Jurnal Rekayasa Industri (JRI), Vol. 7 No. 1 April 2025
p-ISSN: 2714-8882 / e-ISSN: 2714-8874

Analisis Produktivitas Tenaga Kerja dengan Menggunakan *Regression Tree* dan *Classification* C50

Valentine Andreas Manurung *, The Jin Ai
Departemen Teknik Industri, Fakultas Teknologi Industri, Universitas Atma Jaya Yogyakarta
Jl. Babarsari No.43, Janti, Caturtunggal, Kec. Depok, Kabupaten Sleman, Daerah Istimewa Yogyakarta
Email: 225611915@students.uajy.ac.id

** Corresponding Author*

ABSTRAK

Peningkatan produktivitas tenaga kerja merupakan salah satu masalah utama yang dihadapi banyak pelaku bisnis. Dalam upaya mengatasi masalah ini, penelitian sebelumnya telah memberikan berbagai pemahaman dan solusi terkait masalah tersebut. Berbeda dengan penelitian sebelumnya, artikel ini mengusulkan suatu *framework* metodologi untuk melakukan analisis produktivitas tenaga kerja dengan menggunakan teknik *machine learning* melalui dua macam keperluan yaitu regresi dan klasifikasi dengan tujuan agar hasil yang diperoleh lebih komprehensif dan dapat dengan mudah diinterpretasikan sebagai saran manajemen dengan tetap mencapai akurasi prediksi yang relatif baik. Sebagai ilustrasi penggunaan dari *framework* yang diusulkan tersebut, algoritma *Regression Tree* dipilih untuk keperluan regresi dan algoritma *Classification C50* dipilih untuk keperluan klasifikasi, bahasa pemrograman R dalam RStudio digunakan sebagai alat untuk menjalankan *framework*, dan diaplikasikan dengan suatu data sekunder produktivitas tenaga kerja.

Kata kunci: Produktivitas Tenaga Kerja, *Regresion Tree*, *Classification* C50, *Machine Learning*

ABSTRACT

Increasing labor productivity is a primary challenge encountered by numerous corporations. To address this issue, prior research has offered diverse insights and solutions. Different with previous studies, this article introduces a methodological framework for analyzing labor productivity through machine learning techniques. It employs two objectives: regression and classification, aiming for results that are both comprehensive and easily interpretable as management recommendations, while maintaining satisfactory prediction accuracy. For illustrative purposes, the proposed framework is employed the Regression Tree algorithm for regression and the Classification C50 algorithm for classification, implemented using the R programming language in RStudio, and applied to secondary data on labor productivity.

Keywords: *Employee Productivity, Regression Tree, Classification C50, Machine Learning.*

I. PENDAHULUAN

Banyak pelaku bisnis menyadari bahwa upaya untuk meningkatkan produktivitas tenaga kerja merupakan salah satu masalah penting yang harus dihadapi. Dalam upaya mengatasi masalah ini, penelitian sebelumnya telah memberikan berbagai pemahaman dan solusi untuk meningkatkan produktivitas tenaga kerja. Menurut Kementerian Ketenagakerjaan Republik Indonesia (2022) mendefinisikan produktivitas tenaga kerja sebagai rasio antara jumlah tenaga kerja dengan hasil produksi selama periode waktu tertentu. Data dari Kemnaker tersebut menunjukkan bahwa pada tahun 2018 produktivitas pekerja Indonesia adalah Rp82,56 juta per orang per tahun. Meskipun terdapat peningkatan produktivitas pada tahun 2019, pada tahun 2020 produktivitas mengalami penurunan akibat pandemi COVID-19. Namun demikian, produktivitas mulai pulih pada tahun 2021. Pada tahun 2022, nilainya mencapai Rp86,55 juta per orang per tahun. Secara keseluruhan, produktivitas tenaga kerja Indonesia meningkat sebesar 4,8 persen selama periode 2018–2022, yang menunjukkan perbaikan yang signifikan.

Analisis produktivitas tenaga kerja telah menjadi subjek sejumlah besar penelitian. Terdapat berbagai aspek yang berkaitan dengan produktivitas telah diteliti, diantaranya yaitu dampak model kerja *Working form Home* (WFH) terhadap produktivitas karyawan di sektor teknologi informasi (Sungheetha & Sharma, 2021), dampak stress di tempat kerja terhadap produktivitas karyawan dan absensi dalam organisasi (Aristizabal *et al.*, 2021), pengaruh kebisingan terhadap produktivitas (De Salvio *et al.*, 2023), dan pengaruh sentimen terhadap produktivitas (Saxena *et al.*, 2023). Sementara itu, beberapa peneliti telah menggunakan analisis produktivitas untuk keperluan klasifikasi karyawan dalam perusahaan (Fadli *et al.*, 2021) dan optimalisasi

konfigurasi SDM dalam organisasi (Gu, 2022). Selain itu, terdapat juga penelitian yang bertujuan untuk menghasilkan prediksi produktivitas tenaga kerja dalam perusahaan (Sabuj *et al.*, 2022; Obiedat & Toubasi, 2022; Razali *et al.*, 2023).

Dalam beberapa tahun terakhir ini telah terjadi perubahan dalam arah penelitian dalam bidang manajemen sumber daya manusia yang melingkupi analisis produktivitas tenaga kerja, yaitu dari manajemen tradisional yang tidak berbasis pada data dan informasi menjadi manajemen modern yang berbasis pada data dan informasi (Razali *et al.*, 2023). Seturut dengan perkembangan ini, banyak metode analisis data modern, termasuk berbagai teknik *data mining* dan *machine learning*, telah digunakan untuk mendukung analisis produktivitas. Mayoritas dari teknik yang digunakan ini merupakan teknik untuk keperluan regresi atau klasifikasi. Sejumlah teknik *machine learning* yang sudah pernah digunakan untuk analisis produktivitas adalah Random Forest (Sungheetha & Sharma, 2021; De Silva *et al.*, 2022; Sabuj *et al.*, 2022; Obiedat & Toubasi, 2022; Adeniji *et al.*, 2022), Decision Tree (Sungheetha & Sharma, 2021; Sabuj *et al.*, 2022), dan Naïve Bayes (Sungheetha & Sharma, 2021), Support Vector Machine (SVM) (Fadli *et al.*, 2021; Sabuj *et al.*, 2022; Obiedat & Toubasi, 2022) Gradient Boost, XG-Boost (Sabuj *et al.*, 2022), serta berbagai teknik dalam Artificial Neural Network dan Deep Learning (Al Imran *et al.*, 2019; Adeniji *et al.*, 2022; Obiedat & Toubasi, 2022).

