244  313
YA:    392  2908
----AUC (optimistic): 0.958 +/- 0.006 (micro average: 0.958) (positive class: YA)
----AUC: 0.906 +/- 0.017 (micro average: 0.906) (positive class: YA)
----AUC (pessimistic): 0.897 +/- 0.020 (micro average: 0.897) (positive class: YA)
]
k-NN.k = 3
k-NN.weighted_vote = true
Cross Validation.number_of_folds = 10
k-NN.measure_type = NumericalMeasures
    
```

Gambar 10. Best Parameter & Performance Vector Training (k-NN)

D. Random Forest

Pada proses modeling random forest, semua atribut dataset yang sudah dilakukan pre-processing kemudian dilakukan cross validation dan optimasi parameter. Hasil yang didapat dari model random forest menunjukkan bahwa kombinasi parameter terbaik adalah dengan number of fold

cross validation sebesar 10 yang artinya model diuji dengan menggunakan data training dan testing yang berbeda-beda sebanyak 10 kali. Nilai criterion didapatkan gini index (CART), apply pruning dan pre-pruning bernilai false, dan maximal depth sebesar 15 yang artinya model tidak melakukan pemangkasan pohon dan tidak menghapus cabang yang tidak signifikan. Dengan parameter ini, model random forest menghasilkan nilai akurasi sebesar 91,08% dan AUC sebesar 0,962. Hal ini menunjukkan bahwa model dapat memberikan prediksi dengan akurasi yang cukup tinggi dalam memprediksi kelas pada data yang diberikan.

```

Parameter set:

Performance:
PerformanceVector [
----accuracy: 91.08% +/- 1.04% (micro average: 91.08%)
ConfusionMatrix:
True:  TIDAK  YA
TIDAK: 3237  293
YA:    401  3237
----precision: 89.81% +/- 1.54% (micro average: 89.96%) (positive class: YA)
ConfusionMatrix:
True:  TIDAK  YA
TIDAK: 3237  293
YA:    401  3237
----recall: 89.29% +/- 0.94% (micro average: 89.29%) (positive class: YA)
ConfusionMatrix:
True:  TIDAK  YA
TIDAK: 3237  293
YA:    401  3237
----AUC (optimistic): 0.962 +/- 0.003 (micro average: 0.962) (positive class: YA)
----AUC: 0.962 +/- 0.003 (micro average: 0.962) (positive class: YA)
----AUC (pessimistic): 0.942 +/- 0.008 (micro average: 0.941) (positive class: YA)
]
Cross Validation.number_of_folds = 10
Random Forest.criterion = gini_index
Random Forest.apply_pruning = false
Random Forest.apply_prepruning = false
Random Forest.maximal_depth = 15
    
```

Gambar 11. Best Parameter & Performance Vector - Training (RF)

Hasil pengukuran performance vector dari SVM, decision tree, k-NN, random forest selanjutnya dilakukan teknik pengujian dengan confusion matrix, ROC curve dan uji T-Test. Menurut Han and Kamber (2006) dalam [50-57] Evaluasi dilakukan melalui teknik confusion matrix, yaitu salah satu cara untuk menghitung tingkat akurasi dalam penerapan data mining. Hasilnya berupa nilai akurasi, presisi, dan recall. (1) Tingkat akurasi dalam tahapan klasifikasi merupakan presentase dari kebenaran rekaman data yang diklasifikasikan setelah melalui uji coba atas hasil klasifikasi, sedangkan (2) Proporsi dari kasus positif yang diprediksi dan benar-benar terjadi pada data asli disebut precision, dan (3) Proporsi kasus positif yang benar-benar terdeteksi sebagai positif melalui hasil prediksi disebut sebagai recall. Tabel 2 mempresentasikan confusion matrix dengan pengukuran terhadap accuracy dan AUC.

