Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 363−370
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2137

Pemodelan Klasifikasi Gaji Menggunakan Support Vector Machine

Anas Satria Lombu, Syarif Hidayat*, Ahmad Fathan Hidayatullah

Fakultas Teknologi Industri, Program Studi Magister Informatika, Universitas Islam Indonesia, Yogyakarta, Indonesia
Email: [1]anas.lombu@students.uii.ac.id , [2,*]syarif@uii.ac.id, , [3]fathan@uii.ac.id
Submitted: **18/08/2022**; Accepted: **30/08/2022**; Published: **30/08/2022**

Abstrak−Diketahui saat ini telah banyak jenis pekerjaan yang ada di lapangan. Kreatifitas masyarakat serta tekanan ekonomi yang dirasakan membuat masyarakat harus bekerja keras untuk dapat memenuhi kebutuhan hidup. Salah satu cara yang mesti dilakukan untuk dapat terus bertahan hidup dengan cara bekerja. Dengan bekerja seseorang dapat menghasilkan upah atau gaji sehingga kebutuhan hidup seseorang dapat terpenuhi. Beragam pekerjaan yang ada menimbulkan suatu permasalahan. Dalam menentukan gaji atau upah dari suatu pekerjaan. Gaji yang diberikan kepada seseorang harus sesuai dengan kriteria dari pekerja tersebut. Maka diperlukan suatu model machine learning untuk memprediksi gaji seseorang. Pada penelitian ini, dibuat suati model klasifikasi untuk menentukan seseorang dikategorikan ke dalam gaji diatas 7 juta dan gaji dibawah 7 juta berdasarkan kriteria-kriteria atau atribut yang cocok. Penelitian ini menggunakan bahasa pemrograman python dan mengambil 1000 sample dari dataset yang diperoleh dari kaggle. Metode machine learning yang digunakan adalah Support Vector Machine. Kemudian dibandingan dengan metode K-Nearest Neighbors. Pada model SVM diperoleh akurasi model sebesar 87% dan 86% untuk model KNN. Dari hasil akurasi diperoleh bahwa model SVM lebih baik dibanding dengan model KNN dalam melakukan klasifikasi gaji berdasarkan pekerjaan yang ada.

Kata Kunci: Klasifikasi; Gaji; Machine Learning; SVM

Abstract−It is known that there are currently many types of work in the field. Creativity of the community and economic pressure that is felt makes people have to work hard to be able to meet the needs of life. One way that must be done to be able to continue to survive by working. By working someone can produce wages or salaries so that the necessities of life of a person can be met. Various work that exists raises a problem. In determining the salary or wages of a job. The salary given to someone must be in accordance with the criteria of the worker. Then we need a Machine Learning model to predict a person's salary. In this study, a classification model was made to determine a person to be categorized into salaries above 7 million and salaries below 7 million based on suitable criteria or attributes. This study uses the Python programming language and took 1000 samples from the dataset obtained from Kaggle. The Machine Learning method used is the Support Vector Machine. Then compared to the K-Nearest Neighbors method. In the SVM model the model accuracy was obtained of 87% and 86% for the KNN model. From the results of accuracy, it was found that the SVM model was better than the KNN model in conducting salary classifications based on existing jobs.

Keywords: Classification; Wages; Machine Learning; SVM

1. PENDAHULUAN

Gaji merupakan upah kerja yang dibayar dalam waktu yang tetap. Gaji juga dapat diartikan sebagai suatu balas jasa yang diterima pekerja dalam bentuk uang berdasarkan waktu tertentu[1]. Gaji atau upah memiliki peranan yang sangat penting dalam sebuah perusahaan karena upah merupakan salah satu faktor pendukung dalam kinerja para karyawan dalam sebuah perusahaan. Dengan upah yang optimal, dapat menjadi salah satu motivasi kepada karyawan untuk selalu memberikan kemampuan terbaik yang dimiliki untuk perusahaan [2]. Dimana kinerja yang bagus akan menunjang produktivitas suatu perusahaan. Dalam memberikan gaji kepada pegawai setiap perusahaan memiliki sistem yang berbeda-beda. Banyak faktor yang dapat mempengaruhi besar kecil upah yang akan diberikan. Banyak faktor dalam penentuan gaji kepada setiap pegawai, ditambah jumlah pegawai yang selalu bertambah di setiap perusahaan [3]. Akan sangat sulit untuk dikerjakan oleh manusia. Sehingga diperlukan suatu sistem yang dapat menentukan kisaran gaji yang optimal. *Machine Learning* merupakan suatu mesin yang dikembangkan untuk bisa belajar dengan sendirinya tanpa arahan penggunanya. Sistem dapat menentukan gaji seorang pegawai berdasarkan data-data yang telah dipelajari sebelumnya. Banyak nya faktor dalam menentukan besaran gaji akan dipelajari oleh mesin. Dari hasil pembelajaran tersebut mesin akan membuat suatu model yang dapat memprediksi besaran gaji yang sebaiknya diterima oleh setiap pegawai seacara otomatis. Model machine learning yang bisa digunakan dalam melakukan prediksi besaran gaji pegawai bervariasi. Dalam penelitian ini, penulis menggunakan model *machine learning* untuk memprediksi suatu kelas dengan metode *Support Vector Machine* untuk memisahkan pegawai yang memiliki gaji diatas sama dengan 7 juta dan dibawah 7 juta rupiah.

