

Penerapan Teknik Information Mining dengan Algoritma Regresi Linear untuk Mengidentifikasi Potensi Biaya Pajak

Application of Information Mining Techniques Using Linear Regression Algorithm to Identify Tax Cost Potential

Khoiratul Azmi¹, Aninda Evioni², Muhammad Siddik Hasibuan³

^{1,2,3}Sains dan Teknologi, Ilmu Komputer, Universitas Islam Negeri Sumatera Utara, Indonesia

E-mail: ¹miiiazmiirpan@gmail.com, ²anindaevioni1@gmail.com,

³muhammadsiddik@uinsu.ac.id

Abstrak

Peningkatan pendapatan pajak merupakan salah satu fokus utama pemerintah dalam mendukung pembangunan ekonomi. Salah satu langkah strategis untuk memaksimalkan penerimaan pajak adalah dengan mengidentifikasi potensi biaya pajak yang selama ini belum tergali secara optimal. Penelitian ini menggunakan teknik information mining dengan pendekatan algoritma regresi linear untuk memprediksi dan mengelompokkan wajib pajak yang memiliki potensi biaya pajak tinggi. Data historis wajib pajak digunakan sebagai basis analisis untuk membangun model prediktif yang mampu memberikan klasifikasi kategori wajib pajak secara lebih akurat. Hasil analisis menunjukkan bahwa algoritma regresi linear efektif dalam mengidentifikasi pola data dan mengelompokkan wajib pajak berdasarkan tingkat potensi biaya pajaknya. Penelitian ini memberikan rekomendasi berbasis data kepada otoritas pajak untuk meningkatkan efisiensi pengumpulan pajak melalui pengelolaan wajib pajak secara lebih terfokus dan terarah. Dengan implementasi strategi berbasis data, diharapkan penerimaan pajak dapat meningkat secara signifikan dan mendukung pencapaian target pembangunan nasional.

Kata kunci: penerimaan pajak, potensi biaya pajak, information mining, algoritma regresi linear, klasifikasi wajib pajak

Abstract

Increasing tax revenue is one of the government's primary focuses to support economic development. A strategic approach to maximizing tax collection is by identifying untapped tax cost potential. This study employs information mining techniques using the linear regression algorithm to predict and classify taxpayers with high tax cost potential. Historical taxpayer data serves as the basis for analysis to build a predictive model capable of accurately categorizing taxpayers. The analysis results indicate that the linear regression algorithm effectively identifies data patterns and classifies taxpayers based on their tax cost potential. This study provides data-driven recommendations to tax authorities to enhance tax collection efficiency through more targeted and focused taxpayer management. By implementing a data-based strategy, it is expected that tax revenue can increase significantly, contributing to achieving national development goals.

Keywords: tax revenue, tax cost potential, information mining, linear regression algorithm, taxpayer classification

1. PENDAHULUAN

Pajak merupakan salah satu sumber utama pendapatan negara yang digunakan untuk mendukung pembangunan nasional. Namun, pengumpulan pajak sering kali mengalami kendala karena beberapa wajib pajak tidak membayar pajak sesuai dengan kemampuan mereka [1]. Oleh karena itu, diperlukan suatu metode yang dapat membantu mengidentifikasi potensi biaya pajak yang mungkin tersembunyi [2]. Penerapan information mining dengan teknik classification menawarkan solusi yang efisien dan berbasis information untuk melihat potensi biaya pajak dari masing-masing wajib pajak, sehingga pemerintah dapat mengambil tindakan yang lebih tepat dalam rangka memaksimalkan penerimaan pajak [3].

Dengan adanya prediksi yang akurat, diharapkan otoritas perpajakan dapat meningkatkan upaya penegakan pajak dan mendorong wajib pajak untuk memenuhi kewajiban mereka secara tepat waktu [4]. Penelitian ini memberikan kontribusi dalam penggunaan teknologi data mining dalam bidang perpajakan, serta menunjukkan potensi implementasinya dalam pengelolaan data besar untuk mendukung kebijakan publik yang lebih efektif [5].

RapidMiner sebagai perangkat lunak data mining yang user-friendly menyediakan lingkungan yang ideal untuk melakukan klasifikasi, visualisasi data, dan pembuatan model prediktif berdasarkan data yang ada [6]. Dalam penelitian ini, data wajib pajak kendaraan dianalisis untuk menemukan faktor-faktor yang mempengaruhi kepatuhan dan untuk memprediksi kelompok wajib pajak yang berpotensi menunggak atau patuh [7].

