

Prediksi Capaian Bulanan Pajak Daerah Kabupaten Bandung Barat Menggunakan Metode Logistic Regression

Muhammad Ramdhani*, Fajri Rakhmat Umbara, Ridwan Ilyas

Fakultas Sains Dan Informatika, Informatika, Universitas Jenderal Achmad Yani, Cimahi

Jl. Terusan Jend. Sudirman, Cimahi, Jawa Barat, Kota Cimahi, Jawa Barat, Indonesia

Email: ^{1,*}mramdhani5432@gmail.com, ²fajri.rakhmat@lecture.unjani.ac.id, ³ilyas@lecture.unjani.ac.id

Email Penulis Korespondensi: mramdhani5432@gmail.com

Submitted: 12/06/2024; Accepted: 17/07/2024; Published: 20/07/2024

Abstrak– Pajak merupakan sumber pendapatan utama bagi negara, sehingga penerimaan pajak menjadi penyumbang terbesar bagi instansi pemerintah. Namun, Badan Pendapatan Daerah (BAPENDA) kadang mengalami kesulitan dalam memprediksi pendapatan bulanan daerah. Data yang dimiliki BAPENDA sangat penting untuk memperkirakan kenaikan pajak setiap bulannya, namun sering kali perkiraan tersebut meleset. Oleh karena itu, penelitian tentang prediksi CAPAIAN pajak bulanan sangat membantu. Peneliti mengunakan pendekatan metode data mining adalah teknik yang dapat membantu BAPENDA menemukan pola prediksi yang penting untuk membuat keputusan peningkatan pajak. sebumnya telah melakukan prediksi capaian metode Ann dimana untuk neuron 20 menghasilkan prediksi rmse sebesar 0.12. Dalam penelitian ini, menggunakan pendekatan metode algoritma logistic regression untuk memprediksi capaian pajak daerah Kabupaten Bandung Barat. Selain itu, dilakukan eksperimen untuk mengevaluasi variabel apa saja yang mempengaruhi nilai probabilitas.

Kata Kunci: Data Mining; Pajak; Klasifikasi; Logistic Regression; Ekonomi

Abstract–Tax is the main source of income for the state, so that tax revenue is the biggest contributor to government agencies. However, the Regional Revenue Agency (BAPENDA) sometimes has difficulties in predicting regional monthly income. The data owned by BAPENDA is very important for estimating tax increases every month, but often these estimates are wrong. Therefore, research on predicting monthly tax ACHIEVEMENT is very helpful. Researchers consider the data mining method approach is a technique that can help BAPENDA find predictive patterns that are important for making tax increase decisions. sebumnya has predicted the results of the Ann method where for neurons 20 it produces an rmse prediction of 0.12. In this study, the logistic regression algorithm approach was used to predict regional tax achievements in West Bandung Regency. In addition, experiments were carried out to evaluate which variables affect the probability value.

Keywords: Data Mining; Taxes; Classification; Logistic Regression; Economy

1. PENDAHULUAN

Pajak merupakan sumber pendapatan yang penting bagi pemerintah daerah, pada Kabupaten Bandung Barat. Penerimaan pajak yang tinggi dapat meningkatkan kinerja pemerintah daerah dalam menyediakan fasilitas dan pelayanan kepada masyarakat [1].

Di Indonesia, pajak dapat dibedakan menjadi dua jenis, yaitu pajak pusat dan pajak daerah. Pajak pusat adalah pajak yang dikelola oleh pemerintah pusat, seperti Pajak Penghasilan (PPh), Pajak Pertambahan Nilai (PPN), Pajak Penjualan atas Barang Mewah (PPnBM), dan Bea Materai. Sedangkan, pajak daerah dikelola oleh pemerintah daerah tingkat provinsi dan kabupaten/kota, terdiri dari pajak provinsi seperti Pajak Kendaraan Bermotor, Bea Balik Nama Kendaraan Bermotor, Pajak Bahan Bakar Kendaraan Bermotor, Pajak Air Permukaan, dan Pajak Rokok, serta pajak kabupaten/kota seperti Pajak Hotel, Pajak Restoran, Pajak Hiburan, Pajak Reklame, Pajak Penerangan Jalan, Pajak Mineral Bukan Logam dan Batuan, Pajak Parkir, Pajak Air Tanah, Pajak Sarang Burung Walet, Pajak Bumi dan Bangunan Perdesaan dan Perkotaan, serta Bea Perolehan Atas Tanah dan Bangunan[2].

Masalah yang dialami kabupaten bandung barat yaitu pengelolaan target pada hasil pajak Pendapatan Asli Daerah (PAD) yang belum maksimal. Pengelolaan target yang belum maksimal dikarenakan nilai pendapatan selalu berubah-ubah ,jika hasil penerimaan ini stabil maka akan mudah untuk diprediksi untuk bulan-bulan selanjutnya, hal ini akan berimbas pada lambatnya analis potensi pengembangan pajak, proses pembangunan ,dan realisasi suatu kesejahteraan suatu wilayah masyarakat pada kabupaten bandung barat. Degan demikian perlu dilakukan suatu upaya untuk mengestimasi atau memprediksi target pajak secara cepat dan benar .

