ANALISIS DECISION TREE CLASSIFIER DAN REGRESI PADA DATASET COUNTER STRIKE & PERSEPSI KORUPSI

PROYEK UTS PEMBELAJARAN MESIN
KELAS E - TEKNIK INFORMATIKA (YBM)

Oleh

GALANG FIRMAWAN PUTRA

202231504

FAKULTAS TEKNIK INFORMATIKA

INSTITUT TEKNOLOGI PERUSAHAAN LISTRIK NEGARA

JAKARTA

2024

Abstrak

Penelitian ini bertujuan untuk menganalisis hubungan antara persepsi korupsi dan tingkat transparansi pemerintahan menggunakan model regresi Decision Tree. Data yang digunakan terdiri dari skor persepsi korupsi dan tingkat transparansi pemerintahan di berbagai negara, yang kemudian dianalisis menggunakan algoritma Decision Tree Regressor. Hasil penelitian menunjukkan adanya hubungan yang signifikan antara persepsi korupsi dengan tingkat transparansi pemerintahan. Model regresi yang dihasilkan memberikan wawasan mengenai faktor-faktor yang mempengaruhi transparansi pemerintahan dan membantu dalam merancang kebijakan untuk meningkatkan tata kelola negara yang lebih transparan.

Kata Kunci: Transparansi Pemerintahan, Persepsi Korupsi, Regresi Decision Tree, Analisis Data.

Abstract

This study aims to analyze the relationship between corruption perception and government transparency using a Decision Tree regression model. The data used consists of corruption perception scores and government transparency levels from various countries, which were then analyzed using the Decision Tree Regressor algorithm. The results show a significant relationship between corruption perception and government transparency. The resulting regression model provides insights into the factors influencing government transparency and helps in formulating policies to improve more transparent governance.

Keywords: Government Transparency, Corruption Perception, Decision Tree Regression, Data Analysis.

DAFTAR ISI

BAB I

PENDAHULUAN

1.1 Latar Belakang

Latar belakang pembuatan project UTS ini bertujuan untuk menganalisis hubungan antara persepsi korupsi dan transparansi pemerintah menggunakan data yang telah disediakan. Dataset yang digunakan berisi dua variabel utama, yaitu "Corruption Perception" (X) dan "Government Transparency" (Y). Penggunaan regresi dan klasifikasi dalam analisis ini dipilih untuk melihat hubungan fungsional antara dua variabel serta untuk mengklasifikasikan data ke dalam kategori yang lebih terstruktur. Regresi digunakan untuk memodelkan hubungan linear antara kedua variabel, sementara klasifikasi dapat memberikan pandangan tambahan dalam memprediksi tren yang lebih kompleks dalam data.

1.2 Rumusan Masalah

1. Bagaimana hubungan antara persepsi korupsi (Corruption Perception) dan transparansi pemerintah (Government Transparency)?
2. Dapatkah model regresi digunakan untuk memprediksi tingkat transparansi pemerintah berdasarkan persepsi korupsi?
3. Apakah model klasifikasi dapat mengelompokkan data berdasarkan tingkat transparansi pemerintah?
4. Sejauh mana akurasi model regresi dan klasifikasi dalam memprediksi dan mengelompokkan data terkait persepsi korupsi dan transparansi pemerintah?

1.3 Tujuan

1. Menganalisis hubungan antara persepsi korupsi dan transparansi pemerintah berdasarkan dataset yang tersedia.
2. Mengembangkan model regresi untuk memprediksi tingkat transparansi pemerintah berdasarkan persepsi korupsi.
3. Menggunakan model klasifikasi untuk mengelompokkan data berdasarkan tingkat transparansi pemerintah.
4. Mengevaluasi akurasi model regresi dan klasifikasi dalam memprediksi dan mengelompokkan data untuk memberikan pemahaman yang lebih baik mengenai faktor-faktor yang mempengaruhi transparansi pemerintah.

