Technical Report Machine Learning

"Visualization and Exploration of Breast Cancer Data from Sci-Kit Learn"

Oleh:

Nama : Juan Meta Sirgianto

NIM : 1103202092

PROGRAM STUDI TEKNIK KOMPUTER

FAKULTAS TEKNIK ELEKTRO

UNIVERSITAS TELKOM

2023

I. MACHINE LEARNING

Machine learning adalah suatu bidang ilmu komputer yang mempelajari bagaimana suatu program atau sistem dapat belajar dari data dan mengambil keputusan atau tindakan yang sesuai dengan data tersebut tanpa harus secara eksplisit diprogram oleh manusia. Dalam machine learning, suatu algoritma digunakan untuk menganalisis data, mengenali pola-pola dalam data tersebut, dan mempelajari hubungan antara variabel dalam data tersebut. Setelah algoritma tersebut dilatih menggunakan data pelatihan, maka ia dapat digunakan untuk memprediksi atau mengambil keputusan yang tepat ketika diberikan data baru.

Contoh dari aplikasi machine learning termasuk pengenalan suara, pengenalan gambar, penerjemahan bahasa, dan pengambilan keputusan dalam bisnis. Machine learning juga sering digunakan dalam bidang kecerdasan buatan (AI), di mana sistem komputer dapat digunakan untuk melakukan tugas-tugas yang sebelumnya hanya dapat dilakukan oleh manusia.

II. MODEL-MODEL MACHINE LEARNING

Ada beberapa model yang sering digunakan dalam machine learning, di antaranya:

1. **Regresi**: model ini digunakan untuk memprediksi nilai numerik berdasarkan variabel input yang diberikan. Contohnya adalah prediksi harga rumah berdasarkan ukuran, jumlah kamar, dan lokasi.
2. **Klasifikasi**: model ini digunakan untuk mengklasifikasikan data menjadi kelas-kelas tertentu. Contohnya adalah klasifikasi email sebagai spam atau tidak spam, atau klasifikasi gambar sebagai kucing atau anjing.
3. **Clustering**: model ini digunakan untuk mengelompokkan data ke dalam kelompok-kelompok yang memiliki kemiripan berdasarkan fitur-fitur tertentu. Contohnya adalah mengelompokkan konsumen berdasarkan perilaku pembelian mereka.
4. **Jaringan Saraf Tiruan (Neural Network)**: model ini terinspirasi dari cara kerja otak manusia, di mana terdapat banyak neuron yang saling terhubung. Model ini digunakan untuk mempelajari pola-pola yang kompleks pada data. Contohnya adalah pengenalan wajah dalam aplikasi pengenalan suara atau visual.
5. **Pohon Keputusan (Decision Tree)**: model ini menggunakan pohon untuk merepresentasikan keputusan-keputusan yang diambil berdasarkan fitur-fitur tertentu pada data. Contohnya adalah prediksi apakah seseorang akan membeli produk atau tidak berdasarkan umur, jenis kelamin, dan pendapatan.
6. **Metode Markov (Markov Chain)**: model ini digunakan untuk memprediksi kejadian berikutnya berdasarkan kejadian-kejadian sebelumnya dalam suatu urutan. Contohnya adalah memprediksi kata-kata berikutnya dalam sebuah kalimat berdasarkan kata-kata sebelumnya.

Setiap model memiliki kelebihan dan kekurangan tergantung pada jenis data yang digunakan dan tujuan dari machine learning yang ingin dicapai. Oleh karena itu, pemilihan model yang tepat sangat penting untuk mendapatkan hasil yang akurat dan efektif.

III. 3 MODEL MACHINE LEARNING TERBAIK

Tiga model pembelajaran mesin terbaik adalah:

1. **Random Forest**: model ini menggunakan banyak pohon keputusan yang dibangun pada dataset yang berbeda-beda, lalu mengambil hasil voting dari semua pohon tersebut. Random Forest memiliki performa yang sangat baik pada data yang besar dan memiliki banyak fitur.
2. **Support Vector Machine (SVM)**: model ini mencari batas keputusan yang optimal antara dua kelas dengan membuat hyperplane (bidang pemisah) yang maksimal. SVM bekerja sangat baik pada dataset yang memiliki banyak fitur dan relatif sedikit sampel.
3. **Jaringan Saraf Tiruan (Neural Network)**: model ini terdiri dari beberapa lapisan neuron yang saling terhubung dan dapat mempelajari pola-pola yang kompleks pada data. Neural Network sangat baik digunakan pada data yang besar dan memiliki banyak fitur, terutama jika data tersebut bersifat non-linear.