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan metode information mining berbasis algoritma regresi linear untuk mengidentifikasi potensi biaya pajak. Langkah-langkah penelitian meliputi:

1. Pengumpulan Data

Data historis wajib pajak diperoleh dari database otoritas pajak yang mencakup informasi seperti pendapatan, pembayaran pajak sebelumnya, dan kategori wajib pajak. Data ini akan digunakan sebagai dasar untuk membangun model prediktif.[6].

2. Pra-pemrosesan Data

Data yang diperoleh dibersihkan untuk menghilangkan informasi yang tidak relevan atau duplikat. Selain itu, proses normalisasi dilakukan untuk memastikan data berada dalam format yang seragam.[8].

3. Penerapan Algoritma Regresi Linear

Algoritma regresi linear diterapkan untuk memprediksi potensi biaya pajak berdasarkan variabel-variabel yang telah diidentifikasi, seperti pendapatan dan riwayat pembayaran pajak. Model ini digunakan untuk menemukan hubungan antara variabel independen (faktor-faktor wajib pajak) dan variabel dependen (biaya pajak potensial) [9].

4. Klasifikasi Wajib Pajak

Setelah model prediksi selesai, data wajib pajak dikelompokkan ke dalam beberapa kategori berdasarkan tingkat potensi biaya pajaknya, seperti tinggi, sedang, dan rendah.

5. Analisis Hasil

Hasil klasifikasi dianalisis untuk mengidentifikasi pola dan faktor-faktor utama yang memengaruhi potensi biaya pajak. Analisis ini dilakukan dengan menggunakan teknik deskriptif

dan evaluasi performa model regresi linear, termasuk mean squared error (MSE) dan koefisien determinasi (R^2) [10].

6. Validasi Model

Validasi dilakukan dengan membandingkan hasil prediksi model terhadap data aktual menggunakan metode cross-validation. Hal ini bertujuan untuk memastikan akurasi dan keandalan model dalam memprediksi potensi biaya pajak.

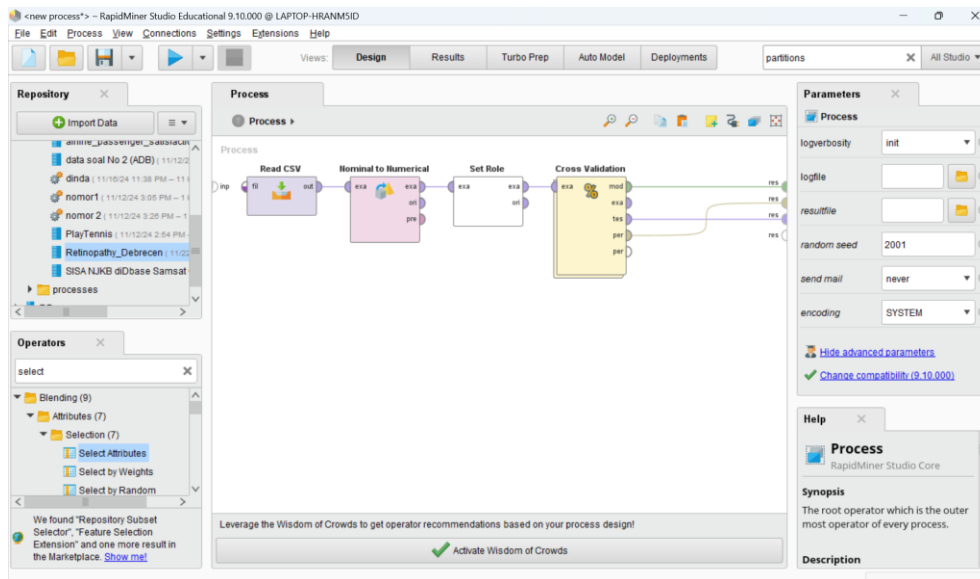
7. Penyusunan Rekomendasi

Berdasarkan hasil analisis, rekomendasi berbasis data disusun untuk membantu otoritas pajak meningkatkan efisiensi pengumpulan pajak dan mengoptimalkan pengelolaan wajib pajak dengan potensi tinggi.

