**Jurnal Teknik Informatika (JUTIF)**
Vol. 5, No. 3, June 2024, pp. 767-775

DOI: https://doi.org/10.52436/1.jutif.2024.5.3.1871
p-ISSN: 2723-3863
e-ISSN: 2723-3871

# *APPLICATION OF ENSEMBLE METHOD FOR EMPLOYEE TURNOVER PREDICTIONS IN FINANCIAL SERVICES COMPANY*

**Muhamad Fadel[*1], Kanasfi[2], Zainal Arifin[3], Gandung Triyono[4]**

[1,2,3,4]Master of Computer Science, Information and Technology Faculty Universitas Budi Luhur, Indonesia
Email: [1]2211600081@student.budiluhur.ac.id, [2]2211601170@student.budiluhur.ac.id,
[3]2111601148@student.budiluhur.ac.id, [4]gandung.triyono@budiluhur.ac.id

(Article received: February 18, 2024; Revision: March 7, 2024; published: May 28, 2024)

***Abstract***

*High employee turnover is a challenge for every company, considering that employees are a valuable asset for the company. A high employee turnover rate indicates the high frequency of employees leaving a company. This will harm the company in terms of time, costs, human resources, and reduce the company's reputation. Low employee turnover is an objective for every company in its efforts to achieve its vision and mission, the employee turnover rate is high at 78.97% at PT. HCI operating in the financial services sector can have a negative impact on the company's reputation. Therefore, there is a need to analyze and predict employee turnover so that company management can take preventive and persuasive actions so as to reduce employee turnover rates. Therefore, a tool is needed to predict whether an employee will leave the company. This paper aims to predict the possibility of employees out of the company using the ensemble method, which is a method that uses a combination of several algorithms consisting of base learners and individual learners, algorithms with the ensemble method used are stacking, random forest, and adaboost, then comparing the result to get the best accuracy. The test results prove that the Stacking algorithm technique is the best model with the highest score in terms of accuracy with a value of 86.84%, while the Random Forest and AdaBoost algorithm techniques have a value of 81.04% and 80.30%. With this high accuracy value, the Stacking model is proven to have better individual performance in analyzing employee turnover predictions in human resource applications in companies.*

**Keywords**: *AdaBoost, Machine Learning, Random Forest, Stacking.*

# PENERAPAN METODE ENSEMBLE UNTUK PREDIKSI TURNOVER KARYAWAN PADA PERUSAHAAN JASA KEUANGAN

**Abstrak**

Turnover karyawan yang tinggi merupakan sebuah tantangan bagi setiap perusahaan, mengingat karyawan adalah asset yang berharga bagi perusahaan. Tingkat turnover karyawan yang tinggi memberikan indikasi tingginya frekuensi karyawan yang keluar dari sebuah perusahaan. Hal ini akan merugikan perusahaan dari sisi waktu, biaya, sumber daya manusia, dan menurunkan reputasi perusahaan. Turnover karyawan yang rendah merupakan objektif bagi setiap perusahaan dalam upaya pencapaian visi misinya, tingkat turnover karyawan yang tinggi sebesar 78.97% pada PT. HCI yang bergerak pada bidang jasa keuangan dapat berdampak negative pada reputasi perusahaan. Oleh karena itu, perlunya untuk menganalisis dan memprediksi turnover karyawan agar manajemen perusahaan dapat melakukan tindakan preventif dan persuasif sehingga dapat mengurangi tingkat turnover karyawan. Oleh karena itu, dibutuhkan sebuah tools atau alat bantu untuk memprediksi apakah seorang karyawan akan keluar dari perusahaan. Paper ini bertujuan untuk memprediksi kemungkinan karyawan keluar dari Perusahaan menggunakan metode *Ensemble*, yang merupakan sebuah metode yang menggunakan kombinasi beberapa algoritma yang terdiri dari *base learner* dan *individual learner*, algoritma dengan metode *Ensemble* yang digunakan adalah *Stacking*, *Random Forest*, dan *AdaBoost*, kemudian dilakukan perbandingan untuk mendapatkan motede akurasi terbaik. Hasil pengujian membuktikan bahwa teknik algoritma *Stacking* merupakan model terbaik dengan nilai tertinggi dalam hal akurasi dengan nilai 86,84%, sementara teknik algoritma *Random Forest* dan *AdaBoost* memiliki nilai masing-masing 81,04% dan 80,30%.

**Kata kunci**: *AdaBoost, Machine Learning, Random Forest, Stacking*.

## 1. PENDAHULUAN

Turnover karyawan adalah proses keluar dan masuknya karyawan dalam sebuah perusahaan secara sukarela atau tidak, keluar sukarela artinya karyawan meninggalkan perusahaan karena resign, keluar tidak sukarela artinya karyawan meninggalkan perusahaan karena pemecatan.

Tingkat turnover karyawan yang tinggi sebesar 78.97% pada data karyawan PT. HCI sejak tahun 2020 sampai dengan 2023 berdampak negatif terhadap perusahaan dalam hal bertambahnya waktu yang dibutuhkan untuk rekrutmen ulang, bertambahnya biaya rekrutmen, hilangnya sumber daya manusia kompeten, dan reputasi perusahaan yang buruk. Tingkat turnover yang rendah berdampak positiv terhadap sebuah perusahaan dalam hal efisiensi waktu, biaya, loyalitas karyawan, dan reputasi perusahaan yang baik.

Mempertahankan karyawan saat ini menjadi objektif dalam setiap perusahaan. Salah satu tujuan utama dari departemen sumber daya manusia adalah mempertahankan karyawan dan menerapkan keahlian karyawannya untuk mencapai visi dan misi sebuah perusahaan. Metode yang dapat digunakan adalah melakukan prediksi resiko turnover karyawan.