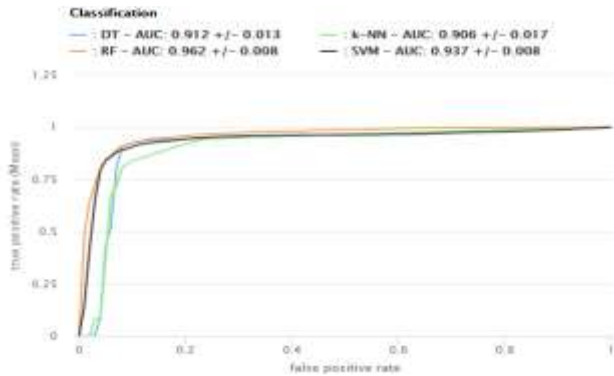
TABEL 2. HASIL PENGUKURAN KINERJA MODEL

No	Model	Accuracy (%)	Area Under the Curve (AUC)
1	SVM	90,07	0,937
2	Decision Tree	91,45	0,912
3	k-NN	87,28	0,906
4	Random Forest	91,08	0,962

Hasil pengukuran kinerja setiap model klasifikasi dengan metode SVM, k-NN, decision tree dan random forest bahwa Hasil pengukuran kinerja model klasifikasi menggunakan metode Random Forest dan Decision Tree menunjukkan

bahwa tingkat nilai akurasi keduanya lebih tinggi dibandingkan metode SVM dan k-NN, Kemudian nilai *Area Under the Curve (AUC) random forest* dan SVM lebih tinggi di dibandingkan dengan *decision tree* dan k-NN.

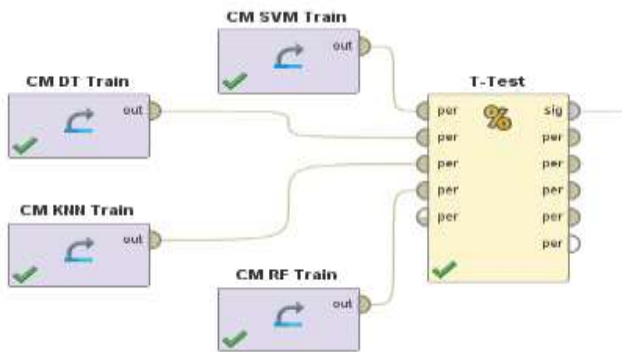
Performa dari masalah klasifikasi dapat diukur dengan menggunakan kurva ROC (*Receiver Operating Characteristics*), yang mengukur performa pada berbagai pengaturan *threshold*. Grafik yang ditempatkan adalah *true positive rate (TPR)* berbanding *false positive rate (FPR)*, dengan *true positive rate* yang diletakkan pada sumbu y dan *false positive rate* pada sumbu x, seperti yang tercantum dalam referensi [58-63], pada gambar 12 merupakan hasil dari ROC terhadap keempat model.



Gambar 12. ROC Curve

Penilaian *classifier* dikatakan baik jika memperlihatkan nilai yang mendekati angka 1 dan rendah jika mendekati angka 0.5. Dengan demikian, prediksi yang dilakukan hampir sama dengan tebakan yang dilakukan secara acak [64]. Berdasarkan visualisasi kurva ROC didapatkan hasil AUC di atas 0,9 dengan nilai tertinggi pada model *random forest* sebesar 0,962 yang mengartikan bahwa kinerja *classifier* bagus karena nilai AUC mendekati 1.

Uji *T-Test* dilakukan untuk mendapatkan model terbaik. Signifikansi uji nilai menggunakan tingkat signifikansi statistik menjadi 0.05. Secara statistik < 0.05 menunjukkan perbedaan signifikan nilai rata-rata, dengan demikian berarti harus menolak hipotesis nol (H_0).



Gambar 13. Proses Uji T-Test

A	B	C	D	E
	0.901 +/- 0.008	0.914 +/- 0.004	0.873 +/- 0.017	0.911 +/- 0.011
0.901 +/- 0.008		0.000	0.000	0.032
0.914 +/- 0.004			0.000	0.342
0.873 +/- 0.017				0.000
0.911 +/- 0.011				

Probabilities for random values with the same result.
