**LAPORAN TUGAS**

**SELEKSI DAN EKSTRAKSI FITUR**

Disusun untuk memenuhi tugas Mata Kuliah Proyek Sains Data

Dosen Pengampu: EKA MALA SARI ROCHMAN, S.Kom.M.Kom



**DISUSUN OLEH:**

SYAYID AL AZIZ 210411100008

**PRODI TEKNIK INFORMATIKA**

**JURUSAN TEKNIK INFORMATIKA**

**FAKULTAS TEKNIK**

**UNIVERSITAS TRUNOJOYO MADURA**

**2023**

1. Perbedaan Seleksi Dengan Ekstraksi Fitur

Feature extraction dan feature selection adalah strategi penting dalam pengolahan data untuk mengurangi kompleksitas dan memfasilitasi analisis yang lebih efektif. Keduanya bertujuan untuk mengurangi jumlah variabel, namun dengan pendekatan yang berbeda. Dalam feature extraction, variabel baru dibentuk melalui kombinasi variabel yang ada, sehingga menghasilkan representasi baru yang lebih kompak namun informatif. Di sisi lain, feature selection memilih variabel dari yang sudah ada tanpa menciptakan variabel baru, memfokuskan pada subset variabel yang dianggap paling relevan atau penting untuk analisis. Baik feature extraction maupun feature selection memiliki peran vital dalam membentuk model yang lebih efisien dan memberikan wawasan yang lebih mendalam dalam pemahaman data. Pemilihan metode tergantung pada konteks dan tujuan analisis yang diinginkan.

1. Metode Seleksi Fitur
2. information gain

Information gain adalah suatu metrik yang digunakan dalam proses feature selection, terutama dalam algoritma pohon keputusan seperti ID3, C4.5, atau Random Forest. Tujuannya adalah untuk menilai sejauh mana suatu fitur dapat memberikan informasi yang relevan dalam memprediksi kelas atau label pada data. Proses penggunaan information gain meliputi perhitungan entropi awal dari data sebelum pemilihan fitur. Selanjutnya, data dipisahkan berdasarkan nilai-fitur tertentu, entropi dari setiap kelompok hasil pemisahan dihitung, dan entropi akhir setelah pemisahan dihitung dengan menggabungkan entropi dari setiap kelompok, yang diponderasi berdasarkan ukuran kelompok tersebut. Information gain adalah perbedaan antara entropi awal dan entropi akhir setelah pemisahan. Semakin tinggi nilai information gain, semakin penting fitur tersebut dalam membedakan kelas atau label pada data. Oleh karena itu, fitur dengan information gain tertinggi dianggap sebagai fitur yang paling informatif dan kemungkinan besar digunakan dalam pembuatan model atau analisis lebih lanjut. Information gain membantu memprioritaskan fitur-fitur yang paling relevan dalam memisahkan data ke dalam kategori yang lebih homogen atau terpisah, sehingga mendukung pembuatan model prediksi yang lebih efisien.

1. chi-square test

Chi-square adalah sebuah metode statistik yang digunakan untuk mengukur hubungan antara dua variabel kategorikal dalam suatu dataset. Metode ini berfokus pada analisis data yang telah diklasifikasikan ke dalam kategori-kategori tertentu. Chi-square menghasilkan nilai statistik yang mengindikasikan sejauh mana perbedaan antara jumlah yang diobservasi dalam setiap kategori dengan jumlah yang diharapkan jika tidak ada hubungan antara variabel-variabel tersebut. Nilai chi-square ini digunakan untuk menentukan apakah ada hubungan yang signifikan antara variabel-variabel tersebut atau tidak. Semakin tinggi nilai chi-square, semakin signifikan hubungan antara variabel-variabel tersebut. Chi-square sering digunakan dalam analisis data kategorikal seperti dalam penelitian perilaku, ilmu sosial, kesehatan, dan bidang-bidang lain di mana data terkait dengan kategori-kategori atau frekuensi observasi.

1. fisher score

Fisher score, juga dikenal sebagai Fisher's discriminant ratio, adalah sebuah metode yang digunakan dalam analisis statistik untuk menilai pentingnya suatu fitur atau variabel dalam membedakan dua kelompok atau kategori yang berbeda dalam data. Metode ini mencari perbedaan yang signifikan antara mean dari setiap kelompok dan variabilitas dalam setiap kelompok. Fisher score dihitung dengan membandingkan varians antar-kelompok (between-group variance) dengan varians dalam-kelompok (within-group variance). Tujuannya adalah untuk memaksimalkan varians antar-kelompok sambil meminimalkan varians dalam-kelompok. Nilai Fisher score yang tinggi menunjukkan bahwa fitur tersebut memiliki kemampuan diskriminatif yang kuat antara kelompok-kelompok yang diamati. Fisher score banyak digunakan dalam berbagai bidang, termasuk analisis data, pengenalan pola, dan pembelajaran mesin untuk memilih fitur-fitur yang paling informatif dan relevan dalam menganalisis atau memprediksi suatu fenomena.