Turnover karyawan dapat dibagi menjadi dua kategori, yaitu keluar secara sukarela, dimana karyawan memilih untuk meninggalkan perusahaan atau pensiun, dan keluar secara tidak sukarela artinya perusahaan melakukan pemutusan hubungan kerja terhadap karyawannya, pensiun tidak perlu dilakukan prediksi karena sudah diatur dalam undang-undang ketenagakerjaan. Focus dari penelitian ini adalah turnover karyawan secara sukarela. Oleh karena itu, turnover karyawan secara tidak sukarela tidak termasuk ke dalam cakupan penelitian ini.

Metode ensemble adalah teknik yang digunakan untuk meningkatkan kinerja dan akurasi model-model pembelajaran mesin dengan menggabungkan hasil prediksi dari beberapa model atau algoritma yang berbeda. Istilah ensemble merujuk pada konsep menggabungkan banyak elemen menjadi satu kesatuan. Konsep penggabungan berbagai pendekatan pembelajaran mesin diharapkan dapat mengatasi kelemahan individu dari masing-masing model dan menciptakan prediksi yang lebih akurat dan stabil. Cara kerja dari metode ensemble melibatkan penggunaan beberapa model yang beragam dalam satu tim. Hasil prediksi dari masing-masing model tersebut digabungkan untuk membuat keputusan akhir.

Penelitian terdahulu mengenai prediksi karyawan yaitu Eko Hardiyanto yang berjudul kombinasi metode ensemble, CFS, dan pohon keputusan untuk prediksi kinerja petugas studi kasus survey podes badan pusat statistic. Dimana teknik decision tree ini mampu untuk memberikan keputusan dengan variasi data pada data nominal dan numerik. Limitasi dari penelitian ini adalah hasil yang bias, yang disebabkan oleh kurangnya data [1]. Pada kasus implementasi data mining untuk pengelompokkan dan prediksi karyawan yang berpotensi PHK dengan algoritma k-means cluster yang dibahas pada penelitian yang dilakukan oleh Windania Purba, menghasilkan bahwa metode k-mean clustering dapat menghasilkan dua kelompok klaster karyawan, yaitu karyawan yang memiliki potensi PHK, dan karyawan yang tidak memiliki potensi PHK. Limitasi dari penelitian ini adalah kurangnya jumlah dataset [2]. Pada kasus komparasi tiga metode algoritma klasifikasi data mining pada prediksi kenaikan jabatan yang dibahas pada penelitian yang dilakukan oleh Jaka Tirta Samudra, menghasilkan algoritma Naïve Bayes memiliki nilai akurasi terbaik sebesar 76.6%, sedangkan algoritma KNN dan Neural Network memiliki nilai akurasi masing-masing sebesar 68.6% dan 72.6%. Limitasi dari penelitian ini adalah hasil yang bias, yang disebabkan oleh kurangnya jumlah dataset [3]. Pada kasus perbandingan metode seleksi fitur untuk mengoptimasi model support vector machine dalam memprediksi turnover pegawai yang dibahas pada penelitian yang dilakukan oleh Ahmad Syafiq Abiyyu, menghasilkan algoritma SVM tanpa dilakukan seleksi fitur memiliki nilai sebesar 0.56 untuk semua metode pengukuran. Metode pengukuran yang digunakan adalah presisi, recall, dan f1-score. Metode seleksi fitur yang menunjukkan hasil evaluasi yang meningkat dibandingkan hasil evaluasi tanpa seleksi fitur adalah wrapper method dengan nilai performa model adalah 0,60. Limitasi dari penelitian ini adalah kurangnya penggalian lebih jauh terhadap parameter yang dapat mempengaruhi hasil nilai evaluasi [4]. Pada kasus prediksi employee attrition menggunakan algoritma support vector machine yang dibahas pada penelitian yang dilakukan oleh Muhammad Abdurrohman Alfatih, menghasilkan bahwa jika dilihat dari nilai akurasi, f1-score, dan rata-rata nilai geometric mean, model SVM lebih akurat dalam melakukan prediksi dibandingkan dengan model KNN. Limitasi dari penelitian ini adalah metode yang digunakan hanya dua jenis [5]. Pada kasus system prediksi awal terhadap atrisi karyawan menggunakan algoritma C4.5 yang dibahas pada penelitian yang dilakukan oleh Tulus Harry Lamramot, menghasilkan algoritma C4.5 merupakan algoritma yang dapat memprediksi kebenaran karyawan lama yang terkena atrisi di PT. Indorama Petrochemicals, hasil yang diperoleh menggunakan metode Confussion Matrix sebesar 0,9466 apabila dihitung dengan persen hasiil yang didapatkan adalah 94.6%. Limitasi dari penelitian ini adalah metode yang digunakan hanya satu jenis [6].

Penelitian terdahulu mengenai *Random Forest* yaitu Shubham Karande yang berjudul prediction of employee turnover using ensemble learning. Penelitian tersebut menghasilkan bahwa ensemble model memiliki tingkat akurasi terbaik sebesar

83.87%, dibandingkan dengan algoritma Support Vector Machine, Logistic Regression, dan Random Forest yang memiliki nilai akurasi masing-masing sebesar 77.65%, 81.77%, dan 82.64%. Limitasi dari penelitian ini adalah hasil yang bias, yang disebabkan oleh data yang sensitif [7]. Pada kasus perbandingan algoritma *KNN, Decision Tree,* dan *Random Forest* pada data *imbalanced class* untuk klasifikasi promosi karyawan yang dibahas pada penelitian yang dilakukan oleh Louis Madaerdo Sotarjua, menghasilkan bahwa model KNN adalah model klasifikasi dengan tingkat akurasi terbaik yaitu sebesar 86.57%, lebih baik dibandingkan dengan model klasifikasi Decision Tree, dan Random Forest dengan tingkat akurasi masing-masing sebesar 85.29% dan 86.37%. Limitasi dari penelitian ini adalah kurangnya jumlah data [8]. Pada kasus ensemble method-based architecture using random forest importance to predict employee turnover yang dibahas pada penelitian yang dilakukan oleh Md. Anwar Hossen, menghasilkan bahwa algoritma Random Forest memiliki nilai akurasi tertinggi sebesar 98.64%, lebih baik dibandingkan algoritma Decision Tree, SVM, Gradient Boosting, dan KNN dengan nilai akurasi masing-masing sebesar 97.93%, 95.27%, 79.33%, dan 95.93%. Limitasi dari penelitian ini adalah kurangnya data [9]. Pada kasus prediksi pengunduran diri karyawan perusahaan "Y" menggunakan random forest yang dibahas pada penelitian yang dilakukan oleh Daniel Dwi Eryanto Manurung, menghasilkan bahwa confusion matrix algoritma Random Forest memiliki nilai akurasi sebesar 0.8775, sehingga algoritma Random Forest memiliki nilai akurasi sebesar 87%, dan error sebesar 13%. Limitasi dari penelitian ini adalah metode yang digunakan hanya satu jenis [10]. Pada kasus early prediction of employee turnover using machine learning algorithms yang dibahas pada penelitian yang dilakukan oleh Markus Atef, menghasilkan bahwa algoritma KNN memiliki nilai akurasi terbaik yaitu sebesar 84%, lebih baik dibandingkan dengan algoritma Random Forest yang memiliki nilai akurasi sebesar 80%. Limitasi dari penelitian ini adalah metode yang digunakan hanya dua jenis [11].

