Perbandingan Metode Seleksi Fitur untuk Mengoptimasi Model *Support Vector Machine* dalam Memprediksi *Turnover* Pegawai

1st Ahmad Syafiq Abiyyu
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
ahmadsyafiq@students.telkomuniversity.ac.id

2nd Kemas Muslim Lhaksmana
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
kemasmuslim@telkomuniversity.ac.id

Abstrak-Seleksi fitur merupakan salah satu proses yang dilakukan untuk mengurangi dimensi data. Pengurangan dimensi bertujuan untuk meningkatkan performa model algoritma pembelajaran mesin. Turnover pegawai adalah suatu fenomena yang merujuk pada tingkat pegawai yang keluar dari suatu perusahaan. Penelitian mengenai implementasi algoritma pembelajaran mesin dalam memprediksi turnover pegawai sudah banyak dilakukan. Namun, performa model algoritma support vector machine (SVM) secara umum tidak menghasilkan performa yang baik. Dengan menggunakan metode seleksi fitur, hasil performa algoritma SVM diharapkan dapat menjadi lebih baik dalam memprediksi pegawai yang hendak melakukan turnover. Seleksi fitur digunakan pada dataset turnover pegawai sebelum dipelajari oleh model SVM yang dibangun. Metode seleksi fitur yang digunakan adalah filter methods, wrapper methods, dan embedded method. Penelitian ini menampilkan metode seleksi fitur mana yang paling baik dalam meningkatkan performa dari algoritma SVM. Matriks evaluasi seperti akurasi, recall, presisi, dan f1-score digunakan untuk menilai hasil akhir performan dari model SVM setelah dilakukan seleksi fitur. Hasil yang didapatkan adalah metode wrapper method meningkatkan performa dengan lebih baik dibandingkan metode lain. Nilai performa secara keseluruhan naik sebesar 4% dari performa sebelum dilakukan seleksi fitur.

Kata kunci - turnover pegawai, pembelajaran mesin, *support vector machine*, seleksi fitur

Abstract-*Feature selection is one of the processes carried out to reduce the dimensions of data. Dimension reduction aims to improve the performance of machine learning algorithm models. Employee turnover is a phenomenon that refers to the level of employees leaving a company. Research on the implementation of machine learning algorithms in predicting employee turnover has been widely carried out. However, the performance of the support vector machine (SVM) algorithm model generally does not produce a good performance. By using the feature selection method, the results of the SVM algorithm's performance are expected to be better at predicting employees who want to make a turnover. Feature selection is used in the employee turnover dataset before being studied by the SVM model. The feature selection methods used are filter methods, wrapper methods, and embedded methods. This study shows which feature selection method is the best for improving the performance of the SVM algorithm. Evaluation matrices such as accuracy, recall, precision, and f1-score are used to assess the final performance of the SVM model after feature selection. The result obtained is that the wrapper method improves performance better than other methods. Overall performance value increased by 4% from performance before feature selection.*

Keywords- employee turnover, machine learning, support vector machine, feature selection

I. PENDAHULUAN

A. Latar Belakang

Permasalahan yag dihadapi oleh perusahaan ketika berhadapan dengan karyawannya sendiri salah satunya adalah masalah tingkat perputaran (*turnover*) pegawai dalam perusahaan tersebut. Penting bagi perusahaan untuk dapat mengelola dan mengontrol tingkat perputaran pegawai dengan baik. Tingkat perputaran pegawai sangat berkaitan dengan tingkat rekruitasi suatu perusahaan. Semakin sering pegawai keluar dari perusahaan maka perusahaan juga harus semakin sering untuk melakukan rekruitasi pegawai baru [1]. Hal tersebut, selain dapat menyebabkan pengeluaran biaya yang tinggi, dapat menyebabkan suasana pekerjaan dalam perusahaan menjadi kurang menyenangkan.

Untuk mengatasi permasalahan tersebut, banyak penelitian yang telah membahas tentang bagaimana prediksi *turnover* pegawai menggunakan algoritma pembelajaran mesin. Penggunaan pembelajaran mesin ini dimaksudkan agar hasil analisis mengenai pegawai yang keluar dapat lebih akurat. Algoritma yang digunakan beraneka ragam contoh beberapa di antaranya adalah *decision tree, logistic regression, support vector machine, naïve bayes*, dan algoritma pembelajaran mesin modern lainnya. Dari beberapa hasil penelitian tersebut, beberapa algoritma memiliki hasil yang baik dan beberapa algoritma lain memiliki

hasil yang kurang baik. Salah satu algoritma pembelajaran mesin yang memiliki hasil performa yang kurang baik adalah algoritma *support vector machine*.

Support vector machine (SVM) merupakan algoritma pembelajaran mesin yang menjadi fokus dalam penelitian ini. Berdasarkan hasil pada beberapa penelitian sebelumnya, SVM cenderung memiliki nilai performansi yang lebih buruk dari nilai performansi algoritma pembelajaran mesin lainnya [2]–[4]. Oleh karena itu, penelitian ini berfokus pada bagaimana memaksimalkan performa dari SVM. Salah satu cara yang digunakan adalah dengan melakukan seleksi fitur kepada dataset sebelum dipelajari oleh algoritma SVM.