Values with a colored background are smaller than alpha=0.050 which indicates a probably significant difference between the actual mean values.

Gambar 14. Hasil Proses Pengujian T-Test

TABEL 3. HASIL PROSES PENGUJIAN T-TEST

Model		Klasifikasi			
		SVM	DT	k-NN	RF
	Accuracy	0,901 +/- 0,008	0,914 +/- 0,004	0,873 +/- 0,017	0,911 +/- 0,011
SVM	0,901 +/- 0,008		0,000	0,000	0,032
DT	0,914 +/- 0,004			0,000	0,342
k-NN	0,873 +/- 0,017				0,000
RF	0,911 +/- 0,011				

V. KESIMPULAN

Dari hasil kajian dan evaluasi menggunakan empat algoritma yakni *SVM*, *k-NN*, *decision tree* dan *random forest* dapat dinyatakan bahwa (1) ketiga algoritma *SVM*, *decision tree* dan *random forest* melakukan prediksi sangat baik dengan nilai akurasi di atas 90% dan nilai AUC di atas 0,90, (2) hasil uji *T-Test*, algoritma *decision tree* dan *random forest* tidak terdapat perbedaan yang signifikan, dengan tingkat akurasi masing-masing adalah 91,45% dan 91,08% dengan nilai AUC masing-masing 0,912 dan 0,962, (3) hasil komparasi *ROC Curve*, *random forest* memperoleh nilai lebih tinggi dari *SVM*, *decision tree* dan k-NN. dengan nilai AUC 0,962, (4) hasil evaluasi *T-Test* dan komparasi AUC *random forest* merupakan algoritma dengan *performance* terbaik yang bisa digunakan untuk melakukan klasifikasi status perpanjangan asuransi, (5) secara umum tingkat *customer retention rate* menunjukkan 47,35% yang berarti cukup rendah di mana setiap perusahaan memiliki tingkat retensi yang berbeda.

REFERENCES

- [1] R. Riduan and S. Dedy, "Analysis of IndiHome Internet Network Usage System at PT. Telkom, Tbk Palembang," Universitas Bina Darma, 2020..
- [2] A. A. Mamusung, N. B. Anshary, and R. A. Sumarni, "Design of Wifi. Id Access Disturbance Monitoring System at PT Telkom East Jakarta Region Based on Netbeans," National Journal of Computing and Information Technology (JNKTI), vol. 3, no. 3, pp. 255-261, 2020.
- [3] M. L. DAMSIR, "Effect of Service Quality and Product Quality on Consumer Loyalty of IndiHome Service in Pekanbaru," State Islamic University Sultan Syarif Kasim Riau, 2019.
- [4] D. Arianty, D. Antoni, and M. Akbar, "Teacher Readiness in Using Information Technology for 2013 Curriculum Learning in SMP Negeri Kota Palembang," Journal of Information System and Informatics Development, vol. 1, no. 1, pp. 1-15, 2020.
- [5] N. A. Aulia, D. Antoni, D. Syamsuar, and W. Cholil, "Information Technology Security Management System Based on COBIT 5 Framework (Case Study: SMA NEGERI 1 PALEMBANG)," Journal

- of Informatics, vol. 9, no. 2, pp. 30-37, 2021.
- [6] E. Retnawiyati, D. Antoni, and M. I. Herdiansyah, "Information Technology Service Management Model Based on ITIL Version 3 in LLDIKTI Region II," *Journal of Computer Information System and Technology (SISKOMTI)*, vol. 4, no. 1, pp. 1-9, 2021.
- [7] H. Budiman, "The role of information and communication technology in education," *Al-Tadzkiyyah: Journal of Islamic Education*, vol. 8, no. 1, pp. 31-43, 2017.