Penelitian terdahulu mengenai *Stacking* yaitu Atik Surmasani yang berjudul algoritma stacking untuk klasifikasi penyakit jantung pada dataset imbalance class. Penelitian tersebut menghasilkan bahwa algoritma Stacking mampu menghasilkan kinerja dari sisi akurasi TPR, TNR, G-Mean, dan AUC yang lebih baik dibandingkan single classifier lainnya, algoritma Stacking menghasilkan nilai akurasi terbaik sebesar 81%, lebih baik dibandingkan dengan algoritma KNN, C4.5, SVM, dan Neural Network dengan nilai akurasi masing-masing sebesar 62%, 78%, 79%, dan 80%. Limitasi dari penelitian ini adalah kurangnya jumlah data [12]. Pada kasus penerapan metode stacking dalam mengklasifikasikan penderita penyakit diabetes yang dibahas pada penelitian yang dilakukan oleh Binti Mamluatul Karomah, menghasilkan bahwa metode Stacking menggunakan C4.5 dan SVM sebagai base model dengan logistic regression sebagai meta model dalam klasifikasi penderita penyakit diabetes, menunjukan bahwa algoritma Stacking mampu meningkatkan performa base/single classifier dari sisi accuracy, recall, dan precision. Algoritma Stacking menghasilkan nilai akurasi terbaik sebesar 97.11%, lebih baik dibandingkan algoritma C4.5 dan SVM dengan nilai akurasi masing-masing sebesar 95.19% dan 96.53%. Limitasi dari penelitian ini adalah kurangnya jumlah data [13]. Pada kasus penerapan metode stacking dan random forest untuk meningkatkan kinerja klasifikasi pada proses deteksi web phishing yang dibahas pada penelitian yang dilakukan oleh Anggit Ferdita Nugraha, menghasilkan bahwa nilai akurasi yang dihasilkan dari penggunaan Random Forest sebesar 96.4% merupakan nilai kinerja terbaik pada dataset dengan imbalanced ratio sebesar 1.25%, sedangkan nilai akurasi sebesar 88.8% yang dihasilkan dari dataset dengan imbalanced ratio sebesar 6.82%. Limitasi dari penelitian adalah kurangnya jumlah data [14]. Pada kasus penerapan ensemble stacking untuk peramalan laba bersih bank syariah Indonesia yang dibahas pada penelitian yang dilakukan oleh Nurfia Oktaviani Syamsiah, menghasilkan bahwa peramalan yang menggunakan ensemble stacking terbukti lebih unggul dari hasil metode lainnya. Limitasi dari penelitian ini adalah kurangnya jumlah data [15]. Pada kasus prediksi retweet berdasarkan user-based dan content-based menggunakan metode ensemble stacking yang dibahas pada penelitian yang dilakukan oleh Muhammad Rizqi Akbar, menghasilkan bahwa algoritma Stacking menghasilkan nilai akurasi terbaik sebesar 85.35%, lebih baik dibandingkan dengan algoritma-algoritma lainnya yaitu Random Forest, Gradient Boosting, dan SVM dengan nilai akurasi masing-masing sebesar 85.34%, 80.95%, dan 73.58%. Limitasi dari penelitian ini adalah kurangnya jumlah data [16].

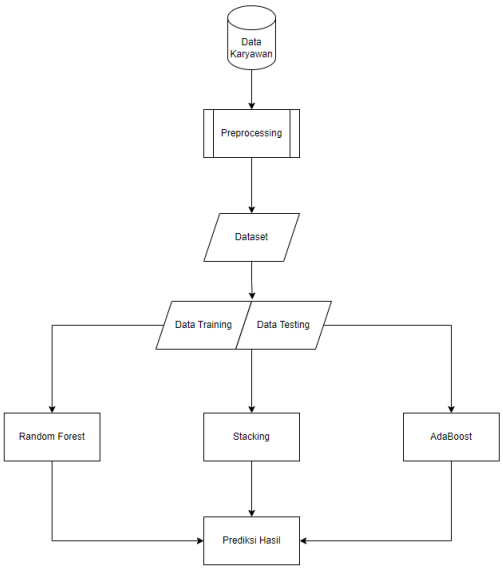
Penelitian terdahulu mengenai *Boosting* yaitu Vasthu Imaniar Ivanoti yang berjudul decision support system for predicting employee leave using the light gradient boosting machine and k-means algorithm. Prediksi yang dihasilkan oleh model ini sangat cocok dengan data asli dalam dua pengujian terpisah. Oleh karena itu, model regresi LightGBM yang memanfaatkan variabel independen seperti jenis kelamin, usia, eselon 1, dan tahun cuti, dapat secara efektif memprediksi jumlah karyawan yang mengambil cuti tahunan bersamaan dengan hari libur. Limitasi dari penelitian ini adalah metode yang digunakan hanya dua jenis [17]. Pada kasus automated prediction of employee attrition using ensemble model based on machine learning algorithm yang dibahas pada penelitian yang dilakukan oleh Fahad Kamal Alsheref, menghasilkan bahwa temuan mengungkapkan bahwa hingga saat ini tidak ada model yang dapat dianggap ideal dan sempurna untuk