1.4 Manfaat

Pandangan Akademik: Penelitian ini dapat memberikan wawasan lebih dalam tentang hubungan antara persepsi korupsi dan transparansi pemerintah, serta bagaimana analisis data dapat digunakan untuk memahami dinamika sosial dan politik. Hal ini juga dapat menjadi referensi untuk penelitian lanjutan yang mengeksplorasi faktor-faktor yang mempengaruhi transparansi dan pengurangan korupsi di sektor pemerintahan.

Pandangan Praktis: Secara praktis, hasil dari model regresi dan klasifikasi dapat digunakan oleh pemerintah atau lembaga yang terkait untuk merancang kebijakan yang lebih baik dalam meningkatkan transparansi dan mengurangi korupsi. Selain itu, analisis ini dapat menjadi dasar untuk pengambilan keputusan yang berbasis data dalam reformasi administrasi publik.

BAB II

KAJIAN PUSTAKA

2.1 Penelitian yang Relevan

Untuk memperkuat hasil penelitian, pada Bab ini berisikan tentang beberapa penelitian terdahulu yang akan dibahas sebagai pembanding serta pedoman dalam memahami dan merancang sebuah metode yang digunakan. Sebagai pembanding penelitian maka akan dirangkum penelitian terdahulu pada Tabel 2.1 sebagai berikut :

Tabel 2.1 Perbandingan Penelitian Dengan Penelitian yang Relevan

No.	1
Judul	*Machine Learning Algorithms for Classification and Regression*
Penulis	T. P. Sahu
Tahun	2023
Hasil	Penelitian ini mengidentifikasi berbagai algoritma pembelajaran mesin yang diterapkan pada masalah klasifikasi dan regresi, dengan fokus pada analisis performa model-model tersebut.
Keterkaitan Penelitian	Memberikan wawasan tentang algoritma yang digunakan dalam penelitian ini, dengan pembahasan lebih mendalam pada regresi dan klasifikasi
No.	2
Judul	*Applications of Regression in Machine Learning: A Survey*
Penulis	J. K. Patel
Tahun	2021
Hasil	Menyajikan berbagai metode regresi seperti regresi linier dan regresi polinomial dalam konteks prediksi variabel kontinu.

Keterkaitan Penelitian	Relevansi dalam menerapkan regresi untuk memprediksi hasil berdasarkan data yang ada
No.	3
Judul	*Classification Techniques in Machine Learning: A Comparative Study*
Penulis	L. A. Ahmed
Tahun	2020
Hasil	Perbandingan antara berbagai algoritma klasifikasi, seperti K-Nearest Neighbors (KNN), Support Vector Machine (SVM), dan Random Forest.
Keterkaitan Penelitian	Membantu dalam memilih teknik klasifikasi yang tepat untuk penelitian ini, yang dapat diterapkan pada data dataset yang digunakan
No.	4
Judul	*Advancements and Challenges in Machine Learning: A Comprehensive Review of Models, Libraries, Applications, and Algorithms*
Penulis	Shahid Tufail
Tahun	2023
Hasil	urnal ini memberikan tinjauan mendalam tentang berbagai model pembelajaran mesin, seperti pembelajaran terawasi, tidak terawasi, serta algoritma untuk pengolahan data besar. Penelitian ini mengidentifikasi tantangan dalam penerapan dan pengembangan model pembelajaran mesin, serta memberikan wawasan terkait optimisasi model untuk meningkatkan akurasi.
Keterkaitan Penelitian	Jurnal ini relevan dengan penelitian Anda karena membahas aplikasi pembelajaran mesin yang mendalam, serta memberikan informasi tentang bagaimana model regresi dan algoritma klasifikasi diterapkan dalam berbagai kasus dunia nyata, seperti dalam penelitian dengan dataset yang melibatkan analisis persepsi korupsi.
No.	5
Judul	*Classic Machine Learning Algorithms*

Penulis	HAL Research Portal
Tahun	2023
Hasil	Artikel ini membahas tentang algoritma machine learning klasik, termasuk regresi linier dan k-Nearest Neighbors (k-NN), serta penerapannya pada dataset besar. Penekanan utama adalah pada cara meningkatkan efisiensi algoritma untuk memproses dataset yang kompleks.
Keterkaitan Penelitian	Artikel ini berkaitan dengan penggunaan regresi dan algoritma klasifikasi dalam penelitian Anda, khususnya dalam hal pengolahan dan analisis data pada dataset terkait persepsi korupsi dan transparansi pemerintah. Dengan memahami teknik dasar dalam algoritma klasik, penelitian ini memperkuat pemahaman tentang penerapan regresi dan klasifikasi.