Seperti telah disebutkan sebelumnya, dalam berbagai teknik *machine learning* yang tersedia tersebut terdapat dua macam keperluan yaitu regresi dan klasifikasi. Perbedaan utama dari keduanya adalah kemampuan dalam prediksi hasil yaitu regresi untuk memprediksi hasil yang berupa nilai numerik atau bilangan riil kontinyu dan klasifikasi untuk memprediksi hasil yang berupa faktor atau label kategori atau bilangan bulat diskret. Beberapa metode hanya mampu untuk menjalankan satu keperluan saja, yaitu regresi saja (misalnya Regresi Linier) atau klasifikasi saja (misalnya Naïve Bayes). Akan tetapi ada juga teknik yang mampu untuk menjalankan dua keperluan sesuai yang diinginkan pengguna (misalnya Random Forest).

Hal yang sering dianggap penting dalam proses membandingkan antar berbagai teknik tersebut adalah keakuratan prediksi dari model yang dihasilkan. Akan tetapi seringkali dijumpai bahwa akurasi dicapai dengan teknik yang sangat kompleks dan pada akhirnya hasil model prediksi yang diperoleh sulit diinterpretasikan sebagai masukan bagi manajemen untuk perbaikan berkelanjutan. Untuk itulah dalam artikel ini diusulkan suatu *framework* metodologi untuk melakukan analisis produktivitas tenaga kerja dengan menggunakan teknik *machine learning* melalui dua macam keperluan yaitu regresi dan klasifikasi dengan tujuan agar hasil yang diperoleh lebih komprehensif dan dapat dengan mudah diinterpretasikan sebagai saran manajemen dengan tetap mencapai akurasi prediksi yang relatif baik. Kontribusi utama dari *framework* yang diusulkan adalah (1) menggunakan berbagai macam data yang tersedia di perusahaan, baik yang bertipe numerik maupun faktor sebagai dasar pembuatan model, (2) menggunakan teknik *machine learning* untuk membuat model berdasarkan data di Perusahaan dengan dua keperluan yaitu regresi dan klasifikasi, (3) menyajikan hasil dalam bentuk grafik sehingga mudah untuk dimengerti dan diinterpretasikan.

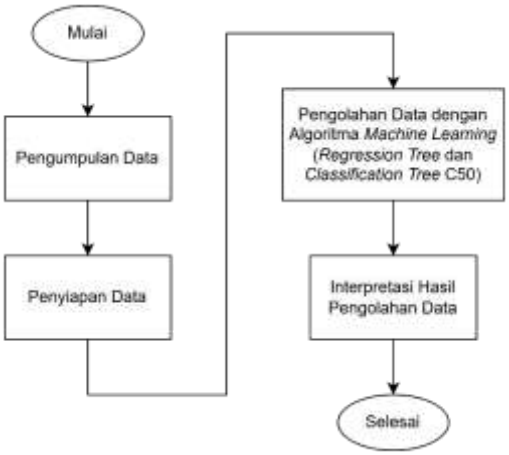
Sebagai ilustrasi penggunaan dari framework yang diusulkan tersebut, digunakan data sekunder yang berasal dari sebuah dataset produktivitas tenaga kerja (UCI Machine Learning Repository, 2020). Algoritma *Regression Tree* dan *Classification C50*, masing-masing dipilih sebagai teknik *machine learning* untuk keperluan regresi dan klasifikasi, karena keduanya dikenal sebagai teknik dasar yang hasilnya mudah untuk diinterpretasikan (Breiman *et al.*, 1984; Prabawati & Ajie, 2019; Benediktus & Oetama, 2020). Sementara itu, bahasa pemrograman R dalam RStudio digunakan sebagai alat untuk menjalankan teknik *machine learning*, terkhusus dengan menggunakan library rpart dan C50 yang sudah tersedia di dalam R.

II. METODE PENELITIAN

Metodologi yang diusulkan untuk melakukan analisis terhadap data produktivitas tenaga kerja dapat dilihat pada Gambar 1. Metodologi usulan terdiri dari 4 tahap yang masing-masing tahap akan dijelaskan pada beberapa sub bab di bawah ini

2.1 Tahap Pengumpulan Data

Pada tahap ini, data mengenai produktivitas tenaga kerja dan faktor-faktor yang diperkirakan dapat mempengaruhi produktivitas dikumpulkan. Data ini dapat berasal dari dikumpulkan melalui survei, wawancara, atau dengan mengakses basis data yang dimiliki perusahaan. Pada sebagian besar kasus, terutama pada perusahaan yang sudah memiliki sistem informasi terintegrasi, data yang berkaitan dengan produktivitas tenaga kerja dapat diakses melalui basis data.



Gambar 1. Usulan Metodologi Analisis Produktivitas Tenaga Kerja

2.2 Tahap Penyiapan Data

Pada tahap ini, data yang sudah dikumpulkan pada tahap sebelumnya disiapkan agar dapat diproses dalam tahap selanjutnya yaitu analisis dengan algoritma *machine learning* tertentu yaitu algoritma *regression tree* dan algoritma *classification tree* C50. Tahap ini biasanya dikenal dengan tahap pra pemrosesan data atau *data preprocessing*. Tahap ini perlu dilakukan karena sering kali data yang dikumpulkan tidak sempurna, tidak seimbang, atau mengandung nilai yang hilang. Hal lain yang juga perlu dilakukan agar data menjadi siap diproses dengan algoritma *machine learning* adalah menyelaraskan data tersebut dengan spesifikasi yang dibutuhkan oleh algoritma yang digunakan.

Hal pertama yang perlu dilakukan dalam tahap ini adalah mengidentifikasi dan menangani adanya data yang tidak lengkap, misalnya hilang atau kosong pada variabel tertentu. Kondisi data yang tidak lengkap ini dapat mengganggu jalannya algoritma atau dapat mempengaruhi hasil analisis. Oleh sebab itu, proses untuk mengidentifikasi adanya data yang tidak lengkap adalah penting. Kemudian, jika ditemukan adanya data yang tidak lengkap, terdapat dua opsi yang dapat dilakukan yaitu menghapus satu baris atau satu kolom data yang mengandung data yang tidak lengkap tersebut atau mengganti data yang hilang dengan suatu nilai perkiraan.

Hal kedua yang dilakukan dalam tahap ini adalah mengidentifikasi adanya data yang tidak wajar atau sering disebut dengan *outlier*. Data yang tidak wajar ini biasanya adalah data yang nilainya sangat berbeda dengan mayoritas data yang ada dan besar kemungkinan merupakan data yang tidak benar, misalnya kesalahan pengukuran atau kesalahan dalam proses memasukkan data dalam basis data. Data yang tidak wajar ini perlu dipertimbangkan akan dimasukkan atau tidak sebagai data yang digunakan dalam analisis, karena akan sangat berpengaruh terhadap hasil analisis.