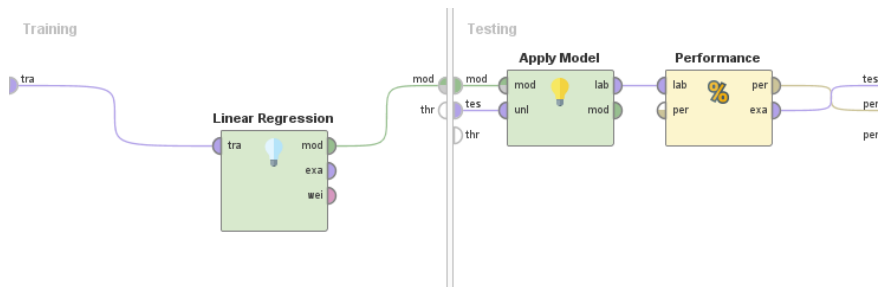
3. HASIL DAN PEMBAHASAN

1. Penerapan Algoritma Regresi Linear Operator yang digunakan sebagai berikut:

- a. Read CSV untuk membaca atau mengakses dataset.
- b. Nominal to Numerical untuk mengonversi data kategori (nominal) ke dalam bentuk angka (numerikal). Dalam kasus ini kolom country berupa teks dan tidak bisa langsung digunakan dalam perhitungan matematis, sehingga harus diubah menjadi numerik.
- c. Set Role untuk menentukan label. Dalam kasus ini labelnya adalah sales atau untuk memprediksi pajak.
- d. Cross Validation untuk mengevaluasi performa model dengan cara membagi data menjadi beberapa subset (folds), melatih model di sebagian data (training set), dan mengujinya di bagian data lainnya (test set). Dalam kasus ini, 10-fold cross-validation digunakan, yang artinya pada tiap iterasi, 9 bagian digunakan untuk training set dan sisa 1 bagian digunakan untuk menguji model. Alasan menggunakan operator ini adalah untuk mengukur overfitting.
- e. Linear Regression untuk melakukan prediksi berdasarkan judul dari penelitian untuk memodelkan hubungan antara satu atau lebih variabel prediktor (independen) dengan variabel target (dependen). Dalam hal ini, mula-mula regresi linear harus mencari persamaan dulu, kemudian melakukan feature selection dengan menggunakan M5 prime untuk memilih fitur yang paling relevan dan mengeliminasi fitur yang tidak signifikan. Feature selection dalam mengurangi resiko overfitting. Kemudian menggunakan metode kuadrat terkecil atau OLS untuk meminimalkan error.
- f. Apply model digunakan memprediksi nilai sales dengan model yang telah dibuat.
- g. Performance untuk mengevaluasi prediksi dengan rmse dan r2.
- h. Rmse: $43,931,567.030 \pm 10,020,047.048$
- i. R2: 0.926 ± 0.025
- j. Artinya model memiliki performa yang sangat baik ($R^2 = 0.926$), yang menunjukkan hubungan kuat antara fitur dan target, lalu r2 memiliki hasil yang bagus.



Gambar 3 Tampilan Desain Proses Data Mining di RapidMiner



Gambar 2 Diagram Alur Model Linear Regression dengan Training dan Testing

T. Perbandingan Algoritma A dan Algoritma B

Proses penerapan algoritma regresi linear dilakukan menggunakan RapidMiner untuk menganalisis data pajak kendaraan bermotor. Langkah pertama adalah mengimpor dataset dari file CSV yang berisi data pajak berdasarkan tahun, kuartal, dan wilayah (county). Dataset terdiri dari atribut year, quarter, county, dan sales, dengan sales sebagai variabel target (label). Data ini diimpor menggunakan operator Read CSV, dengan parameter yang memastikan format data sesuai untuk analisis lebih lanjut.

Selanjutnya, data kategorikal county diubah menjadi bentuk numerik menggunakan metode dummy coding melalui operator Nominal to Numerical. Setiap nilai unik dalam county diubah menjadi kolom dummy dengan nilai 0 atau 1, sehingga memungkinkan model regresi linear untuk memproses variabel ini. Setelah transformasi, atribut sales ditetapkan sebagai label menggunakan operator Set Role, sedangkan atribut lainnya (year, quarter, dan kolom dummy dari county) menjadi prediktor.

Proses validasi model dilakukan menggunakan teknik 10-fold Cross Validation. Data dibagi menjadi 10 bagian, di mana 9 bagian digunakan sebagai data pelatihan dan 1 bagian sebagai data pengujian secara bergantian. Dalam setiap iterasi, model regresi linear dilatih untuk memprediksi nilai pajak berdasarkan data pelatihan, dengan menggunakan parameter seperti

metode seleksi fitur M5 Prime dan regularisasi ridge ($1.0E-8$) untuk mengatasi potensi multikolinearitas.