Beberapa penelitian telah dilakukan untuk menetapkan atau memprediksi capaian pajak daerah, di antaranya menggunakan menggunakan perhitungan metode makro satu titik yaitu dengan menganggap bahwa realisasi penerimaan seperti melalui beberapa indikator meliputi Produk Domestik Regional Bruto (PDRB),pertumbuhan ekonomi, inflasi, pengangguran ,kemiskinan, dan tingkat suku bunga pada tahun terakhir adalah merupakan potensi retribusi untuk tahun selanjutnya [3] , lalu metode ANN [4] dan metode Backpropagation Neural Network [5]. Kedua penelitian tersebut berfokus pada hasil dari Root Mean Squared Error (RMSE) dan Mean Absolute Error (MAE). Dari hasil penelitian tersebut Root Mean Squared Error (RMSE) dari model ANN degan 20 neuron sebesar 0,12 dan Mean Absolute Error (MAE) degan 20 neuron mendapatkan hasil sebesar 0,09[4] .

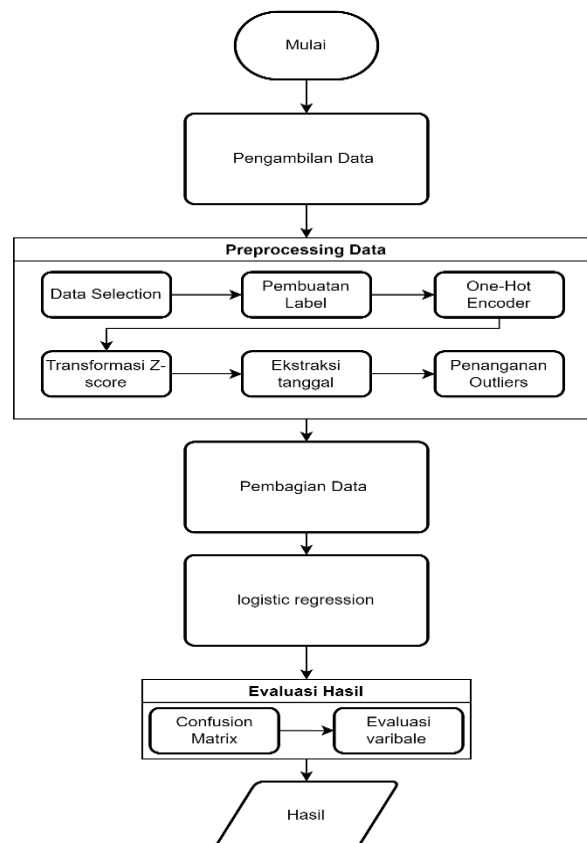
Penelitian akan dilakukan proses prediksi pencapaian bulanan pajak daerah Kabupaten Bandung Barat terdapat pada data (PAD) [5]. Untuk membuat prediksi yang akurat dan cepat, diperlukan suatu pendekatan yang

tepat guna mengurangi kesalahan saat meramalkan data, sehingga hasil yang dihasilkan dapat mendekati ketepatan yang diinginkan.

Metode prediksi yang digunakan yaitu metode prediksi logistic regression dipilih karena Model ini memiliki kemampuan interpretasi yang jelas terhadap pengaruh variabel atau interpretasi koefisien, bahkan pada situasi non-linear[6]. Serta memiliki kemampuannya kinerja yang baik[7] menjadikannya alat penting dalam memperoleh wawasan tentang faktor-faktor yang memengaruhi penerimaan pajak. Dengan Logistic Regression, hasil prediksi dapat memberikan informasi atau sumbangan pemikiran yang berharga bagi upaya perencanaan dan pengelolaan penerimaan pajak daerah.

2. METODOLOGI PENELITIAN

Pada subbab ini, penulis membahas metode desain yang digunakan dan langkah-langkah yang terlibat dalam penelitian ini. Di bawah ini adalah gambar prosedur penelitian berikut :



Gambar 1. Perancangan Umum Tahap Penelitian

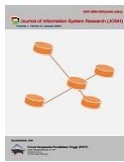
2.1 Pengambilan Data

Pada Tahap penelitian ini ,dilakukan proses pengambilan data yang berkaitan dengan penelitian ini. Pengambilan data ini merupakan tahapan awal agar dapat membentuk dataset Pendapatan Asli Daerah (PAD) kabupaten bandung barat . cara untuk mendapatkan data ini dilakukan melalui mekanisme wawancara kepada kepala bidang pajak.

Hasil dari wawancara tersebut menjadi patokan awal didapatkannya dataset pendapatan daerah. Setelah itu, rincian data tersebut didapatkan dalam bentuk excel, dilanjutkan dengan beberapa tahapan agar dataset dapat representasikan data yang cocok untuk di kelola oleh metode logistic regression.

Selain itu, dari hasil wawancara tersebut , peneliti mendapatkan informasi yang penting bagaimana memilih fitur atau kolom, dan juga bagaimana mengidentifikasi penghasilan pajak bernilai surplus atau defisit untuk membuat suatu pedoman parameter-parameter berbentuk label atau kolom data bersifat biner true (1) atau false (0) dan dapat di representasikan sebagai label klasifikasi tercapai atau tidak tercapai, karna label ini penting untuk pembelajar mesin klasifikasi logistic regression.

Dari hasil wawancara tersebut peneliti mendapatkan rekap data Pendapatan Asli Daerah (PAD) dari tahun 2015 sampai 2022 , seperti yang dijelaskan pada poin 2.2.1 yaitu representasi sampel data excel pada tahun 2015, lalu mendapatkan suatu konsep untuk membuat label untuk merepresentasikan sebagai tercapai atau tidak tercapai seperti pada poin 2.3.2 , dan tak hanya itu peneliti mendapatkan masukan dari narasumber bagi mana memilih kolom yang relevan seperti hasil pada poin 2.3.1 .