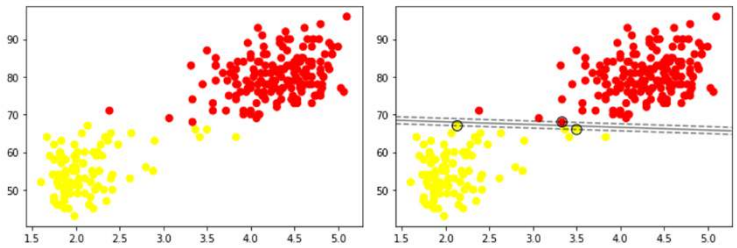
Feature selection atau seleksi fitur adalah suatu metode yang dapat digunakan untuk melakukan pengurangan dimensi terhadap dataset yang diberikan. Seleksi fitur dapat menyebabkan model yang dibangun lebih sederhana dan komprehensif, meningkatkan performa model, dan membantu untuk dapat memahami data [5]. Dalam kasus ini, seleksi fitur diharapkan dapat mengoptimasi hasil evaluasi dari model SVM yang telah dibangun. Sehingga, model SVM memiliki tingkat performansi yang lebih baik dari sebelumnya.

B. Topik dan Batasannya

Turnover pegawai, atau dalam bahasa Indonesia berarti perputaran pegawai, merujuk pada tingkat pergantian atau perputaran pegawai pada suatu perusahaan. Dalam perusahaan, pergantian pegawai merupakan hal yang lumrah terjadi dari waktu ke waktu. Namun, hal yang menjadi masalah adalah ketika pergantian pegawai terjadi secara terus menerus tanpa adanya evaluasi dari pihak perusahaan. Perusahaan tentu memiliki data-data terkait pegawai di perusahaan tersebut baik pegawai dengan status aktif atau pun pegawai dengan status sudah keluar dari perusahaan. Data-data pegawai dapat digunakan sebagai bahan evaluasi oleh pihak pengelola perusahaan, khususnya bagian pengelolaan sumber daya manusia, sehingga perusahaan dapat mengetahui indikasi yang membuat pegawai hendak keluar dari perusahaan.

Dalam teknologi pembelajaran mesin, terdapat berbagai macam metode yang dapat digunakan untuk mengklasifikasi suatu data ke dalam kelompok tertentu sesuai dengan topik data tersebut. Pada data terdapat satu fitur yang mengelompokkan data ke dalam klasifikasi yang diinginkan, fitur tersebut dapat disebut juga dengan variabel terikat. Variabel terikat tersebut dapat dijadikan rujukan dalam mengklasifikasi data ke dalam kelompok tertentu. Terdapat banyak metode pembelajaran mesin yang cukup terkenal untuk dapat mengatasi masalah klasifikasi data di antaranya adalah *naïve bayes*, *random forest*, *decision tree*, dan *support vector machine* (SVM). Untuk penelitian ini, metode pembelajaran mesin yang difokuskan adalah algoritma SVM.

Support vector machine (SVM) merupakan sebuah metode dalam pembelajaran mesin yang dapat digunakan untuk menyelesaikan permasalahan klasifikasi ataupun regresi. Algoritma ini termasuk dalam kategori *supervised learning*. *Supervised learning* berarti data yang akan dijadikan pembelajaran oleh algoritma ini telah memiliki label pada variabel terikat berdasarkan konteks data tersebut. SVM melakukan pendekatan klasifikasi dengan mencari *hyperplane* terbaik dengan jarak antar kelas yang paling optimal. *Hyperplane* adalah sebuah fungsi yang dapat dijadikan pemisah antar kelas dalam suatu data. Gambar 1.1 menunjukkan bagaimana garis *hyperplane* bekerja dalam mengklasifikasi dua jenis data yang berbeda.



GAMBAR 1. 1
CONTOH IMPLEMENTASI SVM DENGAN HYPERPLANE

Seleksi fitur digunakan sebagai metode dalam pengurangan dimensi dari data yang dipelajari oleh model SVM yang dibangun. Dataset dibagi ke dalam beberapa kategori yaitu dataset tanpa seleksi fitur dan dataset dengan seleksi fitur. Dataset dengan kategori seleksi fitur dibagi lagi ke dalam metode seleksi fitur yang digunakan. Hasil evaluasi dari tiap dataset dibandingkan untuk menilai seberapa efektif penggunaan metode seleksi fitur dalam mengoptimasi performa dari pembelajaran mesin.

Untuk menghindari terlalu banyak parameter yang digunakan, penerapan *hyperparameter tuning* tidak digunakan selama proses penelitian. Seleksi fitur dan model yang dibangun hanya menggunakan parameter bawaan yang sudah disediakan oleh *library*. Hal ini bertujuan agar fokus penelitian hanya pada perbandingan metode seleksi fitur mana yang paling baik dalam mengoptimasi model pembelajaran mesin.

C. Tujuan

Tujuan dari penelitian tugas akhir ini adalah untuk dapat membuat model sistem prediksi *turnover* pegawai dengan menggunakan algoritma *support vector machine*, melakukan implementasi metode seleksi fitur ke dalam dataset, dan menganalisis perbandingan hasil performa model algoritma *support vector machine* berdasarkan metode seleksi fitur yang digunakan.

D. Organisasi Tulisan

Penelitian ini disusun berdasarkan lima bab secara berturut-turut adalah Pendahuluan, Studi Terkait, Sistem yang Dibangun, Evaluasi, dan Kesimpulan. Studi Terkait menjelaskan tentang dasar-dasar pemahaman yang perlu diketahui mengenai penelitian yang dilakukan. Subbab Studi Terkait antara lain sebagai berikut: Penelitian Terkait, *Turnover* Pegawai, *Support Vector Machine*, dan *Feature Selection*. Sistem yang Dibangun menjelaskan tentang tahapan dalam membangun sistem dimulai dari dataset yang digunakan sampai implementasi pembelajaran mesin. Evaluasi menjelaskan tentang hasil dari sistem yang dibangun serta hasil analaisis dari matrik evaluasi yang digunakan. Kesimpulan berisi rangkuman hasil dan kaitannya dengan tujuan penulisan serta saran untuk penilitian yang akan datang.