Hasil prediksi diuji menggunakan data pengujian pada setiap fold, dan kinerja model dievaluasi menggunakan metrik seperti Root Mean Squared Error (RMSE) dan Squared Correlation (R^2). Proses ini memastikan bahwa model diuji secara menyeluruh pada berbagai subset data, sehingga menghasilkan evaluasi yang adil dan representatif terhadap kemampuan model untuk memprediksi pajak kendaraan bermotor.

2. Algoritma Linear Regresi Untuk Mengetahui Hasil Tingkat Akurasi Dari Pajak Kendaraan

a. ExampleSet Regresi Linear

Pada tab ExampleSet, hasil prediksi model regresi linear dibandingkan dengan nilai aktual sales. Kolom prediction(sales) menunjukkan hasil prediksi model untuk setiap baris data, sedangkan kolom sales adalah nilai aktual dari dataset. Hasil menunjukkan bahwa prediksi model mendekati nilai aktual pada banyak kasus, namun ada beberapa penyimpangan, termasuk prediksi negatif yang muncul di beberapa baris, seperti pada nilai minimum prediksi -43,783,209.802.

Row No.	sales	prediction(sales)	county = Ad...	county = Ara...	county = Bo...	county = De...	county = Do...	county = El...	county = Fre...	county = Gar...	county = ...
1	277611000	224775034.0...	0	0	0	0	0	0	0	0	1
2	241925000	232372743.6...	1	0	0	0	0	0	0	0	0
3	323709000	321579077.3...	0	0	0	0	0	1	0	0	0
4	8924000	-40583567.776	0	0	0	0	0	0	1	0	0
5	55206000	-2069258.367	0	0	0	0	0	0	0	1	0
6	538296000	584453054.9...	0	1	0	0	0	0	0	0	0
7	299526000	325440097.9...	0	0	0	0	0	1	0	0	0
8	70123000	27099201.889	0	0	0	0	0	0	0	0	0
9	162708000	240094784.7...	1	0	0	0	0	0	0	0	0
10	103053000	121919571.2...	0	0	0	0	0	0	0	0	0
11	162119000	171406495.1...	0	0	0	1	0	0	0	0	0
12	258899000	335398350.7...	0	0	0	0	0	1	0	0	0
13	100387000	129325045.0...	0	0	0	0	0	0	0	0	0
14	194231000	179128536.2...	0	0	0	1	0	0	0	0	0
15	124297000	124576748.1...	0	0	0	0	0	0	0	0	0

Gambar 3 Tampilan Dataset

b. Statistic Regresi Linear

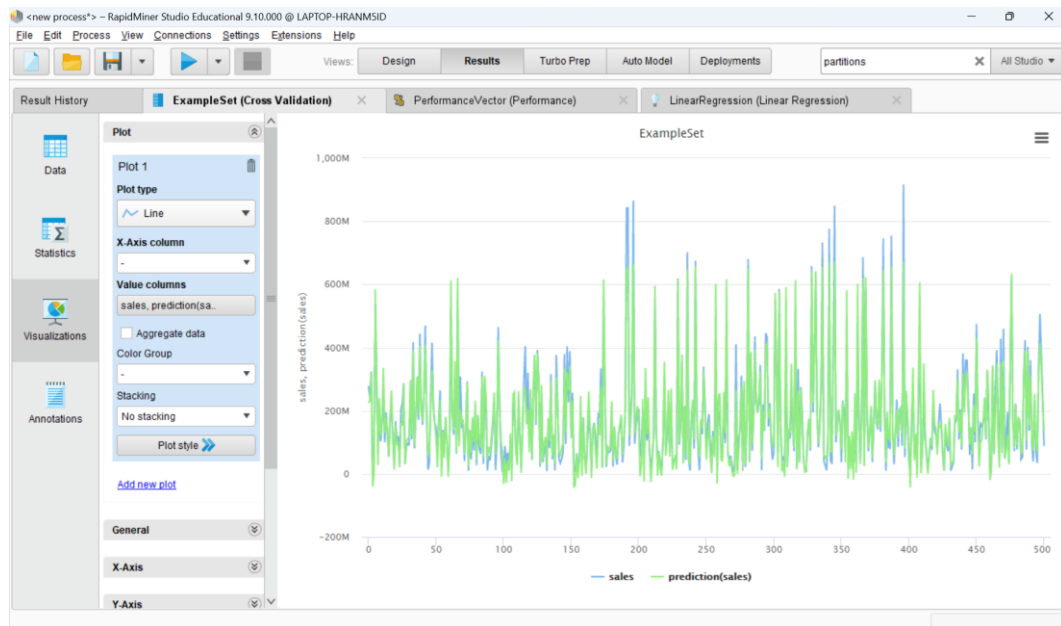
Tab statistik memberikan gambaran distribusi data untuk atribut sales dan prediction(sales). Rata-rata nilai sales adalah 17,605,483.034, sedangkan rata-rata prediksi model adalah 17,583,970.426, yang menunjukkan bahwa model menghasilkan prediksi yang cukup akurat secara global. Namun, nilai minimum prediksi negatif menunjukkan bahwa model tidak memberikan batasan pada hasil prediksi, yang menjadi salah satu kelemahan regresi linear dalam konteks ini.