2.2.1 Representasi Data

Data yang telah di dapatkan peneliti maka selanjutnya melakukan representasi fakta dan menggali wawasan berharga, selanjutnya diolah dan dianalisis untuk mendapatkan pemahaman dan informasi yang lebih baik mengenai suatu fenomena atau masalah yang sedang diteliti, dari hasil wawancara data yang didapat penulis sebanyak 960 data dari tahun 2015 sampai 2022 .Dari data yang telah diambil dari hasil wawancara maka data simpel yang telah didapat sebagai berikut :

Tabel 1. Data Pendapatan Asli Daerah

No	Jenis Pajak	Target	Ketetapan Bulan Lalu	Ketetapan Bulan ini	Ketetapan s/d sekarang	%
1	HOTEL	7.750.000.000,00	-	746.078.199,00	746.078.199,00	9,63
2	RESTORAN	8.675.000.000,00	-	1.238.603.976,00	1.238.603.976,00	14,28
3	HIBURAN	1.500.000.000,00	-	461.456.370,00	461.456.370,00	30,76
4	REKLAME	1.400.000.000,00	-	110.384.593,00	110.384.593,00	7,88
5	PENERANGAN JALAN	36.000.000.000,00	-	3.973.117.823,00	3.973.117.823,00	11,04
6	PARKIR	1.000.000.000,00	-	176.569.452,00	176.569.452,00	17,66
7	AIR TANAH	12.000.000.000,00	-	312.165.052,00	312.165.052,00	2,60
8	MINERAL BUKAN LOGAM & BATUAN	4.000.000.000,00	-	281.331.192,00	281.331.192,00	7,03
9	PBB	59.500.000.000,00	-	515.999.966,00	515.999.966,00	0,87
10	BPHTB	63.000.000.000,00	-	3.132.901.835,00	3.132.901.835,00	4,97

Data ini mencakup informasi tentang tahun (dalam bentuk integer) dan bulan (dalam bentuk integer) ketika data tersebut dihasilkan. Selain itu, terdapat kolom yang mengidentifikasi jenis pajak (dalam bentuk karakter) dan target yang ditetapkan oleh kepala pajak setiap tahun (dalam bentuk kontinu). Data juga mencakup informasi mengenai ketetapan pajak yang dikumpulkan oleh masyarakat pada bulan-bulan sebelumnya (dalam bentuk diskrit), ketetapan pajak yang dikumpulkan saat ini (dalam bentuk kontinu), serta total ketetapan pajak yang dikumpulkan hingga saat ini (dalam bentuk diskrit). Selain itu, data juga mencakup kolom-kolom defisit dan surplus, yang mengindikasikan apakah pajak yang dikumpulkan oleh masyarakat melampaui atau kurang dari target yang ditetapkan (dalam bentuk diskrit).

2.3 Preprocessing Data

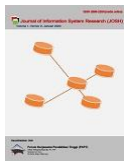
Pre-processing adalah serangkaian langkah atau teknik yang diterapkan pada data mentah sebelum diproses lebih lanjut atau digunakan dalam analisis data. Tujuan dari pre-processing adalah untuk mempersiapkan data agar lebih sesuai dan siap digunakan oleh algoritma pembelajaran mesin atau teknik analisis data lainnya [8].

2.3.1 Data Selection

Selection adalah untuk mengurangi kompleksitas dan volume data yang ada sehingga fokus hanya pada data yang penting dan memiliki nilai informasi yang tinggi yang berkaitan dengan judul [9].Data seleksi ini diperoleh melalui hasil wawancara dan masukan dari narasumber [10]. Dari data tersebut, kita dapat menghasilkan contoh pemilihan data seperti yang terlihat dalam tabel di bawah ini:

Tabel 2. Data Selection

Tanggal	Jenis Pajak	Target/bulan	Bulan ini
2015-01	HOTEL	645.833.333,3	746.078.199
2015-02	RESTORAN	722.916.666,7	1.238.603.976
2015-03	HIBURAN	125.000.000	461.456.370
2015-04	REKLAME	116666666,7	110.384.593,00
2015-05	PENERANGAN JALAN	3.000.000.000	3.973.117.823



Tanggal	Jenis Pajak	Target/bulan	Bulan ini
2015-06	PARKIR	83.333.333,33	176.569.452
2015-07	AIR TANAH	1.000.000.000	312.165.052
2015-08	MINERAL BUKAN LOGAM & BATUAN	333.333.333,3	281.331.192
2015-09	PBB	4958333333	515.999.966
2015-10	BPHTB	5.250.000.000	3.132.901.835

2.3.2 Pembuatan Label

Yaitu proses dimana pembuatan kolom independen atau label yang diambil dari data dependen alasan mengunakan metode ini ialah dari hasil wawancara dan rujukan kantor pajak, dari kasus ini harus membuat label menggunakan algoritma pembuatan label seperti berikut:

Tabel 3. Pembuatan Label

PROGRAM BuatLabel

{kita akan membuat kondisi jika kolom target/bulan kurang dari kolom bulan ini maka masukan Label 0}

DEKLARASI

target/bulan, bulan ini, Label: **integer**

ALGORITMA

read(target/bulan)

read(bulan ini)

if target/bulan < bulan ini **then**

Label <= 0

else

Label<=1

endif

Maka akan menghasilkan gambar kolom sebagai berikut:

Tabel 4. contoh hasil pembuatan label

Tanggal	Jenis pajak	Target/bulan	Bulan ini	Label
2015-01	MINERAL BUKAN LOGAM & BATUAN	333.333.333,3	259.349.485	0
2015-02	PBB	4.958.333.333	843.662.554	0
2015-03	BPHTB	5.250.000.000	5.173.924.920	0
2015-04	HOTEL	645.833.333,3	513.527.590	0
2015-05	RESTORAN	722.916.666,7	824.645.451	1
2015-06	HIBURAN	125.000.000	121.042.745	0
2015-07	REKLAME	116.666.666,7	243.797.135	1
2015-08	PENERANGAN JALAN	3.000.000.000	3.493.987.591	1
2015-09	PARKIR	83.333.333	40.761.536	0
2015-10	AIR TANAH	1.000.000.000	2239835660	1
2015-02	MINERAL BUKAN LOGAM & BATUAN	333.333.333,3	711.886.805	1

2.3.2 One-Hot Encoder

Sebuah teknik dalam pengolahan data yang digunakan untuk mengubah variabel kategorikal menjadi representasi biner yang dapat digunakan dalam pemodelan atau analisis data. Teknik ini digunakan ketika variabel kategorikal tidak memiliki relasi ordinal dan perlu diubah menjadi bentuk numerik agar dapat diproses oleh algoritma machine learning, Berikut adalah contoh sesudah di one hot encoder [9], [11], [12]:

Tabel 5. contoh sesudah di one hot encoder

Target Bulanan	Bulan ini	Label	AIR TANAH	BPHTB	HIBURAN	HOTEL	MINERAL BUKAN LOGAM & BATUAN	PARKIR	PBB	PENERANGAN JALAN	REKLAME	RESTORAN
645.833.333,3	746.078.199	1	0	0	0	1	0	0	0	0	0	0
722.916.666,7	1.238.603.976	1	0	0	0	0	0	0	0	0	0	1
125.000.000	461.456.370	1	0	0	1	0	0	0	0	0	0	0
116666666,7	110.384.593,00	0	0	0	0	0	0	0	0	0	1	0
3.000.000.000	3.973.117.823	1	0	0	0	0	0	0	0	1	0	0
83.333.333,33	176.569.452	1	0	0	0	0	0	1	0	0	0	0
1.000.000.000	312.165.052	0	1	0	0	0	0	0	0	0	0	0
333.333.333,3	281.331.192	0	0	0	0	0	1	0	0	0	0	0
4958333333	515.999.966	0	0	0	0	0	0	0	1	0	0	0
5.250.000.000	3.132.901.835	0	0	1	0	0	0	0	0	0	0	0

2.3.2 Transformasi Z Score

Sebuah metode transformasi data yang mengubah setiap nilai dalam distribusi menjadi nilai yang berada pada skala z-score setelah menggunakan metode skala z-score akan menghasilkan contoh data sebagai berikut [13]:

Tabel 6. Transformasi

Kolom Target Bulanan		Kolom Bulan ini	
Sebelum Transformasi	Sesudah Transformasi	Sebelum Transformasi	Sesudah Transformasi
645.833.333,3	0.882505	746.078.199	0.796697
722.916.666,7	0.874365	1.238.603.976	0.738592
125.000.000	0.932367	461.456.370	0.825210
116666666,7	0.933103	110.384.593,00	0.856744
3.000.000.000	0.509857	3.973.117.823	0.731913

2.3.3 Ekstraksi Tanggal

proses mengambil informasi waktu yang relevan dari suatu nilai datetime dan mengubahnya menjadi variabel terpisah [14] yang dapat digunakan dalam processing data lebih lanjut, dari ekstraksi tanggal menghasilkan contoh tabel sebagai berikut :

Tabel 7. Ekstraksi Tanggal

Ekstraksi tanggal		
Sebelum Transformasi	Sesudah Transformasi	
Tanggal	Tahun	Bulan
2015-01	2015	1
2015-05	2015	5
2015-06	2015	6

2.3.4 Penanganan Outliers

Penanganan outlier adalah proses untuk mengatasi atau memperlakukan nilai-nilai yang ekstrim atau tidak umum dalam suatu kumpulan data. Dalam konteks regresi logistik, outlier dapat memiliki dampak signifikan dan dapat mengacaukan model tersebut [15]. Oleh karena itu, penanganan outlier ini menjadi penting. Salah satu pendekatan yang digunakan adalah menggunakan Interquartile Range (IQR) untuk mengidentifikasi dan mengelola outlier dalam data. IQR adalah jarak antara nilai-nilai kuartil dalam data [16], [17]. penghapusan menguankan kondisi Jika suatu nilai berada dalam rentang antara $Q1 - \text{ambang} * IQR$ dan $Q1 + \text{ambang} * IQR$, nilai tersebut dianggap bukan outlier. Namun, jika tidak, titik tersebut akan dihapus [18].

2.3.4 Pembagian Data

Pembagian dataset yaitu pembagian data menjadi 2 secara acak diantaranya data latih dan data uji, data latih bertujuan untuk melatih model untuk mencoba menyesuaikan pola dalam dataset tersebut sedangkan data uji bertujuan untuk mengukur kinerja model [19].

