J-INTECH (Journal of Information and Technology)
Terakreditasi Kemendikbud SK No. 204/E/KPT/2022
E-ISSN: 2580-720X || P-ISSN: 2303-1425

Model Klasifikasi Penyebab *Turnover* Karyawan Menggunakan Kerangka Kerja CRISP-DM

Daud Fernando[1*], Rangga Gelar Guntara[2]

[1] Rekayasa Perangkat Lunak, Universitas Pendidikan Indonesia, Jalan Pendidikan No 15, Kec. Cileunyi, Kab. Bandung, Jawa Barat, 40625, Indonesia
[2] Bisnis Digital, Universitas Pendidikan Indonesia, Jl. Lingkar Dadaha No.18, Nagarawangi, Kec. Cihideung, Kab. Tasikmalaya, Jawa Barat, 46124, Indonesia

***Email Korespondensi:**
daudfernando@upi.edu

Abstrak

Permasalahan tingkat kemunduran diri (turnover) karyawan yang tinggi di sebuah perusahaan menimbulkan beberapa dampak negatif dari sisi biaya, tenaga, maupun waktu dan salah satunya dirasakan oleh Perusahaan fiktif "XYZ". Tujuan penelitian ini akan mengklasifikasi penyebab turnover karyawan di industri menggunakan model pembelajaran mesin klasifikasi pada dua algoritma berbeda yaitu Random Forest dan Decision Tree. Selain itu, penelitian ini menjawab implikasi dari penelitian klasifikasi sebelumnya, klasifikasi karyawan di industri pendidikan, yang menyarankan untuk mengomparasi evaluasi dua performa model pembelajaran mesin. Terdapat 10 variabel dan 9.540 data historis karyawan yang digunakan dalam penelitian. Teknik atau metode penelitian yang digunakan adalah Cross-industry Standard Process for Data Mining (CRISP-DM). Hasil dari penelitian ini menunjukkan model klasifikasi Random Forest adalah model pembelajaran mesin yang optimal dengan nilai AUC – ROC mencapai 0.9988. RapidMiner digunakan untuk memvalidasi kembali perfora model pembelajaran mesin menggunakan parameter yang sama dan dihasilkan nilai akurasi tertinggi sebesar 85.04% untuk model Random Forest dibandingkan Decion Tree.
Kata Kunci : CRISP-DM; Decision Tree; Klasifikasi; Random Forest; RapidMiner

Abstract

The problem of high employee turnover in a company has several negative impacts in terms of cost, energy, and time and one of them is felt by the fictitious Company "XYZ". The purpose of this research is to classify the causes of employee turnover in the industry using a classification machine learning model on two different algorithms namely Random Forest and Decision Tree. In addition, this study addresses the implications of previous classification research, employee classification in the education industry, which suggests comparing the evaluation of two machine learning model performances. There are 10 variables and 9,540 historical employee data used in the research. The research technique or method used is Cross-industry Standard Process for Data Mining (CRISP-DM). The results of this study show that the Random Forest classification model is the optimal machine learning model with an AUC - ROC value reaching 0.9988. RapidMiner was used to revalidate the performance of the machine learning model using the same parameters and resulted in the highest accuracy value of 85.04% for the Random Forest model compared to the Decion Tree model.
Keywords: Classification; CRISP-DM; Decision Tree; Random Forest; RapidMiner

1. Pendahuluan

Sumber daya manusia bagi sebuah perusahaan adalah aset yang paling berharga. Mereka mengambil peran dalam sebuah perusahaan untuk perkembangan, penguatan, dan juga perubahan kultur untuk meningkatkan keuntungan atau mengurangi pengeluaran (Aris, 2023). Umumnya di sebuah perusahaan memiliki departemen atau penanggung jawab dalam mengelola sumber daya manusia yang dinamakan departemen Human

Resources (HR). Permasalahan yang sering dihadapi oleh departemen HR, berdasarkan sebuah survei terhadap 500 karyawan di beragam lini bisnis yang dilakukan pada tahun 2021, adalah sebanyak 16% (1 dari 6 karyawan) yang tidak senang dengan pekerjaannya dan sedang aktif mencari pekerjaan lainnya (Richardson, 2021). Faktanya, terdapat lima alasan utama mengapa seorang karyawan turnover dari sebuah perusahaan dan berpindah ke perusahaan lainnya, dua di antaranya adalah nominal gaji dan promosi yang diberikan (Holliday, 2021).

Penelitian menggunakan model klasifikasi Logistic Regression dapat membantu mengidentifikasi penyebab turnover karyawan di industri pendidikan. Namun, permasalahan dari penelitian ini adalah memunculkan bias karena hanya menggunakan satu model pembelajaran mesin saja dan juga model logistic regression tidak dapat mengatasi kasus *missing values* (Effendi, 2023). Selain itu, dalam kasus HR analytics lainnya, yang menggunakan model random forest, terdapat kekurangan penelitian mengenai pengimplementasian model machine learning menggunakan decision tree karena meminimalisir kompleksitas prediksi model yang digunakan (Pradana, 2024). Selain itu, algoritma decision tree juga telah bermanfaat untuk mengklasifikasikan keputusan pembelian konsumen di industri farmasi dengan tingkat akurasi sebesar 80% dan terkategorisasi sebagai model yang cukup baik dalam melakukan prediksi dari data historis (Ardhana, 2024).