II. KAJIAN TEORI

A. Penelitian Terkait

Penelitian terkait dengan prediksi turnover pegawai dapat dilihat pada Tabel 2.1. Hasil yang ditampilkan dalam Tabel 2.1 merupakan hasil dari penelitian untuk algoritma SVM.

TABEL 2. 1
PENELITIAN TERKAIT PREDIKSI TURNOVER PEGAWAI

No	Judul	Penulis/Tahun	Tujuan Penelitian	Hasil
1.	Employee Turnover Prediction with Machine Learning: A Reliable Approach [2]	Yue Zhao, Maciej K. Hryniewicki, Fracncesca Cheng, Boyang Fu, dan Xiaoyu Zhu/2019	Menyajikan deskripsi, demonstrasi, dan penilaian yang komprehensif terhadap pendekatan pembelajaran mesin untuk memprediksi *turnover* pegawai.	SVM termasuk dalam kategori metode pembelajaran mesin yang memiliki performa rendah. Nilai performa terendah yang diperoleh adalah 0,55.
2.	A Review On Employee Attrition Using Machine Learning [3]	Vishak Amin, JayantKumar A Rathod, Kshama, Mayuresh Kunder, Pathiksha Patkar/2021	Menyajikan struktur dalam memprediksi *turnover* pegawai menggunakan teknik klasifikasi	Dari tiga metode yang digunakan (*random forest, decision tree,* dan SVM), *random forest* memiliki hasil yang paling baik
3.	Prediction of Employee Turnover in Organizations using Machine Learning Algortihms [4]	Rohit Punnoose, Pankaj Ajit/2016	Mengeksplorasi pengaplikasian metode *extreme gradient boosting* sebagai peningkatan algoritma tradisional	SVM dengan menggunakan RBF kernel memiliki nilai akurasi 0,68, *run-time* 105 menit 30 detik, dan penggunaan kapasitas memori sebesar 12%

Penelitian terkait dengan implemenasi metode seleksi fitur dapat dilihat pada Tabel 2.2. Secara keseluruhan hasil evaluasi dengan menggunakan metode seleksi fitur mengalami peningkatan. Posisi penelitian ini dibandingkan dengan penilitian yang sudah ada adalah penelitian ini berfokus pada bagaimana meningkatakan performa untuk model SVM yang dibangun setelah dilakukan implementasi metode seleksi fitur.

TABEL 2. 2
PENELITIAN TERKAIT IMPLEMENTASI METODE FEATURE SELECTION

No	Judul	Penulis/Tahun	Metode	Hasil
1.	Implementation of the Naïve Bayes with Feature Selection using Genetic Algorithm for Sentiment Analysis of Fashion Online Companies [6]	Siti Ernawati, Eka Rini Yulia, Frieyadie, Samudi/2018	Genetic Algorithm	Hasil perbedaan untuk sebelum dan sesudah dilakukan seleksi fitur memiliki kenaikan yang cukup signifikan. Untuk akurasi memiliki nilai 68,5% menjadi 87,5% setelah seleksi fitur
2.	An Application of Machine Learning with Feature Selection to Improve Diagnosis and Classification of Neurodegenerative Disorders [7]	Josefa Diaz Alvarez, Jordi A. Matias-Guiu, Maria Nieves Cabrera-Martin, Jose L. Risco Martin dan Jose L.Ayala/2019	Cfs, Chi Squeard Attribute, Classifier Attribute, Wrapper Subset, Principal Component Analysis	Setelah fitur yang paling relevan diidentifikasi, jumlah data yang dapat diklasifikasi (benar) meningkat
3.	Implementation of Ensemble Learning and Feature Selection for Performance Improvements in Anomaly-Based Intrusion Detection Systems [8]	Qusyairi Ridho Saeful Fitni, Kalmullah Ramli/2020	Filter Methods (Spearman Correlation Coefficient)	Tingkat akurasi dari model yang dibangun cukup tinggi dengan nilai 98,8%. Waktu menjalankan program berkurang dari 34 menit menjadi 10 menit 54 detik

B. Turnover Pegawai

People Analytics (PA), atau dalam bahasa Indonesia berarti analisis manusia, merupakan istilah yang merujuk pada penelitian di bidang manajemen sumber daya manusia dalam menganalsis dinamika yang terdapat pada lingkungan pekerjaan. PA berfokus pada penggunaan teknologi informasi, analisis data, dan visualisasi data untuk mendapatkan wawasan baru sehingga dapat digunakan untuk mengoptimasi efektivitas, efisiensi, dan keluaran dari perusahaan [9]. Dinamika pekerjaan yang dimaksud mencakup aspek yang berhubungan dengan sumber daya manusia, performa individu atau pun kelompok. Masalah bisnis yang terjadi dalam perusahaan dapat diselesaikan salah satunya dengan menggunakan PA [10]. Hal tersebut karena PA mengidentifikasi pengaruh bisnis dan pengambilan keputusan berdasarkan data terkait dengan proses manajemen sumber daya manusia [11].

Implementasi PA dalam perusahaan dapat mendatangkan keuntungan bagi perusahaan tersebut. Dalam bidang rekruitasi pegawai, PA dapat meningkatkan efisiensi dalam pemilihan pegawai, meningkatkan jumlah pendaftar yang kompeten, dan mengurangi biaya perekruitan [1]. Dalam mengatasi perputaran (*turnover*) pegawai, PA dapat membantu memahami dan mengungkap praktek-praktek yang kurang efisien dalam pekerjaan dan memprediksi kebutuhan dan keahlian pegawai untuk memperoleh tujuan perusahaan [12]. Dua hal tersebut merupakan aspek yang dapat berpengaruh dalam pengambilan keputusan pegawai apakah tinggal dalam perusahaan atau tidak.