2. METODOLOGI PENELITIAN

2.1 Data Set

Dataset yang digunakan dalam penelitian ini merupakan dataset yang diperoleh Kaggle. Kaggle merupakan salah satu repositori terkenal untuk mendapatkan dataset. dataset ini sebenarnya digunakan untuk kompetisi final project di kelas Sanbercode. Dataset memiliki 13 atribut dan 35.994 sample. Atribut-atribut pada dataset diantaranya id unik yang dimiliki setiap sample(ID), umur yang dimiliki setiap *sample*(Umur), kelompok kelas perkerjaan masing-masing *sample* (Kelas Pekerja), berisi nilai akumulasi berdasasarkan populasi, ras, dan *gender* dengan

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 363−370
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2137

umur 16+ suatu wilayah, sample yang diambil dari wilayah yang memiliki karakteristik demografis yang sama akan memiliki nilai berat akhir yang sama (Berat Akhir), tingkat pendidikan terakhir masing-masing *sample* (Pendidikan), Tingkat pendidikan terakhir masing-masing sample (Jmlh Tahun Pendidikan), Status perkawinan masing-masing *sample*(Status Perkawinan), Pekerjaan saat ini masing- masing *sample* (Pekerjaan), Jenis Kelamin masing-masing sample(Jenis Kelamin), Keuntungan yang didapat jika sample menjual semua aset miliknya(Keuntungan Kapital), Kerugian yang didapat jika *sample* menjual semua aset miliknya(Kerugian Capital), Jam kerja masing-masing *sample* setiap minggunnya(Jam per Minggu), Nilai gaji masing-masing *sample* apakah kurang dari sama dengan 7jt atau lebih dari 7jt(Gaji).

2.2 Data Cleaning

Data cleaning ialah proses pembersihan data yang dilakukan pada dataset. Tidak semua dataset yang akan didapatkan dari *repository* yang ada merupakan dataset yang sudah siap pakai. *Data cleaning* sangat mempengaruhi performa dari suatu model. Faktanya, banyak penelitian lain yang menghabiskan hampir sebagian besar waktunya pada tahapan ini. Dikarenakan dalam data cleaning penulis diminta untuk membuang data dan informasi yang tidak dibutuhkan, melihat konsistensi format yang digunakan pada dataset, duplikasi data, dan *missing value* pada dataset.

2.3 Preprocessing Data

Setelah dataset melalui proses *cleaning data*, kemudian dataset telah siap di olah. *Preprocessing* data ialah tahapan dimana data akan di olah sebelum dijadikan sebuah model. Model machine learning hanya memahami data berupa *numeric*. Sedangkan dataset yang dimiliki memuat *data numeric* dan *categorical data.* Sehingga *categorical* data harus diubah dengan menggunakan teknik *One-Hot Encoding*. *One-Hot encoding* adalah salah satu teknik mempresentasikan data bertipe kategori sebagai vektor biner yang bernilai integer 0 dan 1[8]. biasanya digunakan pada nominal data. Nominal data ialah data kategori yang tidak memiliki unsur *ordering*. Dapat dilakukan dengan menggunakan 2 cara yaitu melalui library *Scikit Learn* atau dengan method *Get Dummies* dari pandas. Selain itu, data kategori yang bersifat Ordinal data adalah data kategori yang memiliki sifat *ordering* secara interinsik. Untuk menangani ordinal data digunakan *method Replace* dari pandas atau menggunakan *Sklearn Label Encoder*.

2.4 Model

Model yang digunakan dalam penlitian ini merupakan model klasifikasi Support Vector Machine. SVM merupakan metode machine learning yang bertujuan untuk memaksimalkan margin antara pola pelatihan dan batas keputusan(decision boundary) [9]. SVM bekerja dengan cara menemukan hyperplane terbaik dalam suatu model. Hyperplane dianotasikan sebagai:

$$f(x) = w^T x + b \quad (1)$$

Tahapan awal SVM ialah mencari support vector pada tiap kelas. Support vector ini berupa sampel dari masing-masing kelas yang memiliki jarak paling dekat. Support vector pada penelitian ialah gaji diatas 7 jt dan gaji di bawah 7 jt. Setelah support vector ditemukan. Setelah support vector ditemukan, kemudian menghitung margin. Margin ialah jarak antara support vector yang memisahkan 2 kelas.