Sementara itu terdapat 3 hal yang biasanya perlu dilakukan untuk menyiapkan data siap dianalisis dengan algoritma *machine learning*, yaitu penyesuaian tipe data, pemilihan faktor atau variabel, serta pembagian data menjadi data pelatihan dan data pengujian. Tidak semua faktor atau variabel yang tersedia dapat digunakan secara langsung dalam algoritma *machine learning*, karena setiap algoritma biasanya membutuhkan input dengan format dan tipe data tertentu. Sehingga apabila data dalam faktor atau variabel masih memiliki format atau tipe data yang berbeda dengan yang dibutuhkan oleh algoritma, perlu dilakukan penyesuaian. Hal lain yang perlu diperhatikan adalah tidak semua faktor atau variabel yang tersedia digunakan dalam analisis dengan algoritma *machine learning*. Dalam tahap ini, dilakukan seleksi terhadap faktor atau variabel yang tersedia, sehingga hanya faktor atau variabel yang relevan saja yang akan dimasukkan ke dalam algoritma *machine learning*. Harapannya kompleksitas dari analisis bisa menjadi berkurang dan model yang terbentuk dapat menunjukkan kinerja yang lebih baik. Hal terakhir yang dilakukan dalam tahap ini adalah membagi data menjadi dua bagian. Bagian pertama yang biasa disebut data latihan atau *train*, dipakai untuk melatih atau membentuk model. Sementara bagian kedua yang biasa disebut data uji atau *test*, dimanfaatkan untuk untuk mengevaluasi seberapa baik model yang dibentuk. Dalam studi dengan *machine learning*, hal ini sangat penting dilakukan untuk mengevaluasi model dengan cara yang objektif.

2.3 Tahap Pengolahan Data dengan Algoritma Machine Learning

Pada tahap ini, data yang sudah melalui tahap pra pemrosesan data diolah dengan menggunakan algoritma *machine learning* yang dipilih. Dalam artikel ini dipilih algoritma *regression tree* dan algoritma *classification tree* C50, meskipun terbuka penggunaan algoritma yang lain dengan spesifikasi yang sama yaitu mampu menghasilkan model regresi dan klasifikasi dengan hasil yang mudah diinterpretasikan.

2.4 Tahap Interpretasi Hasil

Pada tahap ini, hasil-hasil yang diperoleh dari tahapan pengolahan data diinterpretasikan, sehingga dapat menghasilkan identifikasi terhadap faktor-faktpr yang paling mempengaruhi produktivitas, mempelajari bagaimana model membuat prediksi, dan menggunakan hasil analisis untuk membuat saran manajemen yang dapat meningkatkan produktivitas karyawan.

III. HASIL DAN PEMBAHASAN

Untuk menguji usulan metodologi analisis produktivitas tenaga kerja yang dijelaskan pada Bab II, usulan metodologi tersebut diterapkan dengan menggunakan data sekunder, yaitu dengan menggunakan dataset yang bisa diakses secara bebas oleh publik. Sebagai alat bantu untuk melakukan analisis digunakan bahasa pemrograman yang difasilitasi oleh IDE RStudio. Penjelasan mengenai setiap tahap yang dijalankan sesuai dengan usulan metodologi tersebut di bawah ini.

3.1 Tahap Pengumpulan Data

Karena penerapan usulan metodologi analisis produktivitas tenaga kerja dilakukan dengan data sekunder, maka proses pengumpulan data yang sesungguhnya tidak dilakukan dalam proses penulisan artikel ini. Data sekunder yang digunakan adalah sebuah dataset produktivitas tenaga kerja, yang merupakan dataset publik tersedia melalui UCI Machine Learning Repository (2020). Data merupakan data produktivitas tenaga kerja dari suatu industri garmen, yang mencakup dua departemen selama 3 bulan kerja. Data tersebut terdiri dari 15 variabel dan 1197 baris data. Gambar 2 menunjukkan tampilan dataset tersebut dan penjelasan masing-masing variabel terdapat pada Tabel 1.

	date	quarter	department	day	team	targeted_productivity	smv	wip	over_time	incentive	idle_time	idle_men	no_of_style_change	no_of_workers	actual_productivity
1	1/1/2015	Quarter1	sewing	Thursday	8	0.80	26.16	1108	7080	98	0	0	0	59.0	0.9407254
2	1/1/2015	Quarter1	finishing	Thursday	1	0.75	3.94	NA	960	0	0	0	0	8.0	0.8865000
3	1/1/2015	Quarter1	sewing	Thursday	11	0.80	11.41	968	3660	50	0	0	0	30.5	0.8005705
4	1/1/2015	Quarter1	sewing	Thursday	12	0.80	11.41	968	3660	50	0	0	0	30.5	0.8005705
5	1/1/2015	Quarter1	sewing	Thursday	6	0.80	25.90	1170	1920	50	0	0	0	56.0	0.8003819
6	1/1/2015	Quarter1	sewing	Thursday	7	0.80	25.90	984	6720	38	0	0	0	56.0	0.8001250
7	1/1/2015	Quarter1	finishing	Thursday	2	0.75	3.94	NA	960	0	0	0	0	8.0	0.7551667
8	1/1/2015	Quarter1	sewing	Thursday	3	0.75	28.08	795	6900	45	0	0	0	57.5	0.7536683
9	1/1/2015	Quarter1	sewing	Thursday	2	0.75	19.87	733	6000	34	0	0	0	55.0	0.7550975
10	1/1/2015	Quarter1	sewing	Thursday	1	0.75	28.08	681	6900	45	0	0	0	57.5	0.7504278
11	1/1/2015	Quarter1	sewing	Thursday	9	0.70	28.08	972	6900	44	0	0	0	57.5	0.7211270
12	1/1/2015	Quarter1	sewing	Thursday	10	0.75	19.31	576	6480	45	0	0	0	54.0	0.7132062
13	1/1/2015	Quarter1	sewing	Thursday	5	0.80	11.41	555	3660	50	0	0	0	30.5	0.7005459
14	1/1/2015	Quarter1	finishing	Thursday	10	0.65	3.94	NA	960	0	0	0	0	8.0	0.7099167
15	1/1/2015	Quarter1	finishing	Thursday	8	0.75	2.90	NA	960	0	0	0	0	8.0	0.6766667
16	1/1/2015	Quarter1	finishing	Thursday	4	0.75	3.94	NA	2160	0	0	0	0	18.0	0.9000556
17	1/1/2015	Quarter1	finishing	Thursday	7	0.80	2.90	NA	960	0	0	0	0	8.0	0.9497292
18	1/1/2015	Quarter1	sewing	Thursday	4	0.65	23.69	581	7200	0	0	0	0	60.0	0.8211600

Showing 1 to 18 of 1,197 entries, 15 total columns

Gambar 2. Tampilan Dataset

Tabel 1. Penjelasan Variabel dalam Dataset Produktivitas (UCI Machine Learning Repository, 2020)