Turnover pegawai, atau dalam bahasa Indonesia berarti perputaran pegawai merupakan istilah yang digunakan untuk menunjukkan tingkat pegawai yang memutuskan hubungan kerja dengan suatu perusahaan. Kata "*turnover*" didefinisikan sebagai rasio jumlah pegawai yang memutuskan keluar dari perusahaan dibagi dengan rata-rata jumlah pegawai dalam rentang waktu tertentu [13]. *Turnover* pegawai juga dapat diartikan sebagai proses dimana seseorang mengambil hasil materi dari perusahaan dengan memutuskan keanggotaan perusahaan [14].

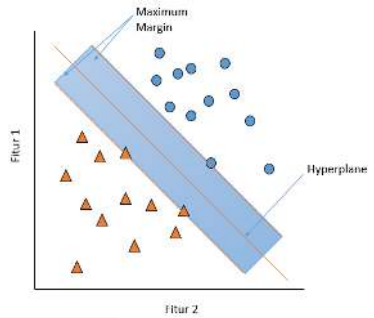
Banyak faktor yang dapat mempengaruhi keinginan pegawai untuk meninggalkan pekerjaannya. Pegawai yang puas dengan pekerjaan yang mereka kerjakan akan memperkecil kemungkinan pegawai tersebut akan menanggalkan posisi yang dikerjakan [15]. Faktor-faktor personal lain seperti usia, status pernikahan, tingkat pendidikan, dan lama bekerja juga secara tidak langsung mempengaruhi keinginan pegawai untuk keluar
dari perusahaan [14].

C. Support Vector Machine

Pembelajaran mesin merupakan suatu algoritma dengan melalui proses komputasi dapat menggunakan input data untuk memperoleh hasil yang diinginkan tanpa melakukan pengkodean secara langsung [16]. Pembelajaran mesin idealnya dapat meniru cara manusia belajar. Berdasarkan input yang diberikan, algoritma digolongkan menjadi dua jenis yaitu *supervised* dan *unsupervised*. Algoritma *supervised* input yang diberikan sudah diberi label kelas lebih dulu. Sedangkan, algoritma *unsupervised* input masih belum memiliki label kelas.

Dengan semakin berkembangnya pembelajaran mesin, pembelajaran mesin dapat diaplikasikan ke berbagai macam bidang dalam industri. Contoh bidang yang saat ini telah diaplikasikan pembelajaran mesin adalah visualisasi komputer, pabrik, keuangan, hiburan, dan biomedis. Hal tersebut salah satunya disebabkan karena, kemampuan dari pembelajaran mesin untuk menghasilkan kecerdasan buatan yang dapat dieksekusi tanpa memerlukan sumber daya yang signifikan [17].

SVM merupakan salah satu metode dalam pembelajaran mesin yang masuk dalam kategori algoritma *supervised*. SVM dapat membantu dalam menyelesaikan permasalahan masalah klasifikasi dalam *big data* [18]. Proses pelatihan algoritma SVM adalah dengan mengidentifikasi garis *hyperplane* yang dapat memaksimalkan jarak antar dua label kelas seperti yang terlihat pada Gambar 2.1 [19]. *Hyperplane* merujuk pada sebuah garis pemisah antar kelas. Dalam kasus tertentu garis *hyperplane* yang digunakan dapat berbentuk linear atau nonlinear.



GAMBAR 2. 1
KLASIFIKASI DALAM SVM

Secara umum Support Vector Machine dibagi menajdi dua bagian berdasarkan kegunaannya yaitu *support vector classification* (SVC) dan *support vector regression* (SVR). SVC bertujuan dalam mengklasifikasikan data ke dalam dua atau lebih kelas berbeda [20]. Sedangkan SVR analisis dilakukan untuk menilai hubungan antar variabel independen/bebas dan variabel dependen/terikat. Karena masalah yang ingin diselesaikan dalam penelitian ini adalah tentang memprediksi status *turnover* pegawai, maka SVM yang akan digunakan adalah SVM klasifikasi atau SVC. Dalam SVC, *hyperplane* linear untuk suatu data dapat ditulis dalam rumus
sebagai berikut,

$$w^T x + b = 0$$

$$w^T x + b \begin{cases} \geq 1 \; for \; y_i = 1 \\ \leq -1 \; for \; y_i = -1 \end{cases} \quad \textbf{2}$$

Untuk penjelasan dari masing-masing variabel sebagai berikut,

1. x_i adalah sekumpulan data latih (i = 1,2,3,…,n)
2. *w* adalah vektor dengan n-dimensi
3. *b* adalah nilai bias

Rumus menunjukkan bagaimana sebua data akan dikelompokkan dalam suatau kelas tertentu. Berdasarkan rumus di atas data tiap kelas hanya dapat diidentifikasi dengan berada di sebelah kiri (y=1) ata di sebelah kanan (y=-1) dari garis *hyperplane*.

D. Feature Selection

Feature selection merupakan salah satu metode dalam proses pengurangan dimensi (*dimensionality reduction*). Pengurangan dimensi merupakan suatu proses untuk mengurangi jumlah fitur awal data tetapi tetap mempertimbangkan fitur awal tersebut. Manfaat utama dari proses pengurangan dimensi adalah peningkatan akurasi dari *classifier* yang digunakan dan mengurangi kompleksitas pemrograman [21]. Oleh karena itu, *feature selection* sering diaplikasikan di berbagai macam bidang contohnya seperti penambangan teks dan analisis genetik [5].