Pada SVM, terdapat *low margin dan high margin*. *Low margin* ialah jarak antara masing-masing *support vector* berdekatan , sedangkan *high margin* ialah jarak antara masing-masing support vector berjauhan. SVM mencari margin terbesar yang mampu memisahkan kedua kelas [9]. SVM juga memiliki keunggulan diantaranya yaitu SVM efektif pada data berdimensi tinggi(data dengan jumlah fitur atau atribut yang sangat banyak), efektif pada kasus dimana jumlah fitur pada data lebih besar dari jumlah sample, dan penggunaan *support vector* sehingga membuat penggunaan memori lebih efisien[10].

Dalam penelitian ini, model SVM menggunakan kernel linear. Kernel merupakan sebuah metode untuk mengubah data pada dimensi tertentu(misal 2D) ke dalam dimensi yang lebih tinggi(3D) sehingga dapat menghasilkan hyperplane terbaik [11]**.** Beberapa macam fungsi kernel diantaranya :

Tabel 1. Kernel

Kernel	Definition
Linear	$K(x_i * x_j) = (x_i * x_j)$
Polynominal	$K(x_i * x_j) = (x_i * x_j + 1)^p$
Gaussian RBF	$K(x_i * x_j) = \exp(-\frac{\|\|x_i - x_j\|\|^2}{2\sigma^2}$

2.5 Evaluasi Model

Evaluasi model bertujuan untuk melihat performa dari model yang dikembangkan sebelumnya. Evaluasi model yang akurat pada saat model yang ada di lakukan evaluasi dengan menggunakan data testing. *Confusion matrix* merupakan salah satu teknik yang dapat dilakukan pada evaluasi model. *Confusion matrix* akan menghasilkan

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 363−370
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2137

berbagai performance *metrix* dari model yang telah dibuat. Performa *matrix* yang sering digunakan ialah *accuracy*, *precision*, dan *recall* [12].

2.6 Literatur Review

Istilah Machine Learning pertama kali dipopulerkan oleh Arthur Samuel, seorang ilmuwan komputer yang memelopori kecerdasan buatan pada tahun 1959, Machine learning adalah suatu cabang ilmu yang memberi komputer kemampuan untuk belajar tanpa diprogram secara eksplisit. Terdapat Aturan dan data yang menjadi masukan atau input bagi sistem. Secara eksplisit, aturan diekspresikan dalam bahasa pemrograman. Kemudian ditambah dengan data yang akan menghasilkan sebuah solusi yang dijadikan sebagai keluaran atau output. Paradigma pemrograman seperti ini sering disebut sebagai pemrograman trasdisional[4].

Machine Learning (ML) merupakan salah satu cabang dari *Artificial Inteligence* (kecerdasan buatan). Proses dalam machine *learning* ialah manusia memerintahkan suatu mesin untuk belajar dan melakukan improvisasi dari pengalaman(data), kemudian membuat keputusan atau prediksi di masa depan. Dalam *Machine Learning* ada beberapa istilah. Dataset yang dimiliki akan dibagi menjadi data *predictor* dan data target. Data *predictor* juga disebut data features dan dalam dunia statistik sebagai *independent variabels*. Sedangkan data target dikenal juga sebagai *label* dan dalam dunia statistik sebagai *dependent variabel*.

Machine Learning memiliki beberapa kategori diantaranya ialah *Supervised Learning*, pembelajaran menggunakan masukan data pembelajaran yang telah memiliki label. Setelah itu membuat prediksi dari data yang telah diberi label. Pada *Unsupervised Learning*, data pembelajaran tidak memiliki label. Model akan mengelompokan data berdasarkan karakteristik data. Reinforcement Learning, model yang belajar menggunakan sistem reward dan pinalti. Teknik yang mempelajari bagaimana membuat keputusan terbaik, secara berututan, untuk memaksimalkan ukuran sukses kehidupan nyata [5]. Entitas pembuat keputusan belajar melalui proses *trial* dan *error*.

Pada penelitian [6], melakukan prediksi dengan menggunakan model *linear regression*. Model ini akan melakukan prediksi dalam bilangan kontinu. Bilangan kontinu adalah bilangan numerik. Model regresi akan memprediksi sebuah nilai berdasarkan atribut atau variabel yang telah tersedia. Dalam penelitian (sitasi adrian), memiliki atribut berupa *position* dan *level*. atribut biasa nya disebut dengan variabel X (*independent variabel*) sedangkan target prediksi adalah variabel y (*dependent variabel*). model regresi akan melihat hubungan linear antara variabel x dan

y. hubungan linear tersebut akan direpresentasikan dengan sebuah garis lurus. Garis regresi adalah sebuah model probabilitas dan prediksi dari perkiraan yang telah dibuat. Ketika garis regresi telah dibuat, terdapat data yang berada dekat dengan garis regresi dan beberapa data berada pada posisi yang jauh dari garis regresi. Data yang berada jauh dari garis regresi biasanya dikenal sebagai *error*. Jika dalam suatu model ditemukan jumlah *error* yang cukup banyak maka model tersebut belum optimal. Sehingga diperlukan model regresi baru untuk mendapatkan model regresi terbaik.