Label	Integer	0	Min	Max	Average
sales			6274000	916910000	176058483.034
Prediction	Integer	0	Min	Max	Average
prediction(sales)			-43783209.802	673364751.028	175837970.426
county = Adams	Integer	0	Min	Max	Average
			0	1	0.064
county = Arapahoe	Integer	0	Min	Max	Average
			0	1	0.064
county = Boulder/Broomfield	Integer	0	Min	Max	Average
			0	1	0.008
county = Denver	Integer	0	Min	Max	Average
			0	1	0.064
county = Douglas	Integer	0	Min	Max	Average
			0	1	0.064
county = El Paso	Integer	0	Min	Max	Average
			0	1	0.064
county = Fremont	Integer	0	Min	Max	Average
			0	1	0.064
county = Garfield	Integer	0	Min	Max	Average
			0	1	0.064
county = Jefferson	Integer	0	Min	Max	Average
			0	1	0.064
county = La Plata	Integer	0	Min	Max	Average
			0	1	0.064
county = Larimer	Integer	0	Min	Max	Average
			0	1	0.064
county = Mesa	Integer	0	Min	Max	Average
			0	1	0.064
county = Pueblo	Integer	0	Min	Max	Average
			0	1	0.064
county = Weld	Integer	0	Min	Max	Average
			0	1	0.064
county = Boulder	Integer	0	Min	Max	Average
			0	1	0.056
county = Broomfield	Integer	0	Min	Max	Average
			0	1	0.056
county = Rest of State	Integer	0	Min	Max	Average
			0	1	0.050
year	Integer	0	Min	Max	Average
			2008	2015	2011.571
quarter	Integer	0	Min	Max	Average
			1	4	2.503

Gambar 4 Tampilan Data Tabel

c. Example Visualizations

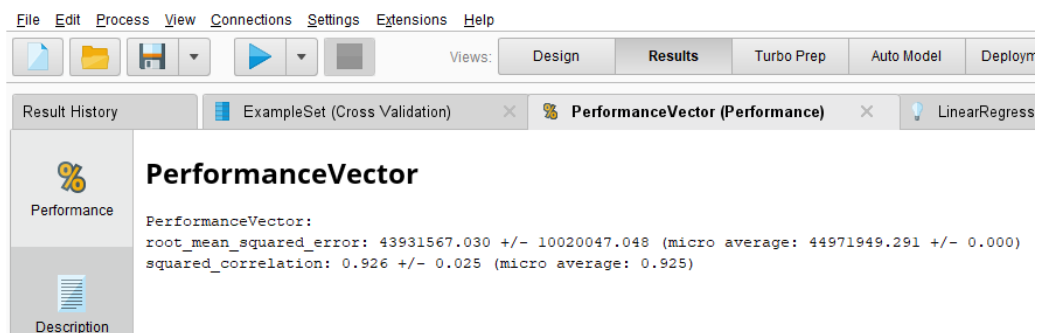
Grafik visualisasi menunjukkan perbandingan antara nilai aktual (sales) dan prediksi (prediction(sales)) untuk setiap baris data. Grafik ini menunjukkan pola bahwa prediksi model cukup konsisten dengan tren data aktual, meskipun ada beberapa deviasi terutama pada nilai-nilai ekstrem. Fluktuasi di grafik juga menunjukkan variasi pajak di antara wilayah dan kuartal yang berbeda.