No	Nama Variabel	Deskripsi
1	date	Tanggal dengan format MM-DD-YYYY
2	quarter	Bagian dari bulan, satu bulan dibagi menjadi lima bagian
3	department	Departemen yang berkaitan dengan obyek pada baris tersebut
4	day	Hari dalam minggu
5	team	Nomor tim yang berkaitan dengan obyek pada baris tersebut
6	targeted_productivity	Target produktivitas yang ditetapkan oleh pihak yang berwenang untuk setiap tim pada hari tersebut
7	smv	*Standard Minute Value* yaitu waktu yang dialokasikan untuk suatu tugas
8	wip	*Work in Progress* yaitu jumlah item yang belum terselesaikan
9	over_time	*Overtime* atau waktu lembur untuk setiap tim dalam satuan menit
10	incentive	Insentif yang diberikan untuk memotivasi aksi tertentu
11	idle_time	Waktu produk menunggu karena berbagai alasan
12	idle_men	Jumlah tenaga kerja yang idle karena disrupsi produksi
13	no_of_style_change	Jumlah perubahan style produk
14	no_of_workers	Jumlah tenaga kerja per tim

15	actual_productivity	Produktivitas aktual tenaga kerja, berkisar antara 0-1

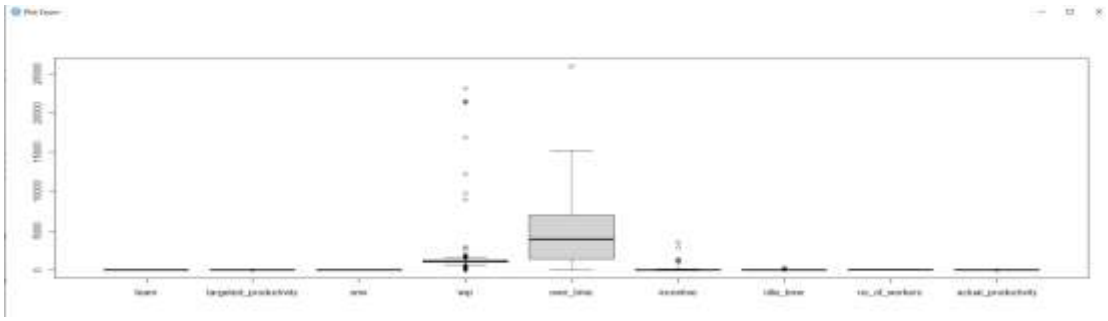
3.2 Tahap Penyiapan Data

Tahap penyiapan data dilakukan sesuai dengan penjelasan pada bagian 2.2 di atas. Hal pertama yang dilakukan adalah mengidentifikasi dan menangani adanya data yang tidak lengkap. Dapat dilihat pada Gambar 3 bahwa dari keseluruhan variabel yang ada, hanya terdapat data yang tidak lengkap pada variabel wip. Karena jumlahnya cukup besar, yaitu sebanyak 506 data, jika baris atau data tersebut dihapus akan sangat mempengaruhi kuantitas data yang dapat dianalisis. Oleh karena itu untuk data yang tidak lengkap pada variabel wip ini dilakukan penggantian nilai yang tidak lengkap tersebut dengan nilai perkiraan lain yaitu rata-rata atau *mean* variable wip tersebut.

```
> missing_count <- colSums(is.na(data_garment))
> print(missing_count)
                date              quarter           department                  day                 team targeted_productivity                  smv
                   0                    0                    0                    0                    0                    0                    0
                 wip           over_time            incentive           idle_time             idle_men    no_of_style_change       no_of_workers
                 506                    0                    0                    0                    0                    0                    0
  actual_productivity
                   0
>
```

Gambar 3. Jumlah missing data tiap variabel

Hal kedua yang dilakukan dalam tahap ini adalah mengidentifikasi adanya data yang tidak wajar dengan melihat pola data dari masing-masing variabel dengan menggunakan diagram *box plot* seperti yang terlihat pada Gambar 4. Terlihat pada gambar tersebut bahwa terdapat data yang berada di luar batas kuartil untuk variabel wip dan incentive. Hal ini menunjukkan bahwa dalam kedua variabel tersebut terdapat indikasi variasi yang tinggi dalam nilai-nilai datanya. Tentunya fakta ini akan menjadi pertimbangan khusus terhadap hasil analisis yang didapatkan nanti.



Gambar 4. Diagram *box plot* seluruh variabel

Langkah selanjutnya yang dilakukan dalam tahap penyiapan data ini adalah penyesuaian tipe data, pemilihan faktor atau variabel, serta pembagian data menjadi data pelatihan dan data pengujian. Seperti terlihat pada Gambar 5 bagian atas, data yang tersedia memiliki tipe data yang bervariasi. Untuk itulah sesuai dengan kebutuhan dari metode *machine learning* yang dipilih, maka terlihat pada Gambar 5 bagian bawah, variabel quarter, department, dan day yang awalnya bertipe data character, harus diubah menjadi tipe data factor, sedangkan variable team yang awalnya bertipe data int diubah menjadi numeric.

```
> str(data_garment)
'data.frame':   1197 obs. of  15 variables:
 $ date                  : chr  "1/1/2015" "1/1/2015" "1/1/2015" "1/1/2015" ...
 $ quarter               : chr  "Quarter1" "Quarter1" "Quarter1" "Quarter1" ...
 $ department            : chr  "sewing" "finishing " "sewing" "sewing" ...
 $ day                   : chr  "Thursday" "Thursday" "Thursday" "Thursday" ...
 $ team                  : int  8 1 11 12 6 7 2 3 2 1 ...
 $ targeted_productivity: num  0.8 0.75 0.8 0.8 0.8 0.8 0.75 0.75 0.75 0.75 ...
 $ smv                   : num  26.16 3.94 11.41 11.41 25.9 ...
 $ wip                   : int  1108 NA 968 968 1170 984 NA 795 733 681 ...
 $ over_time             : int  7080 960 3660 3660 1920 6720 960 6900 6000 6900 ...
 $ incentive             : int  98 0 50 50 50 38 0 45 34 45 ...
 $ idle_time             : num  0 0 0 0 0 0 0 0 0 0 ...
 $ idle_men              : int  0 0 0 0 0 0 0 0 0 0 ...
 $ no_of_style_change    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ no_of_workers         : num  59 8 30.5 30.5 56 56 8 57.5 55 57.5 ...
 $ actual_productivity   : num  0.941 0.886 0.801 0.801 0.8 ...
```

```
#mengubah variable
data_garment$quarter <- as.factor(data_garment$quarter)
data_garment$department <- as.factor(data_garment$department)
data_garment$day <- as.factor(data_garment$day)
data_garment$team <- as.numeric(data_garment$team)
```