setiap kasus dalam konteks bisnis. Namun, model pilihan kami cukup optimal sesuai dengan kebutuhan kami dan cukup memenuhi tujuan yang diinginkan. Limitasi dari penelitian ini adalah kurangnya jumlah data [18]. Pada kasus perbandingan akurasi algoritma adaboost dan algoritma lightgbm untuk klasifikasi penyakit diabetes yang dibahas pada penelitian yang dilakukan oleh Rangga Ahsana, menghasilkan bahwa Metode boosting yang memiliki tingkat akurasi paling tinggi pada algoritma AdaBoost adalah metode boosting SAMME dengan nilai akurasi sebesar 91.14% dibandingkan dengan algoritma AdaBoost yang menggunakan metode boosting SAMME.R yaitu hanya sebesar 89.58%. Limitasi dari penelitian ini adalah metode yang digunakan hanya dua jenis [19]. Pada kasus light gradient boosting machine untuk deteksi penyakit stroke yang dibahas pada penelitian yang dilakukan oleh Felix Indra Kurniadi, menghasilkan bahwa akurasi diberikan dari ketiga metode yaitu LightGDM, SVM, dan Random Forest memiliki nilai yang sama di kedua scenario akan tetapi hasil precision dan recall pada scenario memberikan hasil yang berbeda pada metode Random Forest. Nilai precision dan recall sangat jauh berbeda dengan nilai akurasi ini mengindikasikan bahwa adanya imbalanced data yang tidak ditangani akan membuat bias. Limitasi dari penelitian ini adalah hasil yang bias, yang disebabkan oleh imbalanced data [20]. Pada kasus penerapan teknik random oversampling untuk mengatasi imbalance class dalam klasifikasi website phishing menggunakan algoritma lightgbm yang dibahas pada penelitian yang dilakukan oleh Sri Diantika, menghasilkan bahwa model yang diusulkan lebih baik dari bebarapa model lain yang juga telah diuji dengan nilai akurasi sebesar 96,9%, recal 96,9%, F1-score 96,9% dan ROC 99,7%. Algoritma LightGBM dan Random Forest memiliki nilai akurasi terbaik yaitu masing-masing sebesar 96.90%, lebih baik dibandingkan algoritma Gradient Boosting Classifier, Decision Tree, dan Naïve Bayes dengan nilai akurasi masing-masing sebesar 92.80%, 91.50%, dan 88%. Limitasi dari penelitian ini adalah kurangnya jumlah data [21].

Penelitian ini bertujuan untuk melakukan perbandingan dan memprediksi kemungkinan keluarnya karyawan menggunakan teknik *Machine Learning Random Forest, Stacking*, dan *AdaBoost*. 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 *Random Forest, Stacking*, dan *AdaBoost*.

## 2. METODE PENELITIAN

Alur diagram dan rancangan diawali dengan langkah-langkah secara urut sebagai berikut: identifikasi masalah penelitian, studi literatur, pemodelan, analisa hasil, kesimpulan, seperti yang ditampilkan pada Gambar 1.




Gambar 1. Pemodelan Prediksi Terminasi Karyawan

Pemodelan prediksi yang diajukan menggunakan gabungan teknik *Machine Learning Random Forest, Stacking*, dan *AdaBoost*. Sebelumnya, dataset melalui tahap preprocessing dan pembagian menjadi data latih dan data uji dengan rasio 80:20, seperti yang ditampilkan pada Gambar 1.

Pada penelitian ini data yang dipilih untuk digunakan sebagai atribut yang terkait langsung dengan aktivitas kepegawaian (sesuai yang ada pada dataset) yaitu usia, gaji, jenis kelamin, status pernikahan, dan class sebagai label prediksi terminasi karyawan. Sebelum dimasukkan ke dalam pemodelan, dataset harus melalui tahap persiapan. Diawali dengan pembersihan data dengan cara menghilangkan data redundan dan missing value, mengkonversi dataset ke dalam format csv, sampai dengan membagi dataset menjadi data latih dan data uji.

Masing-masing algoritma dalam pemodelan prediksi ini memproses dataset yang sama dan menghasilkan nilai prediksi yang berbeda-beda berupa klasifikasi terminasi. Hasil pemodelan berupa tiga level klasifikasi prediksi turnover karyawan.

Kinerja algoritma klasifikasi diukur menggunakan acuan *confusion matrix*, yang menampilkan hasil prediksi dan kondisi actual berupa akurasi, presisi, dan recall. Akurasi merupakan rasio prediksi benar positif dan negative, presisi merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif, sedangkan recall merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif.

Tabel 1. Confusion Matrix

| **Klasifikasi Benar** | **Diklasifikasikan Sebagai** | |
|---|---|---|
| | + | - |
| + | TP | FN |
| - | FN | TP |

True Possitive (TP) meruapakan jumlah record positif dalam dataset yang diklasifikasikan positif. True Negative (TN) merupakan jumlat record negatif dalam dataset yang diklasifikasikan negatif. False Possitive (FP) merupakan jumlah record negatif dalam dataset yang diklasifikasikan positif. False Negative (TN) merupakan jumlah record positif dalam dataset yang diklasifikasikan negatif. Berikut ini merupakan persamaan model confusion matrix:

1. *Precision* digunakan untuk mengukur seberapa besar proporsi dari kelas data positif yang berhasil diprediksi dengan benar dari keseluruhan hasil prediksi kelas positif. Adapun rumus untuk menghitung precision adalah:

$$Precision = \frac{tp}{tp+fp} \quad (1)$$

2. *Recall* digunakan untuk menunjukkan presentase kelas data positif yang berhasil diprediksi dengan benar dari keseluruhan data kelas positif. Adapun rumus untuk menghitung recall adalah:

$$recall = \frac{tp}{tp+fn} \quad (2)$$

3. *F1-Measure* merupakan gabungan dari precision dan recall yang digunakan untuk mengukur kemampuan algoritma dalam mengklasifikasikan kelas minoritas. Adapun rumus untuk menghitung f-measure adalah:

$$F1-measure = \frac{2\ x\ recall\ x\ precision}{recall+precision} \quad (3)$$

4. Akurasi adalah jumlah perbandingan data yang benar dengan jumlah keseluruhan data. Adapun rumus untuk menghitung akurasi adalah:

$$akurasi = \frac{tp+tn}{tp+fn+fp+fn} \quad (4)$$

## 3. HASIL DAN PEMBAHASAN

Langkah-langkah untuk melaksanakan proses data mining dimulai dari penarikan data karyawan PT. HCI dari system kepegawaian, diputuskan menggunakan data karyawan perusahaan per bulan Oktober 2023 yang terdiri dari 7,264 data (record) dan 21 atribut (variable).

Tabel 2. Dataset Pribadi Karyawan

| Nomor Karyawan | Jenis Kelamin | Status Pernikahan | Usia |
|---|---|---|---|
| 10000089 | Laki-laki | Single | 33 |
| 10000090 | Wanita | Single | 26 |
| 10000091 | Laki-laki | Married | 44 |
| 10000092 | Wanita | Single | 29 |
| 10000093 | Laki-laki | Married | 39 |
| 10000094 | Laki-laki | Married | 31 |
| 10000095 | Laki-laki | Married | 27 |
| 10000096 | Wanita | Single | 25 |
| 10000097 | Laki-laki | Married | 44 |
| 10000098 | Laki-laki | Married | 47 |

Pada table 2 dapat dilihat contoh dataset pribadi karyawan yang berisi informasi nomor karyawan, jenis kelamin, status pernikahan, dan usia

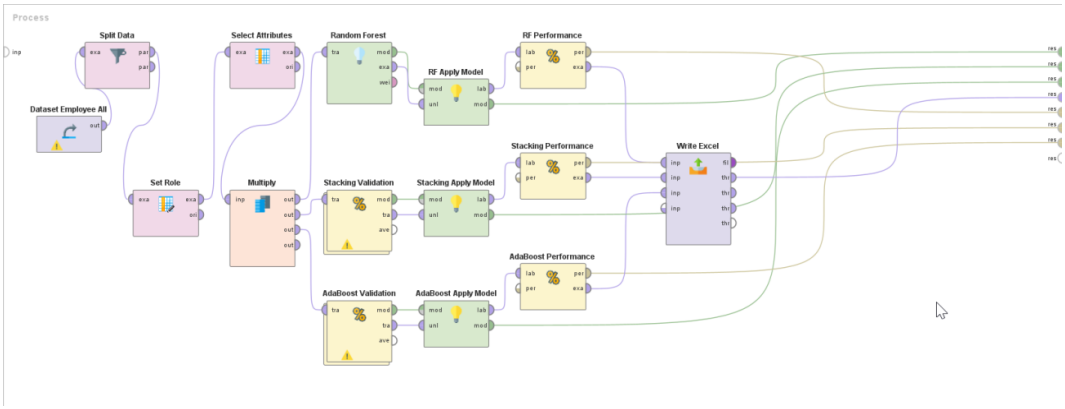
Tabel 3. Variable Data Penelitian

| Nomor | Atribut | Tipe |
|---|---|---|
| 1 | Pernr | Integer |
| 2 | Full Name | Polynomial |
| 3 | Join Date | Polynomial |
| 4 | Termination Date | Polynomial |
| 5 | Termination Reason | Polynomial |
| 6 | Last Position | Polynomial |
| 7 | Grade | Polynomial |
| 8 | Warning Letter 1 | Polynomial |
| 9 | Warning Letter 1 Description | Polynomial |
| 10 | Warning Letter 1 Detail | Polynomial |
| 11 | Warning Letter 2 | Polynomial |
| 12 | Warning Letter 2 Description | Polynomial |
| 13 | Warning Letter 2 Detail | Polynomial |
| 14 | Warning Letter 3 | Polynomial |
| 15 | Warning Letter 3 Description | Polynomial |
| 16 | Warning Letter 3 Detail | Polynomial |
| 17 | Gaji | Real |
| 18 | Jenis Kelamin | Polynomial |
| 19 | Status Pernikahan | Polynomial |
| 20 | Usia | Integer |
| 21 | Terminated | Polynomial |

Pada table 3 dapat dilihat daftar atribut yang digunakan pada penelitian ini yang berisi informasi nomor karyawan, nama, tanggal bergabung, tanggal terminasi, alasan terminasi, posisi terakhir, grade, surat peringatan satu, deskripsi surat peringatan satu, detil surat peringatan satu, surat peringatan dua, deskripsi surat peringatan dua, detil surat peringatan dua, surat peringatan tiga, deskripsi surat peringatan tiga, detil surat peringatan tiga, gaji, jenis kelamin, status pernikahan, usia, dan status terminasi Selanjutnya dilakukan transformasi data menjadi sebuah format yang bisa diterima oleh Rapid Miner untuk kebutuhan analisa.