Penelitian ini akan menggunakan himpunan data dari perusahaan fiktif "XYZ" yang memiliki total data historis karyawan sebanyak 9540 baris data pada 10 variabel. Tujuan dari penelitian ini adalah mengklasifikasi penyebab karyawan turnover dan membandingkannya menggunakan dua algoritma Random Forest dan juga Decision Tree pada nilai akurasinya. Manfaat dari penelitian ini adalah menjadi dasar bagi manjerial perusahaan untuk mengoptimalkan pada faktor penentu model klasifikasi yang diimplementasi agar menekan jumlah karyawan turnover di perusahannya. Metode penelitian yang digunakan adalah kerangka kerja Cross Industry Standard Process Data Mining (CRISP-DM). Penelitian klasifikasi menggunakan kerangka kerja ini berhasil meningkatkan rata-rata akurasi yang berkisar pada persentase 85% karena memiliki pola pengerjaan yang terstruktur dalam implementasi model pembelajaran mesin (Hidayati, 2021).

Penelitian ini akan menggunakan Jupyter Notebook dalam mengeksplorasi data dan juga penyetelan parameter dalam masing – masing algoritma yang sejalan dengan penelitian (Alawi, 2024). Kemudian dilakukan komparasi performa pada metrik akuradi di dua algoritma tersebut. Kebaruan dari penelitian ini adalah penggunaan sebuah alain lain yaitu RapidMiner dalam pembuatan visualisasi model yang parameternya telah disesuaikan sebelumnya di Jupyter Notebook. Kelemahan dari penelitian ini adalah diperlukan komparasi menggunakan model klasifikasi lainnya seperti Logistic Regression untuk memvalidasi akurasi model yang digunakan. Selain itu juga, himpunan data dari Perusahaan fiktif juga dapat membuat bias dalam penelitian klasifikasi yang dilakukan karena tidak terlalu sesuai dengan kondisi industry di dunia nyata.

2. Metode Penelitian

Data mining atau penambangan data adalah serangkaian proses untuk menganalisis pola dan juga nilai – nilai tersembunyi dari sebuah variabel dalam data set yang besar. Dalam menambang sebuah data set, ada lima jenis studi kasus yang dilakukan untuk mencapai pengetahuan yang diinginkan. Lima peran penambangan data di antaranya estimasi, prediksi, klasifikasi, klastering, dan juga asosiasi (Singgalen, 2024).

Penelitian ini akan menggunakan sebuah kerangka kerja yang marak digunakan oleh beragam industri untuk menambang sebuah data dan bahkan telah menjadi standar yang dinamakan dengan CRISP-DM. CRISP-DM ialah metode – metode sekuensial dan dijadikan sebagai standar baku dalam menambang data agar dapat diterapkan ke beragam strategi pemecahan masalah umum di suatu industri guna mencapai sebuah luaran pengetahuan (Shedriko & Firdaus, 2022). Gambar 1 adalah serangkaian proses yang dilakukan untuk menerapkan kerangka kerja CRISP-DM yang diawali semuanya dengan akusisi data yang kemudian dilakukan analisis lanjutan melalui enam tahapan kerja.

Gambar 1. Kerangka kerja CRISP-DM

Tahapan pertama adalah Business Understanding, yaitu mendefinisikan tujuan dari sebuah himpunan data yang dimiliki. Tahapan kedua adalah Data Understanding yaitu memahami bentuk, dimensi, dan bahkan menginterpretasi setiap kolom pada sebuah himpunan data. Tahap ketiga merupakan Data Preparation atau bisa dikenal juga sebagai Data Preprocessing. Tahap keempat ialah Modelling, di tahap ini penentuan peran serta algoritma yang cocok akan mulai dieksekusi untuk mencapai pengetahuan yang diinginkan. Tahap kelima merupakan tahap evaluation yang akan menilai seberapa baik kinerja model yang dibuat dalam mengimplementasi peran penambangan data. Lalu fase terakhir ialah Deployment yakni mengembangkan model yang sudah dibuat agar dapat berintegrasi pada antarmuka lalu bisa digunakan oleh berbagai pengguna (Sutisna, 2022).

Sebagai upaya meningkatan performa dari evaluasi metrik model pembelajaran mesin, maka akan dilakukan konfigurasi hyperparameter untuk algoritma decision tree dan juga random forest. Hyperparameter tuning dapat meningkatkan pengevaluasian pada metrik akurasi keseluruhan model klasifikasi yang berada pada rentang 80% akurasi (Jamiluddin, 2024). Penelitian ini akan menggunakan hyperparameter tuning pada argument *max_depth* dan juga *min_samples_leaf* pada kedua model decion tree dan random forest. Evaluasi data terhadap kondisi real-world perlu dilakukan, terutama untuk membuktikan validitas model dengan cara menggunakan Rapid Miner dan mengomparasi antara kedua model klasifikasinya agar mendapatkan model klasifikasi terbaik.