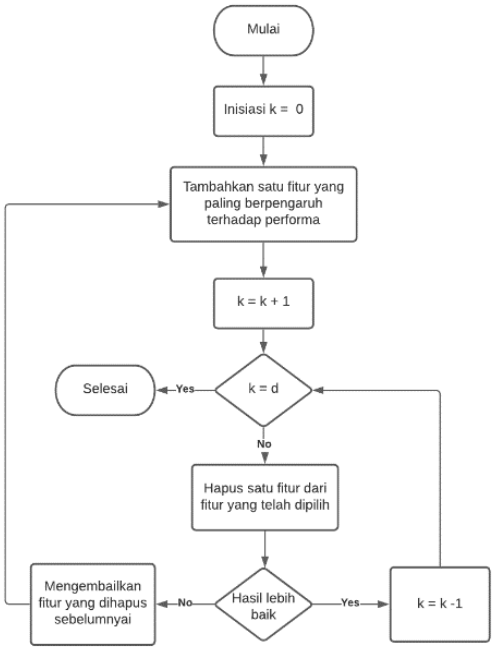
Filter methods merupakan metode untuk memilih fitur dengan cara mengurutkan tingkat hubungan fitur dengan label kelas pada dataset. Metode ini digunakan alasannya karena tidak terlalu kompleks dan berhasil untuk diaplikasikan. Fitur yang tidak berpengaruh terhadap pelabelan kelas dapat disingkirkan. Sehingga, dibutuhkan nilai batasan (*threshold*) yang menjadi acuan tingkat relevansi dengan label kelas. Jika fitur tidak mencpai nilai batas, fitur tersebut tidak akan digunakan. Untuk pengukuran relevansi salah satu cara yang dapat digunakan adalah dengan menggunakan koefisiensi korelasi Pearson [22].

$$R(i) = \frac{cov(x_i, Y)}{\sqrt{var(x_i) * var(Y)}} \quad \textbf{3}$$

Untuk penjelasan dari masing-masing variabel sebagai berikut,

1. x_i adalah fitur ke-i
2. *Y* adalah label kelas
3. *cov()* adalah nilai kovariansi dan *var()* adalah nilai variansi

Wrapper methods merupakan metode yang menyeleksi fitur dengan cara mengevaluasi performansi dari model untuk tiap fitur pada dataset. Contoh algoritma yang dapat digunakan adalah *sequential floating forward selection* (SFFS). Pertama algoritma ini memilih satu fitur yang memiliki pengaruh paling besar terhadap performa model. Kemudian algoritma menghapus fitur yang diperoleh dari tahap pertama dan mengevaluasi penghapusan fitur tersebut. Jika penghapusan fitur yang dipilih meningkatkan performa dari model, fitur yang dihapus dibiarkan tidak terpakai pada tahap selanjutnya. Jika penghapusan fitur yang dipilih menurunkan performa dari model, penghapusan fitur dikembalikan pada kondisi sebelum fitur dihapus dan algoritma mengulang kembali dari tahap pertama hingga jumlah fitur yang diinginkan tercapai. Gambar 2.4 adalah alur dari algoritma SFFS dengan k adalah ukuran subset fitur dan d adalah ukuran dimensi yang diinginkan [22].



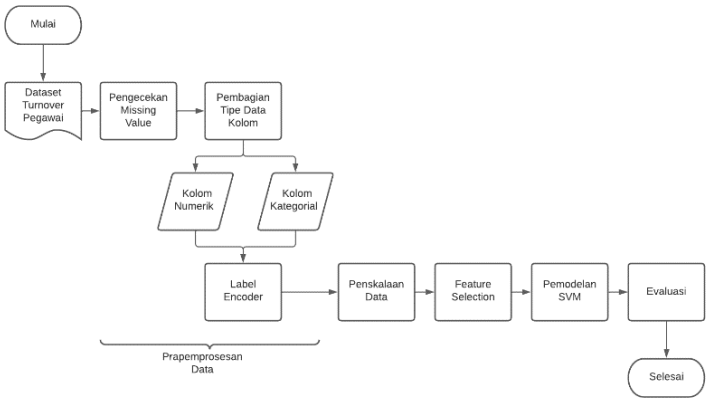
GAMBAR 2. 2
ALUR ALGORITMA SFFS

Embedded methods menggunakan pendekatan dengan cara menggabungkan seleksi fitur sebagai bagian dari pelatian model. Hal ini bertujuan untuk mengurangi waktu yang dibutuhkan untuk mengklasifikasi kembali subset yang berbeda sebagaimana yang dilakukan oleh *wrapper methods* [22]. *Embedded methods* memiliki pendekatan yang sedikit berbeda dari *wrapper methods* yaitu seleksi fitur dilakukan bersamaan dengan pelatihan model sedangkan *wrapper methods* menyeleksi fitur berdasarkan evaluasi matriks yang dihasilkan akibat dari fitur yang sudah diseleksi.

III. METODE

Sistem yang dibangun dalam penelitian ini bertujuan untuk mengimplementasikan salah satu metode pembelajaran mesin, yaitu *support vector machine* (SVM), dalam memprediksi *turnover*/perputaran pegawai di suatu perusahaan. Dalam penelititan ini, analisis dilakukan kepada

dataset terdapat pada internet. Kumpulan data pegawai tersebut berguna sebagai data *training*/latih dan data *test*/uji pada tahap pengujian. Sebelum dilakukan pemodelan algoritma SVM, dataset diproses dalam tahap prapemprosesan data yang mencakup eksplorasi mengenai dataset, pengisian *missing value*, pelabelan tipe data kolom kategorial, dan penskalaan dataset. Pemodelan SVM dilakukan dengan menggunakan *library sklearn*. Metrik evaluasi yang digunakan adalah metrik *confussion matrix* karena metrik evaluasi ini sudah umum digunakan dalam mengevaluasi model pembelajaran mesin dan cocok untuk mengevaluasi permasalahan klasifikasi. Confussion matrix digunakan untuk mengukur nilai akurasi, presisi, *recall*, dan *f1-score*. Skenario keseluruhan sistem dapat dilihat pada Gambar 3.1.