Gambar 5 Grafik Perbandingan

d. Performance Vector

Tab PerformanceVector memberikan hasil evaluasi model menggunakan metrik Root Mean Squared Error (RMSE) dan Squared Correlation (R^2): RMSE: $43,931,567.030 \pm 10,020,047.048$, menunjukkan rata-rata kesalahan prediksi dalam satuan sales; R^2 : 0.926 ± 0.025 , menunjukkan bahwa model berhasil menjelaskan 92.6% variasi dalam data pajak. Evaluasi ini menunjukkan bahwa meskipun model cukup baik dalam menjelaskan hubungan data, ada ruang untuk perbaikan, terutama untuk menangani outlier.



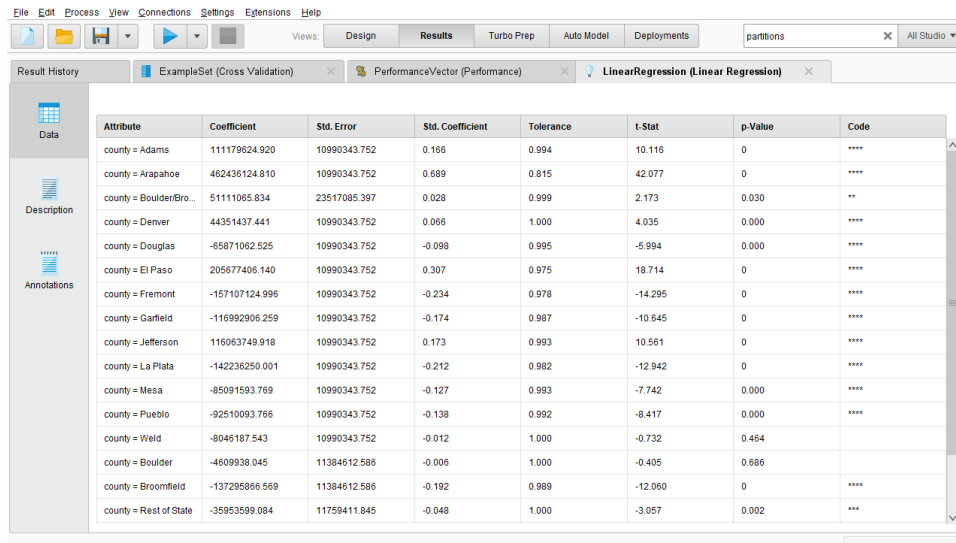
Gambar 6 Performance Vector - Hasil Evaluasi Model

e. Linear Regression

Pada tab Linear Regression, ditampilkan persamaan regresi yang dihasilkan. Koefisien untuk setiap atribut (county, year, quarter) menunjukkan kontribusi masing-masing variabel terhadap nilai prediksi. Contoh: county = Adams memiliki koefisien 11,117,924.920, menunjukkan bahwa wilayah ini memberikan kontribusi positif yang besar terhadap prediksi pajak; Atribut year memiliki koefisien 13,976,241.216, menunjukkan adanya peningkatan pajak seiring bertambahnya tahun; Atribut quarter memiliki koefisien -279,105,164.064, yang menunjukkan adanya tren musiman dalam data.

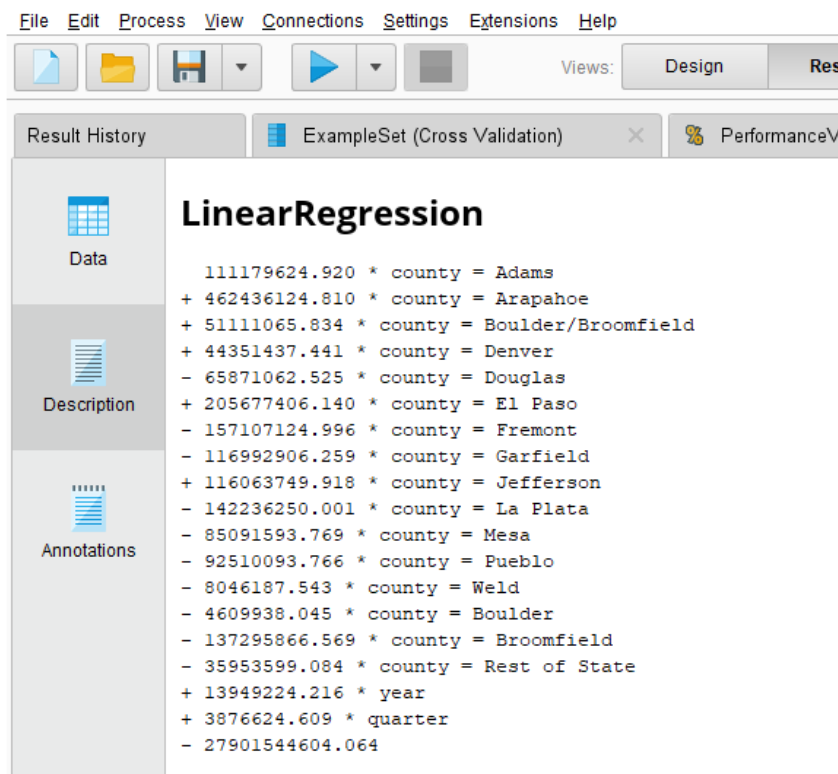
Koefisien ini membantu menjelaskan bagaimana setiap fitur memengaruhi hasil prediksi, yang dapat digunakan untuk interpretasi lebih lanjut terkait tren pajak kendaraan. Namun, p-value