3. Hasil

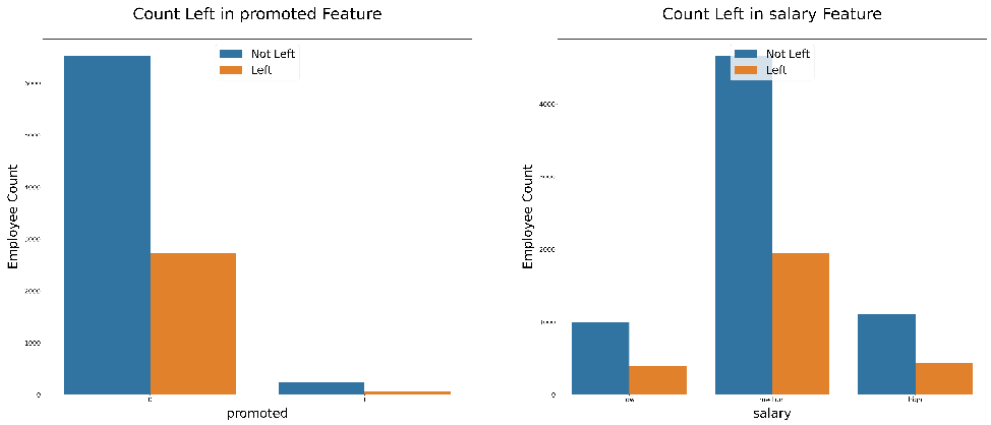
Tahapan pertama yang berhasil didapatkan dari penelitian ini adalah pemahaman studi kasus dari sebuah perusahaan XYZ pada departemen HR mengenai keluar masuknya karyawan yang ada. Pimpinan di perusahaan XYZ mulai merasa khawatir mengenai lonjakan karyawan yang akan meninggalkan perusahaannya karena dapat meningkatkan modal lebih besar lagi untuk merekrut karyawan baru. Selain itu, pimpinan perusahaan juga ingin mengetahui terkait variabel apa yang menjadi penentu seseorang keluar dari perusahaan dan berasal dari departemen mana yang mayoritas keluar dari perusahaan XYZ. Himpunan data yang dikumpulkan oleh departmen HR tersedia pada Tabel 1.

Tabel 1. Lima Himpunan Data Teratas

department	promoted	review	projects	salary	tenure	satisfaction	bonus	avg_hrs_month	left
operations	0	0.57757	3	low	5	0.62676	0	180.866	no
operations	0	0.7519	3	medium	6	0.44368	0	182.708	no
support	0	0.72255	3	medium	6	0.44682	0	184.416	no
logistics	0	0.67516	4	high	8	0.44014	0	188.708	no

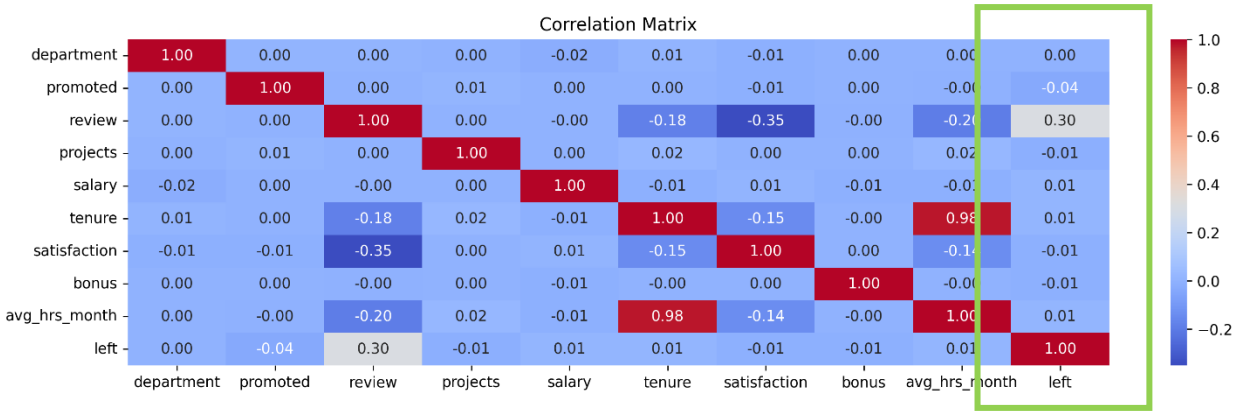
Masing – masing variabel tersebut memiliki artian : "department" - departemen milik karyawan, "promoted" - bernilai 1 jika karyawan dipromosikan sebelum dua tahun, 0 sebaliknya, "review" - penilaian performa kinerja karyawan dengan jangkauan 0 sampai 1, "projects" - total proyek yang terlibat untuk diikuti, "salary" - untuk kerahasiaan, gaji dikelompokkan menjadi low, medium, high, "tenure" - sudah berapa lama karyawan menetap di Perusahaan, "satisfaction" - kepuasan karyawan terhadap perusahan dengan jangkauan 0 sampai 1, "bonus" - nilai 1 untuk karyawan yang menerima bonus, 0 sebaliknya, "avg_hrs_month" - rata-rata jam kerja dalam satu bulan, "left" - "yes" untuk karyawan yang mengundurkan diri, "no" untuk sebaliknya.