- [8] R. Hutami and E. Z. Astuti, "Implementation of K-Nearest Neighbor Method for Furniture Sales Prediction at CV. Octo Agung Jepara," *Univ. Dian Nuswantoro Semarang*, vol. 3, no. 2, pp. 40-51, 2016.
- [9] J. Avegad and A. Wibowo, "Data Mining Klasifikasi Untuk Memprediksi Status Keberlanjutan Polis Asuransi Kesehatan Dengan Algoritma Naïve Bayes," in *Prosiding Seminar Nasional Teknologi Informasi dan Komunikasi (SeNTIK)*, vol. 3, no. 1, pp. 1-6, 2019.
- [10] D. R. Utari and A. Wibowo, "Pemodelan Prediksi Status Keberlanjutan Polis Asuransi Kendaraan dengan Teknik Pemilihan Mayoritas Menggunakan Algoritma-Algoritma Klasifikasi Data Mining," in *Jurnal Sistem Informasi dan Teknologi Informasi (JSITI)*, vol. 5, no. 1, pp. 19-24, 2020.
- [11] S. Susanto and D. Suryadi, "Introduction to data mining: extracting knowledge from data chunks," Ed: Andi Publisher, 2010.
- [12] A. M. Siregar, S. Kom, M. K. D. A. Puspabhuana, S. Kom, and M. Kom, *Data Mining: Processing Data into Information with RapidMiner*. CV Kekata Group, 2017.
- [13] A. P. Natasuwarna, "Challenges facing the 4.0 revolution era-big data and data mining," *SINDIMAS*, vol. 1, no. 1, pp. 23-27, 2019.
- [14] E. Prasetyowati, *DATA MINING: Clustering Data for Information and Evaluation*. Duta Media Publishing, 2017.
- [15] M. Fauzy and I. Asror, "Penerapan metode association rule menggunakan algoritma apriori pada simulasi prediksi hujan wilayah kota bandung," *Jurnal Ilmiah Teknologi Infomasi Terapan*, vol. 2, no. 3, pp. 221-227, 2016.
- [16] W. Alimuddin, E. Tungadi, and Z. Saharuna, "Analisis Tingkat Kecelakaan Lalu Lintas Dengan Metode Association Rule Menggunakan Algoritma Apriori Analisis Tingkat Kecelakaan Lalu Lintas Dengan Metode Association Rule, January, 2018.
- [17] O. Pahlevi, A. Mulyani, and M. Khoir, "Sistem Informasi Inventori Barang Menggunakan Metode Object Oriented di PT. Livaza Teknologi Indonesia Jakarta," *PROSISKO: Jurnal Pengembangan Riset dan Observasi Sistem Komputer*, vol. 5, no. 1, pp. 27-35, March 31, 2018.
- [18] I. P. Astuti, "Algoritma Apriori Untuk Menemukan Hubungan Antara Jurusan Sekolah Dengan Tingkat Kelulusan Mahasiswa," *Jurnal Teknik Informatika*, vol. 12, no. 1, pp. 69-78, 2019.
- [19] I. D. Id, *Machine Learning: Teori, Studi Kasus dan Implementasi Menggunakan Python*. Unri Press, 2021.
- [20] B. Santoso and A. I. Azis, *Machine Learning & Reasoning Fuzzy Logic Algoritma, Manual, Matlab, & Rapid Miner*. Deepublish, 2020.
- [21] A. M. Simanullang, "Makalah Machine Learning," 2021.
- [22] Y. Heryadi and T. Wahyono, *Konsep Dan Implementasi*, 2020.
- [23] R. Viryawan, A. Suroso, and R. Hasbullah, "Integrasi Dan Pengembangan Sistem Machine Learning Pada Kegiatan Maintenance Unit Bgmf Pt. Fi," *Jurnal Aplikasi Bisnis dan Manajemen (JABM)*, vol. 7, no. 3, pp. 787-787, 2021.