GAMBAR 3. 1
ALUR PEMBANGUNAN SISTEM

A. Pengambilan dan Pelabelan Data

Data yang digunakan merupakan dataset asli yang dibagikan oleh blog Edward Babushkin. Pada blog Edward Babushkin, dataset digunakan untuk memprediksi risiko pegawai yang melakukan *turnover* dengan menggunakan *survival analysis model*. Dataset memiliki jumlah fitur sebanyak 16 fitur dan baris data sebanyak 1129 data pegawai yang berbeda. Tabel 3.1 berisi penjelasan mengenai fitur apa saja yang terdapat pada dataset. Fitur '*event*' merupakan variabel terikat atau dapat disebut juga label kelas sedangkan lima belas fitur lainnya merupakan variabel bebas.

TABEL 3. 1
PENJELASAN FITUR PADA DATASET

No	Nama Fitur	Deskripsi
1.	*"stag"*	Pengalaman yang dimiliki oleh pegawai
2.	*"event"*	Label kelas yang menunjukkan pegawai tersebut *turnover* atau tidak
3.	*"gender"*	Jenis kelamin pegawai
4.	*"age"*	Umur pegawai
5.	*"industry"*	Jenis industri dimana tempat pegawai bekerja
6.	*"profession"*	Jenis profesi atau jabatan pegawai
7.	*"traffic"*	Bagaimana pegawai datang ke perusahaan
8.	*"coach"*	Kehadiran pelatih saat masa percobaan pegawai
9.	*"head_gender"*	Jenis kelamin dari atasan pegawai tersebut
10.	*"greywage"*	Pengurangan gaji terhadap otoritas pajak
11.	*"way"*	Transportasi pegawai menuju tempat kerja
12-16.	*"extraversion", "independend", "selfcontrol", "anxiety", "novator"*	Skala pengukuran kepribadian pegawai meliputi ekstraversi, kemandirian, kontrol atas diri sendiri, kecemasan, dan novator.

B. Prapemrosesan Data

Tahap prapemrosesan data ini berfungsi untuk menyiapkan data sebelum nanti masuk ke dalam proses pemodelan. Tahap ini mencakup pengecekan *missing-value*, penentuan tipe data kolom, pengubahan tipe data kategorial menjadi numerik. Tahap ini menghasilkan data dengan semua kolom berisi nilai numerik atau tidak ada nilai untuk kategorial.

1. Pengecekan *missing-value*:
   Proses ini merupakan tahapan untuk mengisi data yang masih kosong dalam dataset. Metode pengisian *missing value* di antaranya adalah pengisian dengan nilai rata-rata kolom, pengisian dengan nilai terbanyak (modus/mode), atau dihapus dari dataset. Pertama, untuk baris dengan *missing value*-nya lebih dari dua maka baris akan langsung dihapus. Selanjutnya, untuk missing value pada kolom dengan tipe data numerik maka *missing-value* diisi dengan nilai rata-rata kolom tersebut. Sedangkan, untuk *missing-value* pada kolom dengan tipe data kategorial maka *missing value* diisi dengan nilai terbanyak dalam kolom tersebut. Pemilihan keputusan pengisian metode *missing value* berkaitan dengan jumlah dataset yang diperoleh dari tahap sebelumnya.
2. Penentuan tipe data kolom:
   Proses ini merupakan kelanjutan dari identifikasi kategori fitur/kolom dalam dataset. Kolom

dengan tipe data numerik merupakan kolom yang memiliki nilai berupa bilangan bulat atau pun *real*. Kolom dengan tipe data kategorial merupakan kolom selain berisi nilai yang bersifat kategori seperti contoh paling umum adalah kolom jenis kelamin

3. Pengubahan tipe data kategorial menjadi numerik
   Proses ini akan mengubah tipe data kategorial yang masih berupa karakter string akan diubah menjadi numerik menggunakan *encoder*. Metode *encoder* yang digunakan adalah motede *label encoder*. Hal ini bertujuan agar setiap nilai dalam kolom dapat terbaca oleh algoritma pembelajaran mesin.

C. Eksplorasi Data

Tahap eksplorasi data ini berfokus dalam menampilkan persebaran data pada masing-masing fitur/kolom. Hasil persebaran data ditampilkan secara visual agar lebih memudahkan pembacaan kode yang dihasilkan. Eksplorasi selanjutnya berhubungan dengan korelasi dari tiap fitur dengan fitur yang lain. Penampilan korelasi ini bertujuan untuk melihat keterhubungan pada masing-masing fitur. Korelasi juga digunakan dalam penggunaan seleksi fitur *filter methods*.

D. Penskalaan Data

Tahap penskalaan (*scaling*) data ini berfungsi untuk menyamakan rentang nilai dari masing-masing kolom. Teknik penskalaan yang digunakan dalam penelitian ini adalah teknik penskalaan *StandarScaller*. Pensklaan ini dilakukan sebelum pembelajaran model pembelajaran mesin agar model dapat belajar dengan lebih cepat dan nilai akurasi dapat meningkat. Tahap ini menghasilkan data dengan peresebaran rentang nilai yang merata.

E. Seleksi Fitur

Pada tahap ini, dataset akan dibagi menjadi beberapa kategori. Kategori yang pertama adalah dataset tanpa menggunakan seleksi fitur. Kategori selanjutnya merupakan dataset yang telah dilakukan proses seleksi fitur berdasarkan metode-metode yang digunakan. Metode seleksi fitur yang digunakan adalah *filter method*, *wrapper method,* dan *embedded method*.

F. Pemodelan SVM

Model yang akan digunakan dalam penelitian ini adalah SVM. Sebelum dilakukan pemodelan data akan dibagi menjadi data latih dan data uji. Metode yang digunakan untuk pemecahan data latih dan data uji ini adalah menggunakan metode *train-test-split* yang sudah disediakan oleh *library scikit-learn*. Model SVM akan diuji menggunakan data uji untuk melihat performansi model tersebut.