Exploratory Data Analysis (EDA) digunakan untuk menganalisis dan juga menginvestigasi sebuah himpunan data untuk dilakukan penemuan wawasan baru mengenai data tersebut. Umumnya dapat dilakukan dalam beberapa bentuk, seperti univariate analysis, bivariate analysis, dan juga multivariate analysis. Setelah dilakukan proses EDA pada himpunan data perusahaan XYZ didapatkan sebuah grafik beberapa variabel dengan nilai dari kelasnya atau EDA berjenis bivariate yang ditunjukkan pada Gambar 3.



Gambar 2. EDA Bivariate

Gambar 2 menunjukkan masing – masing pengelompokkan nilai beberapa variabel dengan kelas di himpunan data karyawan. Karyawan yang meninggalkan perusahaan ditandai dengan balok berwarna oranye, sedangkan yang berwarna biru adalah jumlah karyawan yang tidak meninggalkan perusahaan. Gambar 2 sebelah kiri merupakan keterhubungan dengan variabel "promoted" yang artinya karyawan tanpa ada promosi akan cenderung meninggalkan perusahaan. Sedangkan Gambar 2 sebelah kanan menunjukkan nilai variabel "salary" medium atau rataan tengah dari semua karyawan rentan sekali turnover di perusahaan fiktif ini.



Gambar 3. Matriks Korelasi Himpunan Data Karyawan

Gambar 3 adalah proses analisis untuk menentukan variabel mana yang berkorelasi dengan turnover-nya suatu karyawan. Dapat dilihat pada kotak hijau, variabel "review" cukup berkorelasi dengan turnover karyawan sebesar 30% kemudia di susul pada faktor dipromosi atau tidaknya karyawan tersebut di variable "promoted" sebesar 4%. Tabel 2 merupakan analisis risk ratio yang digunakan terutama dalam masalah klasifikasi dan analisis statistik untuk mengevaluasi kemungkinan suatu peristiwa terjadi pada satu kelompok relatif terhadap kelompok lainnya (Shafie, 2024).

Tabel 2. Risk ratio berdasarkan variabel Review per 4 Quartile

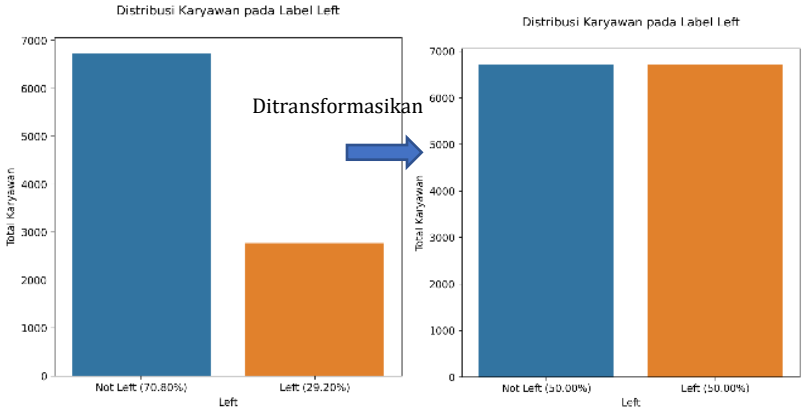
Rentang Nilai Variabel Review	Tidak Turnover	Turnover
(0.309, 0.593]	80.8	19.2
(0.593, 0.647]	80.55	19.45
(0.647, 0.708]	75.64	24.36
(0.708, 1.0]	46.29	53.71

Tabel 2 menunjukkan karyawan yang diberikan nilai "review" pada rentang 70% - 100% cenderung melakukan turnover dibandingkan yang tidak. Hal ini dapat menjadi basis investigasi lanjutan oleh perusahaan untuk melakukan survei kembali pada karyawan yang diberikan review relatif bagus dibandingkan kategori karyawan lain.

Proses CRISP-DM dilanjutkan dengan pra-pemrosesan yang dilakukan meliputi penanganan Data Outlier, Pengkodean Variabel, Transformasi Data Statistik, Penanganan Ketidakseimbangan Kelas dengan teknik SMOTE, dan Pemodelan Pembelajaran Mesin. Data outlier merupakan nilai dalam sebuah variabel yang langka atau terlalu jauh dalam jangkauan kuartil yang ada (Faradisa et al., 2021). Hal ini akan memperburuk kualitas sebuah model pembelajaran mesin karena ada kemungkinan akan mencapai kondisi overfitting terhadap data latihnya saja. Penanganan masalah data outlier melalui penggunaan z-score yang akan mereduksi rentang nilai lebih dari 3 atau kurang dari -3. Hal tersebut didasari dari aturan empiris yang akan mendapatkan persentase data sebanyak 99.7% dengan tiga standar deviasi dari nilai rataannya (Marques, 2023). Pada penelitian ini data outlier dari keseluruhan nilai di variabel berhasil disesuaikan dan terdapat 50 data outlier yang berhasil dihilangkan.