Pada penelitian [7] memprediksi apakah pendapatan seseorang berada di atas kurang dari sama dengan 50k dan lebih besar dari 50k. dataset yang digunakan diperoleh dari UCI machine learning repository. Model *machine learning* yang digunakan ialah *Support Vector Machine*. Dataset memiliki 14 atribut dan sample berjumlah 48.842. kemudian dilakukan *preprocessing data* pada dataset. Dalam *preprocessing data*, diperoleh terdapat *categorical data* dan *numeric data. Machine learning* tidak dapat mengenali data berbentuk *categorical data* sehingga data yang awalnya berbentuk *categorical* perlu diubah menjadi data numeric terlebih dahulu. Kemudian data-data yang telah siap, digunakan untuk menentukan atribut dan class nya. Selanjutnya, dilakukan split data untuk membagi data menjadi data *training* dan data *testing*. data training digunakan untuk melatih model svm. Sedangkan data testing digunakan untuk melakukan pengujian dari model yang telah dibuat sebelumnya. Diperoleh hasil akurasi sebesar 0,74 untuk kernel ‘rbf’ dan 0,77 untuk kernel ‘linear’.

3. HASIL DAN PEMBAHASAN

Pada penelitian ini dilakukan klasifikasi untuk menentukan pegawai dengan gaji yang berada di atas 7jt dan gaji yang berada di bawah 7jt. Metode machine learning yang digunakan dalam penelitian ialah menggunakan SVM. Untuk mendapatkan model SVM terbaik telah dilalui tahapan-tahapan adalah sebagai berikut.

3.1 Karakteristik *Data Set*

Tahapan awal dalam membuat machine learning dimulai dari melakukan *import* dataset. Setalah melakukan *import* dataset, kemudian penulis harus mengenali karakteristik dari data-data dan mengenali nilai apa yang terkandung dalam data tersebut. Pada Gambar 1 merupakan dataset yang akan digunakan dalam penelitian ini.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 363−370
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2137

	id	Umur	Kelas Pekerja	Berat Akhir	Pendidikan	Jmlh Tahun Pendidikan	Status Perkawinan	Pekerjaan	Jenis Kelamin	Keuntungan Kapital	Kerugian Capital	Jam per Minggu	Gaji
0	0	21	Wiraswasta	242912	SMA	9	Belum Pernah Menikah	Servis Lainnya	Perempuan	0.0	0.0	35.0	<=7jt
1	1	49	Wiraswasta	140782	10th	6	Cerai	Eksekutif Managerial	Perempuan	0.0	0.0	40.0	<=7jt
2	2	44	Wiraswasta	120057	D3	12	Menikah	Eksekutif Managerial	Laki2	61404000.0	0.0	45.0	>7jt
3	3	24	Wiraswasta	194630	Sarjana	13	Belum Pernah Menikah	Spesialis	Laki2	0.0	0.0	35.0	<=7jt
4	4	33	Wiraswasta	219619	Master	14	Menikah	Spesialis	Laki2	210336000.0	0.0	40.0	>7jt
...	...	...	...	...	...	...	...	...	...	...	...	...	...
35989	35989	47	Pekerja Bebas Bukan Perusahan	148169	SMA	9	Menikah	Perbaikan Kerajinan	Laki2	0.0	0.0	40.0	<=7jt
35990	35990	69	Pekerja Bebas Perusahaan	264722	D3	12	Menikah	Sales	Laki2	0.0	0.0	40.0	>7jt
35991	35991	24	Pekerja Bebas Bukan Perusahan	31606	Sarjana	13	Menikah	Spesialis	Perempuan	0.0	0.0	20.0	>7jt
35992	35992	47	Wiraswasta	197836	SMA	9	Menikah	Sales	Laki2	0.0	0.0	45.0	<=7jt
35993	35993	45	Wiraswasta	243743	Sarjana	13	Menikah	Eksekutif Managerial	Laki2	0.0	0.0	60.0	>7jt

35994 rows × 13 columns

Gambar 1. Karakteristik *Data Set*

Dataset yang digunakan memuat 13 kolom dan 35994 baris. Selain julah baris dan kolom dalam dataset, perlu diketahui tipe data apa saja yang ada pada dataset tersebut. Tipe data yang digunakan dalam dataset tersebut adalah tipe data *int64*, *object*, dan *float64* seperti yang ditunjukkan pada gambar 2. dilihat dari tipe data yang terkandung pada dataset. Jenis data dibagi menjadi 2 yaitu data kategori dan data numerik. Tipe data object termasuk ke dalam tipe data kategori sedangkan tipe data *int64* dan *float64* termasuk ke tipe data numerik.