2.3.4.1 Train/Test Split

Train/Test split digunakan untuk membagi dataset menjadi subset data pelatihan (training set) dan subset data pengujian (test set). Alasan utama membagi latih sebanyak 80% dan data uji sebanyak 20% adalah untuk membangun model yang memiliki kemampuan generalisasi yang baik.

Oleh karena itu, penting sehingga dapat memastikan bahwa model tidak hanya belajar dari pola data pelatihan tetapi juga dapat memprediksi dengan baik pada data baru [8]. Peneliti menguankan acuan latih sebanyak 80% dan data uji sebanyak 20% dari kesimpulan papir terdahulu [20], [21].

2.3.4 Logistic Regression

logistic regression, yang juga dikenal sebagai logistik regresi, merupakan metode statistik yang digunakan untuk memodelkan hubungan antara variabel dependen dengan satu atau lebih variabel independen, khususnya dalam kasus ketika variabel dependen memiliki dua kategori atau kelas yang bersifat saling eksklusif, seperti "Ya" atau "Tidak," "Positif" atau "Negatif" [22]. Metode ini dipilih karena memiliki kinerja yang baik [7] dan dapat digunakan bahkan dalam situasi data yang bersifat non-linear [6]. Rumus dasar dari regresi logistik biner adalah sebagai berikut:

$$\pi(X) = 1 / (1 + \exp(-z)) \quad (1)$$

Keterangan:

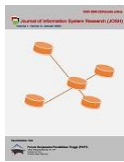
$\pi(X)$:probabilitas terjadinya kejadian (kelas positif)

X :vektor fitur (variabel independen)

z :kombinasi linear antara X dan koefisien β

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p \quad (2)$$

Keterangan :



- β_0 :intercept atau bias
 $\beta_1, \beta_2, \dots, \beta_p$:koefisien atau bobot yang menghubungkan setiap variabel independen dengan probabilitas kejadian
 $X_1, X_2, \dots, X_p$:nilai input prediksi

2.3.4 Evaluasi Hasil

Dalam tahap evaluasi ini, peneliti akan mengungkapkan temuan yang muncul dari analisis Confusion Matrix serta evaluasi koefisien variabel. Evaluasi ini menjadi sangat penting untuk memahami sejauh mana model regresi logistik telah berhasil dan seberapa baik variabel-variabel yang digunakan dalam model tersebut memengaruhi hasil prediksi.

2.3.4.1 Confusion Matrix

Evaluasi yang akan digunakan yaitu Confusion Matrix adalah salah satu cara untuk mengukur kinerja model klasifikasi [13]. Confusion matrix adalah tabel yang digunakan untuk menganalisis kinerja model berdasarkan jumlah data yang diklasifikasikan dengan benar dan salah. Confusion matrix memiliki empat elemen utama:

Tabel 8. Nilai Confusion Matrix

True Positive (TP): Jumlah data positif yang diklasifikasikan dengan benar	False Positive (FP): Jumlah data negatif yang salah diklasifikasikan sebagai positif,
False Negative (FN): Jumlah data positif yang salah diklasifikasikan sebagai negatif	True Negative (TN): Jumlah data negatif yang diklasifikasikan dengan benar,

Dari informasi yang diberikan pada tabel 2x2 penulis dapat menghitung seperti

1. Precision (Presisi): Persentase data positif yang diklasifikasikan dengan benar dari semua data yang diklasifikasikan sebagai positif. Rumus: $TP / (TP + FP)$ [13].
2. Accuracy (Akurasi): Ini mengukur sejauh mana model dapat memprediksi secara benar, baik positif maupun negatif. Rumus: $(TP + TN) / (TP + TN + FP + FN)$ [13].
3. Recall :seberapa baik model dapat mengidentifikasi semua data positif yang sebenarnya. Rumus: $TP / (TP + FN)$ [13].
4. F1-Score: Menggabungkan precision dan recall dalam satu metrik memberikan gambaran keseluruhan tentang performa model. Rumus: $2 * (Precision * Recall) / (Precision + Recall)$ [13].

2.3.4.2 Evaluasi Variabel

Evaluasi ini akan mencakup analisis koefisien variabel dalam model. Ini membantu dalam memahami sejauh mana setiap variabel independen berkontribusi terhadap hasil prediksi. Koefisien ini akan diinterpretasikan untuk mengevaluasi arah hubungan (positif atau negatif), besaran pengaruh, dan signifikansi statistiknya [23].

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

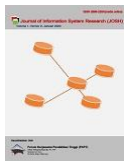
Pengumpulan data adalah tahap awal untuk membentuk data set Pendapatan Asli Daerah (PAD) kabupaten bandung barat, menguankan pendekatan wawancara kepada kepala bidang pajak, hasil dari wawancara tersebut, menjadi patokan awal didapatkannya dataset pendapatan daerah. Setelah itu, rincian data tersebut didapatkan dalam bentuk excel yang di rekap dari tahun 2015 sampai 2022 degan jumlah data sebanyak 960 data contoh data seperti di point 2.1 , dilanjutkan degan beberapa tahapan agar datasetset dapat representasikan data yang cocok untuk di kelola oleh metode logistik regression.