Penelitian yang telah dilakukan, variabel – variabel nominal tersebut dapat dilakukan pra-prosesing berupa pengkodean untuk mengabstraksi satu variabel berdasarkan kelompok nilai nominal yang ada dengan nilai biner berupa 0 atau 1 supaya dapat digunakan ke dalam pemodelan pembelajaran mesin (Maylani et al., 2022). Teknik standarisasi data menggunakan StandardScaler akan merubah distribusi data numerik dengan nilai rataan 0 dan nilai standar deviasinya menjadi 1 (Ihsani et al., 2020). Terbukti dapat meningkatkan nilai akurasi dari sebuah model hingga mencapai nilai maksimal 90,2%.

Ketidakseimbangan kelas adalah permasalahan yang ditemui dalam penelitian ini, di mana jumlah data observasi terlalu timpang, sehingga kuantitas antar kategori dalam kelas tidak seimbang. Kondisi seperti ini akan cenderung menihilkan kelompok minoritas dan mengutamakan kelompok mayoritas dalam pengklasifikasian modelnya nanti (Purwa, 2019). Untuk menanggapi ketakseimbangan kelas pada himpunan data, akan digunakan teknik Synthetic Minority Oversampling Technique (SMOTE) yang pendekatannya dengan pembuatan replikasi data dari data minoritas berbasis nilai ketetanggan terdekat menggunakan algoritma *k-nearest neighbor* (Jungryeol, 2023). Gambar 4 menunjukkan ketidakseimbangan berdasarkan label yang ada dari rasio 70:30 menggunakan teknik SMOTE berhasil direplikasi data dari kelompok minoritas supaya menjadi 50:50 dengan masing – masing kelompok nilai labelnya sebanyak 6719.

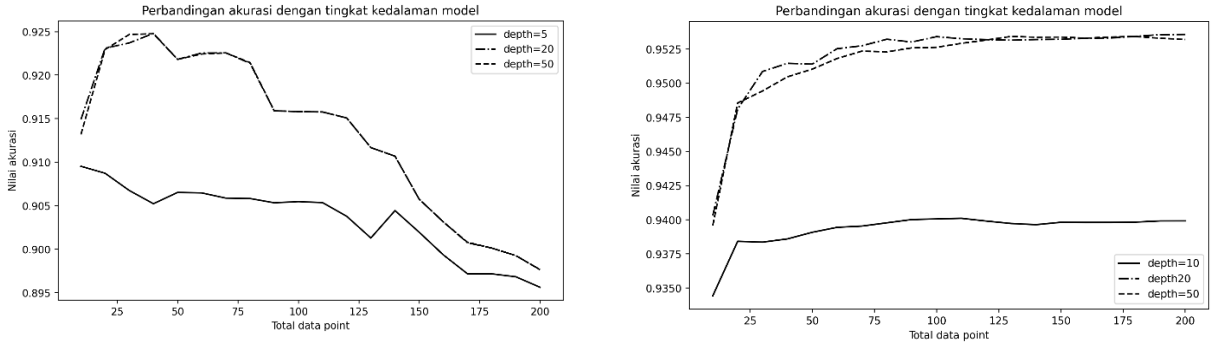


Gambar 4. Proses Sebelum dan Sesudah Penyeimbangan SMOTE

Pembuatan model akan menggunakan modul DecisionTreeClassifier dari paket sklearn.tree untuk pembuatan model dari algoritma Decision Tree, kemudian modul RandomForestClassifier dari paket sklearn.ensemble untuk pembuatan model Random Forest.

Himpunan data akan dibagi terlebih dahulu menjadi data latih dan data validasi. Rasio pembagian data di antara keduanya adalah 80% untuk data latih dan 20% untuk data uji. Dari data latih sendiri, kemudian dibagi kembali menjadi dua bagian. Bagian pertama sebagai data latih yang sebesar 67% dan data 33% untuk data validasi. Secara umum, kedua model tersebut memiliki parameternya masing – masing yang dapat disesuaikan dengan kondisi data himpunan yang ada. Parameter – parameter tersebut dapat disetel secara manual maupun otomatis menggunakan beberapa library yang tersedia.

Perbedaan performa antara kedua algoritma yang digunakan ditentukan dalam menentukan parameter modelnya atau disebut dengan hyperparameter tuning. Keterbatasan GPU dalam penelitian ini maka akan melakukan pada parameter depth saja untuk di kedua model dan mengambil nilai parameter yang kerap menghasilkan evaluasi akurasi yang optimal (Patange, 2023). Gambar 5 sebelah kiri adalah hyperparameter tuning untuk menentukan ordo parameter depth model dari Decision Tree, sedangkan Gambar 6 sebelah kanan digunakan untuk hyperparameter tuning dari model Random Forest.