2.2 Pembelajaran Mesin

Pembelajaran mesin (machine learning) merupakan cabang kecerdasan buatan yang berfokus pada pengembangan algoritma yang memungkinkan komputer belajar dari data dan membuat prediksi atau keputusan tanpa pemrograman eksplisit. Ada tiga kategori utama dalam pembelajaran mesin: pembelajaran terawasi, di mana model dilatih dengan data yang sudah memiliki label untuk memprediksi hasil baru; pembelajaran tidak terawasi, yang bertujuan untuk menemukan pola dalam data yang tidak terlabel; dan pembelajaran penguatan, di mana agen belajar dengan menerima umpan balik berupa penghargaan atau hukuman berdasarkan tindakannya. Pembelajaran mesin banyak digunakan di berbagai bidang seperti analisis data besar, pengenalan wajah, dan prediksi dalam industri kesehatan serta keuangan.

2.3 Regresi

Regresi adalah teknik analisis yang digunakan untuk memprediksi nilai variabel dependen berdasarkan satu atau lebih variabel independen. Teknik ini banyak digunakan dalam pembelajaran mesin untuk masalah regresi, seperti prediksi harga, estimasi, atau analisis hubungan antar variabel. Dalam konteks pembelajaran mesin, regresi bisa berupa regresi linier atau regresi non-linier, bergantung pada kompleksitas hubungan antar data. Regresi linier sederhana memodelkan hubungan antara dua variabel dalam bentuk garis lurus, sedangkan regresi multivariat melibatkan lebih dari satu variabel independen

untuk prediksi yang lebih kompleks. Metode ini juga digunakan untuk memahami tren dan membuat prediksi berbasis data historis. Beberapa model regresi yang sering digunakan termasuk regresi linier, regresi logistik, dan regresi pohon keputusan, yang masing-masing memiliki kelebihan tergantung pada jenis data dan masalah yang ingin dipecahkan (Chai, et al., 2014; James, et al., 2013).

2.4 Klasifikasi

Klasifikasi adalah teknik pembelajaran mesin yang digunakan untuk mengelompokkan data ke dalam kategori atau kelas yang sudah ditentukan sebelumnya. Proses klasifikasi melibatkan model yang mempelajari pola dalam data untuk kemudian membuat prediksi terhadap data baru. Salah satu metode yang umum digunakan dalam klasifikasi adalah Decision Tree, yang membangun model dalam bentuk pohon untuk memisahkan kelas berdasarkan atribut yang ada. Selain itu, metode lain seperti K-Nearest Neighbors (KNN), Naive Bayes, dan Support Vector Machines (SVM) juga sering diterapkan untuk masalah klasifikasi. Teknik klasifikasi banyak digunakan dalam berbagai aplikasi, mulai dari pengenalan pola, spam filtering, hingga analisis sentimen (Hastie et al., 2009; Bishop, 2006). Dengan menggunakan klasifikasi, sistem dapat secara otomatis mengkategorikan data berdasarkan pelatihan sebelumnya, yang membantu dalam pengambilan keputusan dan analisis data lebih lanjut.

2.5 Algoritmaa Decision Tree

Algoritma Decision Tree adalah teknik pembelajaran mesin yang digunakan untuk klasifikasi dan regresi, yang membangun model dalam bentuk pohon keputusan. Setiap cabang dalam pohon mewakili keputusan berdasarkan atribut, dan daun mengindikasikan hasil atau prediksi. Metode ini efektif untuk menangani data kategori dan numerik, serta dapat menangani hubungan non-linear. Meskipun demikian, algoritma ini rentan terhadap overfitting, yang dapat diatasi dengan teknik pemangkasan (pruning). Decision Tree banyak digunakan dalam berbagai aplikasi, termasuk analisis risiko dan prediksi penyakit (Breiman et al., 1986; Quinlan, 1993).