3.2 Preprocessing

Preprocessing adalah serangkaian langkah atau teknik yang diterapkan pada data mentah sebelum diproses lebih lanjut atau digunakan dalam analisis data. Menggunakan beberapa tahap seperti yang dijelaskan pada point 2.3 dan menghasilkan data sebagai berikut:

Tabel 9 Hasil Preprocessing

Bulan	Target Bulann	Bulan Ini	AIR TANA H	BP HT B	HIB URA N	HO TE L	MINERAL BUKAN LOGAM & BATUAN	PAR KIR	P B B	PENERANG AN JALAN	REK LAM E	REST ORA N	la be l
1	0,5215 96817	0,3059 51175	0	0	0	1	0	0	0	0	0	0	1



Bulan	Target Bulann	Bulan Ini	AIR TANA H	BP HT B	HIB URA N	HO TE L	MINERAL BUKAN LOGAM & BATUAN	PAR KIR	P B B	PENERANG AN JALAN	REK LAM E	REST ORA N	la bel
1	0,4724 25258	0,0773 47139	0	0	0	0	0	0	0	0	0	1	1
1	0,8538 37085	0,5274 52408	0	0	1	0	0	0	0	0	0	0	1
1	0,8591 52929	0,8006 66981	0	0	0	0	0	0	0	0	1	0	0
1	0,9801 29192	2,2054 27727	0	0	0	0	0	0	0	1	0	0	1
1	0,8804 16306	0,7491 59941	0	0	0	0	0	1	0	0	0	0	1
1	0,2956 73435	0,6436 35384	1	0	0	0	0	0	0	0	0	0	0
1	0,7209 40978	0,6676 31217	0	0	0	0	1	0	0	0	0	0	0
1	2,2293 52598	0,4850 04948	0	0	0	0	0	0	1	0	0	0	0
1	2,4154 07148	1,5515 46476	0	1	0	0	0	0	0	0	0	0	0
2	0,5215 96817	0,3021 05825	0	0	0	1	0	0	0	0	0	0	1
2	0,4724 25258	0,1365 20432	0	0	0	0	0	0	0	0	0	1	1
2	0,8538 37085	0,7681 08443	0	0	1	0	0	0	0	0	0	0	1
2	0,8591 52929	0,7479 81808	0	0	0	0	0	0	0	0	1	0	1
2	0,9801 29192	2,1886 51046	0	0	0	0	0	0	0	1	0	0	1
2	0,8804 16306	0,8419 16606	0	0	0	0	0	1	0	0	0	0	0
2	0,2956 73435	0,3093 21892	1	0	0	0	0	0	0	0	0	0	1
2	0,7209 40978	0,5133 34635	0	0	0	0	1	0	0	0	0	0	1
2	2,2293 52598	0,5288 92444	0	0	0	0	0	0	1	0	0	0	0
2	2,4154 07148	2,4798 71923	0	1	0	0	0	0	0	0	0	0	0
3	0,5215 96817	0,5400 68991	0	0	0	1	0	0	0	0	0	0	0
3	0,4724 25258	0,2197 63376	0	0	0	0	0	0	0	0	0	1	1
3	0,8538 37085	0,7846 04776	0	0	1	0	0	0	0	0	0	0	1
3	0,8591 52929	0,8576 30704	0	0	0	0	0	0	0	0	1	0	0
3	0,9801 29192	2,0124 00424	0	0	0	0	0	0	0	1	0	0	1
3	0,8804 16306	0,7554 85748	0	0	0	0	0	1	0	0	0	0	1
3	0,2956 73435	0,1811 50706	1	0	0	0	0	0	0	0	0	0	0
3	0,7209 40978	0,6847 3804	0	0	0	0	1	0	0	0	0	0	0
3	2,2293 52598	0,2300 08107	0	0	0	0	0	0	1	0	0	0	0
4	0,5215 96817	0,4869 29025	0	0	0	1	0	0	0	0	0	0	0
4	0,4724 25258	0,2448 07786	0	0	0	0	0	0	0	0	0	1	1
4	0,8538 37085	0,7923 72488	0	0	1	0	0	0	0	0	0	0	0
4	0,8591 52929	0,6968 41345	0	0	0	0	0	0	0	0	1	0	1

3.3 Pembagian Data

Pembagian dataset yaitu pembagian data menjadi 2 secara acak diantaranya data latih dan data uji, Peneliti ini mengunakan acuan latih sebanyak 80% atau 628 data dan data uji sebanyak 20% atau 157 data dari kesimpulan papar terdahulu [20], [21] hasil sebagai berikut:

```
X_train, X_test, y_train, y_test = train_test_split(x,y,test_size = 0.2,random_state=30,train_size=None,
shuffle=True,stratify=None)#

print(f'dimensi dari X_train : {len(X_train)}')
print(f'dimensi dari y_train : {len(y_train)}')
print(f'dimensi dari X_test : {len(X_test)}')
print(f'dimensi dari y_test : {len(y_test)}')

dimensi dari X_train : 628
dimensi dari y_train : 628
dimensi dari X_test : 157
dimensi dari y_test : 157
```

Gambar 2. Hasil Pembagian Data Train/Test split

3.4 Logistic Regression

penelitian ini bertujuan untuk memperlihatkan hasil implementasi teknik klasifikasi menggunakan data sebanyak 628 data latih degan metode logistic regression hasil sebagai berikut:

```
model_Lr.fit(X_train,y_train)

LogisticRegression
LogisticRegression(max_iter=1000, random_state=25)
```

Gambar 3. Implementasi Logistic Regression

3.5 Evaluasi Hasil

Pada evaluasi ini peneliti akan menampilkan hasil dari Confusion Matrix dan Evaluasi koefisien Variabel