- [24] T. Wahyono, *Python for Machine Learning*. Gava Media Yogyakarta, 2018.
- [25] R. S. Bahri and L. Lahindah, "Cross Channel Integration Dalam Meningkatkan Retensi Pelanggan Pada Industri Ritel," *Ekonomi, Keuangan, Investasi Dan Syariah (EKUITAS)*, vol. 3, no. 3, pp. 495-501-495- 501, 2022.
- [26] R. S. Bahri, "Kajian Literature Strategi Ritel Omnichannel Sebagai Pengembangan Dari Strategi Ritel Multichannel Pada Industri Ritel," *Journal of Accounting and Business Studies*, vol. 5, no. 2, pp. 69-81, 2020.
- [27] N. S. Ying, et al., "The Antecedents and Outcome of Perceived Social Media Advertising Value (PSMAV) on Smartphone: Evidence from Malaysian Consumer Perspective," *Journal of Xi'an University of Architecture & Technology*, vol. 12, no. 3, pp. 3428-436, 2020.
- [28] R. Bramulya, I. Primiana, E. Febrian, and D. Sari, "Impact of Relationship Marketing, Service Quality and Customer Orientation of Service Employees on Customer Satisfaction and Behavioral Intentions and its Impact on Customer Retention," *International Journal of Economics, Commerce and Management*, vol. 4, no. 5, pp. 151-163, 2016.
- [29] M. J. Sembiring, A. Fattah, and S. Wahyudi, "Kapabilitas Manajemen Hubungan Pelanggan Offline to Online Sebagai Pondasi Pemberdayaan Pengrajin Hasil Olahan Ikan Laut Masyarakat Pesisir Jawa Timur," 2019.
- [30] K. Gengeswari, P. Padmashantini, and S. Sharmaela-Banu, "Impact of Customer Retention Practices on Firm Performance," *International Journal of Academic Research in Business and Social Sciences*, vol. 3, no. 7, pp. 68, 2013.
- [31] F. Rangkuti, "Customer Care Excellence: Meningkatkan Kinerja Perusahaan Melalui Pelayanan Prima Plus Analisis Kasus Jasa Raharja," *Gramedia Pustaka Utama*, 2017.
- [32] E. Winata, "Pengaruh Bauran Pemasaran Jasa Terhadap Loyalitas Konsumen Pada Grand Serela Hotel & Convention Medan," *Jurnal Mutiara Manajemen*, vol. 2, no. 1, pp. 104-117, 2017.
- [33] A. Handijono, R. I. Gunarto, and S. Marpitasa, "Menjaga Loyalitas Pelanggan Dengan Strategi CRM pada PT. Desalite, Pamulang," *RESWARA: Jurnal Pengabdian Kepada Masyarakat*, vol. 2, no. 1, pp. 64-71, 2021.
- [34] D. C. Y. Simanjuntak and P. Y. Purba, "Peran Mediasi Customer Satisfaction dalam Customer Experience Dan Loyalitas Pelanggan," *Jurnal Bisnis dan Manajemen*, vol. 7, no. 2, pp. 2020-11-30, 2020, doi: 10.26905/jbm.v7i2.4795.
- [35] R. Nurfajriah and K. Widiawati, "Strategi Pemasaran Produk Wheater Strip Pada PT Toyota Auto Body Tokai Extrusion Bekasi," *Jurnal Mahasiswa Bina Insani*, vol. 5, no. 1, pp. 11-20, Aug. 2020.
- [36] N. A. Lilawangsa and A. Triyono, "Strategi Komunikasi Pemasaran Terpadu PT. Rumawan Pusaka Negeri Dalam Mempertahankan Loyalitas Pelanggan," *Universitas Muhammadiyah Surakarta*, 2022.
- [37] R. Munandar, N. Afrilla, and R. Y. Septa, "Strategi Customer Retention Marketing (CRM) Pada Distributor Duta Parfum Center Dalam Mempertahankan Pelanggan," *Universitas Sultan Ageng Tirtayasa*, 2018.