IV. HASIL DAN PEMBAHASAN

Metrik evaluasi yang akan digunakan untuk mengevaluasi model pembelajaran mesin adalah *confusion matrix*. *Confusion matrix* mengelompokkan data hasil prediksi dan realita ke dalam empat kategori. Empat kategori tersebut yaitu: hasil prediksi benar realita benar (*true positive*/TP), hasil prediksi benar realita salah (*false positive*/FP), hasil prediksi salah realita benar (*false negative*/FN), dan hasil prediksi salah realita salah (*true negative*/TN). Pada topik TA, nilai positif atau benar (dinotasikan dengan angka 1) diberikan pada kelas atau label pegawai yang melakukan *turnover*. Sedangkan, nilai negatif atau salah (dinotasikan dengan angka 0) diberikan pada kelas atau label pegawai yang tidak melakukan *turnover*. Kelas pegawai turnover dipilih sebagai kelas positif karena merupakan fokus utama yang diuji dalam penelitian ini. *Confussion matrix* dipilih karena metrik evaluasi tersebut merupakan metrik evaluasi yang umum cocok untuk digunakan dalam mengevaluasi permasalahan klasifikasi. Kemudian, hasil dari *confussion matrix* digunakan untuk menghitung performa model berdasarkan empat metode pengukuran yaitu akurasi, presisi, *recall*, dan *f1-score*.

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} * 100\% \quad \textbf{4}$$

$$Presisi = \frac{TP}{FP + TP} * 100\% \quad \textbf{5}$$

$$Recall = \frac{TP}{FN + TP} * 100\% \quad \textbf{6}$$

$$F1\ Score = \frac{2 * Recall * Presisi}{(Recall + Presisi} \quad \textbf{7}$$

A. Hasil Pengujian

Model pembelajaran mesin dievaluasi menggunakan empat metode pengukuran yaitu akurasi, presisi, *recall*, dan *f1-score*. Evaluasi dilakukan kepada empat jenis dataset yaitu dataset tanpa dilakukan seleksi fitur, dataset menggunakan seleksi fitur *filter method*, dataset menggunakan seleksi fitur *wrapper method*, dan dataset menggunakan seleksi fitur *embedded method*. Tabel 4.1 menunjukkan hasil evaluasi dari keempat dataset dan dihitung berdasarkan keempat metode pengukuran.

TABEL 4. 1
HASIL EVALUASI PENGUJIAN

Dataset	Akurasi	Presisi	*Recall*	*F1-Score*
Tanpa Seleksi Fitur	0,56	0,56	0,56	0,56
Filter Method	0,55	0,55	0,55	0,55
Wrapper Method	**0,60**	**0,60**	**0,60**	**0,60**

Embedded Method	0,55	0,55	0,55	0,55

Pada Tabel 4.1, hasil evaluasi paling baik ditunjukkan oleh dataset setelah dilakukan seleksi fitur dengan metode *wrapper method*. Nilai evaluasi yang diperoleh oleh dataset setelah dilakukan seleksi fitur *wrapper method* naik sebesar 0,4 untuk semua metode pengukuran. Sedangkan, dua metode lain, *filter method* dan *embedded method*, mengalami penurunan hasil evaluasi dari sebelum dilakukan seleksi fitur.

B. Analisis Hasil Pengujian

Perbedaan metode dari seleksi fitur yang digunakan menghasilkan pemilihan fitur yang akan dipelajari oleh pembelajaran mesin berbeda pula. *Filter method* memilih fitur dengan membandingkan fitur mana yang memiliki tingkat korelasi paling besar dengan fitur variabel terikat atau dalam dataset berarti kolom '*event*'. *Wrapper method* memilih fitur berdasarkan hasil evaluasi paling baik dengan model pembelajaran mesin yang digunakan. *Embedded method* melakukan seleksi fitur bersamaan dengan proses pembelajaran model yang dipilih. Tabel 4.2 menunjukan fitur apa saja yang mempengaruhi seseorang melakukan *turnover* berdasarkan metode seleksi fitur yang digunakan.

TABEL 4. 2
FITUR YANG DISELEKSI SETELAH SELEKSI FITUR TIAP METODE

Metode Seleksi Fitur	Fitur yang Diseleksi	Jumlah Fitur
Filter Method	*Coach, independend, anxiety, industry, way*	5
Wrapper Method	*Stag, gender, age, industry, profession, head_gender, greywage, way, independend, anxiety*	10
Embedded Method	*Stag, gender, age, independend, extraversion, anxiety, way, selfcontrol*	8

V. KESIMPULAN

Model pembelajaran mesin SVM dapat melakukan prediksi dengan cukup baik terhadap data *turnover* pegawai yang diberikan. Hasil evaluasi yang diperoleh oleh model SVM untuk dataset tanpa dilakukan seleksi fitur memiliki nilai sebesar 0,56 untuk semua metode pengukuran. Metode pengukuran yang digunakan adalah akurasi, presisi, *recall*, dan *f1-score*.

Seleksi fitur yang dilakukan terhadap dataset dapat menaikkan dan juga menurunkan hasil evaluasi yang diperoleh tergantung pada metode yang digunakan. 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. Sedangkan dua metode lain, *filter method* dan *embedded method*, mengalami penurunan performa dengan nilai performa yang sama yaitu 0,55. Hal tersebut dapat disebabkan karena pemilihan fitur yang berbeda dari ketiga metode yang digunakan menghasilkan hasil performa yang berbeda pula.