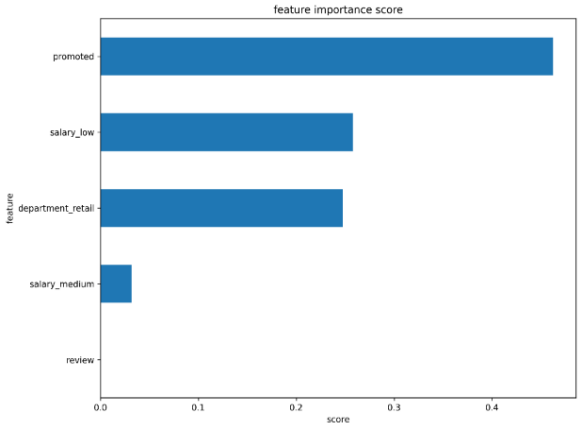


Gambar 5. Hyperparameter tuning model Decision Tree (Kiri) dan model Random Forest (kanan)

Gambar 5 sebelah kiri menunjukkan parameter depth bernilai 20 di model Random Forest menghasilkan nilai akurasi tertinggi dibandingkan model Decision Tree. Hal ini sejalan dengan penelitian yang menggunakan Random Forest untuk melakukan klasifikasi di bidang kesehatan (Ordila et al., 2020). Penelitian ini akan memberikan kebaruan dalam menggunakan model Decision Tree untuk kasus HR analytics.

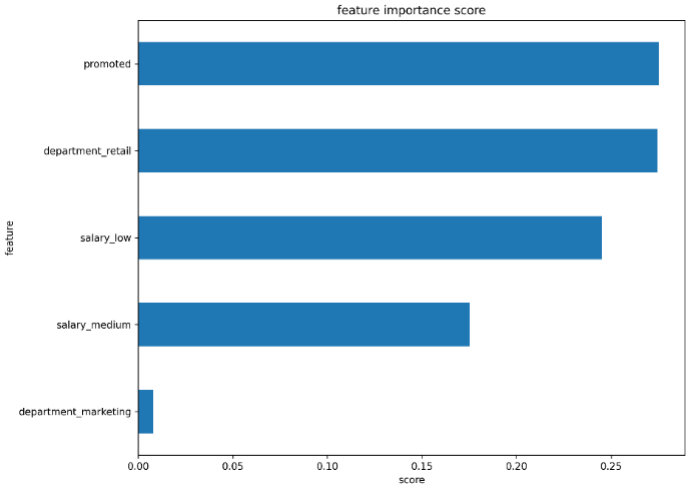
4. Pembahasan

Keterbatasan penelitian ini adalah hanya mengguanakan dua model klasifikasi untuk menentukan komparasi model terbaik dalam prediksi karyawan turnover atau tidak. Pada bagian pembahasan, penulis menginterpretasikan hasil penelitian berdasarkan apa yang telah diketahui. Ketika telah tercipta model dari masing – masing algoritma, saatnya mengevaluasi performansi dari tiap model yang ada. Pengevaluasian berdasarkan nilai dari AUC – ROC sebuah model dan juga nilai confusion matrixnya. Sebelum mendapatkan nilai AUC – ROC-nya seluruh data latih dan juga data validasi dilakukan transformasi kembali ke dalam bentuk vektor menggunakan modul DictVectorizer yang terbukti dapat meningkatkan setiap variabel yang melatih modelnya (Maulana et al., 2019). Pada proses pengevaluasian model Decision Tree (DT) didapatkan sebuah nilai AUC – ROC, dengan penyesuaian argumen pada dengan kedalaman atau ketinggian pohon bernilai 6 dan masing – masing daun atau node harus memiliki 6 buah sampel juga, sebesar 0.937 untuk nilai AUC – ROC yang didapatkan pada data latihnya. Gambar 6 menunjukkan beberapa variabel yang menentukan seseorang akan keluar dari perusahaan.



Gambar 6. Variabel penentu utama algoritma DT

Didapatkan bahwa variabel promoted, salary_low, departement_retail, dan juga salary_medium merupakan empat variabel tertinggi dalam pembuatan model ini. Lain halnya dengan model pada Random Forest (RF) yang memiliki pengaturan argumen terhadap parameternya berupa nilai maksimal kedalaman pepohonanannya adalah 50 dan juga minimal kuantitas di setiap daunnya sebanyak 20 sampel menghasilkan nilai AUC – ROC sebesar 0.948 yang didapatkan dari data latihnya. Dengan variabel yang sangat berpengaruh pada model ini ditunjukkan pada Gambar 7.



Gambar 7. Variabel penentu utama algoritma RF

Gambar 8 adalah evaluasi kinerja dari model klasifikasi yang dibentuk menggunakan AUC-ROC pada data ujinya. Setelah melakukan proses hyperparameter tuning dan juga pra pemrosesan data di awal penelitian menunjukkan keberhasilan model yang diciptakan. Kedua model cenderung menghasilkan nilai evaluasi pada 99%, namun model random forest memberikan evaluasi yang lebih besar dibandingkan model decision tree sebanyak 0,005% dan berpengaruh kepada hasil dari keakuratan prediksi karyawan turnover ke depannya.

```
rf.fit(X_train,y_train)
y_pred = rf.predict_proba(X_test)[:,1]
roc_auc_score(y_test, y_pred)
```

```
0.9988492966200995
```