Gambar 5. Perubahan tipe data

Selanjutnya dilakukan pemilihan faktor atau variabel yang akan dimasukkan ke dalam algoritma *machine learning* dengan bantuan analisis korelasi. Dapat dilihat pada Gambar 6 bahwa variabel yang bukan merupakan variabel numerik, yaitu date, quarter, department, dan day, tidak disertakan dalam analisis ini untuk melihat korelasi antar variabel yang memiliki hubungan dengan nilai produktivitas yang terdapat pada variabel actual_productivity. Pada Gambar 6 juga terlihat bahwa beberapa variabel memiliki nilai korelasi yang relatif besar terhadap variabel actual_productivity seperti variabel targeted_productivity, team, dan smv. Sebetulnya untuk analisis bisa saja variabel yang nilai korelasinya sangat kecil tidak dimasukkan ke dalam algoritma *machine learning*, akan tetapi dalam penelitian ini semua dimasukkan untuk melihat pengaruhnya seperti apa terhadap hasil. Tentunya fakta ini juga akan menjadi pertimbangan khusus terhadap hasil analisis yang didapatkan nanti.

```
> #cek korelasi data numeric
> correlation <- cor(data_garment[, c('team','targeted_productivity','smv', 'wip', 'over_time', 'incentive', 'idle_time', 'no_of_workers', 'actual_productivity')])
> correlation
                        team targeted_productivity         smv         wip    over_time    incentive    idle_time no_of_workers actual_productivity
team                1.000000000          0.03027435 -0.11001074 -0.025384135 -0.096736886 -0.007673930  0.003796181  -0.075113389        -0.14875331
targeted_productivity 0.030274346          1.00000000 -0.06948887  0.049114361 -0.088556693  0.032767903 -0.056180897  -0.084287865         0.42159388
smv                -0.110010740         -0.06948887  1.00000000 -0.018322003  0.674887440  0.032628856  0.056862784   0.912176312        -0.12208884
wip                -0.025384135          0.04911436 -0.01832200  1.000000000  0.034489835  0.021880730 -0.026267497   0.009790782         0.08836461
over_time          -0.096736886         -0.08855669  0.67488744  0.034489835  1.000000000 -0.004793252  0.031037596   0.734164174        -0.05420584
incentive          -0.007673930          0.03276790  0.03262886  0.021880730 -0.004793252  1.000000000 -0.012023621   0.049222218         0.07653763
idle_time           0.003796181         -0.05618090  0.05686278 -0.026267497  0.031037596 -0.012023621  1.000000000   0.058049300        -0.08085081
no_of_workers      -0.075113389         -0.08428787  0.91217631  0.009790782  0.734164174  0.049222218  0.058049300   1.000000000        -0.05799059
actual_productivity -0.148753311          0.42159388 -0.12208884  0.088364609 -0.054205837  0.076537627 -0.080850810  -0.057990592         1.00000000
```

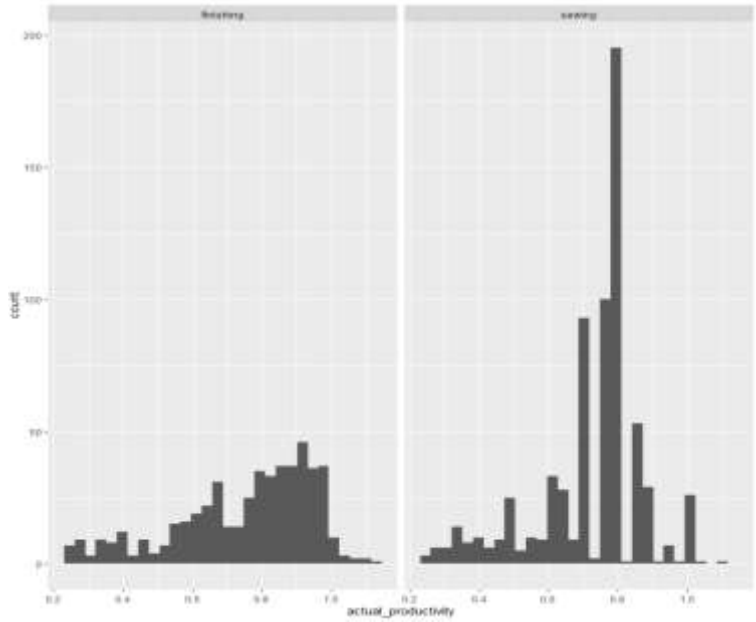
Gambar 6. Korelasi Variabel

Untuk menyiapkan data agar sesuai dengan algoritma *machine learning* yang digunakan, dalam data tersebut ditambahkan satu variabel baru yang diberi nama status. Variabel ini menunjukkan jika nilai produktivitas aktual sudah melebih dari nilai produktivitas yang ditargetkan, atau dapat ditunjukkan dengan kondisi nilai variabel actual_productivity > targeted_productivity. Jika kondisi tersebut terpenuhi, nilai variabel status diatur sebagai "OK", sementara jika kondisi tersebut tidak terpenuhi nilai variabel status diatur sebagai "Tidak". Variabel status ini disiapkan sebagai variabel dependen untuk algoritma classification C50, mengingat pada algoritma tersebut diperlukan variabel dependen yang bertipe *factor*, sementara untuk algoritma *regression tree* yang membutuhkan variabel dependen yang bertipe *numeric* telah tersedia variabel actual_productivity yang dapat digunakan sebagai variabel dependen.

Hal terakhir yang dilakukan dalam tahap ini adalah membagi data menjadi data latihan dan data uji. Terdapat sebanyak 837 data yang digunakan sebagai data pelatihan dan 360 data digunakan sebagai data uji, yang dipisahkan secara acak dengan fungsi bawaan yang tersedia di bahasa pemrograman R.

3.3 Tahap Pengolahan Data dengan Algoritma Machine Learning

Setelah data telah melalui tahap penyiapan data, sebelum dilakukan pengolahan dengan algoritma *machine learning*, yaitu dengan menggunakan *regression tree* dan *classification* C50, dilakukan analisis dengan menggunakan grafik histogram untuk memberikan informasi awal tentang pola data produktivitas actual antara departemen *sewing* dan departemen *finishing* seperti yang terlihat pada Gambar 7. Ditemukan bahwa pola produktivitas dari kedua departemen berbeda, di mana produktivitas aktual departemen *finishing* cukup merata pada rentang nilai 0,75 sampai 0,95, sementara produktivitas actual departemen *sewing* terlihat dominan di sekitar nilai 0,8.