Kekurangan pada penelitian ini adalah tidak menggali lebih jauh mengenai parameter yang dapat mempengaruhi hasil nilai evaluasi. Penggunaan parameter bawaan dari *library* yang digunakan mungkin saja dapat menghasilkan hasil yang berbeda dengan pemilihan sendiri parameter-parameter yang cocok kepada model atau disebut juga *hyperparameter tuning*. Pada penelitian ini, *hyperparameter tuning* tidak dilakukan karena untuk menghindari terlalu banyak parameter yang harus diuji.

REFERENSI

[1] A. Karmańska, "The benefits of HR analytics," *Pr. Nauk. Uniw. Ekon. we Wrocławiu*, vol. 64, no. 8, pp. 30–39, 2020, doi: 10.15611/pn.2020.8.03.

[2] Y. Zhao, M. K. Hryniewicki, F. Cheng, B. Fu, and X. Zhu, *Employee turnover prediction with machine learning: A reliable approach*, vol. 869. Springer International Publishing, 2018.

[3] V. Amin, J. A. Rathod, M. Kunder, and P. Patkar, "A review on employee attrition using machine learning," no. 05, pp. 1237–1241, 2021.

[4] R. Punnoose and P. Ajit, "Prediction of Employee Turnover in Organizations using Machine Learning Algorithms," *Int. J. Adv. Res. Artif. Intell.*, vol. 5, no. 9, pp. 22–26, 2016, doi: 10.14569/ijarai.2016.050904.

[5] S. García, J. Luengo, and F. Herrera, "Feature selection," *Intell. Syst. Ref. Libr.*, vol. 72, no. 6, pp. 163–193, 2015, doi: 10.1007/978-3-319-10247-4_7.

[6] S. Ernawati, E. R. Yulia, Frieyadie, and Samudi, "Implementation of the Naïve Bayes Algorithm with Feature Selection using Genetic Algorithm for Sentiment Review Analysis of Fashion Online Companies," *2018 6th Int. Conf. Cyber IT Serv. Manag. CITSM 2018*, no. Citsm, pp. 1–5, 2019, doi: 10.1109/CITSM.2018.8674286.

[7] J. D. Álvarez, J. A. Matias-Guiu, M. N. Cabrera-Martín, J. L. Risco-Martín, and J. L. Ayala, "An application of machine learning with feature selection to improve diagnosis and classification of neurodegenerative disorders," *BMC Bioinformatics*, vol. 20, no. 1, pp. 1–12, 2019, doi: 10.1186/s12859-019-3027-7.

[8] Q. R. S. Fitni and K. Ramli, "Implementation of ensemble learning and feature selection for performance improvements in anomaly-based intrusion detection systems," *Proc. -*

2020 IEEE Int. Conf. Ind. 4.0, Artif. Intell. Commun. Technol. IAICT 2020, pp. 118–124, 2020, doi: 10.1109/IAICT50021.2020.9172014.

[9] A. Tursunbayeva, S. Di Lauro, and C. Pagliari, "People analytics—A scoping review of conceptual boundaries and value propositions," *Int. J. Inf. Manage.*, vol. 43, no. July, pp. 224–247, 2018, doi: 10.1016/j.ijinfomgt.2018.08.002.

[10] T. Peeters, J. Paauwe, and K. Van De Voorde, "People analytics effectiveness: developing a framework," *J. Organ. Eff.*, vol. 7, no. 2, pp. 203–219, 2020, doi: 10.1108/JOEPP-04-2020-0071.

[11] J. H. Marler and J. W. Boudreau, "An evidence-based review of HR Analytics," *Int. J. Hum. Resour. Manag.*, vol. 28, no. 1, pp. 3–26, 2017, doi: 10.1080/09585192.2016.1244699.

[12] C. Varma and C. Chavan, "A Case of HR Analytics–to Understand Effect on Employee Turnover," *J. Emerg. Technol. ...*, vol. 6, no. 6, pp. 781–787, 2019, [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3619634.

[13] V. S. Rangel, "A review of the literature on principal turnover," *Rev. Educ. Res.*, vol. 88, no. 1, pp. 87–124, Feb. 2018, doi: 10.3102/0034654317743197.

[14] Y. Zhang, "A Review of Employee Turnover Influence Factor and Countermeasure," *J. Hum. Resour. Sustain. Stud.*, vol. 04, no. 02, pp. 85–91, 2016, doi: 10.4236/jhrss.2016.42010.

[15] B. J. Ali and G. Anwar, "Employee Turnover Intention and Job Satisfaction," *Int. J. Adv. Eng. Manag. Sci.*, vol. 7, no. 6, pp. 22–30, 2021, doi: 10.22161/ijaems.76.3.

[16] I. El Naqa and M. J. Murphy, "What Is Machine Learning?," in *Machine Learning in Radiation Oncology*, Springer International Publishing, 2015, pp. 3–11.

[17] R. Rai, M. K. Tiwari, D. Ivanov, and A. Dolgui, "Machine learning in manufacturing and industry 4.0 applications," *International Journal of Production Research*, vol. 59, no. 16. Taylor and Francis Ltd., pp. 4773–4778, 2021, doi: 10.1080/00207543.2021.1956675.

[18] S. Suthaharan, *Machine Learning Models and Algorithms for Big Data Classification*, vol. 36. Boston, MA: Springer US, 2016.

[19] D. A. Pisner and D. M. Schnyer, "Support vector machine," in *Machine Learning: Methods and Applications to Brain Disorders*, Elsevier, 2019, pp. 101–121.

[20] R. Gholami and N. Fakhari, "Support Vector Machine: Principles, Parameters, and Applications," in *Handbook of Neural Computation*, Elsevier Inc., 2017, pp. 515–535.

[21] R. Kavitha and E. Kannan, "An Efficient Framework for Heart Disease Classification using Feature Extraction and Feature Selection Technique in Data Mining."

[22] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 16–28, 2014, doi: 10.1016/j.compeleceng.2013.11.024.