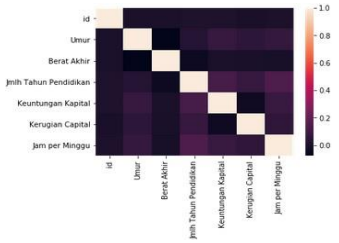
```
Data columns (total 13 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   id                     35994 non-null  int64
 1   Umur                   35994 non-null  int64
 2   Kelas Pekerja          35994 non-null  object
 3   Berat Akhir            35994 non-null  int64
 4   Pendidikan             35994 non-null  object
 5   Jmlh Tahun Pendidikan  35994 non-null  int64
 6   Status Perkawinan      35994 non-null  object
 7   Pekerjaan              35994 non-null  object
 8   Jenis Kelamin          35994 non-null  object
 9   Keuntungan Kapital     35994 non-null  float64
 10  Kerugian Capital       35994 non-null  float64
 11  Jam per Minggu         35994 non-null  float64
 12  Gaji                   35994 non-null  object
dtypes: float64(3), int64(4), object(6)
```

Gambar 2. Karakteristik *Data Set*

Penulis juga perlu melihat korelasi pada setiap atribut. Korelasi menunjukkan hubungan atau kedekatan yang terjadi di setiap atribut. Jadi suatu atribut yang ada di dataset akan dibandingkan dengan atribut lainnya yang berada dalam satu dataset.dapat dilihat pada Gambar 3 bahwa korelasi yang terjadi antar atribut. Korelasi ditunjukkan dalam rentang nilai 0 sampai dengan 1. jika nilai korelasi antar atribut mendekati angka 1 dapat dikatakan adanya korelasi diantara kedua atribut. Sedangkan jika nilai korelasi mendekati angka 0 maka dapat dikatakan tidak adanya hubungan antar atribut yang terjadi.

	id	Umur	Berat Akhir	Jmlh Tahun Pendidikan
id	1.000000	-0.001326	0.000503	0.007473
Umur	-0.001326	1.000000	-0.077357	0.031740
Berat Akhir	0.000503	-0.077357	1.000000	-0.040715
Jmlh Tahun Pendidikan	0.007473	0.031740	-0.040715	1.000000
Keuntungan Kapital	0.009452	0.078759	0.001944	0.121350
Kerugian Capital	-0.002831	0.052011	-0.003807	0.081421
Jam per Minggu	0.004061	0.070684	-0.009896	0.142300

	Keuntungan Kapital	Kerugian Capital	Jam per Minggu
id	0.009452	-0.002831	0.004061
Umur	0.078759	0.052011	0.070684
Berat Akhir	0.001944	-0.003807	-0.009896
Jmlh Tahun Pendidikan	0.121350	0.081421	0.142300
Keuntungan Kapital	1.000000	-0.031582	0.083304
Kerugian Capital	-0.031582	1.000000	0.060385
Jam per Minggu	0.083304	0.060385	1.000000



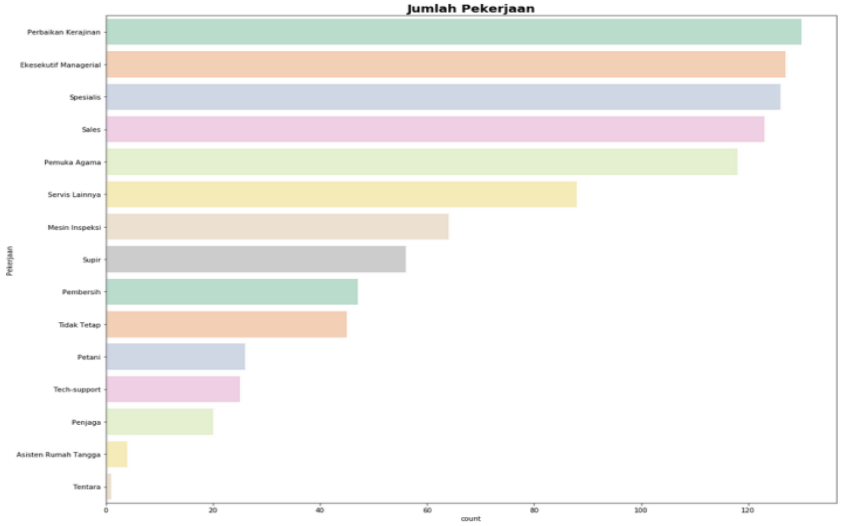
Gambar 3. Korelasi antar atribut

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 363−370
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2137

3.2 Hasil *Data Cleaning*

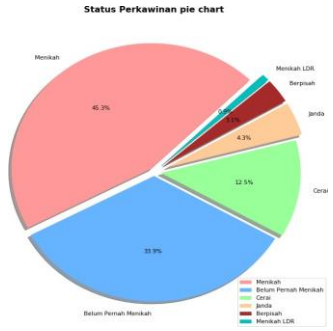
Dalam penelitian ini dilakukan pembersihan data agar data yang digunakan untuk dapat menghasilkan model terbaik. setelah *import* dataset berhasil, penulis melihat karakteristik data yang telah diambil dari sumber dataset. Dalam pengujian, Data akan dipangkas menjadi 1000 sample saja. Mengingat terlalu banyak nya sample yang ada di dalam dataset akan mempengaruhi hasil dan komputasi. Sehingga untuk hasil yang optimal data akan dipangkas. Kemudian dilakukan pengecekan missing value pada dataset. Dataset yang digunakan ini tidak memiliki nilai NULL (*missing value*).