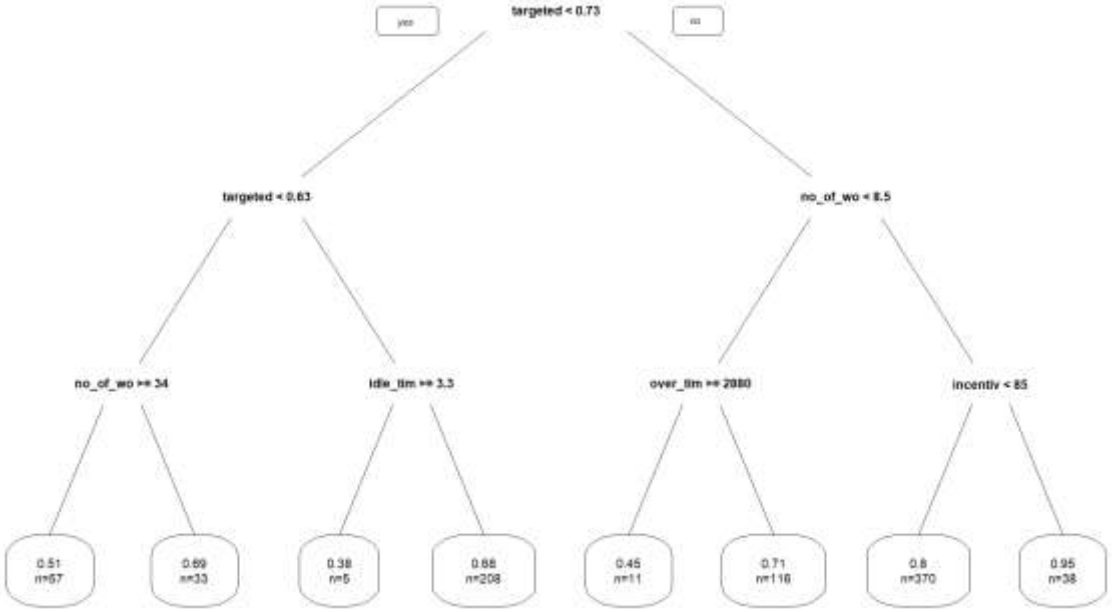


Gambar 7. Histogram Produktivitas Aktual antar Departemen

Setelah itu dilakukan analisis lanjutan dengan menggunakan metode *regression tree*. Pada analisis ini, digunakan variabel actual_productivity sebagai variabel dependen dan variable team, targeted_productivity, smv, wip, over_time, incentive, idle_time, dan no_of_workers sebagai variabel independen.

Terlihat pada Gambar 8, nilai akurasi prediksi dari metode *regression tree* ini adalah 0,8976 yang sudah merupakan nilai akurasi yang relatif tinggi pada penerapan algoritma *machine learning*. Dapat dibaca dari Gambar 8, bahwa kelompok yang mempunyai nilai variabel actual_productivity yang tertinggi, yaitu dengan nilai rata-rata 0,95, adalah kelompok dengan nilai variabel targeted_productivity lebih dari 0,73, nilai variabel no_of_worker lebih dari 8,5, dan nilai variabel incentive lebih dari 85. Kelompok ini terdiri dari 38 data. Sementara itu, kelompok yang dengan nilai variabel actual_productivity yang terendah, yaitu dengan nilai rata-rata 0,38, adalah kelompok dengan nilai variabel targeted_productivity dalam rentang 0,63 sampai 0,73, dan idle_time lebih besar dari 3,3. Hasil analisis juga menunjukkan bahwa tidak semua variabel independen yang dipertimbangkan dalam proses pembuatan model akhirnya masuk ke dalam model yang dibuat. Variabel yang masuk ke dalam model adalah targeted_productivity, over_time, incentive, idle_time, dan no_of_workers. Sementara variabel team, smv, dan wip tidak masuk ke dalam model. Jika hasil ini disandingkan dengan analisis korelasi yang terdapat pada Gambar 6 di atas, variabel targeted_productivity yang mempunyai nilai korelasi tertinggi, juga menjadi variabel pembeda yang utama pada model *regression tree* yang terbentuk.

```
> #analisis regression tree
> library(rpart)
> library(rpart.plot)
> model_regression_tree <- rpart(formula = actual_productivity~team + targeted_productivity
+ smv + wip + over_time + incentive + idle_time + no_of_workers, data = train_data, control
= rpart.control(maxdepth = 3, minbucket = 1, cp = 0))
> accuracy <- 1 - mean(abs(test_data$actual_productivity - predictions))
> accuracy
[1] 0.8975937
```