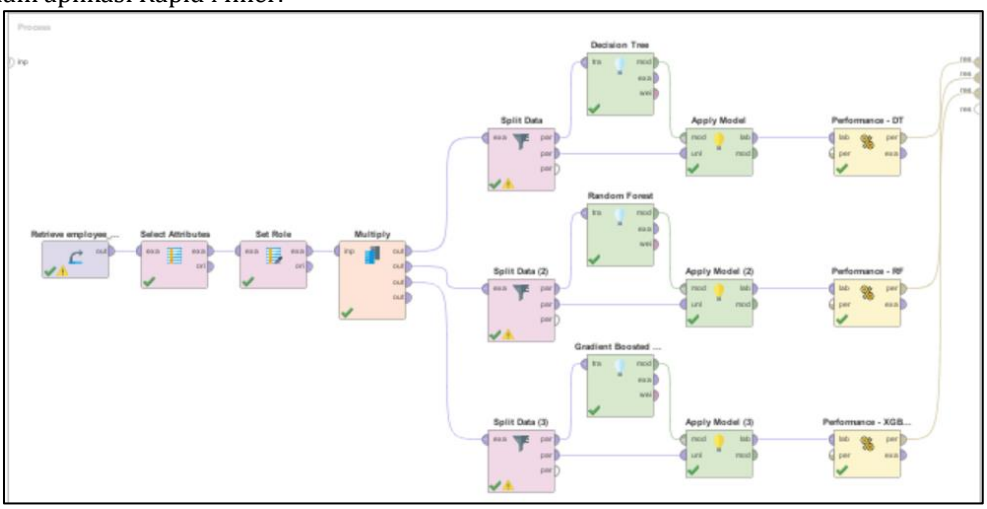
```
dt.fit(X_train,y_train)
pred = dt.predict_proba(X_test)[:,1]
roc_auc_score(y_test,pred)
```

```
0.9983877416310971
```

Gambar 8. Nilai AUC – ROC data uji

Berlandaskan hal ini, maka dapat ditarik sebuah pernyataan bahwa model yang menggunakan algoritma Random Forest adalah yang terbaik, khususnya untuk menggunakan himpunan data dan juga teknik pra-pemrosesan yang sama.

Evaluasi yang terakhir ialah menggunakan nilai confusion matrix menggunakan aplikasi penambangan data yaitu Rapid Miner. Secara praktikal akan dibandingkan nilai confusion matrix dengan data himpunan yang tidak dilakukan transformasi DictVectorizer pada Rapid Miner. Berikut panel kerja untuk pembentukan evaluasi dalam aplikasi Rapid Miner.



Gambar 9. Panel kerja di Rapid Miner

Berdasarkan proses yang ditampilkan Gambar 9, alur dalam Rapid Miner sama saja seperti dalam pengimplementasian menggunakan Python karena menggunakan data hasil pra-pemrosesan yang sama baik untuk data latih maupun data uji. Dengan nilai standar di setiap parameternya, nilai Confusion Matrix dari kedua model ini di Tabel 3.

Tabel 3. Komparasi Performa Model Algoritma Klasifikasi

Model Algoritma	Akurasi	Presisi	Recall
DT	81.54%	80.08%	83.97%
RF	85.04%	86.72%	82.77%

Dari pengevaluasian yang berhasil dilakukan baik menggunakan nilai AUC – ROC maupun confusion matrix, keduanya menunjukkan bahwa algoritma RandomForest tetap yang menjadi model terbaik dalam mengklasifikasi apakah karyawan akan meninggalkan perusahaan XYZ atau tidak. Keterbatasan dalam model yang dibuat ini akan memberikan probabilitas pada data historis yang digunakan dan tidak mencerminkan atau berkesesuaian dengan data karyawan di masa depan.

5. Kesimpulan

Kerangka kerja CRISP – DM sangat membantu dalam pengerjaan penambangan data yang cukup besar. Mulai dari pendefinisian bisnis hingga tahap evaluasi menjadikan proyek penambangan data mendapatkan hasil yang optimal. Sebagai buktinya, kini departemen HC di perusahaan XYZ dapat mengetahui bahwa variabel review yang memiliki korelitas dengan turnover tertinggi berkat perhitungan nilai rasio risiko dan juga matriks korelasi. Komparasi model klasifikasi menggunakan Decision Tree dan Random Forest menghasilkan model Random Forest yang memiliki akurasi tertinggi melalui evaluasi AUC-ROC dan juga Rapid Miner. Penilaian AUC – ROC terbaik didapatkan sebesar 0.9988 dan nilai akurasi sebesar 85.04% melalui penggunaan model Random Forest. Variabel penentu utamanya ialah variabel promoted, departement_retail, salary_low, dan juga salary_medium. Rekomendasi untuk penelitian selanjutnya adalah menggunakan hyperparameter tuning yang lebih ekstensif lagi pada parameter selain depth dan juga mengombinasikan kedua algoritma yang digunakan dalam algorimta Gradient Boosting sebagai upaya untuk meningkatan akurasi klasifikasi model di HR analytics. Pandangan secara praktis, penelitian ini menganjurkan pada Perusahaan fiktif XYZ untuk mengoptimasi pemberian review setiap karyawan dan juga variabel penentu model perlu ditingkatkan oleh departemen HR.

Referensi

Alawi, A. I. (2024). Machine Learning in Human Resource Analytics: Promotion Classification using Data Balancing Techniques. *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems*, *10*(1), 1001–1021.

Ardhana, V. Y. P. (2024). Analysis Of Medicine Sales Classification Using Decision Tree Method. *Jurnal Teknologi Informasi, Komputer Dan Aplikasinya*, *6*(1), 376–383.