Namun, pemilihan model yang tepat tergantung pada karakteristik dan kondisi dataset yang digunakan. Oleh karena itu, sebaiknya dilakukan evaluasi kinerja dari beberapa model sebelum memilih model yang tepat untuk suatu tugas klasifikasi.

IV. KUMPULAN DATA PUBLIK YANG TERSEDIA UNTUK KANKER PAYUDARA

Ada beberapa kumpulan data publik yang tersedia untuk kanker payudara, di antaranya:

1. **Breast Cancer Wisconsin (Diagnostic) Data Set**: kumpulan data ini berisi hasil diagnosis tumor payudara yang bersifat jinak atau ganas berdasarkan citra digital dan fitur-fitur lain yang diekstraksi dari citra tersebut. Kumpulan data ini tersedia di UCI Machine Learning Repository.
2. **Wisconsin Breast Cancer Dataset**: kumpulan data ini juga mengandung informasi mengenai tumor payudara yang bersifat jinak atau ganas berdasarkan pengukuran sel-sel pada citra. Kumpulan data ini tersedia di Kaggle.
3. **Digital Database for Screening Mammography (DDSM)**: kumpulan data ini berisi citra digital dari payudara yang digunakan untuk skrining kanker payudara. Kumpulan data ini tersedia di National Cancer Institute.
4. **Genomic Data Commons (GDC)**: kumpulan data ini berisi data genomik dari pasien kanker payudara, termasuk data RNA-seq, exome, dan whole-genome sequencing. Kumpulan data ini tersedia di National Cancer Institute.
5. **Cancer Imaging Archive (TCIA)**: kumpulan data ini berisi citra medis dari pasien kanker payudara, termasuk citra MRI, CT scan, dan PET scan. Kumpulan data ini tersedia di National Cancer Institute.

Kumpulan data ini dapat digunakan untuk mempelajari dan mengembangkan model-machine learning untuk deteksi dan prediksi kanker payudara, sehingga dapat membantu dokter dalam diagnosis dan pengobatan pasien. Namun, sebelum menggunakan kumpulan data tersebut, pastikan untuk memeriksa lisensi penggunaan dan privasi data yang berlaku.

V. KONTEN DARI DATASET

Konten dari data pada kumpulan data kanker payudara dapat bervariasi, tergantung pada sumber data dan jenis informasi yang disimpan. Beberapa jenis informasi yang biasanya terdapat dalam kumpulan data kanker payudara adalah:

1. **Citra medis**: kumpulan data ini dapat berisi citra medis, seperti mammogram, ultrasound, MRI, CT scan, dan PET scan. Citra medis ini dapat digunakan untuk melihat kondisi kanker payudara pada pasien.
2. **Fitur-fitur medis**: kumpulan data ini juga dapat berisi fitur-fitur medis yang terkait dengan kanker payudara, seperti ukuran tumor, lokasi tumor, jenis tumor, dan riwayat keluarga terkait kanker payudara.
3. **Data genomik**: beberapa kumpulan data kanker payudara juga dapat berisi data genomik, seperti data RNA-seq, exome sequencing, dan whole-genome sequencing. Data genomik ini dapat digunakan untuk mempelajari perubahan genetik pada pasien kanker payudara.
4. **Data klinis**: kumpulan data kanker payudara juga dapat berisi data klinis, seperti usia pasien, jenis kelamin, riwayat medis, dan hasil tes laboratorium.

Data pada kumpulan data kanker payudara biasanya berisi informasi yang penting untuk diagnosis dan pengobatan kanker payudara. Dengan menggunakan kumpulan data tersebut, para peneliti dan dokter dapat mengembangkan model machine learning yang dapat membantu dalam deteksi dini, diagnosis, dan pengobatan kanker payudara.

VI. KESIMPULAN

Dalam Technical Report ini kita mengetahui bahwa Machine Learning dapat digunakan untuk mengetahui diagnose kanker payudara melalui analisis dataset. Pada percobaan kali ini menggunakan dataset Breast Cancer Wisconsin (Diagnostic). Setelah melalui eksplorasi data, kita menemukan bahwa Random Forest adalah cara yang paling efektif untuk memprediksi diagnosis kanker payudara.