Tabel 4. Dataset Pegawai Setelah dilakukan Transformasi

| Nomor Karyawan | Tanggal Bergabung | Tanggal Terminasi | Alasan Terminasi | Surat Peringatan | Gaji |
|---|---|---|---|---|---|
| 10000089 | 10/09/2017 | 11/08/2021 | Org. Optimization | | 3,296,131 |
| 10000090 | 05/01/2018 | 11/16/2020 | Family.reason | | 2,041,382 |
| 10000091 | 05/02/2017 | 02/01/2021 | Org. Optimization | SP1 | 3,510,000 |
| 10000092 | 06/01/2018 | | | | 1,945,824 |
| 10000093 | 06/19/2019 | | | SP1 | 5,460,000 |
| 10000094 | 04/01/2018 | | | | 2,358,081 |
| 10000095 | 02/01/2019 | 09/07/2021 | Family.reason | SP2 | 2,007,966 |
| 10000096 | 06/01/2019 | | | | 2,202,934 |
| 10000097 | 11/27/2017 | 02/01/2021 | Org. Optimization | SP3 | 3,900,000 |
| 10000098 | 06/08/2015 | | | | 10,918,301 |



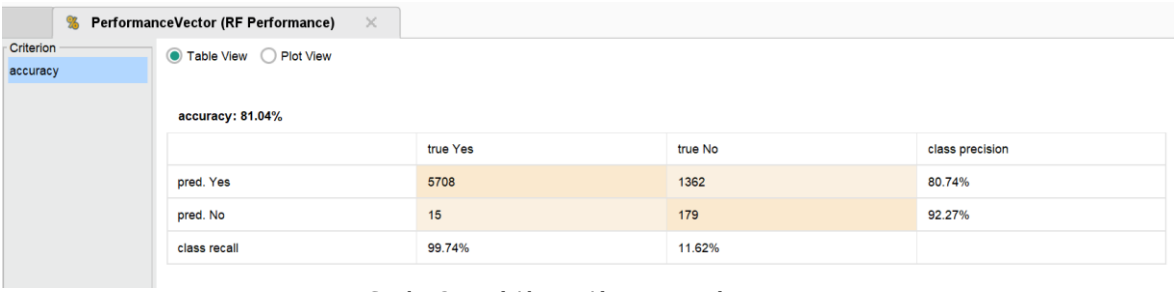

Gambar 2. Desain Rapid Miner

Pada tabel 4 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 pemodelan, 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 *Random Forest, Stacking*, dan *AdaBoost.*

Pada gambar 2 dapat dilihat bahwa rancangan Rapid Miner dimulai dari proses *retrieve dataset employee all*, lalu data dibagi menjadi data latih dan data uji dengan komposisi 80-20, selanjutnya ditentukan *set role*, yaitu target yang ingin diprediksi "terminated", kemudian ditentukan attribute yang mempengaruhi yaitu usia, gaji, jenis kelamin, dan status pernikahan, lalu data akan dihubungkan ke setiap algoritma *Machine Learning* yang digunakan, yaitu algoritma *Random Forest, Stacking,* dan *AdaBoost*.

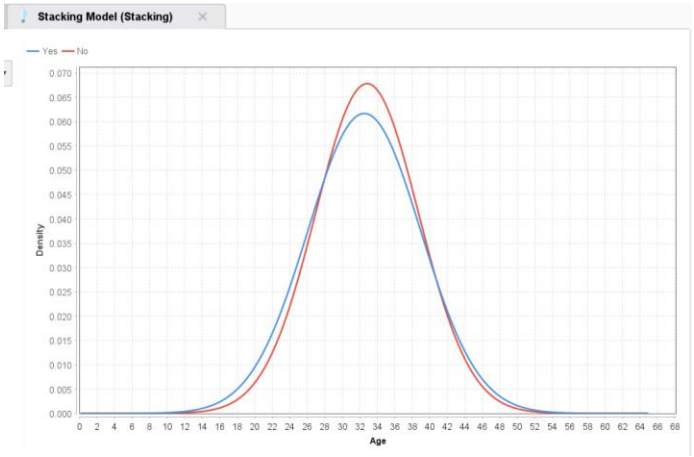
Pada tahap evaluasi, dilakukan pembahasan tentang hasil dan temuan percobaan, kemudian dilakukan pengukuran kinerja menggunakan operator *performance classification*.




| | true Yes | true No | class precision |
|---|---|---|---|
| pred. Yes | 5708 | 1362 | 80.74% |
| pred. No | 15 | 179 | 92.27% |
| class recall | 99.74% | 11.62% | |

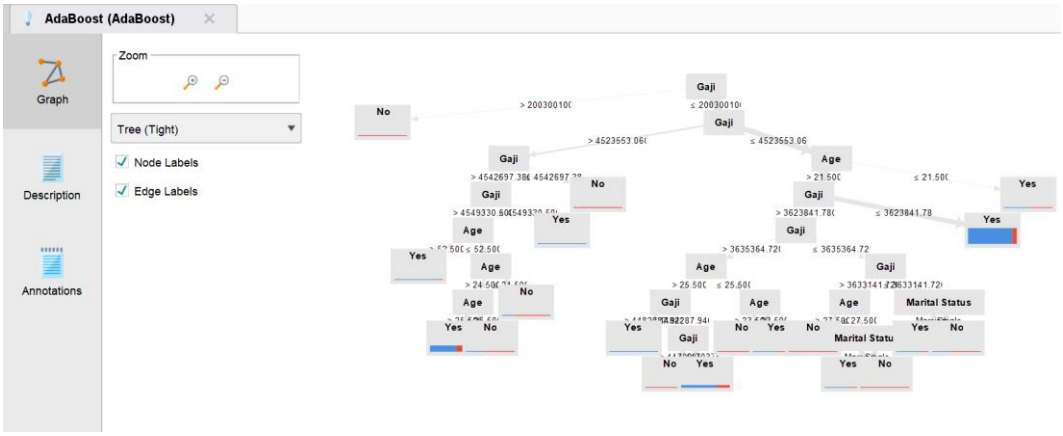
Gambar 3. Hasil Akurasi Algoritma Random Forest

Pada gambar 3 dihasilkan *scoring class recall* pada nilai "Yes" sebanyak 99,74%, *class recall* untuk nilai "No" sebesar 11,62%. *Class precision* untuk nilai "Yes" sebesar 80,74%, *class precision* untuk nilai "No" sebesar 92,27%. Dan diperoleh nilai akurasi secara keseluruhan sebesar 81,04%.




Gambar 4. Hasil Akurasi Algoritma Stacking Usia

Pada gambar 4 dapat dilihat ternyata latar belakang usia karyawan yang diprediksi mendapatkan terminasi mayoritas berusia antara 30-34 tahun, sementara itu latar belakang usia karyawan yang diprediksi paling tinggi tidak mendapatkan terminasi adalah berusia 32-34 tahun.