- [38] C. Thendywinaryo, A. Sidik, and F. Goenawan, "Analisis Strategi Customer Relationship Management Untuk Mempertahankan Loyalitas Pelanggan Umkm Retail Non-Makanan Di Indonesia Timur (Studi Kasus Pada New Em Collection)," *Jurnal e-Komunikasi*, vol. 9, no. 2, 2021.
- [39] S. Nasir, "Customer retention strategies and customer loyalty," in *Advertising and Branding: Concepts, Methodologies, Tools, and Applications*, IGI Global, 2017, pp. 1178-1201.
- [40] H. Haron, S. Abu Bakar, H. Haron, and H. K. Hui, "Critical success factors on delivery preferences of e-commerce in Malaysia," *International Journal of Academic Research in Business and Social Sciences*, vol. 7, no. 11, pp. 1383-1393, 2017.
- [41] G. Adnan, D. Rukminingsih, M. Adnan Latief, E. Munastiwi, and H. Ardi (Eds.), *Metode Penelitian Pendidikan: Penelitian Kuantitatif, Penelitian Kualitatif, Penelitian Tindakan Kelas, Erhaka Utama*, 2020, p. 170.
- [42] A. Masri, "Persepsi siswa terhadap kompetensi profesional guru agama hubungannya dengan intensitas belajar siswa pada mata pelajaran PAI: penelitian pada siswa kelas VIII di SMP PGRI 10 Kota Bandung," *UIN Sunan Gunung Djati Bandung*, 2010.
- [43] H. A. Imron, "Peran Sampling dan Distribusi Data dalam Penelitian Komunikasi Pendekatan Kuantitatif," *Jurnal Studi Komunikasi dan Media*, vol. 21, no. 1, pp. 111-126, 2017.
- [44] Y. Kartika and N. Husna, "Meningkatkan Pemahaman Konsep Siswa Pada Materi Sifat-Sifat Bangun Ruang Dengan Model Pembelajaran Project Based Learning Di Kelas VI SD Negeri 1 Sawang," *Jurnal*

- Pendidikan Almuslim.
- [45] E. Barlian, "Metodologi penelitian kualitatif & kuantitatif," 2018.
- [46] R. Handayati, "Pengaruh Karakteristik Individu terhadap Kinerja Karyawan di Bank Jatim Cabang Lamongan (Suatu Studi pada Bank Jatim Cabang Lamongan)," *Jurnal Penelitian Ekonomi dan Akuntansi (JPENSI)*, vol. 1, no. 2, pp. 14 Halaman, 2016.
- [47] R. Sharda, D. Delen, and E. Turban, "Business intelligence analytics and data science: A managerial perspective," in *Business Intelligence Analytics and Data Science: A Managerial Perspective*, 2022.
- [48] Salesforce, "How to Improve Your Customer Retention Rate," <https://www.salesforce.com/resources/articles/customer-retention-rate/>, accessed 21, 2022.
- [49] R. Ahmad and F. Buttle, "Customer retention management: a reflection of theory and practice," *Marketing Intelligence & Planning*, vol. 20, no. 3, pp. 149-161, 2002, doi: 10.1108/02634500210428003.
- [50] N. W. Wardani, G. R. Dantes, and G. Indrawan, "Prediksi customer churn dengan algoritma decision tree C4. 5 berdasarkan segmentasi pelanggan untuk mempertahankan pelanggan pada perusahaan retail," in *Jurnal RESISTOR (Rekayasa Sistem Komputer)*, vol. 1, no. 1, pp. 16-24, 2018.
- [51] D. Gunawan, D. Riana, D. Ardiansyah, F. Akbar, and S. Alfarizi, "Komparasi Algoritma Support Vector Machine Dan Naive Bayes Dengan Algoritma Genetika Pada Analisis Sentimen Calon Gubernur Jabar 2018-2023," in *Seminar Nasional Informatika (SNIf)*, 2020, vol. 1, pp. 135-138.