3.5.1 Confusion Matrix

Pengujian Confusion Matrix ini peneliti mengunakan data testing sebanyak 157 data testing, berikut ini adalah hasil pengujian:

```
[[66 21]
 [52 18]]

True Positives(TP) = 66
True Negatives(TN) = 18
False Positives(FP) = 21
False Negatives(FN) = 52
```

Gambar 4. Hasil Implementasi Confusion Matrix

Pada hasil Confusion Matrix yang telah di implementasikan maka peneliti dapat menghitung Precision, Recall, Accuracy dan F1-Score berdasarkan poin 2.3.4.1 sebagai berikut

1. Precision = $66 / (66 + 21) = 0.7586$ (sekitar 75.86%)
2. Recall = $66 / (66 + 52) = 0.5593$ (sekitar 55.93%)
3. Accuracy = $(66 + 18) / (66 + 18 + 21 + 52) = 0.5733$ (sekitar 57.33%)
4. F1-Score = $2 * (0.7586 * 0.5593) / (0.7586 + 0.5593) = 0.6413$ (sekitar 64.13%)

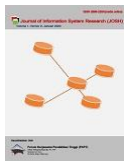
3.5.2 Evaluasi Variabel

ini penting karena memberikan pemahaman yang lebih dalam tentang bagaimana setiap variabel independen mempengaruhi variabel dependen (target)

```
print("Koefisien (B1 sampai B14):", model_Lr.coef_)
print("Intercept (a):", model_Lr.intercept_)

Koefisien (B1 sampai B14): [[-1.11126679e-04  4.22272447e-02 -1.49329299e+00  8.82361729e-01
 2.02039730e-02 -1.04531647e-01 -4.42436683e-01 -1.07363861e-01
 2.24044980e-01 -2.37578112e-01 -6.97636197e-01  7.66792285e-01
 2.84587194e-01  3.02571276e-01]]
Intercept (a): [0.00874311]
```

Gambar 5. Hasil koefisien



Berikut adalah hasil penjabaran dari hasil implementasi nilai koefisien:

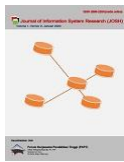
1. Kolom 'Tahun' nilai koefisien = -1.11126679e-04
2. Kolom 'Bulan' nilai koefisien = 4.22272447e-02
3. Kolom 'Target_Bulannan' nilai koefisien = -1.49329299e+00
4. Kolom 'Bulan_Ini' nilai koefisien = 8.82361729e-01
5. Kolom 'AIR TANAH' nilai koefisien = 2.02039730e-02
6. Kolom 'BPHTB' nilai koefisien = -1.04531647e-01
7. Kolom 'HIBURAN' nilai koefisien = -4.42436683e-01
8. Kolom 'HOTEL' nilai koefisien = -1.07363861e-01
9. Kolom 'MINERAL BUKAN LOGAM & BATUAN' nilai koefisien = 2.24044980e-01
10. Kolom 'PARKIR' nilai koefisien = -2.37578112e-01
11. Kolom 'PBB' nilai koefisien = -6.97636197e-01
12. Kolom 'PENERANGAN JALAN' nilai koefisien = 7.66792285e-01
13. Kolom 'REKLAME' nilai koefisien = 2.84587194e-01
14. Kolom 'RESTORAN' nilai koefisien = 3.02571276e-01

4. KESIMPULAN

Hasil penelitian ini menunjukkan bahwa penerapan metode logistic regression pada data Pendapatan Asli Daerah (PAD) dari tahun 2015 hingga 2022, dengan pemrosesan sebanyak 785 data, menghasilkan beberapa temuan utama. Variabel 'Bulan_Ini', 'PENERANGAN JALAN', 'REKLAME', dan 'RESTORAN' terbukti memiliki pengaruh positif yang signifikan terhadap variabel target, ditunjukkan oleh koefisien positif yang cukup besar pada variabel-variabel tersebut. Sebaliknya, variabel 'Target_Bulannan', 'BPHTB', 'HIBURAN', 'HOTEL', 'PARKIR', dan 'PBB' memiliki pengaruh negatif yang signifikan terhadap variabel target, dengan koefisien negatif yang cukup besar. Selain itu, hasil evaluasi Confusion Matrix menunjukkan bahwa model penelitian ini memiliki tingkat akurasi sekitar 57.33%, nilai presisi sekitar 75.86%, nilai recall sekitar 55.93%, dan F1-Score sekitar 64.13%. Ini mencerminkan keseimbangan antara presisi dan recall dalam model ini. Secara keseluruhan, penelitian ini menunjukkan bahwa model regresi logistik memiliki kemampuan yang baik dalam memprediksi variabel target berdasarkan pengaruh variabel independen yang telah diidentifikasi. Meskipun demikian, masih terdapat potensi untuk meningkatkan performa Confusion Matrix pada data Pendapatan Asli Daerah (PAD) dengan menggunakan metode classification algoritma lain selain logistic regression, seperti Decision Tree, SVM, Naive Bayes, dan lainnya.