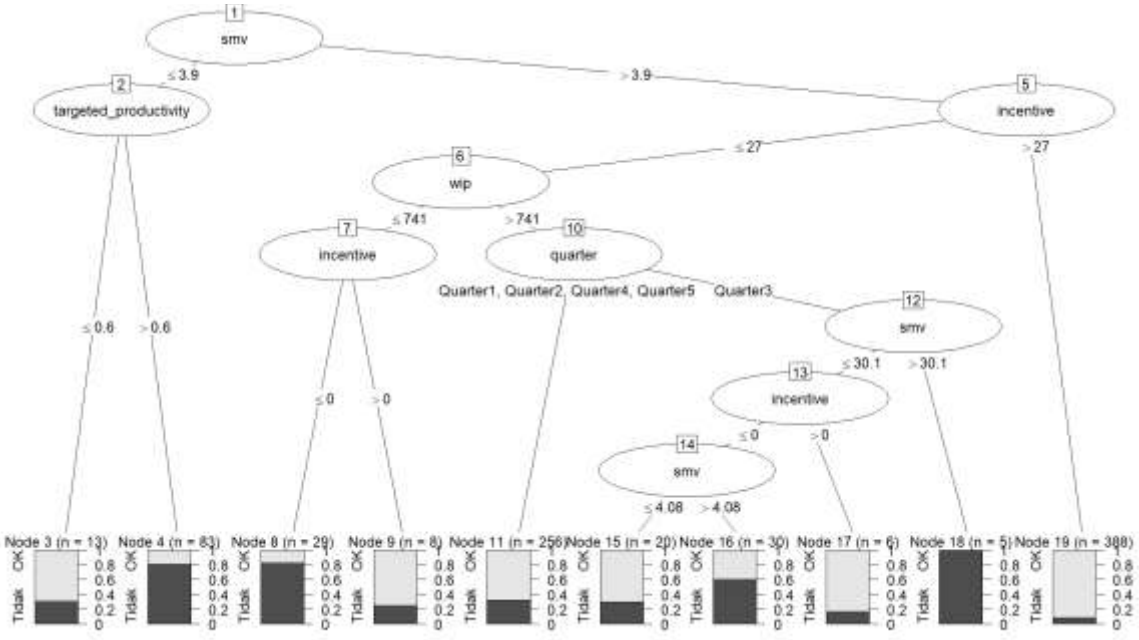


Gambar 8. Analisis Regression Tree

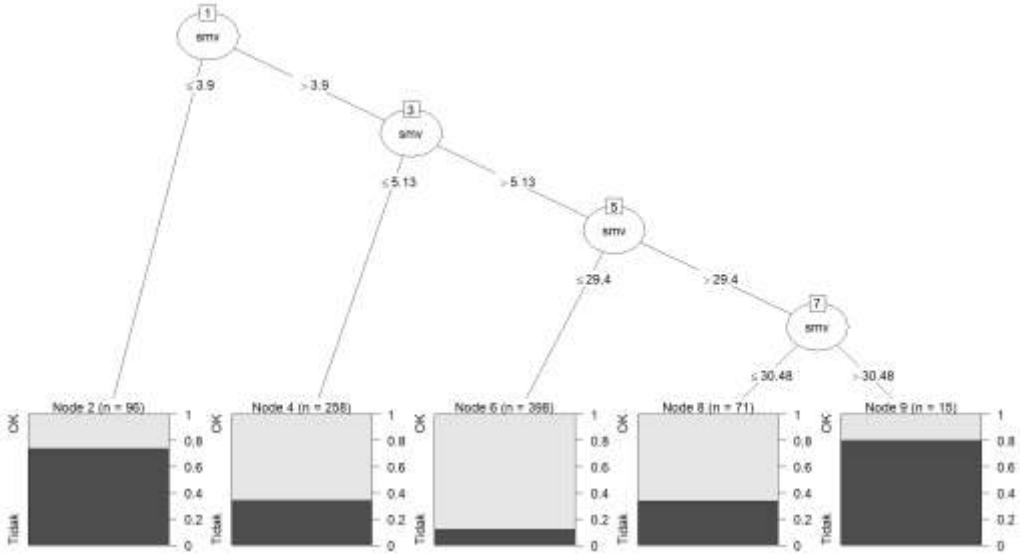
Selanjutnya dilakukan analisis dengan menggunakan metode *classification tree* yang mampu memasukkan variabel independen yang bertipe faktor. Pada analisis ini, mula-mula digunakan variabel status sebagai variabel dependen dan variable team, targeted_productivity, department, smv, wip, no_of_workers, over_time, incentive, quarter, idle_time, day, dan no_of_style_change sebagai variabel independen. Seperti terlihat pada Gambar 9, dengan menggunakan *library* C50, didapatkan model dengan nilai akurasi prediksi sebesar 0,7910. Terlihat juga pada gambar tersebut, bahwa tingkat kepentingan masing-masing variabel independen bervariasi mulai dari variabel smv yang memiliki tingkat kepentingan terbesar, sampai dengan variabel department yang memiliki tingkat kepentingan terkecil. Oleh karena itu, selanjutnya analisis diulang dengan variabel independen yang direduksi menjadi hanya 6 variabel dengan tingkat kepentingan terbesar, yaitu smv, incentive, wip, quarter, targeted_productivity, dan no_of_style_change.

```
> #analisis classification tree
> library(C50)
> library(caret)
> C50_model_1 <- C5.0(status ~ team + targeted_productivity + department + smv + wip
+ no_of_workers + over_time + incentive + quarter + day + no_of_style_change, data =
train_data_1)
> predictions <- predict(C50_model_1, test_data_1)
> confusion_matrix <- table(Predicted = predictions, Actual = test_data_1$status)
> accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
> print(accuracy)
[1] 0.7910864
> importance <- varImp(C50_model_1)
> print(importance)
                     Overall
smv                   100.00
incentive              88.54
wip                    44.87
quarter                38.31
targeted_productivity  35.44
no_of_style_change     34.96
no_of_workers          20.53
over_time              15.39
team                   11.69
day                    11.10
department              6.92
```

Gambar 9. Analisis Classification Tree tahap 1



Gambar 10. Hasil *Classification Tree* tahap 2



Gambar 11. Hasil *Classification Tree* tahap 3

Hasil *classification tree* pada tahap 2 dapat dilihat pada Gambar 10. Dari gambar tersebut bisa dilihat bahwa variabel independent no_of_style_change tidak muncul sebagai variabel penentu yang dapat membuat cabang dalam *classification tree*. Sementara itu dari analisis tingkat kepentingan, variabel targeted_productivity memiliki nilai yang sangat kecil. Kemudian dipertimbangkan bahwa insentif biasanya merupakan imbalan yang diberikan kepada karyawan jika target terpenuhi. Oleh karena itu dalam analisis tahap berikutnya ketiga variabel tersebut dihilangkan dari variabel independen. Hasil *classification tree* tahap 3 ini terdapat pada Gambar 11, dengan nilai akurasi yang sama dengan nilai akurasi pada tahap 1. Berdasarkan hasil tersebut dapat disimpulkan bahwa variabel status yang bernilai OK, yaitu nilai yang menunjukkan kondisi saat produktivitas aktual melebihi target, sangat dipengaruhi oleh variabel smv. Pada saat nilai smv lebih dari 5,13 dan kurang dari 29,4, peluang variabel status bernilai OK sangat tinggi yaitu sekitar 90% (Node 6). Sementara itu pada saat nilai smv sangat kecil, yaitu kurang dari 3,9, atau sangat besar, yaitu lebih dari 30,48, peluang variabel status bernilai OK sangat rendah atau dengan kata lain peluang variabel status bernilai Tidak sangat tinggi, yaitu sekitar 80% (Node 2 dan Node 9).

3.4 Tahap Interpretasi Hasil

Dua hasil analisis yang dilakukan pada tahap sebelumnya memiliki potensi untuk digunakan dalam dua tujuan, yaitu prediksi nilai produktivitas dan perbaikan nilai produktivitas. Untuk mencapai tujuan yang pertama, model *machine learning* yang telah diperoleh pada tahap sebelumnya dapat digunakan secara langsung untuk memprediksi nilai produktivitas jika nilai-nilai variabel independen yang dibutuhkan dalam model tersebut sudah diketahui. Karena terdapat dua model yang dihasilkan dengan variabel dependen yang berbeda, maka nilai produktivitas juga dapat dinyatakan dengan dua output yang berbeda, masing-masing yaitu nilai produktivitas aktual yang bertipe numerik dan nilai status ketercapaian produktivitas yang bertipe faktor (target produktivitas tercapai atau tidak).

Sementara itu untuk mencapai tujuan yang kedua, faktor-faktor dominan yang mempengaruhi produktivitas dapat diinterpretasikan berdasarkan model yang terbentuk. Sebagai contoh, berdasarkan model *Regression Tree* dapat diinterpretasikan bahwa nilai produktivitas aktual sangat dipengaruhi oleh nilai target produktivitas (yang dinyatakan dalam variabel independen targeted_productivity), karena baik dalam kelompok yang memiliki nilai produktivitas aktual tertinggi maupun kelompok yang memiliki nilai produktivitas aktual terendah, variabel targeted_productivity muncul sebagai pembeda utama. Sementara itu nilai jumlah tenaga kerja dan insentif merupakan faktor pendukung lain yang mendukung tingginya nilai produktivitas aktual, sedangkan waktu produk menunggu (*idle time*) merupakan faktor pendukung lain yang menyebabkan rendahnya nilai produktivitas actual.