- [52] S. Hendrian, "Algoritma Klasifikasi Data Mining Untuk Memprediksi Siswa Dalam Memperoleh Bantuan Dana Pendidikan," in *Faktor Exacta*, vol. 11, no. 3, 2018.
- [53] N. Nurajijah, D. A. Ningtyas, and M. Wahyudi, "Klasifikasi Siswa SMK Berpotensi Putus Sekolah Menggunakan Algoritma Decision Tree, Support Vector Machine dan Naive Bayes," in *Jurnal Khatulistiwa Informatika*, vol. 7, no. 2, pp. 11-2019, doi: 10.31294/jki.v7i2.6839.
- [54] M. Yusa and W. Sindu, "Evaluasi Model Decision Tree C4.5 Guna Prediksi Possibilitas Resiko Obesitas," in *Seminar Nasional Informatika (SNIf)*, vol. 1, no. 1, pp. 147-152, 2017.
- [55] A. M. Siregar, "Klasifikasi Algoritma TF dan Neutral Network dalam Sentimen Analisis," *Jurnal Accounting Information System (AIMS)*, vol. 1, no. 2, pp. 93-100, 2018.
- [56] G. A. Panhahesi and I. G. Anugrah, "Klasifikasi Waktu Penyelesaian Skripsi Mahasiswa Menggunakan Metode Weighted Naive Bayes (Studi Kasus: Program Studi Teknik Informatika Universitas Muhammadiyah Gresik)," 2022.
- [57] A. Kurniawan and S. A. Indriati, "Analisis Sentimen Opini Film Menggunakan Metode Naive Bayes dan Lexicon Based Features," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN*, vol. 2548, p. 964X, 2019.
- [58] I. Daqiqil, Y. Andriani, and E. Mahdiyah, "Strategi TIK untuk Meningkatkan Peringkat Webometric Universitas Riau Menggunakan Metode GAP Analysis," *Sistemasi: Jurnal Sistem Informasi*, vol. 10, no. 2, pp. 300-322, 2021.
- [59] R. B. Ugra Wijaya St, "Analisis Stabilitas Lereng Berdasarkan Receiver Operating Characteristic (ROC) Daerah Pekandangan, Kecamatan Banjarmangu, Kabupaten Banjarnegara, Provinsi Jawa Tengah," UPN "Veteran" Yogyakarta, 2022.
- [60] B. T. R. Doni, S. Susanti, and A. Mubarak, "Penerapan Data Mining untuk Klasifikasi Penyakit Hepatocellular Carcinoma Menggunakan Algoritma Naive Bayes," *Jurnal Responsif: Riset Sains dan Informatika*, vol. 3, no. 1, pp. 12-19, 2021.
- [61] A. Darmawan, N. Kustian, and W. Rahayu, "Implementasi Data Mining Menggunakan Model SVM untuk Prediksi Kepuasan Pengunjung Taman Tabebuya," *STRING (Satuan Tulisan Riset dan Inovasi Teknologi)*, vol. 2, no. 3, pp. 299-307, 2018.
- [62] W. Nengsih, M. M. Zein, and N. Hayati, "Coarse-Grained Sentiment Analysis Berbasis Natural Language Processing-Ulasan Hotel," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 10, no. 1, pp. 41-48, 2021.
- [63] L. A. Utami, "Analisis Sentimen Opini Publik Berita Kebakaran Hutan Melalui Komparasi Algoritma Support Vector Machine Dan K-Nearest Neighbor Berbasis Particle Swarm Optimization," *Jurnal Pilar Nusa Mandiri*, vol. 13, no. 1, pp. 103-112, 201.
- [64] M. R. Faisal and D. T. Nugrahadi, *Belajar Data Science: Klasifikasi dengan Bahasa Pemrograman R*, 2017.