1. correlation coefficient

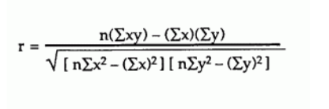
Koefisien korelasi adalah ukuran statistik yang mengukur kekuatan dan arah hubungan antara dua variabel. Rentang nilai koefisien korelasi adalah -1 hingga 1, yang menunjukkan sifat dan kekuatan hubungan. Koefisien korelasi positif (antara 0 dan 1) menunjukkan hubungan langsung atau positif: ketika satu variabel meningkat, variabel lainnya cenderung meningkat juga. Sebaliknya, koefisien korelasi negatif (antara -1 dan 0) menunjukkan hubungan terbalik atau negatif: ketika satu variabel meningkat, variabel lainnya cenderung menurun. Koefisien korelasi 0 menunjukkan tidak adanya hubungan linier antara variabel. Namun, penting untuk diingat bahwa korelasi 0 tidak selalu berarti tidak ada hubungan; bisa jadi terdapat hubungan non-linier. Besarnya koefisien korelasi menunjukkan kekuatan hubungan, dengan nilai 1 atau -1 menunjukkan hubungan linier sempurna dan nilai 0 menunjukkan tidak ada hubungan linier. Semakin dekat koefisien korelasi dengan 1 atau -1, semakin kuat hubungan antar variabel. Korelasi adalah alat yang banyak digunakan di berbagai bidang, termasuk statistik, ekonomi, psikologi, dan bidang lainnya, untuk menganalisis hubungan dan membuat prediksi berdasarkan data yang diamati.

1. variance threshold

Variance threshold merupakan salah satu teknik preprocessing dalam analisis data yang digunakan untuk seleksi fitur atau variabel yang memiliki variasi yang cukup besar. Varians adalah ukuran sebaran data, dan dengan menggunakan variance threshold, kita membatasi fitur-fitur yang memiliki variasi di bawah suatu ambang batas yang ditetapkan sebelumnya. Fitur-fitur dengan variasi rendah dianggap kurang informatif dan dapat dianggap sebagai noise dalam data. Oleh karena itu, dengan menerapkan variance threshold, kita dapat memilih fitur-fitur yang memiliki variasi yang cukup besar, yang diperkirakan lebih relevan dalam mewakili pola-pola penting dalam dataset. Dengan cara ini, proses ini membantu mengurangi dimensi dataset, meningkatkan efisiensi komputasi, dan dapat meningkatkan performa model analisis data yang akan dibangun. Variance threshold menjadi salah satu langkah awal yang berguna dalam preprocessing data untuk mempersiapkan data sebelum digunakan dalam berbagai tugas analisis data seperti klasifikasi, regresi, atau clustering.

1. Penerapan Seleksi Fitur Pada Dataset

Penerapan korelasi pada data juga membantu dalam memilih fitur yang saling berkaitan dan relevan dalam pengembangan model prediktif. Korelasi yang kuat antara dua fitur dapat mengindikasikan bahwa satu fitur dapat digunakan untuk memprediksi atau menjelaskan variabilitas dalam fitur lainnya. Dengan memahami korelasi antar fitur, kita dapat memadukan variabel-variabel yang saling mempengaruhi untuk mengoptimalkan model analisis data. Dengan kata lain, korelasi membantu mengidentifikasi interaksi antar fitur yang dapat membimbing keputusan analisis yang lebih baik dan model prediktif yang lebih akurat. Penerapan seleksi fitur ini menggunakan dataset pada tugas sebelumnya yaitu penyakit hepatitis di laman berikut ini: <https://www.kaggle.com/datasets/codebreaker619/hepatitis-data>. Pada penerapan metode ini menggunakan perhitungan manual dengan rumus:



*Gambar 1. Rumus Metode Correlation Coefficient*

Keterangan:

Σx : Total Nilai x

Σy :Total Nilai Y

Σxy :Total x Dikali Dengan y

Σx2 : Total Akar Kuadrat x2

Σy2 : Total Akar Kuadrat y2

n : Jumlah Data

Dengan menggunakan rumus tersebut didapatkan nilai r atau korelasi dari semua fitur (x) dengan target (y), berikut merupakan nilai r yang disajikan pada tabel dibawah ini:

| No | Fitur (x) | Target (y) | Nilai R | Hubungan |
| --- | --- | --- | --- | --- |
| 1 | Age | Class | -0,218434 | Negatif |
| 2 | Sex | Class | 0,173051 | Positif |
| 3 | Setroid | Class | 0,137413 | Positif |
| 4 | Antiviral | Class | 0,130196 | Positif |
| 5 | Fatigue | Class | -0,306049 | Negatif |
| 6 | Malaise | Class | -0,339530 | Negatif |
| 7 | Anoxeria | Class | -0,133638 | Negatif |
| 8 | Liver Big | Class | -0,093665 | Negatif |
| 9 | Liver Firm | Class | -0,020057 | Negatif |
| 10 | Spleen | Class | -0,234261 | Negatif |
| 11 | Spiders | Class | -0,389137 | Negatif |
| 12 | Ascites | Class | -0,469334 | Negatif |
| 13 | Varices | Class | -0,362385 | Negatif |
| 14 | Billirubin | Class | -0,388131 | Negatif |
| 15 | Alk | Class | -0,093803 | Negatif |
| 16 | Sgot | Class | -0,074160 | Negatif |
| 17 | Albumin | Class | 0,076168 | Positif |
| 18 | Protime | Class | 0,307212 | Positif |
| 19 | Histology | Class | -0,337856 | Positif |

REFERENSI

Kuhn, M., Johnson, K., Kuhn, M., & Johnson, K. (2013). An introduction to feature selection. Applied predictive modeling, 487-519.

Data Learn. Mengenal Feature Selection dalam Machine Learning. Diakses pada 21 September 2023 dari https://www.datalearns247.com/mengenal-feature-selection-dalam-machine-learning-69

LAMPIRAN

Perhitungan Manual Excel:

https://docs.google.com/spreadsheets/d/16U5nKJtgQtz0JA8CCslhpErww8idhddYJ2Tc68hd2tI/edit?usp=sharing