REFERENCES

- [1] O. Arabella dan M. Yenni, "Pengaruh kualitas pelayanan petugas pajak, sanksi perpajakan dan biaya kepatuhan pajak terhadap kepatuhan wajib pajak UMKM," 2013.
- [2] B. E. Widari, "ANALISIS PENERIMAAN PAJAK BUMI DAN BANGUNAN TERHADAP PENDAPATAN DAERAH PEMERINTAH KOTA SURABAYA Sutjipto Ngumar Sekolah Tinggi Ilmu Ekonomi Indonesia (STIESIA) Surabaya," 2016.
- [3] F. Ayu, "Analisis Penetapan Target Penerimaan Pajak Daerah (Studi Pada Dispenda Kota Malang)," 2016.
- [4] C. Ivo, "Model Artificial Neural Network Berbasis Particle Swarm Optimization Untuk Prediksi Pendapatan Asli Daerah (PAD)," 2019.
- [5] E. Shaputra dan B. Serasi Ginting, "PREDIKSI PENDAPATAN ASLI DAERAH (PAD) KABUPATEN LANGKAT MENGGUNAKAN METODE BACKPROPAGATION NEURAL NETWORK," *Jurnal Teknik Informatika Kaputama (JTik)*, vol. 5, no. 1, 2021.
- [6] A. A. T. Fernandes, D. B. F. Filho, E. C. da Rocha, dan W. da Silva Nascimento, "Read this paper if you want to learn logistic regression," *Revista de Sociologia e Politica*, vol. 28, no. 74, hlm. 1/1-19/19, 2020, doi: 10.1590/1678-987320287406EN.
- [7] H. Rianto, S. Tinggi, M. Informatika, D. Komputer, dan N. Mandiri, "Resampling Logistic Regression untuk Penanganan Ketidakseimbangan Class pada Prediksi Cacat Software," *Journal of Software Engineering*, vol. 1, no. 1, 2015, [Daring]. Tersedia pada: <http://journal.ilmukomputer.org>
- [8] J. Paul Mueller dan L. Massaron, "Data Science Programming All-in-One For Dummies ® WARRANTIES OF FITNESS FOR A PARTICULAR PURPOSE. NO WARRANTY MAY BE CREATED OR EXTENDED BY SALES OR PROMOTIONAL MATERIALS. THE ADVICE AND STRATEGIES CONTAINED HEREIN MAY NOT BE SUITABLE," 2020. [Daring]. Tersedia pada: <http://www.wiley.com/go/permissions>.
- [9] paul,mueller john dan massaron luca, "Python ® for Data Science For Dummies ® , 2nd Edition," 2019. [Daring]. Tersedia pada: <http://www.wiley.com/go/permissions>.
- [10] S. Hasuna dkk., "IMPLEMENTASI ASSOCIATION RULE & FREQUENT PATTERN GROWTH UNTUK PENENTUAN SISTEM REKOMENDASI KEPUTUSAN PEMBELIAN," *Seminar Nasional Teknologi Informasi dan Komunikasi STI&K (SeNTIK)*, vol. 4, no. 1, 2020.
- [11] C. D. Larose dan D. T. Larose, *Data science using Python and R*. 2019.
- [12] F. Cady, *Data science : the executive summary : a technical book for non-technical professionals*. 2021.
- [13] Rohan. Chopra, Aaron. England, dan M. Noordeen. Alaudeen, *Data Science with Python : Combine Python with Machine Learning Principles to Discover Hidden Patterns in Raw Data*. Packt Publishing, Limited, 2019.



- [14] M. Bilal, G. Ali, M. W. Iqbal, M. Anwar, M. S. A. Malik, dan R. A. Kadir, “Auto-Prep: Efficient and Automated Data Preprocessing Pipeline,” *IEEE Access*, vol. 10, hlm. 107764–107784, 2022, doi: 10.1109/ACCESS.2022.3198662.
- [15] A. M. Bianco, G. Boente, dan G. Chebi, “Penalized robust estimators in logistic regression with applications to sparse models,” Nov 2019, [Daring]. Tersedia pada: <http://arxiv.org/abs/1911.00554>
- [16] R. Siringoringo, R. Perangin Angin, dan B. Rumahorbo, “MODEL KLASIFIKASI GENETIC-XGBOOST DENGAN T-DISTRIBUTED STOCHASTIC NEIGHBOR EMBEDDING PADA PERAMALAN PASAR,” 2022. [Daring]. Tersedia pada: <https://archive.ics.uci.edu/ml/datasets/online+retail>
- [17] S. Erpolat Tasabat dan O. Aydin, “Outlier detection method by using deep neural networks,” *Pressacademia*, vol. 5, no. 1, hlm. 96–101, Jun 2017, doi: 10.17261/pressacademia.2017.577.
- [18] H. Benallal *dkk.*, “A new approach for removing point cloud outliers using box plot A new approach for removing point cloud outliers using box plot 3-D watermarking View project Deceptive opinion spam detection View project A New Approach for Removing Point Cloud Outliers using Box plot,” 2023, doi: 10.1117/12.2618842i.
- [19] H. Liu dan M. Cocca, “Semi-random partitioning of data into training and test sets in granular computing context,” *Granular Computing*, vol. 2, no. 4, hlm. 357–386, Des 2017, doi: 10.1007/s41066-017-0049-2.
- [20] A. Gholamy, V. Kreinovich, dan O. Kosheleva, “A Pedagogical Explanation A Pedagogical Explanation Part of the Computer Sciences Commons,” 2018. [Daring]. Tersedia pada: https://scholarworks.utep.edu/cs_techrephttps://scholarworks.utep.edu/cs_techrep/1209
- [21] G. S. Luis, *Grokking Machine Learning*. 2021.
- [22] A. Alan dan finlay Barbara, “AGRSTI_Statistical methods for the social sciences,” 2008.
- [23] J. M. Hilbe, “Practical Guide to Logistic Regression,” 2015.