Sedangkan berdasarkan model *Classification Tree*, dapat diinterpretasikan bahwa status ketercapaian target produktivitas sangat dipengaruhi oleh waktu standar penyelesaian pekerjaan yang direpresentasikan variabel independen smv selain beberapa variabel independen lain seperti incentive, wip, quarter, dan targeted_productivity. Hal menarik yang diperoleh adalah terdapat rentang nilai batas bawah dan batas atas smv yang membuat target produktivitas tercapai. Jika variabel independent smv di bawah nilai batas bawah dan di atas nilai batas atas tersebut, besar peluangnya bahwa target produktivitas tidak tercapai.

Kombinasi analisis dengan menggunakan dua model ini memberikan peluang eksplorasi yang lebih luas terhadap analisis produktivitas secara keseluruhan.

IV. SIMPULAN

Berdasarkan contoh penerapan yang ada dalam bab 3 di atas, dapat disimpulkan bahwa usulan metodologi untuk melakukan analisis produktivitas tenaga kerja dengan menggunakan metode regression tree and classification C50 secara umum dapat menghasilkan output yang diharapkan yaitu menganalisis dan memprediksi faktor-faktor yang mempengaruhi produktivitas tenaga kerja pada konteks yang spesifik. Disamping itu hasil yang diperoleh dapat dengan mudah diinterpretasikan dan dijelaskan kepada berbagai pemangku kepentingan. Dengan mengikuti tahapan metodologi usulan tersebut, metode yang sama tentunya dapat diterapkan pada dataset produktivitas tenaga kerja yang lain, misalnya pada perusahaan yang berbeda atau pada kurun waktu yang berbeda. Harapannya adalah output yang dihasilkan bisa bermanfaat bagai perusahaan tersebut terkait upaya untuk meningkatkan produktivitas tenaga kerjanya, misalnya untuk mengidentifikasi faktor-faktor kunci yang mempengaruhi produktivitas tenaga kerja perusahaan tersebut sebagai dasar untuk pengambilan keputusan dan kebijakan manajerial.

UCAPAN TERIMA KASIH

Terima kasih disampaikan kepada Program Studi Magister Teknik Industri dan Departemen Teknik Industri Universitas Atma Jaya Yogyakarta atas dukungan terhadap penelitian dan prose penulisan artikel ini. Terima kasih juga disampaikan kepada para reviewer atas masukan yang membangun dalam proses finalisasi artikel ini.

DAFTAR PUSTAKA

UCI Machine Learning Repository (2020). *Productivity Prediction of Garment Employees [Dataset]*. https://doi.org/10.24432/C51S6D.

Kementerian Tenaga Kerja (2022). *Ketenagakerjaan dalam Data (Edisi 6)*. Jakarta: Satu Data Ketenagakerjaan. https://satudata.kemnaker.go.id/publikasi/90.

Adeniji, O. D., Adeyemi, S. O., & Ajagbe, S. A. (2022). An improved bagging ensemble in predicting mental disorder using hybridized random forest-artificial neural network model. *Informatica*, *46*(4).

Al Imran, A., Amin, M. N., Rifat, M. R. I., & Mehreen, S. (2019, April). Deep neural network approach for predicting the productivity of garment employees. In *2019 6th International Conference on Control, Decision and Information Technologies (CoDIT)* (pp. 1402-1407). IEEE.

Aristizabal, S., Byun, K., Wood, N., Mullan, A. F., Porter, P. M., Campanella, C., Jamrozik, A., Nenadic, I. Z., & Bauer, B. A. (2021). The Feasibility of Wearable and Self-Report Stress Detection Measures in

a Semi-Controlled Lab Environment. *IEEE Access*, *9*, 102053–102068. https://doi.org/10.1109/ACCESS.2021.3097038

Benediktus, N., & Oetama, R. S. (2020). The decision tree c5. 0 classification algorithm for predicting student academic performance. *Ultimatics: Jurnal Teknik Informatika*, *12*(1), 14-19.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

De Salvio, D., Bianco, M. J., Gerstoft, P., D'Orazio, D., & Garai, M. (2023). Blind source separation by long-term monitoring: A variational autoencoder to validate the clustering analysis. *The Journal of the Acoustical Society of America*, *153*(1), 738-750.

De Silva, T. R. S., Dayananda, K. Y., Arachchi, R. G., Amerasekara, M. K. S. B., Silva, S., & Gamage, N. (2022, December). Solution to Measure Employee Productivity with Employee Emotion Detection. In *2022 4th International Conference on Advancements in Computing (ICAC)* (pp. 210-215). IEEE.

Fadli, S., Ashari, M., Studi Sistem Informasi, P., & Lombok, S. (2021). *JISA (Jurnal Informatika dan Sains) Optimization of Support Vector Machine Method Using Feature Selection to Improve Classification Results*.

Gu, J. (2022). Image Model and Algorithm of Human Resource Optimal Configuration Based on FPGA and Microsystem Analysis. *Wireless Communications and Mobile Computing*, *2022*. https://doi.org/10.1155/2022/7911419

Obiedat, R., & Toubasi, S. (2022). A Combined Approach for Predicting Employees' Productivity based on Ensemble Machine Learning Methods. *Informatica (Slovenia)*, *46*(5), 49–58. https://doi.org/10.31449/inf.v46i5.3839

Prabawati, N. I., & Ajie, H. (2019). Kinerja Algoritma Classification And Regression Tree (Cart) dalam Mengklasifikasikan Lama Masa Studi Mahasiswa yang Mengikuti Organisasi di Universitas Negeri Jakarta. *PINTER: Jurnal Pendidikan Teknik Informatika dan Komputer*, *3*(2), 139-145.

Razali, M. N., Ibrahim, N., Hanapi, R., Zamri, N. M., & Manaf, S. A. (2023). Exploring Employee Working Productivity: Initial Insights from Machine Learning Predictive Analytics and Visualization. *Journal of Computing Research and Innovation (JCRINN*, *8*(2), 1–10. https://doi.org/10.24191/jcrinn.v8i2.362

Sabuj, H. H., Nuha, N. S., Gomes, P. R., Lameesa, A., & Alam, M. A. (2022, December). Interpretable Garment Workers' Productivity Prediction in Bangladesh Using Machine Learning Algorithms and Explainable AI. In *2022 25th International Conference on Computer and Information Technology (ICCIT)* (pp. 236-241). IEEE.

Saxena, S., Deogaonkar, A., Pais, R., & Pais, R. (2023). Workplace Productivity Through Employee Sentiment Analysis Using Machine Learning. *International Journal of Professional Business Review: Int. J. Prof. Bus. Rev.*, *8*(4), 14.

Sungheetha, A., & Sharma R, R. (2021). A Comparative Machine Learning Study on IT Sector Edge Nearer to Working From Home (WFH) Contract Category for Improving Productivity. *Journal of Artificial Intelligence and Capsule Networks*, *2*(4), 217–225. https://doi.org/10.36548/jaicn.2020.4.004