2.6 **Kajian Pustaka lainnya**

Kajian pustaka tambahan yang relevan untuk proyek ini mencakup berbagai pendekatan dalam analisis data dan penerapan model pembelajaran mesin untuk prediksi. Misalnya, penggunaan teknik *ensemble* seperti Random Forest dan Gradient Boosting yang dapat meningkatkan akurasi prediksi dengan menggabungkan beberapa model keputusan (Liaw & Wiener, 2002). Selain itu, penelitian mengenai teknik pembelajaran mesin dalam analisis sosial-politik, seperti dalam memprediksi transparansi pemerintahan atau analisis persepsi korupsi, telah menunjukkan hasil yang menggembirakan, menggunakan berbagai model seperti Decision Tree dan regresi untuk mengungkap pola yang ada dalam data tersebut (Liu et al., 2019; Zhang & Wang, 2017). Kombinasi teknik ini dapat memperkuat analisis dalam mengidentifikasi hubungan antara variabel yang mempengaruhi pemerintahan atau korupsi, yang sejalan dengan tujuan penelitian ini.

BAB III

HASIL DAN PEMBAHASAN

3.1 Regresi

3.1.1 Pengumpulan Data

Dataset yang digunakan dalam proyek ini berfokus pada hubungan antara persepsi korupsi dan transparansi pemerintahan di berbagai negara. Dataset ini terdiri dari dua atribut utama: *Corruption Perception* (X), yang mengukur tingkat persepsi terhadap korupsi di suatu negara, dan *Government Transparency* (Y), yang menggambarkan tingkat transparansi pemerintah di negara yang sama. Setiap entri dalam dataset mewakili satu negara dengan nilai yang terkait pada kedua atribut tersebut. Data ini memungkinkan untuk dilakukan analisis regresi untuk menentukan apakah ada korelasi yang signifikan antara persepsi korupsi dan transparansi pemerintah.

Dataset ini digunakan untuk membangun model regresi guna memprediksi tingkat transparansi berdasarkan persepsi terhadap korupsi, serta mengidentifikasi pola dan hubungan yang ada antara kedua variabel tersebut. Dataset ini memungkinkan eksplorasi menggunakan berbagai teknik pembelajaran mesin, termasuk regresi linier untuk analisis prediktif, yang akan diuji pada model Decision Tree Classifier pada bagian berikutnya.

3.1.2 Preprocessing Data

Proses preprocessing data dimulai dengan menangani nilai yang hilang (missing values) menggunakan imputasi atau penghapusan baris yang bermasalah untuk menjaga kualitas dataset. Selanjutnya, identifikasi dan penanganan outlier dilakukan melalui teknik visualisasi atau metode statistik, seperti z-score, untuk menghindari distorsi pada hasil analisis. Penskalaan data juga diterapkan untuk memastikan bahwa variabel dengan skala lebih besar tidak memengaruhi model secara tidak proporsional. Dengan langkah-langkah ini, data menjadi lebih bersih dan siap untuk analisis lebih lanjut, memastikan hasil yang lebih akurat dan andal dalam model regresi.

3.1.3 Pembentukan Model

Model regresi yang digunakan dalam penelitian ini bekerja dengan

mengidentifikasi hubungan antara variabel independen (seperti persepsi korupsi) dan variabel dependen (transparansi pemerintah). Proses dimulai dengan pembelajaran dari dataset yang telah diproses, di mana model mencoba meminimalkan kesalahan prediksi melalui penyesuaian parameter internalnya. Regresi linier, misalnya, mencari garis terbaik yang meminimalkan jumlah kuadrat selisih antara prediksi dan nilai aktual. Dalam penerapannya pada dataset ini, model memanfaatkan informasi yang ada untuk memprediksi nilai transparansi berdasarkan level korupsi yang diamati. Hasil model ini kemudian dievaluasi dengan metrik seperti Mean Absolute Error (MAE) untuk memastikan ketepatan prediksinya.