Gambar 5. Hasil Akurasi Algoritma AdaBoost

Pada gambar 5 dapat dilihat ternyata karyawan yang diprediksi mendapatkan terminasi adalah karyawan dengan gaji <= 4,523,553, dan usia > 21 tahun.

Tabel 5. Hasil Perbandingan Akurasi Algoritma

| Algoritma | Akurasi |
|---|---|
| Random Forest | 81.04% |
| Stacking | 86.84% |
| AdaBoost | 80.30% |

Pada tabel 5 dapat dilihat perbandingan nilai akurasi dari masing-masing algoritma. Algoritma Stacking menghasilkan nilai akurasi terbaik sebesar 86.84%, sedangkan algoritma Random Forest dan Adaboost menghasilkan nilai akurasi masing-masing sebesar 81.04% dan 80.30%.

Turnover karyawan merupakan hal penting dalam sebuah perusahaan, tingkat turnover yang rendah akan membantu perusahaan dalam menggapai objektif dan target yang telah ditetapkan oleh manajemen. Namun, tidak bisa dipungkiri bahwa ada karyawan yang mengundurkan diri untuk meninggalkan perusahaan, oleh karena itu menjadi hal yang sangat esensial bagi perusahaan untuk melakukan tindakan preventif dan persuasif untuk menjaga karyawan yang ada, sehingga tingkat turnover karyawan dapat ditekan, dan operasional perusahaan dapat berjalan dengan baik.

Penelitian sebelumnya yang membahas beberapa metode seperti *Decision Tree, Random Forest, dan AdaBoost.* Namun, terdapat limitasi penelitian pada area dataset yang digunakan, secara spesifik terkait dengan kurangnya jumlah dataset, dan penggunaan dataset yang bersifat public.

Penelitian ini memanfaatkan data private yang diperoleh dari sistem HR perusahaan PT. HCI yang bergerak di bidang jasa keuangan. Teknik algoritma yang digunakan melibatkan prediksi *Random Forest*, *Stacking*, dan *AdaBoost*, untuk memperkirakan kemungkinan turnover karyawan, dan teknik *Majority Voting* digunakan untuk membandingkan nilai akurasi masing-masing algoritma, dan mendapatkan nilai akurasi total. Dataset awal 7,264 baris data dan 22 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 *Random Forest*, *Stacking*, *AdaBoost*, dan *Majority Voting*, diperoleh prediksi untuk turnover karyawan berdasarkan usia, gaji, jenis kelamin, dan status pernikahan.

Hasil prediksi teknik *Stacking* memiliki nilai akurasi tertinggi 86.84%, lebih baik dibandingan dengan nilai akurasi *Random Forest* dan *AdaBoost* dengan nilai masing-masing sebesar 81.04% dan 80.30%.

## 4. DISKUSI

Mempertahankan karyawan yang ada saat ini menjadi objektif dalam setiap perusahaan. Salah satu tujuan utama dari departemen sumber daya manusia adalah mempertahankan karyawan dan menerapkan keahlian karyawannya untuk mencapai visi dan misi sebuah perusahaan.

Penelitian sebelumnya yang membahas turnover karyawan seperti yang sudah dijelaskan di bagian pendahuluan, telah mengeksplorasi beberapa metode seperti *Random Forest, Stacking, Boosting, SVM, C4.5*. Namun, terdapat limitasi penelitian pada jumlah metode yang digunakan, secara spesifik terkait dengan prediksi employee attrition menggunakan algoritma *Support Vector Machine (SVM)*, dengan hanya menggunakan satu metode yaitu *Support Vector Machine (SVM)*.

Penelitian ini mengadopsi metode *Ensemble* yang terdiri dari beberapa algoritma yaitu *Random Forest, Stacking (Decision Tree, KNN, Gradient Boosted Trees* sebagai *base learner*, dan *Naïve Bayes* sebagai *model learner*), dan *AdaBoost (Decision Tree).* Dataset awal 9080 baris data dan 22 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 *Random Forest, Stacking (Decision Tree, KNN, Gradient Boosted Trees sebagai base learner, dan Naïve Bayes sebagai model learner), dan AdaBoost (Decision Tree)*, diperoleh prediksi untuk turnover karyawan berdasarkan usia, gaji, jenis kelamin, dan status pernikahan.

Hasil prediksi teknik *Stacking* memiliki nilai akurasi tertinggi 86.84%, lebih tinggi dibandingkan dengan teknik *Random Forest* dan *AdaBoost* yang memiliki nilai akurasi masing-masing sebesar 81.04% dan 80.30%..

## 5. KESIMPULAN

Penerapan metode *Ensemble* menggunakan 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 Random Forest* sebanyak 81,04%, *Stacking* sebanyak 86,84%, dan *AdaBoost* sebanyak 80,30%.

Berdasarkan hasil pengujian, dapat disimpulkan bahwa:

- Algoritma *Stacking* adalah metode *Ensemble* terbaik karena menghasilkan nilai akurasi tertinggi sebanyak 86,84% dibandingkan dengan algoritma *Ensemble* lainnya yaitu *Random Forest* dan *AdaBoost*
- Peluang untuk penelitian lebih lanjut 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 tindakan preventif dan persuasif untuk menjaga karyawan yang ada, sehingga resiko turnover karyawan yang tinggi dapat dihindari, dan perusahaan dapat berjalan dengan optimal.

## DAFTAR PUSTAKA

[1] E. Hardiyanto, "Kombinasi Metode Ensemble, Cfs Dan Pohon Keputusan Untuk Prediksi Kinerja Petugas Studi Kasus: Survey Podes Badan Pusat Statistik," *Joutica*, vol. 5, no. 1, p. 337, 2020, doi: 10.30736/jti.v5i1.390.

[2] W. Purba, W. Siawin, and . H., "Implementasi Data Mining Untuk Pengelompokkan Dan Prediksi Karyawan Yang Berpotensi Phk Dengan Algoritma K-Means Clustering," *J. Sist. Inf. dan Ilmu Komput. Prima(JUSIKOM PRIMA)*, vol. 2, no. 2, pp. 85–90, 2019, doi: 10.34012/jusikom.v2i2.429.