Aris, A. A. (2023). The Role of Management of Human Resources in Enhancing The Quality of Schools. *Journal Of Social Science Research*, *3*(3), 11012–11023. https://www.breathehr.com/en-gb/blog/topic/business-process/why-is-human-resources-important#:~:text=HR plays a key role,business culture covered by HR.

Effendi, M. E. (2023). Prediksi Guru Kemungkinan Tetap Bekerja di Sekolah Al Uswah Surabaya MenggunakanMachine Learning. *Jurnal Informasi Dan Teknologi*, *5*(1), 129–137.

Faradisa, S. M., Nugrahadi, T. D., Muliadi, Budiman, I., & Kartini, D. (2021). *Implementasi IQR-SMOTE Untuk Mengatasi Ketidakseimbangan Kelas Pada Klasifikasi Diabetes menggunakan K-Nearest Neighbors*. *15*, 48–60.

Hidayati, N. (2021). Perbandingan Algoritma Klasifikasi untuk Prediksi Cacat Software dengan Pendekatan CRISP-DM. *Jurnal Sains Dan Informatika*, *7*(2), 117–126.

Holliday, M. (2021). *What Is Employee Turnover & Why It Matters for Your Business*. Netsuite.Com. https://www.netsuite.com/portal/resource/articles/human-resources/employee-turnover.shtml?mc24943=v2

Ihsani, D. A., Arifin, A., & Fatoni, M. H. (2020). Klasifikasi DNA Microarray Menggunakan Principal Component Analysis (PCA) dan Artificial Neural Network (ANN). *Jurnal Teknik ITS*, *9*(1). https://doi.org/10.12962/j23373539.v9i1.51637

Jamiluddin, F. (2024). Implementasi Hyperparameter Tuning Grid Search CV Pada Prediksi Produksi Padi Menggunakan Algoritma Linear Regresi. *Journal of Information System Research (JOSH)*, *6*(1), 490–498.

Jungryeol, P. (2023). A study on improving turnover intention forecasting by solving imbalanced data problems: focusing on SMOTE and generative adversarial networks. *Journal of Big Data*, *10*(1).

Marques, H. O. (2023). On the evaluation of outlier detection and one-class classification: a comparative study of algorithms, model selection, and ensembles. *Data Mining and Knowledge Discovery*, *37*(4).

Maulana, M. A., Bijaksana, M. A., & Huda, A. F. (2019). *Entity Recognition for Quran English Version with Supervised Learning Approach*. *4*, 77–86. https://doi.org/10.21108/indojc.2019.4.3.362

Maylani, I., Rochman, F., Kurniasari, N. D., & Timur, J. (2022). Seleksi Fitur pada Klasifikasi K-Nearest Neighbors untuk Data Churn for Bank Customers dengan Analisis Korelasi. *SNESTIK*.

Ordila, R., Wahyuni, R., Irawan, Y., & Yulia Sari, M. (2020). Penerapan Data Mining Untuk Pengelompokan Data Rekam Medis Pasien Berdasarkan Jenis Penyakit Dengan Algoritma Clustering (Studi Kasus : Poli Klinik PT. Inecda). *Jurnal Ilmu Komputer*, *9*(2), 148–153. https://doi.org/10.33060/jik/2020/vol9.iss2.181

Patange, A. D. (2023). Augmentation of decision tree model through hyper-parameters tuning for monitoring of cutting tool faults based on vibration signatures. *Journal of Vibration Engineering & Technologies*, *11*(8), 3759–3777.

Pradana, R. Y. (2024). Machine Learning Pengklasifikasikan Performa Karyawan Direct Sales Force Kartu Prabayar Menggunakan Metode Random Forest Classifier. *Jurnal Teknik Informatika*, *4*(3).

Purwa, T. (2019). Perbandingan Metode Regresi Logistik dan Random Forest untuk Klasifikasi Data Imbalanced (Studi Kasus: Klasifikasi Rumah Tangga Miskin di Kabupaten Karangasem, Bali Tahun 2017). *Jurnal Matematika, Statistika Dan Komputasi*, *16*(1), 58. https://doi.org/10.20956/jmsk.v16i1.6494

Richardson, B. (2021). *Employee Happiness Statistics & Facts – What Makes Employees Happy? New Research For Q2 2021*. Development-Academy.Co.Uk. https://development-academy.co.uk/news-tips/employee-happiness-statistics-2021/

Shafie, M. R. (2024). A cluster-based human resources analytics for predicting employee turnover using optimized Artificial Neural Networks and data augmentation. *Decision Analytics Journal 11*, *11*(1).

Shedriko, & Firdaus, M. (2022). Penentuan Klasifikasi Dengan Crisp-Dm. *The Indonesian Journal of Computer Science*, *10*(11), 826–831.

Singgalen, Y. A. (2024). Sentiment Classification of The Capsule Hotel Guest Reviews using Cross-Industry Standard Process for Data Mining (CRISP-DM). *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *8*(1), 632–643.

Sutisna, L. A. (2022). Using Feature Engineering In Logistic Regression And Random Forest Methods To Improve Employee Attrition Prediction In Kimia Farma. *INFOKUM*, *10*(5), 1421–1439.