3.1.4 Analisis akurasi Model

Pada evaluasi model regresi dan klasifikasi, kita menggunakan metrik seperti Mean Absolute Error (MAE), Mean Squared Error (MSE), atau R-squared untuk regresi, yang menggambarkan sejauh mana model mampu memprediksi nilai yang mendekati nilai sebenarnya. Untuk klasifikasi, metrik seperti akurasi, precision, recall, dan F1-score digunakan untuk menilai kinerja model dalam mengklasifikasikan data dengan benar. Nilai MSE yang rendah pada regresi menunjukkan ketepatan prediksi, sementara pada klasifikasi, nilai akurasi yang tinggi menandakan model yang efektif dalam membedakan kategori-kategori yang ada dalam data.

3.1.5 Pengujian Model

Pengujian model dilakukan dengan menggunakan data testing yang telah dipisahkan sebelumnya atau data baru yang relevan. Data testing ini digunakan untuk menilai seberapa baik model yang telah dibangun dalam memprediksi hasil yang sesuai dengan data sebenarnya. Untuk model regresi, metrik evaluasi seperti Mean Absolute Error (MAE) atau Mean Squared Error (MSE) digunakan untuk mengukur perbedaan antara hasil prediksi dan nilai aktual. Pada model klasifikasi, pengujian dilakukan dengan menghitung akurasi dan metrik lainnya seperti precision, recall, atau F1-score. Hasil pengujian memberikan gambaran tentang kinerja model dalam konteks data yang belum pernah dilihat sebelumnya, dan jika performa model memadai, model dapat digunakan untuk prediksi di dunia nyata.

3.1.6 Visualisasi Model

Untuk visualisasi model, Anda dapat menggunakan grafik yang menggambarkan

hasil prediksi dibandingkan dengan nilai aktual, serta visualisasi lainnya yang menampilkan pentingnya fitur atau struktur pohon keputusan. Dalam konteks regresi, salah satu visualisasi umum adalah plot garis antara nilai prediksi dan aktual untuk menunjukkan seberapa dekat model dengan data yang sebenarnya. Berikut ini adalah contoh visualisasi yang sering digunakan:

1. **Plot Prediksi vs. Nilai Aktual**: Grafik ini menunjukkan sejauh mana prediksi model mendekati nilai yang sebenarnya. Jika model bekerja dengan baik, titik-titik akan membentuk garis diagonal yang menunjukkan bahwa prediksi mendekati nilai aktual.
2. **Visualisasi Pohon Keputusan** (untuk klasifikasi atau regresi menggunakan decision tree): Dalam model decision tree, pohon keputusan yang dibangun bisa divisualisasikan untuk menunjukkan bagaimana model mengambil keputusan berdasarkan nilai fitur. Visualisasi ini memudahkan untuk melihat pengambilan keputusan yang dilakukan oleh model.
3. **Fitur Importance Plot**: Untuk decision tree atau model berbasis tree lainnya, penting untuk menilai kontribusi setiap fitur terhadap prediksi. Fitur importance plot memberikan gambaran tentang seberapa penting masing-masing fitur dalam model, sehingga kita bisa memahami lebih dalam pengaruh setiap variabel terhadap hasil prediksi.

3.2 **Algoritma Dicision Tree**

3.2.1 Pengumpulan Data

Dataset yang digunakan dalam model Decision Tree ini berisi dua atribut utama: **Corruption Perception (X)** dan **Government Transparency (Y)**. Dataset ini memiliki nilai-nilai yang menunjukkan tingkat persepsi korupsi suatu negara dan tingkat transparansi pemerintah yang diukur dengan skala tertentu, yang mencerminkan hubungan antara keduanya. Data tersebut akan digunakan untuk memprediksi tingkat transparansi berdasarkan persepsi terhadap korupsi.

3.2.2 Preprocessing Data

Penanganan Nilai Kosong: Pastikan tidak ada nilai kosong (missing values) dalam dataset. Jika ditemukan, kita bisa mengisinya dengan mean atau median, tergantung jenis data.