Pada atribut/kolom pekerjaan dan kelas pekerja ditemui bahwa dataset tidak mengetahui jenis pekerjaan dan kelas dari sample. Kemudia penulis mengubah value yang berada di atribut Pekerjaan semula '?' menjadi Pekerjaan Tidak Tetap'. sama hal nya dengan atribut pada Kelas Pekerja, dilakukan perubahan pada value semula '?' menjadi Kelas Pekerjaan 'Tidak Tetap'. Gambar 4 adalah bentuk visualisasi data secara bar chart horizontal hasil dari cleaning data. Terlihat pekerjaan Tidak Tetap menduduki urutan ke 10 dari jumlah diantara pekerjaan lainya.Sebanyak 3 pengujian dilakukan untuk mencari nilai terbaik dari parameter jumlah layer konvolusi. Gambar 10 menunjukkan bahwa jumlah layer terbaik untuk melakukan klasifikasi terhadap ujaran kebencian pada teks didapatkan pada angka 3 layer. Nilai loss dari training dan testing terakhir pada epoch ke 200 dari layer berjumlah 3 memiliki nilai yang paling kecil dibandingkan dengan loss ketika menggunakan layer kovolusi sebanyak 5 dan 7. Sehingga dapat ditarik kesimpulan bahwa pada pengujian bagian ini parameter dengan jumlah layer konvolusi terbaik didapatkan pada layer konvolusi berjumlah 3.



Gambar 4. Visualiasi Atribut Pekerjaan

Hasil lainnya dari pemangkasan dataset dapat dilihat pada gambar 5 merupakan bentuk visualisasi data pada atribut status perkawinan yang disajikan dengan menggunakan pie chart. Diketahui bahwa bagian potongan yang paling besar pada status Menikah dan bagian potongan paling kecil berada di status Menikah LDR. *Pie chart* ini juga dilengkapi dengan legenda agar memudahkan dalam memahami informasi yang terkandung di dalamnya. Diperlukan juga normalisasi, nilai-nilai agar tidak terjadi inkonsistensi data. Pada tahap ini, dilakukan pengubahan nilai pada seluruh atribut agar dalam nilai skala yang akan digunakan sama. Tujuan normalisasi data agar tidak rentan terhadap pencilan [13].



Gambar 4. Visualiasi Atribut Status Perkawinan

3.3 Hasil *Preprocessing Data*

Tahapan sleanjutnya iyalah preprocessing data. Dimana data yang telah dibersihkan sebelumnya akan diolah lebih lanjut. Seperti yang dibahas pada tahapan sebelumnya. Dataset yang dimiliki terdapat data kategori dan juga data

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 363−370
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2137

numerik. Dalam pemodelan machine learning, data kategori tidak dapat dikenali oleh model sehingga data tersbut harus diubah terlebih dahulu. Atribut yang harus diubah yaitu atribut pendidikan, status perkawinan, kelas pekerja, pekerjaan, jenis kelamin, dan gaji. Dalam penelitian ini digunakan 2 teknik dalam mengubah data kategori yaitu dengan menggunakan function Label Encoder dan function replace. Dalam penelitian ini digunakan function replace sehingga urutan variabel yang akan diganti harus di tentukan dengan sendirinya. dibutuhkan dict variabel baru yang memuat nilai- nilai yang akan di replace nantinya. Contohnya pada atribut pendidikan akan dibuat variabel yang berisi key '1st-4th' hingga 'Doktor' dan juga values '1' hingga '16' secara berurut. Begitu juga pada atribut gaji. Dibuat variabel gaji dengan gaji dibawah 7jt akan di deskripsikan dengan nilai 0 kemudian gaji diatas 7 jt akan bernilai 1.Jika telah membuat dict variabel kemudian lakukan replace pada atribut pendidikan dengan *dict variabel* yang baru dibuat.Penggunaan function replace dilakukan pada data kategori yang bersifat ordinal data. Teknik lainnya dalam mengubah data kategori yaitu dengan menggunakan *function get_dummies*. Dalam dataset function ini digunakan pada atribut kelas pekerja, status perkawinan, jenis kelamin, dan pekerjaan. Function ini akan mengubah nilai yang awalnya berbentuk baris berubah menjadi kolom. Pada Gambar 6 merupakan hasil preprocessing data kategori. Kolom yang ada pada dataset telah bertambah yang awalnya 13 kolom menjadi 41 karena hasil *preprocessing* data kategori.