pada beberapa atribut menunjukkan signifikansi statistiknya, sehingga beberapa atribut mungkin kurang relevan dalam model.



Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
county = Adams	111179624.920	10990343.752	0.166	0.994	10.116	0	****
county = Arapahoe	462436124.810	10990343.752	0.689	0.815	42.077	0	****
county = Boulder/Bro...	51111065.834	23517085.397	0.028	0.999	2.173	0.030	**
county = Denver	44351437.441	10990343.752	0.066	1.000	4.035	0.000	****
county = Douglas	-65871062.525	10990343.752	-0.098	0.995	-5.994	0.000	****
county = El Paso	205677406.140	10990343.752	0.307	0.975	18.714	0	****
county = Fremont	-157107124.996	10990343.752	-0.234	0.978	-14.295	0	****
county = Garfield	-116992906.259	10990343.752	-0.174	0.987	-10.645	0	****
county = Jefferson	116063749.918	10990343.752	0.173	0.993	10.561	0	****
county = La Plata	-142236250.001	10990343.752	-0.212	0.982	-12.942	0	****
county = Mesa	-85091593.769	10990343.752	-0.127	0.993	-7.742	0.000	****
county = Pueblo	-92510093.766	10990343.752	-0.138	0.992	-8.417	0.000	****
county = Weld	-8046187.543	10990343.752	-0.012	1.000	-0.732	0.464	
county = Boulder	-4609938.045	11384612.586	-0.006	1.000	-0.405	0.686	
county = Broomfield	-137295866.569	11384612.586	-0.192	0.989	-12.060	0	****
county = Rest of State	-35953599.084	11759411.845	-0.048	1.000	-3.057	0.002	***

Gambar 7 Tabel Koefisien Regresi Linear per County



LinearRegression

```

111179624.920 * county = Adams
+ 462436124.810 * county = Arapahoe
+ 51111065.834 * county = Boulder/Broomfield
+ 44351437.441 * county = Denver
- 65871062.525 * county = Douglas
+ 205677406.140 * county = El Paso
- 157107124.996 * county = Fremont
- 116992906.259 * county = Garfield
+ 116063749.918 * county = Jefferson
- 142236250.001 * county = La Plata
- 85091593.769 * county = Mesa
- 92510093.766 * county = Pueblo
- 8046187.543 * county = Weld
- 4609938.045 * county = Boulder
- 137295866.569 * county = Broomfield
- 35953599.084 * county = Rest of State
+ 13949224.216 * year
+ 3876624.609 * quarter
- 27901544604.064

```

Gambar 9 Hasil Koefisien Linear Regression

4. KESIMPULAN DAN SARAN

Dengan akurasi yang tinggi dan evaluasi yang komprehensif, penerapan data mining menggunakan klasifikasi untuk potensi biaya pajak memberikan hasil yang sangat baik. Model ini dapat digunakan sebagai alat bantu dalam pengambilan keputusan terkait pengelolaan pajak, baik untuk meningkatkan efisiensi pengumpulan pajak maupun untuk mendeteksi potensi wajib pajak yang belum memenuhi kewajiban. Kombinasi pendekatan analitis dan visualisasi yang mendalam memastikan bahwa hasil penelitian dapat diterapkan secara praktis di lapangan. Selain itu, disarankan agar implementasi model ini disesuaikan dengan kebutuhan spesifik pemerintah daerah untuk mengidentifikasi potensi pendapatan pajak yang belum teroptimalkan, serta melakukan pelatihan kepada tenaga ahli pajak guna memastikan pemanfaatan hasil analisis secara maksimal.