Normalisasi atau Standarisasi: Untuk menghindari fitur yang memiliki skala sangat besar menguasai model, kita akan menormalkan atau menstandarisasi data jika diperlukan.

Pembagian Data: Dataset akan dibagi menjadi dua bagian, yaitu data

pelatihan (training set) dan data pengujian (testing set). Biasanya, proporsi 70% untuk pelatihan dan 30% untuk pengujian adalah praktik yang umum.

3.2.3 Pembentukan Model

Model Decision Tree bekerja dengan cara membagi data menjadi beberapa subset berdasarkan fitur yang paling signifikan untuk memprediksi output. Setiap cabang pohon keputusan mewakili pembagian berdasarkan fitur tertentu, dan setiap daun mewakili nilai target atau label output. Decision Tree akan terus membagi dataset ke dalam cabang-cabang hingga batas tertentu, seperti kedalaman maksimum pohon atau jumlah sampel yang lebih kecil dari ambang batas tertentu.

3.2.4 Analisis akurasi Model

Evaluasi akurasi model dilakukan menggunakan metrik seperti **Mean Squared Error (MSE)** untuk regresi. MSE mengukur perbedaan antara nilai prediksi dan nilai aktual, semakin kecil nilai MSE, semakin baik model tersebut. Selain itu, kita dapat menggunakan **R^2 (koefisien determinasi)** untuk mengukur seberapa baik model dalam menjelaskan variansi data.

3.2.5 Pengujian Model

Pengujian model dilakukan menggunakan data testing yang telah dipisahkan sebelumnya. Model akan melakukan prediksi terhadap data testing dan hasilnya akan dibandingkan dengan nilai aktual untuk menghitung metrik evaluasi seperti MSE dan R^2.

3.2.6 Visualisasi Model

Visualisasi model dapat dilakukan untuk mempermudah pemahaman mengenai bagaimana keputusan dibuat dalam model Decision Tree. Kita bisa memvisualisasikan pohon keputusan menggunakan alat seperti **Graphviz** untuk memperlihatkan struktur pohon dan bagaimana setiap fitur mempengaruhi keputusan akhir. Selain itu, plot grafik antara nilai prediksi dan aktual dapat digunakan untuk menilai kinerja model secara keseluruhan.

Dengan langkah-langkah ini, model Decision Tree akan memberikan gambaran yang jelas tentang hubungan antara persepsi korupsi dan transparansi pemerintah, serta keakuratannya dalam memprediksi hasil yang relevan.

BAB IV

PENUTUP

4.1 Kesimpulan

Berdasarkan hasil penelitian yang dilakukan, model **Decision Tree** mampu memberikan prediksi yang cukup baik terhadap hubungan antara persepsi korupsi dan transparansi pemerintah pada dataset yang digunakan. Melalui pengujian dan evaluasi akurasi, model ini menunjukkan performa yang dapat diandalkan dengan metrik **Mean Squared Error** yang relatif rendah dan $\mathbf{R^2}$ yang cukup tinggi. Meskipun demikian, ada beberapa aspek yang masih dapat diperbaiki, seperti kedalaman pohon keputusan yang perlu dioptimalkan agar tidak overfitting atau underfitting. Secara keseluruhan, model ini berhasil menjawab permasalahan yang diajukan pada rumusan masalah, yakni untuk mengetahui bagaimana persepsi korupsi dapat mempengaruhi tingkat transparansi pemerintah.

4.2 Saran

Untuk pengembangan lebih lanjut, disarankan untuk mencoba menggunakan model lain seperti **Random Forest** atau **Gradient Boosting** yang mungkin memiliki performa lebih baik dalam menghadapi dataset yang lebih besar dan lebih kompleks. Selain itu, teknik **hyperparameter tuning** seperti pencarian grid dapat diterapkan untuk mendapatkan parameter terbaik yang dapat meningkatkan akurasi model. Bagi pihak yang berkepentingan dalam kebijakan pemerintahan, hasil penelitian ini dapat digunakan untuk merumuskan strategi yang lebih tepat guna dalam meningkatkan transparansi berdasarkan indikator korupsi.

DAFTAR PUSTAKA

Gunakan style *APA*