	id	Umur	Berat Akhir	Pendidikan	Jmlh Tahun Pendidikan	Keuntungan Kapital	Kerugian Capital	Jam per Minggu	Gaji	Kelas Pekerja_Pekerja Bebas Bukan Perusahan	...	Pekerjaan_Penjaga	Pekerjaan_Perbaikan Kerajinan	Pek
0	0	21	242912	9	9	0.0	0.0	35.0	0	0	...	0	0	
1	1	49	140782	6	6	0.0	0.0	40.0	0	0	...	0	0	
2	2	44	120057	12	12	61404000.0	0.0	45.0	1	0	...	0	0	
3	3	24	194630	14	13	0.0	0.0	35.0	0	0	...	0	0	
4	4	33	219619	15	14	210336000.0	0.0	40.0	1	0	...	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
35989	35989	47	148169	9	9	0.0	0.0	40.0	0	1	...	0	1	
35990	35990	69	264722	12	12	0.0	0.0	40.0	1	0	...	0	0	
35991	35991	24	31606	14	13	0.0	0.0	20.0	1	1	...	0	0	
35992	35992	47	197836	9	9	0.0	0.0	45.0	0	0	...	0	0	
35993	35993	45	243743	14	13	0.0	0.0	60.0	1	0	...	0	0	

41 columns

Gambar 9. Hasil Preprocessing Data

Pengujian terhadap ukuran kernel dari layer konvolusi terlihat memiliki perbedaan nilai loss di setiap iterasi yang tidak terlalu signifikan. Line chart pada test loss untuk 3 parameter pengujian menunjukkan garis yang saling berhimpitan satu sama lain. Namun pada epoch terakhir, parameter dengan kernel berukuran 2 berhasil memisahkan diri dan memiliki nilai loss yang paling kecil dibandingkan dengan 2 parameter lain yaitu 3 dan 5 kernel. Sehingga parameter terbaik dalam pengujian kali ini adalah parameter dengan ukuran kernel sebesar 2.

3.4 Hasil Pemodelan

Pembuatan model dalam penelitian ini menggunakan metode klasifikasi machine learning yaitu *Support Vector Machine* dan juga KNN (*K-Nearest Neighbors*). untuk melakukan klasifikasi pada SVM, penulis melakukan *drop* atribut yang tidak masuk kedalam kriteria dalam pembuatan model SVM. Ada tiga atribut yang di *drop* pada *independent variabel* diantaranya id, berat akhir, dan gaji. Kemudian membagi dataset menjadi *data training* dan juga *data test*. Dengan jumlah *data training* sebesar 80% sedangkan data test sebesar 20% dari jumlah dataset. Sehingga diperoleh hasil sebanyak 800 sample untuk *data training* dan 200 sample untuk *data test*. Setelah dilakukan *split data*, menentukan parameter terbaik untuk mendapatkan model yang optimal. Kemudian dibuat model dengan menggunakan beberapa kernel. Setelah model di buat kemudian dilakukan prediksi terhadap data testing yang sebelumnya telah disediakan. Untuk melihat performa dari model yang telah dibuat. Model dengan parameter mana yang paling optimal dalam melakukan prediksi.
Pada Tabel 2 diperoleh hasil dari beberapa percobaan dengan menggunakan kernel yang berbeda-beda. Hasil akurasi yang paling optimal dari ketiga percobaan tersebut yaitu dengan menggunakan kernel linear. Dengan score mencapai 82% dibandingkan dengan kernel rbf dan poly yang hanya mendapatkan nilai akuraasi sebesar 85% dan 80%.

Tabel 2. *Model Score*

Kernel	Score
poly	0.80
rbf	0.85
linear	0.87

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 363−370
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2137

Kemudian dilakukan perbandingan dalam penelitian ini dengan metode klasifikasi lainnya yaitu metode KNN. Model knn akan melakukan klasifikasi pada *dependent variabel* berdasarkan *independent variabel* yang sama. Pada model knn ini, telah ditentukan jumlah tetangga sebanyak 5 *(default)* dan menggunakan paramater *weights* yang bernilai *distance.* Model KNN mendapatkan *score* sebesar 82,5%. sedangkan untuk parameter *weights* diubah menjadi *uniform* diperoleh score sebesar 86%.

3.5 Hasil Evaluasi Model

Tahap akhir dari penelitian ini yaitu melakukan evaluasi terhadap model yang telah dikembangkan. Evaluasi bertujuan untuk mengukur performa dari model. Model dapat dikatakan optimal dengan melihat hasil evaluasi model yang akan disajikan. Dalam penelitian ini, digunakan teknik *confusion matrix* dalam melakukan evaluasi model. *Confusion matrix* akan memberikan informasi perbandingan hasil klasifikasi yang dihasilkan oleh model dengan klasifikasi yang sebenarnya. Pada Tabel 3 menunjukkan hasil dari evaluasi menggunakan *confusion matrix*.