DAFTAR PUSTAKA

- [1] P. M. Aritonang, "Pajak penerangan jalan," *Indones. J. Bus. Law*, vol. 1, no. 1, pp. 25–31, 2021, doi: 10.47709/ijbl.v1i1.1296.
- [2] K. P. P. P. Tabanan, "Pengaruh Kesadaran Wajib Pajak, Kegiatan Sosialisasi Perpajakan, dan Pemeriksaan Pajak Terhadap Penerimaan Pajak Penghasilan Di KPP Pratama Tabanan," vol. 17, no. 1, 2022.
- [3] U. W. Karlina and M. H. Ethika, "Pengaruh Pengetahuan Wajib Pajak, Kesadaran Wajib Pajak, dan Sanksi Perpajakan terhadap Kepatuhan Wajib Pajak Kendaraan Bermotor," *J. Kaji. Akunt. dan Audit.*, vol. 15, no. 2, pp. 143–154, 2021, doi: 10.37301/jkaa.v15i2.30.
- [4] B. Setyawan and G. M. Sabrie, "Kajian Potensi Kendaraan Bermotor Menjadi Barang Kena Cukai," *J. Perspekt. Bea Dan Cukai*, vol. 6, no. 2, pp. 365–385, 2022, doi: 10.31092/jpbc.v6i2.1778.
- [5] T. D. W. BAJ and S. A. BAJ, "Pengaruh Program Pemutihan Pajak Kendaraan Bermotor, Kesadaran Wajib Pajak, Sosialisasi Pajak, Dan Pelayanan Terhadap Kepatuhan Wajib Pajak Kendaraan Bermotor (Studi Pada Kantor Bersama Samsat Surabaya Selatan)," *Behav. Account. J.*, vol. 3, no. 2, pp. 129–143, 2023, doi: 10.33005/baj.v3i2.103.
- [6] Francky William Leonardo Lona and Dwikora Harjo, "Analisis Potensi Pemungutan Pkb Dan Bbnkb Dalam Upaya Meningkatkan Penerimaan Pajak Daerah," *J. Ilmu Adm. Publik*, vol. 1, no. 4, pp. 333–343, 2021.
- [7] H. Amri and D. I. Syahfitri, "Pengaruh Pengetahuan Perpajakan, Sosialisasi Perpajakan, Kesadaran Pajak, dan Sanksi Pajak Terhadap Kepatuhan Wajib Pajak Dalam Membayar Pajak Kendaraan Bermotor di Kabupaten Sumbawa," *J. Accounting, Financ. Audit.*, vol. 2, no. 2, pp. 108–118, 2020, [Online]. Available: <http://jurnal.uts.ac.id/index.php/jafa/article/view/542>
- [8] F. Ristanti, Uswatun Khasanah, and Cris Kuntadi, "Literature Review Pengaruh Penerapan Pajak UMKM, Sosialisasi Perpajakan dan Sanksi Perpajakan terhadap Kepatuhan Wajib Pajak," *J. Ilmu Multidisplin*, vol. 1, no. 2, pp. 380–391, 2022, doi: 10.38035/jim.v1i2.49.
- [9] N. L. G. M. D. Ni Komang Ayu Juliantari, I Made Sudiartana, "Pengaruh kesadaran wajib pajak, kualitas pelayanan, kewajiban moral, sanksi pajak, dan sosialisasi pajak terhadap kepatuhan wajib pajak dalam membayar pajak kendaraan bermotor di kantor samsat gianyar," *J. Kharisma*, vol. 3, no. 1, pp. 128–139, 2021.
- [10] P. Isnaini and A. Karim, "Pengaruh Kesadaran Wajib Pajak dan Sanksi Perpajakan Terhadap Kepatuhan Wajib Pajak Kendaraan Bermotor (Studi Kasus pada kantor Samsat Kabupaten Gowa)," *J. Keuang. dan Perbank.*, vol. 3, no. 1, pp. 27–37, 2021, [Online]. Available: chrome-extension://efaidnbmninnibpcapjpcgclclefindmkaj/https://digilibadmin.unismuh.ac.id/upload/29858-Full_Text.pdf