[3] J. T. Samudra, B. H. Hayadi, and P. S. Ramadhan, "Komparasi 3 Metode Algoritma Klasifikasi Data Mining Pada Prediksi Kenaikan Jabatan," *J-SISKO TECH (Jurnal Teknol. Sist. Inf. dan Sist. Komput. TGD)*, vol. 5, no. 2, p. 127, 2022, doi: 10.53513/jsk.v5i2.5642.

[4] A. S. Abiyyu, "Perbandingan Metode Seleksi Fitur untuk Mengoptimasi Model Support Vector Machine dalam Memprediksi Turnover Pegawai," vol. 10, no. 2, p. 1921, 2023.

[5] M. A. Al Fatih and K. M. Lhaksmana, "Prediksi Employee Attrition menggunakan Algoritma Support Vector Machine (SVM)," *J. Prediksi Empl. Attrition*, vol. 10, no. 2, p. 1930, 2023.

[6] T. H. L. Tulus, A. Id Hadiana, and I. Santikarama, "Sistem Prediksi Awal Terhadap Atrisi Karyawan Menggunakan Algoritma C4.5," *Informatics Digit. Expert*, vol. 4, no. 1, pp. 18–24, 2022, doi: 10.36423/index.v4i1.882.

[7] L. S. Shubham Karande, "Prediction of Employee Turnover Using Ensemble Learning," vol. 904, no. July, pp. 339–351, 2019, doi: 10.1007/978-981-13-5934-7.

[8] D. B. S. Louis Madaerdo Sotarjua, "Perbandingan Algoritma Knn, Decision Tree,*Dan Random*Forest Pada Data Imbalanced Class Untuk Klasifikasi Promosi Karyawan," *J. INSTEK (Informatika Sains dan Teknol.*, vol. 7, no. 2, pp. 192–200, 2022, doi: 10.24252/instek.v7i2.31385.

[9] M. A. Hossen, E. Hossain, A. K. Z. Ishwar, and F. Siddika, "Ensemble method based architecture using random forest importance to predict employee's turn over," *J. Phys. Conf. Ser.*, vol. 1755, no. 1, 2021, doi: 10.1088/1742-6596/1755/1/012039.

[10] D. Manurung, F. Sandi, F. Akinardipura, H. AShfahan, and D. Prasvirta, "Prediksi Pengunduran Diri Karyawan Perusahaan 'Y' Menggunakan," *Semin. Nas. Mhs. Imu Komput. dan Apl.*, vol. 2, no. 2, pp. 202–213, 2021.

[11] M. Atef, D. S. Elzanfaly, and S. Ouf, "Early Prediction of Employee Turnover Using Machine Learning Algorithms," *Int. J. Electr. Comput. Eng. Syst.*, vol. 13, no. 2, pp. 135–144, 2022, doi: 10.32985/IJECES.13.2.6.

[12] A. Nurmasani and Y. Pristyanto, "Algoritme Stacking Untuk Klasifikasi Penyakit Jantung Pada Dataset Imbalanced Class," *Pseudocode*, vol. 8, no. 1, pp. 21–26, 2021, doi: 10.33369/pseudocode.8.1.21-26.

[13] B. M. Karomah, "Penerapan Metode

Stacking Dalam Mengklasifikasikan Penderita Penyakit Diabetes," *Jupikom*, vol. 1, no. 3, 2022.

[14] A. Ferdita Nugraha, R. F. A. Aziza, and Y. Pristyanto, "Penerapan metode Stacking dan Random Forest untuk Meningkatkan Kinerja Klasifikasi pada Proses Deteksi Web Phishing," *J. Infomedia*, vol. 7, no. 1, p. 39, 2022, doi: 10.30811/jim.v7i1.2959.

[15] N. O. Syamsiah and I. Purwandani, "Penerapan Ensemble Stacking untuk Peramalan Laba Bersih Bank Syariah Indonesia (BSI)," *Build. Informatics, Technol. Sci.*, vol. 3, no. 3, pp. 295–301, 2021, doi: 10.47065/bits.v3i3.1017.

[16] M. R. Akbar *et al.*, "Prediksi Retweet Berdasarkan User-Based dan Content - Based Menggunakan Metode Ensemble Stacking," vol. 10, no. 2, pp. 1950–1962, 2023.

[17] V. I. Ivanoti, Megananda Hervita P., Gandung Triyono, and Dyah Puji Utami, "Decision Support System for Predicting Employee Leave Using the Light Gradient Boosting Machine (Lightgbm) and K-Means Algorithm," *J. Tek. Inform.*, vol. 4, no. 3, pp. 657–667, 2023, doi: 10.52436/1.jutif.2023.4.3.1084.

[18] F. K. Alsheref, I. E. Fattoh, and W. Mead, "Automated Prediction of Employee Attrition Using Ensemble Model Based on Machine Learning Algorithms," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/7728668.

[19] R. Ahsana, R. Rohmat Saedudin, and V. P. Widartha, "Perbandingan Akurasi Algoritma Adaboost Dan Algoritma Lightgbm Untuk Klasifikasi Penyakit Diabetes," *e-Proceeding Eng.*, vol. 8, no. 5, pp. 9738–9748, 2021.

[20] F. I. Kurniadi and P. D. Larasati, "Light Gradient Boosting Machine untuk Deteksi Penyakit Stroke," *J. SISKOM-KB (Sistem Komput. dan Kecerdasan Buatan)*, vol. 6, no. 1, pp. 67–72, 2022, doi: 10.47970/siskom-kb.v6i1.328.

[21] S. Diantika, "Penerapan Teknik Random Oversampling Untuk Mengatasi Imbalance Class Dalam Klasifikasi Website Phishing Menggunakan Algoritma Lightgbm," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 1, pp. 19–25, 2023, doi: 10.36040/jati.v7i1.6006.