Table 3. *Confusion Matrix*

	Actual Value	
Predicti	143	9
	17	31

Dalam evaluasi menggunakan *confusion matrix*, model dapat klasifikasi pada gaji diatas 7jt secara akurat sesuai dengan faktanya sebanyak 143 sedangkan klasifikasi untuk gaji di bawah 7jt sebanyak 31. klasifikasi ini biasanya dikenal dengan istilah *True Positive* dan *True Negative*. Model juga dapat melakukan klasifikasi pada kesalahan dalam melakukan klasifikasi. Dari hasil evaluasi yang telah dilakukan, diperoleh 17 termasuk kedalam *False Negative* dan 9 sample termasuk dalam *False Positive. False Negative* menunjukkan bahwa model mengklasifikasi Gaji dibawah 7jt akan tetapi dalam kenyataan nya termasuk dalam gaji diatas 7jt. Begitu pula sebaliknya pada False Positive, model memprediksi bahwa sample memiliki gaji diatas 7jt namun dalam kenyataan nya sample memiliki gaji dibawah 7t. sehingga tingkat akurasi dari model yang dikembangkan masih memiliki error. Dibuktikan dengan nilai False Negative dan False Positive yang dihasilkan.

4. KESIMPULAN

Hasil dari penelitian dengan menggunakan metode machine learning Support Vector Machine didapatkan akurasi sebesar 87%. hasil akurasi yang diperoleh berdasarkan dataset yang diperoleh dari kaggle dan telah melalui beberapa tahapan dalam pembuatan model. Penelitian ini juga telah melakukan pengujian dengan melakukan beberapa parameter untuk mendapatkan performa yang terbaik dalam pembuatan pemodelan. Mulai dari kernel yang digunakan secara variasi. Diperoleh kernel linear yang memiliki tingkat akurasi yang cukup tinggi dalam metode Support Vector Machine. Dalam penelitian ini, model juga dibandingakan dengan model lain yaitu metode KNN. Namun metode KNN hanya mampu memperoleh nilai akurasi sebesar 86%. dimana metode SVM lebih unggul 1% dibanding dengan metode KNN. Dengan hasil akurasi yang diperoleh dengan menggunakan metode SVM, model tersebut sudah cukup baik untuk melakukan klasifikasi untuk menentukan bahwa sesorang berpenghasilan diatas atau dibawah dari 7 juta rupiah.

REFERENCES

[1] Kemendikbud, “Kamus Besar Bahasa Indonesia,” 2020. [Online]. Available: https://kbbi.web.id/gaji.
[2] K. Karyawan and B. P. Transjakarta, “Pengaruh Gaji Dan Bonus Terhadap Presensi,” 2006.
[3] E. Suheny, R. R. Kusumawati, and I. Handayani, “Pengaruh Beban Gaji , Upah Dan Kesejahteraan,” *J. Revenue*, vol. 01, no. 02, pp. 171–181, 2020.
[4] L. Monorey, “Course: Introduction to Tensorflow for Artificial Intelligence,” 20 Nov, 2020. .
[5] P. Winder, Reinforcement Learning. .
[6] A. M. Muhammad, “Memprediksi Gaji dengan Polynomial Regression,” 27 Jul, 2019. [Online]. Available: https://medium.com/@adriantoto/memprediksi-gaji- dengan-polynomial-regression-5eb665063da7.
[7] A. Chiu, “Using Support Vector Machine to Classify Income Levels (> 50k or <= 50k)) Packaged SVM Classifier,” 27 Dec, 2018. [Online]. Available: https://annettechiu.medium.com/using-support-vector- machine-to-classify-income-levels-50k-or-50k-part-5-5- e592f67c0814.
[8] G. Aurelien, Hands-On Machine Learning with Scikit- Learn & TensorFlow : concepts, tools, and techniques to build intelligent systems. 2017.
[9] B. E. Boser, T. B. Laboratories, I. M. Guyon, T. B. Laboratories, and V. N. Vapnik, “SVM-A training algorithm for optimal margin classifiers.pdf.”
[10] M. Hachimi, G. Kaddoum, G. Gagnon, and P. Illy, “Multi- stage jamming attacks detection using deep learning combined with kernelized support vector machine in 5G cloud radio access networks,” 2020 Int. Symp. Networks, Comput. Commun. ISNCC 2020, no. October, pp. 20–22, 2020, doi: 10.1109/ISNCC49221.2020.9297290.
[11] O. Natan, A. I. Gunawan, and B. S. B. Dewantara, “Grid SVM: Aplikasi Machine Learning dalam Pengolahan Data Akuakultur,” J. Rekayasa Elektr., vol. 15, no. 1, 2019, doi: 10.17529/jre.v15i1.13298.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 363−370
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2137

[12] K. S. Nugroho, "Confusion Matrix untuk Evaluasi Model pada Supervised Learning," 13 Nov, 2019. [Online]. Available: https://medium.com/@ksnugroho/confusion- matrix-untuk-evaluasi-model-pada-unsupervised- machine-learning-bc4b1ae9ae3f.

[13] U. Wake, "Kesalahan Scaling Data di Machine Learning Menggunakan Scikit-Learn," 1 Sep, 2019. [Online]. Available:\https://medium.com/@uulwake/kesalahan- scaling-data-di-machine-learning-menggunakan-scikit- learn-7b88f2fbaec.